Supplementary Materials of ′A sensitive repeat identification framework based on short and long reads′

Xingyu Liao[1,2], Min Li[1], Kang Hu[1], Fang-Xiang Wu[3], Xin Gao[2](✉) and Jianxin Wang[1](✉)

[1] Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, ChangSha, 410083, CHINA.
liaoxingyu@csu.edu.cn, limin@mail.csu.edu.cn, Kanghu@csu.edu.cn, jxwang@mail.csu.edu.cn

[2] Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology(KAUST), Thuwal 23955, Saudi Arabia.
xin.gao@kaust.edu.sa

[3] Division of Biomedical Engineering, University of Saskatchewan, Saskatchewan, S7N 5A9, Canada.
faw341@mail.usask.ca

1 Introduction

The genomes of all eukaryotes contain a certain proportion of repetitive elements, particularly mammalians in which repeats account for 25-50% of their entire genomes[1],[2]. Repetitive regions can be caused by various mechanisms, such as chromosome translocations, transposons, errors in replication and recombination, etc[3]. Numerous studies have shown that the repetitive elements in the genome play indispensable roles in the evolution, inheritance, variation, gene expression, transcriptional regulation, chromosome construction, and physiological metabolism of living organisms[4],[5],[6],[7], and they are one of the principal causes of genomic instability[8]. How to quickly, accurately and completely identify repetitive regions in genomes has become an important research topic in bioinformatics.

1.1 Classification of the Repetitive DNA Sequence

According to the arrangement, the repetitive regions in eukaryotic genomes can be divided into two types: tandem repeats and interspersed repeats[9], just as shown in Table S1. Tandem repeats are arrays in which repeating elements consisting of 1 to 500 bp sequences are connected end to end to form multiple repeats. They are arranged in clusters in the telomere, the centromere peripheral region and the heterochromatin region on the chromosome arm[10]. They are widely found in the genomes of eukaryotes and certain prokaryotes. The tandem repeats are distributed in both coding regions and non-coding regions[11]. For example, rRNA genes, tRNA genes and histone genes appear in the coding regions in the form of tandem repetitive sequences. Satellite (satellite DNA), small Satellite (minisatellite DNA) and microsatellite DNA (microsatellite DNA, Simple sequence repeats, SSR) are three common types of tandem repeats in non-coding regions.

The repeating elements of interspersed repeats are not connected, but are doped with other unrelated repeats or single copy sequences. They are dispersed throughout the genome and usually refer to transposons, including retrotransposons and DNA transposons[12]. There are two main types of retrotransposons. 1) long-terminal repeat retrotransposons (LTRs). The length of LTR retrotransposons generally ranges from 100 bp to 25 kb [13][14] and 2) non-long terminal repeat retrotransposons (Non-LTRs), which are divided into long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs)[15]. LINE is a type of reverse transposon that can transpose spontaneously, which is the transcription product of RNA polymerase II. LINE does not contain long terminal repeats and can be up to several thousand base pairs in length. It present in all eukaryotic genomes, but only in a small percentage of plant genomes. The length of SINE is between 80 bp to 500 bp. It cannot be transposed autonomously, and there is no open reading frame. It is a product formed by RNA polymerase III and tends to be inserted into gene-rich regions. The DNA transposon is transposed by a cut-paste or copy-paste mechanism[16]. Such elements are excised from the chromosome by interleaved double-stranded cleavage and reinserted into other locations in the genome without the involvement of RNA. They contain a transcriptase and the inverted repeats at both ends, which are transposed between the chromosome autochromatin segments. DNA transposons can be divided into three categories according to the inverted repeat sequences, which are AC/Ds (11bp), CATA (up to 15.2 bp) and small reverse repeat factor MITE (8bp to 12bp terminal reverse repeat).

Table S1. Repetitive sequence classification.

Type	Subtype			Length(bp)	Distribution region
Interspersed	retrotransposons	Non-LTRs	LINE	500∼4,000	Scattered distribution
			SINE	< 500	Scattered distribution
		LTRs	LTR	100∼5,000	Two ends of retrovirus
	DNA transposons		AC/Ds	11bp	Scattered distribution
			CATA	≤15.2bp	Bacterial, plant, animal genes
			MITE	8 to 12bp	Bacterial, plant, animal genes
Tandem	Satellite			150∼500	Heterochromatin
	Minisatellite			10∼100	Euchromatin
	Microsatellite			2∼10	Non-coding region, introns

′*Non-LTRs*′ Non-long terminal repeat retrotransposons. ′*LTRs*′ indicates Long-terminal repeat retrotransposons. ′*LINE*′ indicates long interspersed nuclear element. ′*SINE*′ indicates short interspersed nuclear element. ′*MITE*′ indicates miniature inverted repeat element.

1.2 Classification of existing detection methods

Many computational methods have been proposed to identify repetitive regions in genomes, which can be classified into three categories including homology-based, structure-based and *de novo* methods. The general classification of detection methods is shown in Fig. S1. The homology-based identification methods are based on a certain database for the homology search, so as to find and mask the repetitive sequences. RepeatMasker[17] is a representative method of this category, which performs a similarity search based on the local alignment with AB-BLAST[18] or Crossmatch[19]. RepeatMasker has its own library of repetitive sequences, and has become a gold standard of this field in terms of accuracy. Most other similar methods use RepeatMasker as the main reference library. The homology-based methods have the high search efficiency and can be used to discover families with small numbers of copies. However, everything has two sides, such methods can only be used to search for known repetitive sequences, and can not be used to discover new repetitive sequences. Typical methods based on the homologous search are also include: Censor[20], TESeeker[21], Greedier[22], Greedier[23] and T-lex[24]. Among them, CENSOR is a program designed to identify and eliminate fragments of DNA sequences homologous to any chosen reference sequences, in particular to repetitive elements, which uses RepBase as a homologous database; TESeeker implements an automated homology-based approach for identifying transposable elements, which uses Tefam and RepBase as homologous databases; Greedier effectively solves the problem of embedded duplications by using greedy algorithms and local alignment methods; and T-lex uses high-throughput sequencing data to find tandem repetitive flanking regions, non-LTR and fragment repeat regions.

The structure-based identification methods are based on the prior information of the sequence structure features, using a heuristic algorithm to find and identify the repeated sequences, the types of repeat sequences that can be found are determined by the sequence characteristics they have mastered[25]. Sequence characteristics refer to the sequence structure features obtained through biological means, such as the structure of the translocation factor, and the conserved protein functional domain. The structure-based identification methods also need the support of a known library which was originally constructed by biologists through manual annotation. With the rapid development of machine learning and deep learning technologies, researcher have begun to propose some automatic labeling methods, such as using SVM to train a classification model using a standard structure-based repeated sequence library. The repeat type in here refers to the repeated fragments of different structural types in the known library. If the structural features of the query sequence cannot be found in any type of repeated fragments in the known library, the structure-based identification method cannot complete the detection of this query sequence. Typical structure-based methods include: LTRharvest[26], MASiVE[27], MGEScan-LTR[28], SINEDR[29], MITE-Hunter[30], detectMITE[31], FINDMITE[32], MUST[33], MITE-Digger[34] and MITE Tracker[35]. Among them, LTRharvest implements several steps of filtering based on structural features of sequences, determines the boundary position of the LTR, and annotates the LTR with LTRdigest. MASiVE is a tool specifically designed to analyze specific LTR transposons in plant genomes. MGEScan-LTR uses approximate string matching and protein domain analysis methods to determine intact LTR retrotransposons. Transposable elements (TEs) are a type of repeat sequences abundant in eukaryotic genomes. TEs play important roles in genome organization and evolution. Commonly, TEs in genomes can be classified into two major categories retrotransposons (Class I) and DNA transposons (Class II). Miniature inverted repeat transposable elements(MITEs) are a special type of DNA transposons. MITE-Hunter, detectMITE, FINDMITE, MUST, MITE-Digger and MITE-Hunter are six typical structure-based methods for MITEs identification, among which FINDMITE requires users to predefine the TSD sequences, TIR length and the minimum and maximum distances between the TIRs, MITE-Hunter is a program pipeline that can be used to identify MITEs as well as other small Class II non-autonomous TEs from genomic DNA datasets, compared to FINDMITE and MUST, MUST-Hunter has a much lower false-positive rate and the output is easier to be checked and classified. Both MITE-Hunter and MITE Digger utilized a mixture of both de novo and structural-based methods in MITE detection. Although they have successfully decreased false positive rates in MITE detection, both of them cannot detect all MITEs hidden in the genomes.

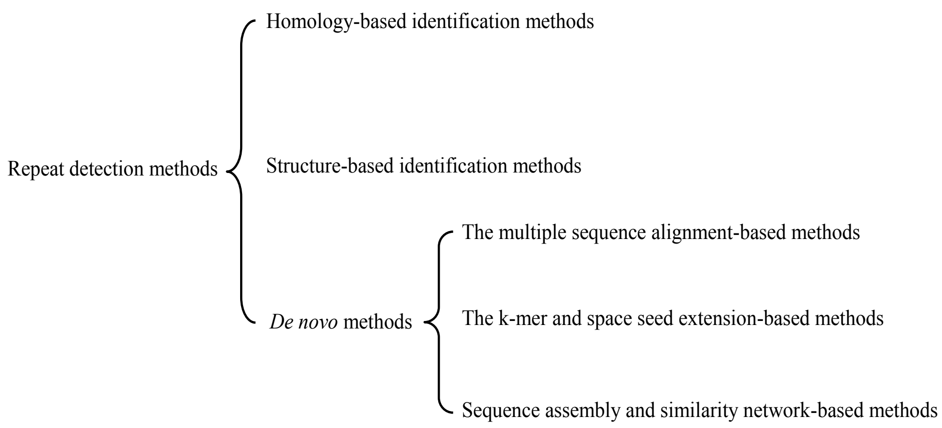

Repeat detection methods
Homology-based identification methods
Structure-based identification methods
De novo methods
The multiple sequence alignment-based methods
The k-mer and space seed extension-based methods
Sequence assembly and similarity network-based methods

Fig. S1. The general classification of detection methods.

The *de novo* methods need no prior information of the repeat structure or similarity to the known repeat sequences, and tend to be more flexible than the other two methods[36]. The *de novo* methods can also be divided into three categories. The first category relies on the multiple sequence alignment to identify repeats, which mainly includes Repeat Pattern Toolkit[37], RECON[38], PILER[39], LTRdigest[40]. Among them, the tools in RPT (repeat pattern toolkit) include identifying significant local alignments (utilizing both two-way and three-way alignments), dividing the set of alignments into connected components (signifying repeat families), computing evolutionary distances between repeat family members, constructing minimum spanning trees from the connected components, and visualizing the evolution of the repeat families[41]. RECON uses WU-BLAST[42] as the initial alignment tool, which resolves the differentiation between segmental duplication and typical interspersed repeat such as TEs, by using the different distribution graphics of the two types of repeats upon the multiple sequence alignment. In the past few years, RECON has become the dominant tool for the *de novo* repeat family identification in newly sequenced genomes. For example, it has been used to construct a library of C.briggsae repeat families, making this an ideal test bed[43]. PILER adopts the pairwise alignment of long sequences to discover and distinguish between different types of repetitions,

Table S2. The detailed classification of existing detection methods.

Types	Tools	Website
homology-based	RepeatMasker	http://www.repeatmask.org
	Censor	http://www.girinst.org/censor/download.php
	TESeeker	http://repository.library.nd.edu/view/16/index.html
	Greedier	https://omictools.com/greedier-tool
	T-lex	http://petrov.standord.edu/cgi-bin/Tlex_manual.html
	Rteclass1	http://www.girinst.org/RTphylogeny/RTclass1
	RetroSeq	http://githup.com/tk2/RetroSeq
	SINEBase	http://sines.eimb.ru
	TRANSPO	http://alggen.lsi.upc.es/recerca/search/transpo/transpo.html
	PLOTREP	http://repeats.abc.hu/cgi-bin/plotrep.pl
	TinT	http://www.bioinformatics.uni-muenster.de/tools/tint/index.hbi?lang=en&mscl=0&cscl=0
structure-based	TE Displayer	http://labs.csb.utoronto.ca/yang?TE_Displayer/TE_Displayer_1.0.2
	FINDMITE	http:yywww.biochem.vt.eduyaedes
	detectMITE	https://sourceforge.net/projects/detectmite
	LTR_STRUC	http://www.mcdonaldlab.biology.gatech.edu/ltr_struc.htm
	LTR_FINDER	http://tlife.fudan.edu.cn/tlife/ltr_finder/
	Retro Tector	http://retrotector.neuro.uu.se/pub/queue.php?show=submit
	RTAnalyzer	http://www.riboclub.org/cgi-bin/RTAnalyzer/index.pl
	P-MITE	http://pmite.hzau.edu.cn/django/mite
	TSDfinder	http://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/TSDfinder
	MAK	http://perl.idmb.tamu.edu/mak.htm
	LTRharvest	http://www.zbh.uni-hamburg.de/ltrharvest
	MASiVE	http://bat.ina.certh.gr/tools/masive/
	MGEScan-LTR	https://bio.tools/mgescan-ltr
	MITE-Hunter	http://target.iplantcollaborative.org/mite_hunter.html
	MITE-Digger	http://labs.csb.utoronto.ca/yang/MITEDigger
de novo	Repeat Pattern Toolkit	https://www.aaai.org/Papers/ISMB/1994/ISMB94-001.pdf
	RECON	http://eddylab.org/software/recon/
	PILER	http://www.drive5.com/piler
	BLASTER suite	https://urgi.versailles.inra.fr/Tools/Blaster
	REPuter	http://bibserver2.cebitec.uni-bielefeld.de/reputer
	RepSeek	http://www.abi.snv.jussieu.fr/public/RepSeek
	Repeat-match	http://mummer.sourceforge.net
	LTRdigest	https://omictools.com/ltrdigest-tool
	LTR_retriever	https://github.com/oushujun/LTR_retriever
	REPET	https://urgi.versailles.inra.fr/Tools/REPET
	RepeatScout	https://omictools.com/repeatscout-tool
	Repeatmodelerl2	http://www.repeatmasker.org/RepeatModeler/
	Generic Repeat Finder	https://github.com/bioinfolabmu/GenericRepeatFinder
	EDTA	https://github.com/oushujun/EDTA
	RepARK	https://github.com/PhKoch/RepARK
	REPdenovo	https://github.com/simoncchu/REPdenovo
	RepLong	https://github.com/ruiguo-bio/replong

including tandem arrays, scattered families, terminal repeats, and pseudo-satellites. LTRdigest uses the local alignment and hidden Markov model(HMM) to identify families based on sequence characteristics and annotates the internal structure of long terminal repetitive retrotransposons (It can be used to identify protein coding regions associated with the retrotransposition process as well as internal regulatory features using local alignment and hidden markov model-based algorithm).

The methods in the second category rely on *k-mer* and space seed extension strategies to identify repetitive sequences. These methods convert the sequences in the genome into *k-mers* of a certain length, and then select the *k-mers* whose frequency exceeds a certain threshold as a seed, search for the locations of these seeds in the genome, and perform the sequence extension to both ends of the genome and get the expended sequences. During the extension process, it always judges whether the extended sequences are consistent at multiple locations in the genome. If yes, it continues the extension, otherwise stops the extension. EDTA[44], RepeatFinder[45], RepeatScout[46], ReAS[47], Generic Repeat Finder (GRF)[48] and Repeatmodelerl2[49] are typical of such kind of tool, they start with a library of high-frequency *k-mers* that are used in initial identification, alignment and extension of sequence substrings. For example, ReAS is designed to use multiple sequence reads as a substrate, RepeatScout is developed for identification of repeats in assembled genomic regions, builds a library of high-frequency *k-mers* and retrieves substrings of the input sequence containing a specific *k-mer* in a manner similar to that of ReAS. Generic Repeat Finder (GRF) is a tool for genome-wide repeat detection based on fast, exhaustive numerical calculation algorithms integrated with optimized dynamic programming strategies, which is the latest representative of the second category of tools. GRF uses a new scoring system to calculate cumulative scores for subsequences of certain length for all chromosomes, and groups sequences with the same scores together using a hash table, and finds candidate repeats (seed regions) with each group by comparing nucleotide sequences, and extends the seed regions by allowing mismatches in both upstream and downstream flanking regions. RepeatModeler2 is a pipeline for automated *de novo* identification of TEs that employs four distinct discovery algorithms (RepeatScout, RECON, LTRharvest[50] and LTR_retriever[51]), which takes advantages of the unique strengths of each tools as well as providing a tractable solution to analyzing large datasets such as whole-genome assemblies. Extensive *de-novo* TE Annotator (EDTA) combines the best-performing programs (LTR_FINDER[52], LTR_retriever, GRF, TIR-Learner[53], HelitronScanner[54] and RepeatModeler) and subsequent filtering methods for *de novo* identification of each TE subclass and compiles the results into a comprehensive non-redundant TE library. Since the tools called in RepeatModeler2 and EDTA are more or less related to the seed-extension strategy, so we classify them as the second category of tool.

The methods in the third category rely on sequence assembly and similarity network to identify repeats, which mainly includes RepARK [55], REPdenovo[56] and RepLong[57]. Among them, the first two methods obtain the repetitive sequences by assembly of the high frequency *k-mers*, and the last method is currently the only detection method that is suitable for the third generation sequencing reads. RepARK is a *de novo* repeat assembly method which avoids potential biases by using abundant *k-mers* of NGS WGS reads without requiring a reference genome. RepARK is orders of magnitude faster than the other methods and generates libraries that are composed almost entirely of repetitive motifs, more comprehensive and almost completely annotated by TEclass[58]. REPdenovo is designed based on the idea of frequency *k-mer* assembly for constructing repeats directly from sequencing reads, which provides many functionalities, and can generate much longer repeats than existing tools. RepLong is a novel *de novo* repeat elements identification method based on PacBio long reads. RepLong can handle lower coverage data and serve as a complementary solution to the existing methods to promote the repeat identification performance on long read sequencing data. The detailed classification of existing detection methods is shown in Table S2.

In the process of NGS sequence assembly, the paired-end reads with large insert size are mainly used to resolve the ambiguity paths generated by the repeated regions in assembly graph and determine the successive positions of contigs in the process of scaffolding[59]. The assembly-based detection methods are based on the high-frequency *k-mers* assembly to obtain repetitive sequences. Due to the lack of support for long sequence fragments that can span the repetitive regions, the assembler will inevitably make misassemblies when processing these short and highly repetitive sequences[60]. On the other hand, they depend too much on the threshold of the high frequency *k-mer*, which is difficult to obtain accurately due to the sequencing bias. The long reads are more likely to cover repetitive regions completely, which is more favorable for recognizing long repeats[61]. However, the high error rate of long reads has a great impact on the accuracy of this method. In addition, the method constructs the similarity network by comparing the long reads, and then uses the community discovery algorithm to get the detection results, which has higher computational complexity when processing large datasets. In summary, it is often difficult for existing *de novo* detection methods to achieve satisfactory results in terms of both accuracy and size.

1.3 Detailed introduction of the working modes of LongRepMarker

1.3.1 reference-assisted mode

The labeling of repetitive regions in large genomes (such as mammalian and plant genomes) plays an important role in understanding the evolution, inheritance, gene expression and other life processes of complex organisms. In addition, labeling the repetitive regions in large genomes is also helpful to improve the quality of sequence assembly of corresponding species. However, due to the large amount of sequencing data, it is very difficult for *de novo* methods to deal with these massive sequencing data, and the detection results are often poor. In order to solve this problem, LongRepMarker provides a reference-assisted mode. The repetitive sequences are a special kind of overlap sequences, and the overlap sequences occupy only a small part of the overall sequences. By finding the overlap sequences between assemblies or chromosomes, the algorithm locates the repetitive regions faster and more accurately. If there is a reference sequence or a rough assembly of a species or a reference sequence of similar species, LongRepMarker can quickly and accurately derive a repeat library for that species. The input of this mode is the reference sequences or the rough assembly of a species or the reference sequences of a similar species, and the output is the detection results which include the repeat library and several detailed reports.

1.3.2 *de novo* mode based on only NGS short reads

In the process of NGS sequence assembly, the paired-end reads with large insert size are mainly used to resolve the ambiguity paths generated by the repeated regions in the assembly graph and determine the successive positions of contigs in the process of scaffolding. The assembly-based detection methods are based on the high-frequency *k-mer* assembly to obtain repetitive sequences. Due to the lack of support for long sequence fragments that can span the repetitive regions, the assembler will inevitably make misassemblies when processing these short and highly repetitive sequences. On the other hand, they depend too much on the threshold of the high frequency *k-mers*, which is difficult to obtain accurately due to the sequencing bias. In this mode, the proposed algorithm produces assemblies by assembling all paired-end reads instead of assembling the high frequency *k-mers*, the algorithm can identify the repeats in the genomes to a greater extent. Due to sequencing bias, the high frequency threshold is often difficult to obtain accurately, which has a great impact on the range of the high frequency *k-mers*. By using the multi-alignment unique *k-mers* to identify repeats in overlap sequences, the algorithm can obtain the repeats in the genomes more comprehensively and stably.

1.3.3 *de novo* mode based on NGS short reads + barcode linked reads / SMS long reads

Repetitive elements have played, and are continuing to play, critical roles in genome evolution. Prokaryotic genomes contain a variety of large size and low copy number repeated sequences, these sequences may contribute to the evolution of chromosome structure through DNA rearrangements such as chromosome deletions, duplications and inversions. However, most existing *de novo* identifications (such as RepARK, Repdenovo, RepLong, etc.) cannot achieve satisfactory results for marking repetitive regions in both accuracy and size as the NGS reads are too short to identify long repeats and SMS long reads are with the high error rate. In this mode, the proposed algorithm produces assemblies based on Illumina short paired-end reads and barcode linked reads or SMS long reads, and calls a NGS-based assembler called SPAdes[62] which adopts some better repeat processing strategies and has superior performance than other similar tools(such as SOAPdenovo2[63], Abyss[64], Velvet[65] and IDBA-UD[66]). The reasons for choosing SPAdes as the assembler and the performance comparison analysis of SPAdes and other similar tools are shown in **section 1.5 of the supplementary**, and the advantages of using SPAdes to assemble Illumina short paired-end reads and barcode linked reads or corrected SMS long reads in repetitive sequences identification are described in **section 1.4 of the supplementary**.

1.3.4 *de novo* mode based on only SMS long reads

As the development of the third generation sequencing, the SMS long reads have been widely applied in various fields of bioinformatics. In order to better comply with market demand and further expand the application scope of this system, we have developed a detection mode only based on the SMS long reads under the LongRepMarker framework. The input of this mode is only SMS long reads (Pacbio, CCS and HiFi reads) and the output is the detection results which include the repeat library and several detailed reports. The workflow of this mode can be divided into the following steps: 1) getting the overlap sequences between long reads. 2) estimating the average coverage of the overlap sequences. 3) filtering the overlap sequences with low coverage. 4) getting the filtered overlap sequences with the high copy number in long reads(for example, the copy number is more than 1.5 × AverageCoverage). 5) identifying the genetic variations existing in the detected repetitive regions and 6) generating the final detection results. It can be seen from the related test results displayed in the experimental section(**Setction S3.3.6**) that compared with the existing long reads-based detection methods, this mode has the advantages of low memory consumption, high speed and high detection accuracy. The working principle of this mode is shown in FigS2.

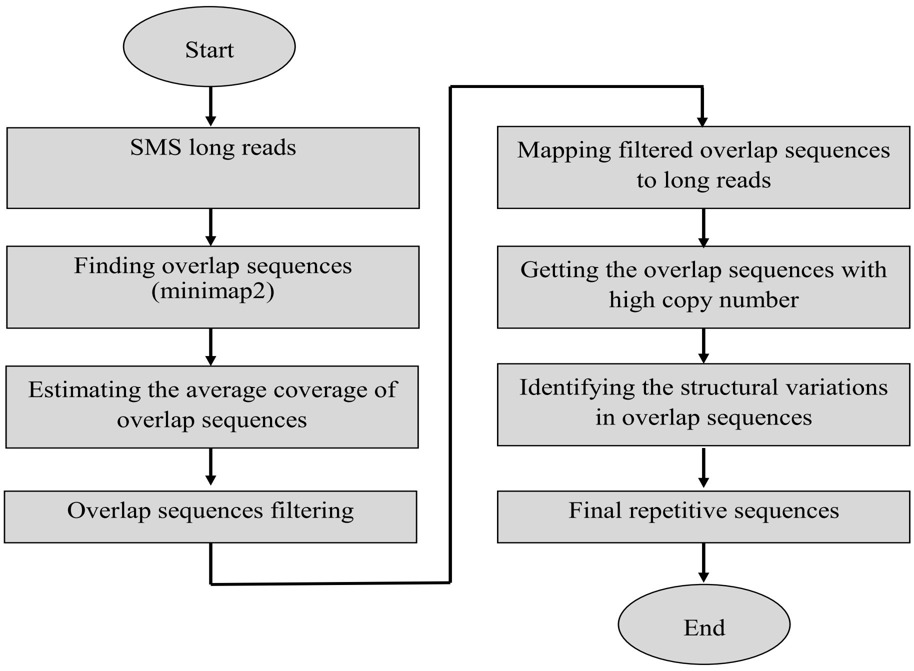

Start
SMS long reads
Finding overlap sequences (minimap2)
Estimating the average coverage of overlap sequences
Overlap sequences filtering
Mapping filtered overlap sequences to long reads
Getting the overlap sequences with high copy number
Identifying the structural variations in overlap sequences
Final repetitive sequences
End

Fig. S2. The working principle of the detection mode based on only the SMS long reads.

1.4 Other issues that users care about

1.4.1 The main differences between the reference-assisted mode and the *de novo* mode

The main differences between the reference-assisted mode and the *de novo* mode are as follows:

1) The reference-assisted mode does not need sequence assembly in the pipeline. Due to this reason, the reference-assisted mode is much faster than the *de novo* mode, and it can be used to handle some large genomes, such as mammalian and plant genomes. On the contrary, since the *de novo* mode needs to rely on the results of sequence assembly for detection, and sequence assembly often requires a lot of time and huge computing resources when processing the large datasets. Therefore, the *de novo* mode is often slow and consumes a lot of memory when processing the large datasets;

2) Since the main processing object of the reference-assisted mode is the reference genome. The reference genome is the consensus genome of the species, which has high accuracy and low specificity. The processing object of the de novo mode is the sequencing reads which has high specificity and relatively low accuracy. The characteristics of these two types of data are quite different, among which the error rate and abundance of the sequencing reads is significantly higher than the reference sequences, so the *de novo* mode sets a large error tolerance rate. The reference-assisted mode is often used to quickly and accurately obtain the repetitive regions in the reference genome of large species, and the error rate it can accommodate is very small. Further more, the pipeline of reference-assisted mode is quite different from that of *de novo* mode. For example, in *de novo* mode, after the preliminary repetitive sequences are obtained based on the assembly of overall reads (NGS short paired-end reads + barcode linked reads / SMS long reads), the paired-end reads with high *k-mer* coverage are selected to enter the scaffolding process[67] together with the preliminary repetitive sequences, which can make the final detection results have better integrity and higher coverage;

3) The reference-assisted mode does not consider the genetic variations that occurred in the repetitive regions, because the structural variations usually occur in sequencing reads, while they are usually not considered in the reference genome. Based on this common sense, we can think that the reference-assisted mode is mainly used to detect the common repetitive sequences of species, while the *de novo* mode can be used to well identify the repetitive sequences existing in individual genomes of species, which may not appear in the reference genome.

1.4.2 The advantages of barcode linked reads and long reads in assisting the assembly of Illumina short paired-end reads

Linked-reads provide the long range information missing from standard approaches[68], which builds on the Illumina sequencing technology to provide indexing and barcoding information along with short reads to localize the latter on long DNA fragments (linked-reads), thus benefiting the economies of a high throughput platform. As sequencing reads from 20 to 200kb are barcoded/linked, applications of the technology mainly focused on phasing variant bases in human genomes[69]. Because the length of linked-reads can reach to 20 to 200kb, they can easily span most of repetitive regions in genome. In addition, because the sequencing accuracy of linked-reads is the same as that of Illumina sequencing (the error rate is about 0.2% to 0.5%), they have large size and high sequencing accuracy. These two characteristics of linked-reads make them very suitable for the repetitive sequence identification.

Although barcode linked reads have a long span and high sequencing accuracy, there are not many species libraries measured by this technology, which severely limits the usability of the proposed tool. In order to solve this problem, we have introduced the rapidly developing and widely used third generation sequencing reads into this framework. The third generation sequencing technologies produce long reads with an average length of 10 kbp to several hundreds kbp. For example, there are two widely used third-generation single molecule sequencing platforms: Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (Nanopore). The average length of PacBio long reads is more than 10kbp, while the average length of the Nanopore long reads can reach several hundreds kbp, and statistics show that the longest read length is more than 1 million bp[70]. Longer lengths are attractive because they enable disambiguation of repetitive regions in a genome or a set of genomes. These long reads however the extremely noisy, and display error rates of 10% to 20% (the average error rate of PacBio long reads is between 10% to 15%, while the average error rate of Nanopore long reads is between 15% and 20%). Moreover, long reads's errors are mainly composed of insertions and deletions[71]. Error correction algorithms (such as HECIL[72], LoRDEC[73] and HALC[74]) have been designed to identify and fix sequencing errors, thereby benefiting re-sequencing or *de novo* sequencing analysis. Although the third-generation sequencing has the defect of high error rate, it is currently the mainstream of the development of sequencing technologies, and the available sequencing data is abundant. In addition, with the development of sequencing and error correction technologies, the error rate of the current third-generation sequencing has a significant downward trend. Introducing the corrected SMS long reads into this framework can greatly expand the application range of it. The principle of barcode linked reads and SMS long reads in assisting short reads to resolve the ambiguous paths caused by repetitive regions in the assembly graph is shown in Fig. S3.

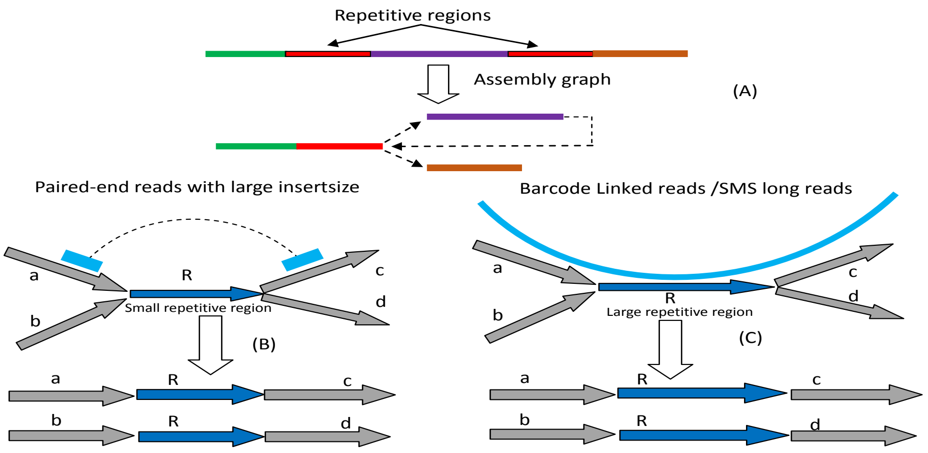


Fig. S3. The principle of barcode linked reads and SMS long reads in assisting short reads to resolve the ambiguous paths caused by repetitive regions in the assembly graph.

1.4.3 The main reasons for choosing SPAdes as assembler

SPAdes is one of the Eulerian *de Bruijn* graph based assemblers, it was originally designed for prokaryotic genomes (Such as microbial and single-cell sequencing) but was later developed to accommodate large eukaryotic genomes. This program uses the paired-end de Bruijn graphs, which is a kind of double-layered *de Bruijn* graphs. The *k-mers* from DNA fragment reads build the inner *de Bruijn* graph, which is used for the contig assembly. On the other hand, the paired *k-mers* with the large insert size build the outer *de Bruijn* graph, which is used for repeat resolving or scaffolding(Fig. S4). ExSPAnder[75] is a universal repeat resolver for short DNA fragment assembly, which uses a simple path extension approach for repeat resolution. ExSPAnder is a separate solution proposed by the SPAdes research group to solve the problem of repetitive regions in the short sequence assembly(Fig. S4), it has been incorporated into the new version of SPAdes (After version 3.0). In order to verify the effectiveness of SPAdes, we tested the performance of it on $B.faecium$ dataset, the detailed information of this dataset is shown in Table S3. The records without an asterisk after the tool name in Tables S4 and S5 are the assemblies generated by the tools under the condition of single paired-end reads (only one group of paired-end reads). The records with an asterisk after the tools name are the assemblies generated by the tools under the condition of multiple sets of paired-end reads(both paired-end reads and jumping reads). The different sets of paired-end reads of $B.faecium$ can be downloaded from GAGE-B (http://ccb.jhu.edu/gage_b/) and GAGE (http://gage.cbcb.umd.edu/) websites, respectively. The results shown in Tables S4 and S5 have fully proved that the repeat solution strategies adopted in ExSPAnder are of great significance to improve the integrity of the assembly. In addition, the new version of SPAdes (After version 3.0) not only incorporates ExSPAnder, but also makes many new optimizations and improvements, so that the new version of SPAdes has more superior performance in solving the repetitive regions.

The advantages of SPAdes are as follows: 1) SPAdes is an iterative short read genome assembly tool, which produces longer and more accurate contigs than other similar tools, for detailed proof please review the following literatures [76], [77], [78], [79] and [80]; 2) SPAdes works with many data types such as Illumina paired-end reads (Illumina paired-end / high-quality mate-pairs / unpaired reads), IonTorrent paired-end and PacBio CCS reads, and can also make full use of the support information provided by Oxford Nanopore long reads, PacBio long reads, and Sanger sequencing reads for high-quality hybird assembly (Note: Oxford Nanopore long reads, PacBio long reads, and Sanger sequencing reads can only be used in hybird assemblies with higher quality short read data).

Table S3. Details of the $B.faecium$ isolate dataset.

Library	Number of reads	Average coverage	Coverage span	Insert size	Insert span	Chimeric read-pairs(%)	Unaligned read-pairs(%)
Paired-end reads	13M	400 ×	210-570	270bp	150-400bp	1	16

$'Insert\,span'$ indicates the shortest insert size interval that contains at least 95% of properly aligned read-pairs. $'Unaligned\,read-pairs'$ indicates the percentage of read-pairs that have at least one read unaligned. $'Chimeric\,read-pairs'$ indicates the percentage of chimeric read-pairs among all read-pairs. All statistics was obtained using bowtie2. $'Coverage\,span'$ is the smallest coverage interval that includes a least 95% of all genomic positions.

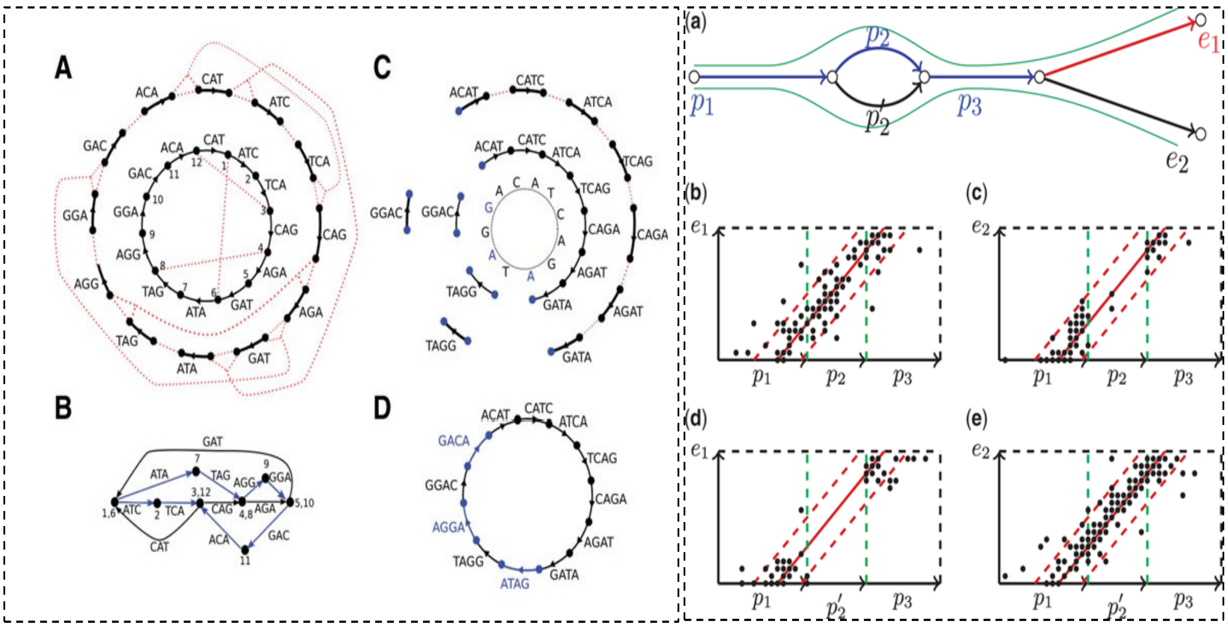


Fig. S4. The left sub-graph shows the standard and multiplied *de Bruijn* graph. SPAdes take advantage of multiple *de Bruijn* graphs to solve the problems caused by sequencing bias and repetitive regions. This program uses the paired-end de Bruijn graphs which is a kind of double-layered *de Bruijn* graphs to construct the contigs. The *k-mers* from DNA fragment reads build the inner de Bruijn graph, which is used for the contig assembly. On the other hand, the paired k-mers with the large insert size build the outer *de Bruijn* graph, which is used for repeat resolving or scaffolding. This program provide a better solution which, instead of using a single k (*k-mer* size), iterates from k=k_{min} to k=k_{max}. At each iteration, the constructed contigs are used as reads for the next iteration. These contigs carry the *k-mers* of the current iteration, which may be missing in the next iteration, to the next iteration, thus solving some of the gap problems. In addition, the large k is used to resolve the branches caused by repetitive regions. The right sub-graph shows the principle of ExSPAnder resolving repeats using either a single or multiple paired-end reads with different insert sizes. ExSPAnder uses a simple path extension approach for repeat resolution that was originally proposed in the Ray assembler [and later used in Telescoper] and combines it with some ideas from the Rectangle Graph approach. Given a set of paths in the assembly graph (i.e. simplified de Bruijn graph of k-mers in reads after removal of bulges, tips and chimeric edges), ExSPAnder attempts to extend each path with the goal to generate longer paths. ExSPAnder uses a decision rule Extend(P) that either chooses one of the extension edges to extend the path P or makes the decision to stop growing this path beyond the ending vertex of P. The procedure is iterated over all the paths until no path can be further extended. To initiate this algorithm one can start with a set of single-edge paths formed by all sufficiently long edges in the assembly graph. The resulting paths are output as contigs after removing the paths that are contained within other paths as well as removing non-informative overlaps (i.e. suffixes of paths that represent prefixes of other paths).

Table S4. Comparison of contigs for the $B.faecium$ isolate dataset.

Assembler	NG50	Num	Max(bp)	MA	GF(%)
ABySS	203	40	**672**	**0**	**99.9**
Ray	114	51	436	1	98.9
SOAP2	20	333	61	**0**	98.9
Velvet	144	47	550	**0**	99.4
Velvet-SC	163	46	550	0	99.4
IDBA-UD	202	39	483	**0**	99.4
SPAdes2.4	361	24	635	1	99.7
ExSPAnder	**380**	**22**	**672**	1	99.5
ABySS *	203	40	672	**0**	**99.9**
ALLPATHS-LG*	313	21	686	**0**	99.5
Ray *	87	88	416	2	96.8
SOAP2 *	87	50	500	23	98.8
Velvet *	103	75	242	11	99.0
Velvet-SC *	253	40	545	15	99.8
IDBA-UD *	207	41	483	**0**	99.4
ExSPAnder *	**3268**	**2**	**3268**	1	**99.9**

$'Num'$ indicates the number of contigs. $'Max(bp)'$ indicates the length of the largest contig. $'MA'$ indicates the number of misassembly. $'GF(\%)'$ indicates the genome fraction(%). The records with an asterisk after the tool name is the assemblies generated by the tool under the condition of multiple groups of paired-end reads.

Table S5. Comparison of scaffolds for the $B.faecium$ isolate dataset.

Assembler	NG50	Num	Max(bp)	MA	GF(%)
ABySS	383	24	676	**0**	**99.9**
Ray	204	31	553	1	98.9
SOAP2	**477**	26	**724**	**0**	99.3
Velvet	**477**	28	**724**	**0**	99.4
Velvet-SC	**477**	28	671	**0**	99.4
IDBA-UD	250	30	671	**0**	99.4
SPAdes2.4	361	**22**	671	1	99.7
ExSPAnder	380	**22**	672	1	99.5
ABySS *	250	30	739	1	**99.9**
ALLPATHS-LG*	3610	7	**3610**	1	99.5
Ray *	106	75	416	2	96.8
SOAP2 *	480	28	810	2	96.4
Velvet *	2651	14	2651	78	99.1
Velvet-SC *	945	102	1381	500	98.9
IDBA-UD *	1002	9	1692	**0**	99.4
ExSPAnder *	**3268**	**2**	3268	1	**99.9**

$'Num'$ indicates the number of contigs. $'Max(bp)'$ indicates the length of the largest contig. $'MA'$ indicates the number of misassembly. $'GF(\%)'$ indicates the genome fraction(%). The records with an asterisk after the tool name is the assemblies generated by the tool under the condition of multiple groups of paired-end reads.

1.4.4 The community discovery algorithm of RepLong

In the paper corresponding to the tool RepLong[57], the authors proposed that the overlaps between the repeat reads are more intensive than that of the other reads, which in the constructed network is characterized with some topology structures of denser intra-connectivity and meanwhile sparse inter-connectivity. Such topology structures are also known as community in graph theory. Thus repeat identification can be transformed into a community identification problem when the network is constructed. For example, in Fig. S5, two piles of reads after sequence alignment correspond to two repeats. In each pile, the coverage inside is much higher than outside, so there are corresponding community structures in the network of read overlaps thanks to the dense intra-connections in the read piles.

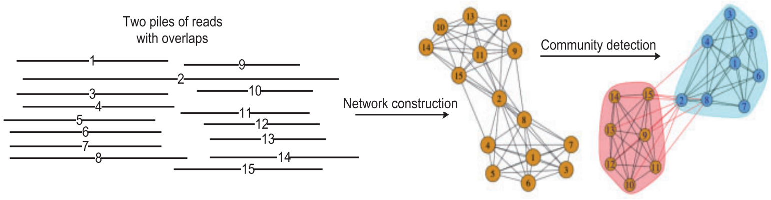


Fig. S5. Two piles of reads after sequence alignment correspond to two repeats.

The authors define a vector $C = [c_1, c_2, c_i, , c_{|v|}]$ to indicate the community labels of all nodes. The *i-th* component c_i of C means that the *i-th* node belongs to c_i-*th* community. Communities can be detected based on modularity optimization[81] for community labels in C. Modularity reflects the concentration of edges within communities compared with random distribution of links between all notes regardless of communities. Particularly, the modularity $Q(C)$ with respect to a node label vector C is defined as follows[82]:

$$Q(C) = \frac{1}{2m} \sum_{i,j=1}^{|V|} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j) \tag{1}$$

Where k_i and k_j are the degrees of the *i-th* and *j-th* nodes with their corresponding community labels c_i and c_j in the network, respectively, and A_{ij} is the element in the adjacency matrix of the network. The delta function is defined as $(c_i, c_j) = 1$, if $c_i = c_j$, otherwise $(c_i, c_j) = 0$. $Q(C)$, in the range of [-1/2, 1), actually measures the strength of division of a network into communities. A larger $Q(C)$ indicates a better grouping of communities. To find the best community structures, one can maximize $Q(C)$ by the following constrained optimization problem:

$$\begin{cases} max_C = [c_1, ..., c_{|V|}]Q(C) \\ subject\,to : 1 \leq c_i \leq |V|, i = 1, ..., |V| \end{cases} \tag{2}$$

The Louvain method[82], which has shown promising performance in terms of accuracy and computing time[83], is utilized to solve the modularity optimization model in (Equation (2)) for community identification. The Louvain method initially assigns each network node as a single community. Afterward, the following two steps are iteratively proceeded until a maximum modularity is reached: (i) each node is removed from its own community and placed in the community of its neighbor node such that the modularity gain is maximized, and (ii) nodes in the same community are aggregated to form a super-node and a new smaller scale network is generated.

1.4.5 How to get the barcode Linked reads

Linked-reads provide the long range information missing from standard approaches[84], which builds on the Illumina sequencing technology to provide indexing and barcoding information along with short reads to localize the latter on long DNA fragments (linked-reads), thus benefiting the economies of a high throughput platform. As sequencing reads from 20 to 200kb are barcoded/linked, applications of the technology mainly focused on phasing variant bases in human genomes[85]. Because the length of linked-reads can reach to 20 to 200kb, they can easily span most of repetitive regions in genome. In addition, because the sequencing accuracy of linked-reads is the same as that of Illumina sequencing (the error rate is about 0.2% to 0.5%), they have large size and high sequencing accuracy. These two characteristics of linked-reads make them very suitable for the repetitive sequence identification.

There are some research institutions have published barcode linked genomics datasets. For example, we can download some real barcode linked genomics reads of human (HG003_NA24149_father, HG004_NA24143_mother and HG002_NA24385_son) from website ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/. In order to verify the detection effect of LongRepMarker on real barcode linked datasets, we also tested the performance of it on HG003_NA241 49_father, HG004_NA24143_mother and HG002_NA24385_son these datasets. The specific detection results are shown in section S1.7 of the supplementary materials.

Although there are some public barcode linked datasets that can be used for free. However, from an overall point of view, the available barcode linked data is still very scarce. In order to fully verify the performance of LongRepMarker on the barcode linked datasets, we can also use the following method to simulate the required barcode linked reads. In the following section, we take the drosophila genome as an example to explain the process of simulating barcode linked reads in detail.

1) Simulation tool selection

LRSIM[86] is a linked-reads simulator, which is published on the computational and structural biotechnology journal. The tool is public available at https://github.com/aquaskyline/LRSIM. Details of the command and parameters for calling LRSIM is shown in Table S6.

Table S6. Details of the command and parameters for calling LRSIM.

Usage: ./simulateLinkedReads.pl -r/-g <reference/haplotypes> -p <output prefix> [options]		
Reference genome and variants:		
-d	INT	Haplotypes to simulate [2]
-g	STRING	Haploid FASTAs separated by comma. Overrides -r and -d.
-1	INT	1 SNP per INT base pairs [1000]
-2	INT	Minimum length of Indels [1]
-3	INT	Maximum length of Indels [50]
-4	INT	# of Indels [1000]
-5	INT	Minimum length of Duplications and Inversions [1000]
-6	INT	Maximum length of Duplications and Inversions [10000]
-7	INT	# of Duplications and # of Inversions [100]
-8	INT	Minimum length of Translocations [1000]
-9	INT	Maximum length of Translocations [10000]
-0	INT	# of Translocations [100]
Illumina reads characteristics:		
-e	FLOAT	Per base error rate of the first read [0.0001, 0.0016]
-E	FLOAT	Per base error rate of the second read [0.0001, 0.0016]
-i	INT	Outer distance between the two ends for pairs [350]
-s	INT	Standard deviation of the distance for pairs [35]
Linked reads parameters:		
-b	STRING	Barcodes list
-x	INT	# million reads pairs in total to simulated [600]
-f	INT	Mean molecule length in kbp [100]
-c	STRING	Input a list of fragment sizes. Overrrides -f.
-t	INT	n*1000 partitions to generate [1500]
-m	INT	Average # of molecules per partition [10]
Miscellaneous:		
-u	INT	Continue from a step [auto]
	1	Variant simulation
	2	Build fasta index
	3	DWGSIM
	4	Simulate reads
	5	Sort reads extraction manifest
	6	Extract reads
-z	INT	# of threads to run DWGSIM [8]
-o		Disable parameter checking
-h		Show this help

2) Taking drosophila genome as an example to illustrate the method of simulating barcode linked reads

```
Installing LRSIM:
git clone –recursive https://github.com/aquaskyline/LRSIM.git
cd LRSIM
sh make.sh
cd test
sh test.sh
```

It is worth noting that the libraries that LRSIM relies on can be found in the lib folder under its installation directory(e.g., Parse, Math, Inline and auto). Users only need to copy them to the corresponding directory in the linux system before installation (For example, copy them to /usr/local/lib/ x86_64-linux-gnu/perl/5.22.1) to successfully install the tool. A screenshot of the working directory after LRSIM is successfully installed is shown in Figure S6.

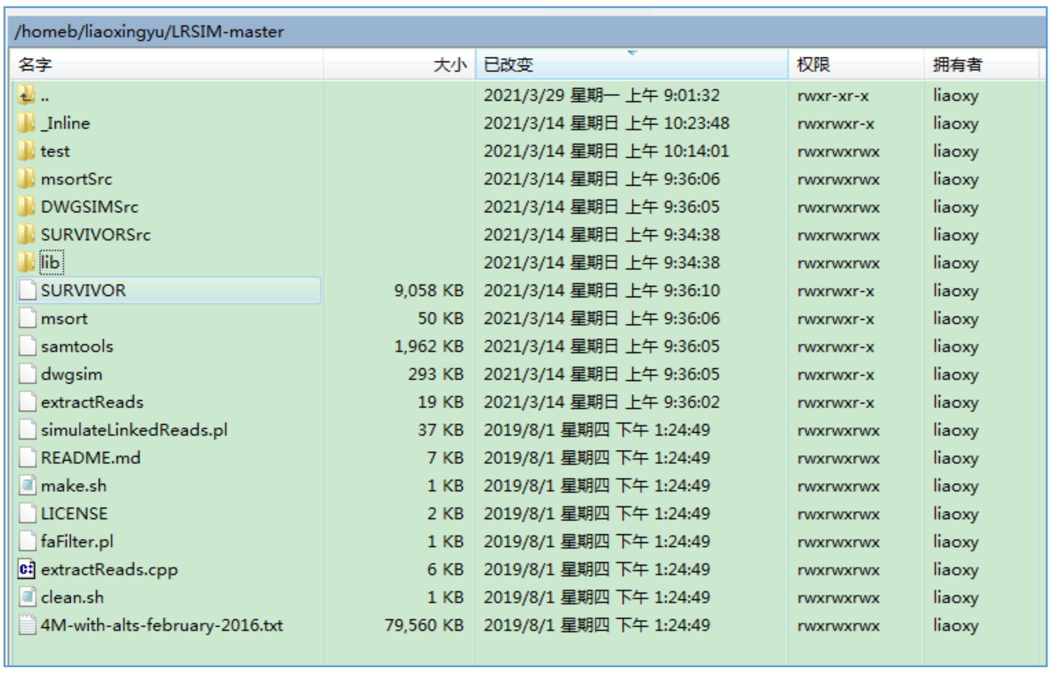
Specific command and parameters:

```
./simulateLinkedReads.pl
-r /home/liaoxingyu/Repeats/reference/dmel-all-chromosome-r5.43.fasta
-p /homeb/liaoxingyu/10X_linked_reads/dmel/10x_dmel
-c /homeb/liaoxingyu/10X_linked_reads/dmel/fragmentSizesList -x 1 -f 50 -t 1 -m 10 -o -4 1 -7 1
```

As can be seen from the above command, the directory for storing simulation data is set to ′/homeb/liaoxin gyu/10X_linked_reads/dmel′. When we run the above command, the simulation data generated by LRSIM will be stored in this directory, just as shown in Figure S7. It is worth noting that the simulation reads generated by LRSIM are still in a paired-end format (e.g., 10x_dmel_S1_L001_R1_001.fastq and 10x_dmel_S1_L001_R2_001.fastq are paired-end reads). Reads in this format can be directly input as data to assembler (such as SPAdes) to participate in the construction of assemblies.

1.4.6 Presentation of detailed reports contained in the detection results of LongRepMarker

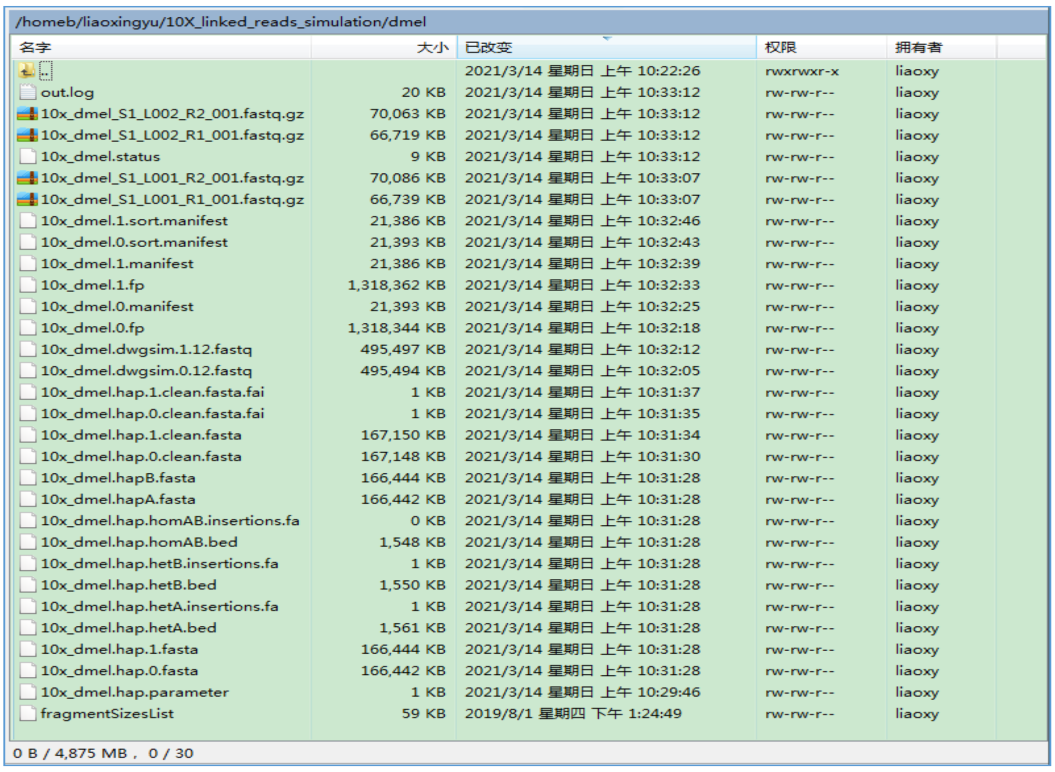
There are some detailed detection reports of repetitive sequences generated by LongRepMarker, just as the shown in Figures S8, S9, S10 and S11(All screenshots shown are based on the drosophila genome). First of all, LongRepMarker generates a repetitive sequence library with annotation information, as shown in Figure S8. In this file, the first line starting with the angle bracket records the fragment ID and the repeat



/homeb/liaoxingyu/LRSIM-master

名字	大小	已改变	权限	拥有者
..		2021/3/29 星期一 上午 9:01:32	rwxr-xr-x	liaoxy
_Inline		2021/3/14 星期日 上午 10:23:48	rwxrwxr-x	liaoxy
test		2021/3/14 星期日 上午 10:14:01	rwxrwxrwx	liaoxy
msortSrc		2021/3/14 星期日 上午 9:36:06	rwxrwxrwx	liaoxy
DWGSIMSrc		2021/3/14 星期日 上午 9:36:05	rwxrwxrwx	liaoxy
SURVIVORSrc		2021/3/14 星期日 上午 9:34:38	rwxrwxrwx	liaoxy
lib		2021/3/14 星期日 上午 9:34:38	rwxrwxrwx	liaoxy
SURVIVOR	9,058 KB	2021/3/14 星期日 上午 9:36:10	rwxrwxr-x	liaoxy
msort	50 KB	2021/3/14 星期日 上午 9:36:06	rwxrwxr-x	liaoxy
samtools	1,962 KB	2021/3/14 星期日 上午 9:36:05	rwxrwxr-x	liaoxy
dwgsim	293 KB	2021/3/14 星期日 上午 9:36:05	rwxrwxr-x	liaoxy
extractReads	19 KB	2021/3/14 星期日 上午 9:36:02	rwxrwxr-x	liaoxy
simulateLinkedReads.pl	37 KB	2019/8/1 星期四 下午 1:24:49	rwxrwxrwx	liaoxy
README.md	7 KB	2019/8/1 星期四 下午 1:24:49	rwxrwxrwx	liaoxy
make.sh	1 KB	2019/8/1 星期四 下午 1:24:49	rwxrwxrwx	liaoxy
LICENSE	2 KB	2019/8/1 星期四 下午 1:24:49	rwxrwxrwx	liaoxy
faFilter.pl	1 KB	2019/8/1 星期四 下午 1:24:49	rwxrwxrwx	liaoxy
extractReads.cpp	6 KB	2019/8/1 星期四 下午 1:24:49	rwxrwxrwx	liaoxy
clean.sh	1 KB	2019/8/1 星期四 下午 1:24:49	rwxrwxrwx	liaoxy
4M-with-alts-february-2016.txt	79,560 KB	2019/8/1 星期四 下午 1:24:49	rwxrwxrwx	liaoxy

Fig. S6. A screenshot of the working directory after LRSIM is successfully installed.



/homeb/liaoxingyu/10X_linked_reads_simulation/dmel

名字	大小	已改变	权限	拥有者
..		2021/3/14 星期日 上午 10:22:26	rwxrwxr-x	liaoxy
out.log	20 KB	2021/3/14 星期日 上午 10:33:12	rw-rw-r--	liaoxy
10x_dmel_S1_L002_R2_001.fastq.gz	70,063 KB	2021/3/14 星期日 上午 10:33:12	rw-rw-r--	liaoxy
10x_dmel_S1_L002_R1_001.fastq.gz	66,719 KB	2021/3/14 星期日 上午 10:33:12	rw-rw-r--	liaoxy
10x_dmel.status	9 KB	2021/3/14 星期日 上午 10:33:12	rw-rw-r--	liaoxy
10x_dmel_S1_L001_R2_001.fastq.gz	70,086 KB	2021/3/14 星期日 上午 10:33:07	rw-rw-r--	liaoxy
10x_dmel_S1_L001_R1_001.fastq.gz	66,739 KB	2021/3/14 星期日 上午 10:33:07	rw-rw-r--	liaoxy
10x_dmel.1.sort.manifest	21,386 KB	2021/3/14 星期日 上午 10:32:46	rw-rw-r--	liaoxy
10x_dmel.0.sort.manifest	21,393 KB	2021/3/14 星期日 上午 10:32:43	rw-rw-r--	liaoxy
10x_dmel.1.manifest	21,386 KB	2021/3/14 星期日 上午 10:32:39	rw-rw-r--	liaoxy
10x_dmel.1.fp	1,318,362 KB	2021/3/14 星期日 上午 10:32:33	rw-rw-r--	liaoxy
10x_dmel.0.manifest	21,393 KB	2021/3/14 星期日 上午 10:32:25	rw-rw-r--	liaoxy
10x_dmel.0.fp	1,318,344 KB	2021/3/14 星期日 上午 10:32:18	rw-rw-r--	liaoxy
10x_dmel.dwgsim.1.12.fastq	495,497 KB	2021/3/14 星期日 上午 10:32:12	rw-rw-r--	liaoxy
10x_dmel.dwgsim.0.12.fastq	495,494 KB	2021/3/14 星期日 上午 10:32:05	rw-rw-r--	liaoxy
10x_dmel.hap.1.clean.fasta.fai	1 KB	2021/3/14 星期日 上午 10:31:37	rw-rw-r--	liaoxy
10x_dmel.hap.0.clean.fasta.fai	1 KB	2021/3/14 星期日 上午 10:31:35	rw-rw-r--	liaoxy
10x_dmel.hap.1.clean.fasta	167,150 KB	2021/3/14 星期日 上午 10:31:34	rw-rw-r--	liaoxy
10x_dmel.hap.0.clean.fasta	167,148 KB	2021/3/14 星期日 上午 10:31:30	rw-rw-r--	liaoxy
10x_dmel.hapB.fasta	166,444 KB	2021/3/14 星期日 上午 10:31:28	rw-rw-r--	liaoxy
10x_dmel.hapA.fasta	166,442 KB	2021/3/14 星期日 上午 10:31:28	rw-rw-r--	liaoxy
10x_dmel.hap.homAB.insertions.fa	0 KB	2021/3/14 星期日 上午 10:31:28	rw-rw-r--	liaoxy
10x_dmel.hap.homAB.bed	1,548 KB	2021/3/14 星期日 上午 10:31:28	rw-rw-r--	liaoxy
10x_dmel.hap.hetB.insertions.fa	1 KB	2021/3/14 星期日 上午 10:31:28	rw-rw-r--	liaoxy
10x_dmel.hap.hetB.bed	1,550 KB	2021/3/14 星期日 上午 10:31:28	rw-rw-r--	liaoxy
10x_dmel.hap.hetA.insertions.fa	1 KB	2021/3/14 星期日 上午 10:31:28	rw-rw-r--	liaoxy
10x_dmel.hap.hetA.bed	1,561 KB	2021/3/14 星期日 上午 10:31:28	rw-rw-r--	liaoxy
10x_dmel.hap.1.fasta	166,444 KB	2021/3/14 星期日 上午 10:31:28	rw-rw-r--	liaoxy
10x_dmel.hap.0.fasta	166,442 KB	2021/3/14 星期日 上午 10:31:28	rw-rw-r--	liaoxy
10x_dmel.hap.parameter	1 KB	2021/3/14 星期日 上午 10:29:46	rw-rw-r--	liaoxy
fragmentSizesList	59 KB	2019/8/1 星期四 下午 1:24:49	rw-rw-r--	liaoxy

0 B / 4,875 MB， 0 / 30

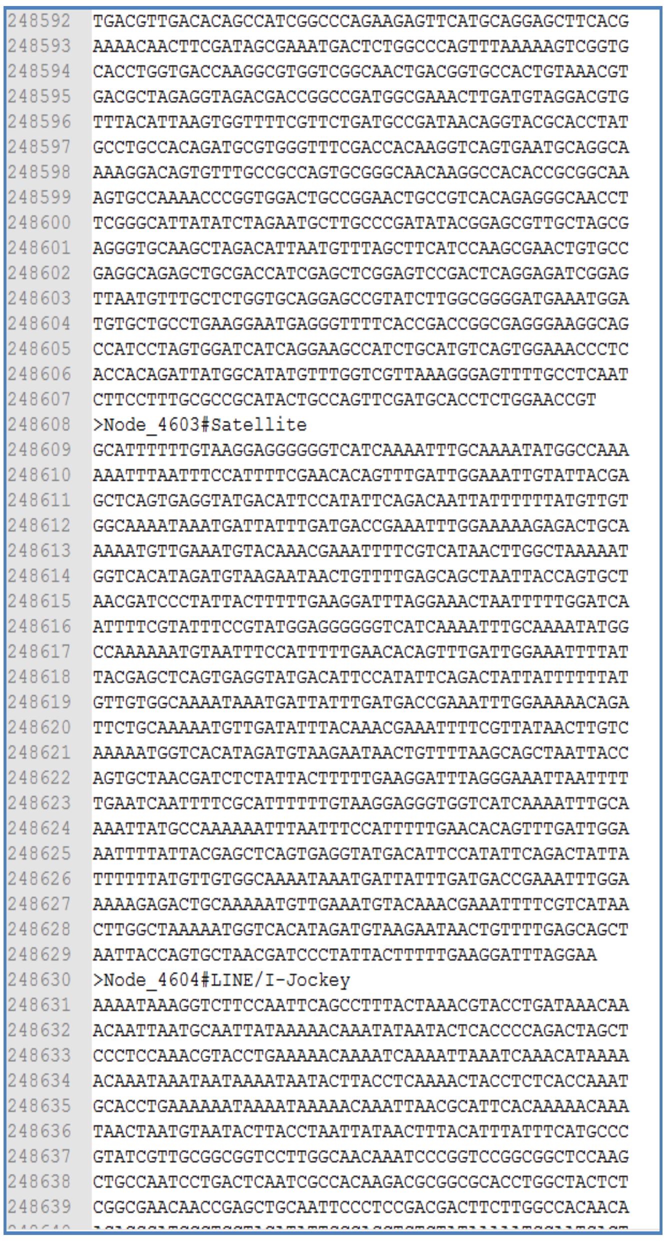
Fig. S7. A screenshot of the working directory after LRSIM is successfully installed.

type of this fragment (e.g. the repeat type of the 4603-th fragment is satellite DNA). The second line is composed of A-T-G-C bases, which records the specific repetitive sequence.

Secondly, LongRepMarker generates a report that detailed records the distribution of repetitive sequences in the genome, as shown in Figure S9. The report includes the fragment ID, the starting position and ending position of the repetitive region on the fragment, the starting position and ending position of the repetitive region aligned to the reference sequence, the detailed alignment (cigar string), and the identity value of the alignment (The higher the identity, the better the quality of the alignment). The occurrence of the same fragment ID multiple times in the report indicates that there are multiple copies of the fragment in the genome, and the number of occurrences is the number of copies.

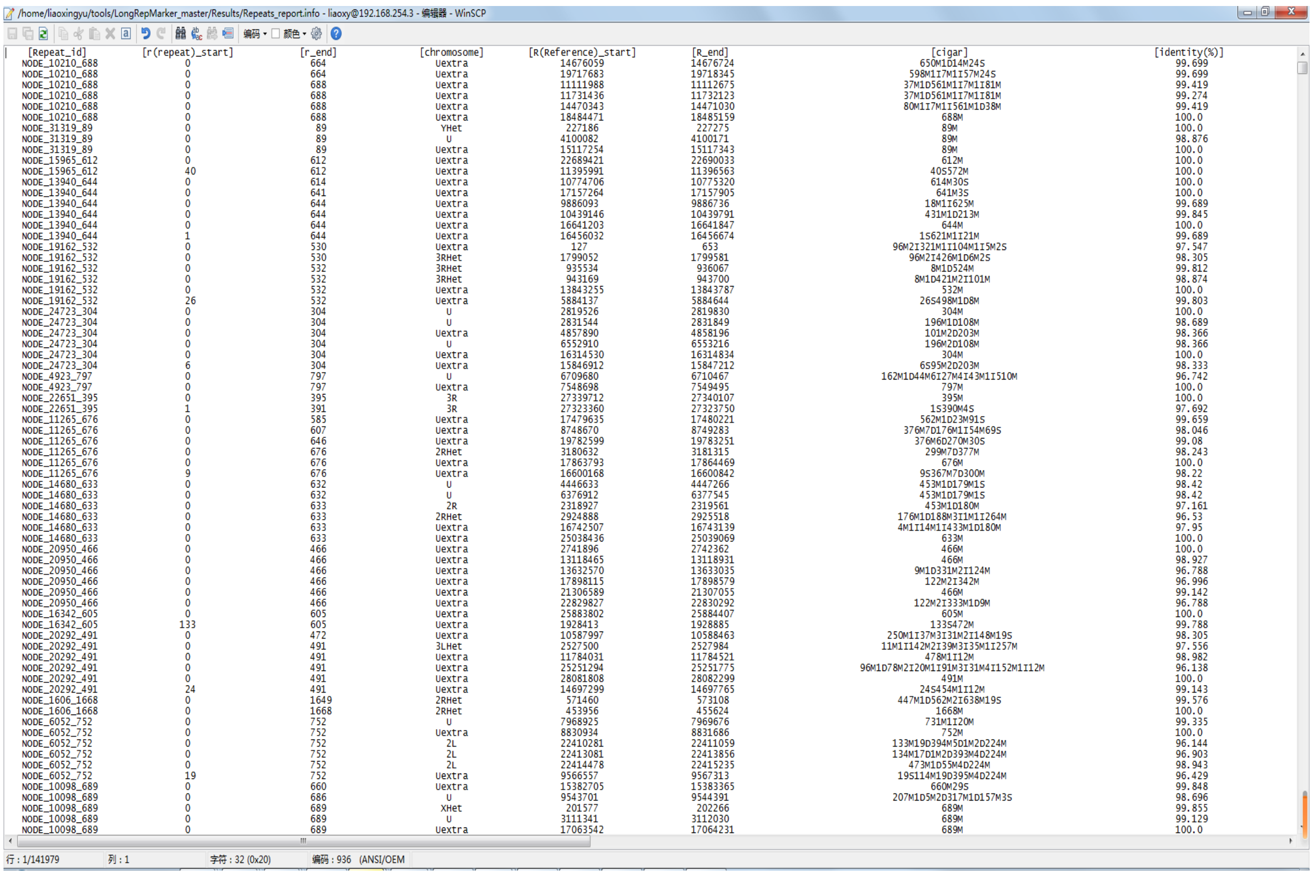
Thirdly, LongRepMarker generates a statistical report which detailed records the number of repeats, the proportion and detailed classification of the repetitive sequences in RepBase library or reference genome that can be covered by each type of repeats generated from LongRepMarker, as shown in Figure S10. This report is obtained by mapping the records in RepBase to the detection results generated by LongRepMarker through the RepeatMasker.

Finally, LongRepMarker generates a VCF format structural variation statistical report in the detection results, as shown in Figure S11. VCF (Variant Calling Format) is a tab-delimited text file that is used to describe single nucleotide variants (SNVs) as well as insertions, deletions, and other sequence variations. This is a bit limiting as it is only tailored to show variations and not genetic features. In addition, INDEL (insertion and deletion) mutations within repetitive sequences are usually easy to be found, but it is very difficult for LongRepMarker to find and determine the inversion and translocation mutations that occur within the repetitive sequences. In order to solve this problem, LongRepMarker calls two professional detection tools, ngmlr (https://github.com/philres/ngmlr) and Sniffles (https://github.com/fritzsedlazeck/Sniffles)[87], to discover and identify the inversion and translocation mutations that occur within the repetitive sequences. The structural variation detection report of LongRepMarker contains the information of SNPs, INDELs, inversions and translocations.



```
TGACGTTGACACAGCCATCGGCCCAGAAGAGTTCATGCAGGAGCTTCACG
AAAACAACTTCGATAGCGAAATGACTCTGGCCCAGTTTAAAAAGTCGGTG
CACCTGGTGACCAAGGCGTGGTCGGCAACTGACGGTGCCACTGTAAACGT
GACGCTAGAGGTAGACGACCGGCCGATGGCGAAACTTGATGTAGGACGTG
TTTACATTAAGTGGTTTTCGTTCTGATGCCGATAACAGGTACGCACCTAT
GCCTGCCACAGATGCGTGGGTTTCGACCACAAGGTCAGTGAATGCAGGCA
AAAGGACAGTGTTTGCCGCCAGTGCGGGCAACAAGGCCACACCGCGGCAA
AGTGCCAAAACCCGGTGGACTGCCGGAACTGCCGTCACAGAGGGCAACCT
TCGGGCATTATATCTAGAATGCTTGCCCGATATACGGAGCGTTGCTAGCG
AGGGTGCAAGCTAGACATTAATGTTTAGCTTCATCCAAGCGAACTGTGCC
GAGGCAGAGCTGCGACCATCGAGCTCGGAGTCCGACTCAGGAGATCGGAG
TTAATGTTTGCTCTGGTGCAGGAGCCGTATCTTGGCGGGGATGAAATGGA
TGTGCTGCCTGAAGGAATGAGGGTTTTCACCGACCGGCGAGGGAAGGCAG
CCATCCTAGTGGATCATCAGGAAGCCATCTGCATGTCAGTGGAAACCCTC
ACCACAGATTATGGCATATGTTTGGTCGTTAAAGGGAGTTTTGCCTCAAT
CTTCCTTTGCGCCGCATACTGCCAGTTCGATGCACCTCTGGAACCGT
>Node_4603#Satellite
GCATTTTTTGTAAGGAGGGGGGTCATCAAAATTTGCAAAATATGGCCAAA
AAATTTAATTTCCATTTTCGAACACAGTTTGATTGGAAATTGTATTACGA
GCTCAGTGAGGTATGACATTCCATATTCAGACAATTATTTTTTATGTTGT
GGCAAAATAAATGATTATTTGATGACCGAAATTTGGAAAAAGAGACTGCA
AAAATGTTGAAATGTACAAACGAAATTTTCGTCATAACTTGGCTAAAAAT
GGTCACATAGATGTAAGAATAACTGTTTTGAGCAGCTAATTACCAGTGCT
AACGATCCCTATTACTTTTTGAAGGATTTAGGAAACTAATTTTTGGATCA
ATTTTCGTATTTCCGTATGGAGGGGGGTCATCAAAATTTGCAAAATATGG
CCAAAAAATGTAATTTCCATTTTGAACACAGTTTGATTGGAAATTTTTAT
TACGAGCTCAGTGAGGTATGACATTCCATATTCAGACTATTATTTTTTAT
GTTGTGGCAAAATAAATGATTATTTGATGACCGAAATTTGGAAAAACAGA
TTCTGCAAAAATGTTGATATTTACAAACGAAATTTTCGTTATAACTTGTC
AAAAATGGTCACATAGATGTAAGAATAACTGTTTTAAGCAGCTAATTACC
AGTGCTAACGATCTCTATTACTTTTTGAAGGATTTAGGGAAATTAATTTT
TGAATCAATTTTCGCATTTTTTGTAAGGAGGGTGGTCATCAAAATTTGCA
AAATTATGCCAAAAAATTTAATTTCCATTTTTGAACACAGTTTGATTGGA
AATTTTATTACGAGCTCAGTGAGGTATGACATTCCATATTCAGACTATTA
TTTTTTATGTTGTGGCAAAATAAATGATTATTTGATGACCGAAATTTGGA
AAAAGAGACTGCAAAAATGTTGAAATGTACAAACGAAATTTTCGTCATAA
CTTGGCTAAAAATGGTCACATAGATGTAAGAATAACTGTTTTGAGCAGCT
AATTACCAGTGCTAACGATCCCTATTACTTTTTGAAGGATTTAGGAA
>Node_4604#LINE/I-Jockey
AAAATAAAGGTCTTCCAATTCAGCCTTTACTAAACGTACCTGATAAACAA
ACAATTAATGCAATTATAAAAACAAATATAATACTCACCCCAGACTAGCT
CCCTCCAAACGTACCTGAAAAACAAAATCAAAATTAAATCAAACATAAAA
ACAAATAAATAATAAAATAATACTTACCTCAAAACTACCTCTCACCAAAT
GCACCTGAAAAAATAAAATAAAAACAAATTAACGCATTCACAAAACAAA
TAACTAATGTAATACTTACCTAATTATAACTTTACATTTATTTCATGCCC
GTATCGTTGCGGCGGTCCTTGGCAACAAATCCCGGTCCGGCGGCTCCAAG
CTGCCAATCCTGACTCAATCGCCACAAGACGCGGCGCACCTGGCTACTCT
CGGCGAACAACCGAGCTGCAATTCCCTCCGACGACTTCTTGGCCACAACA
```

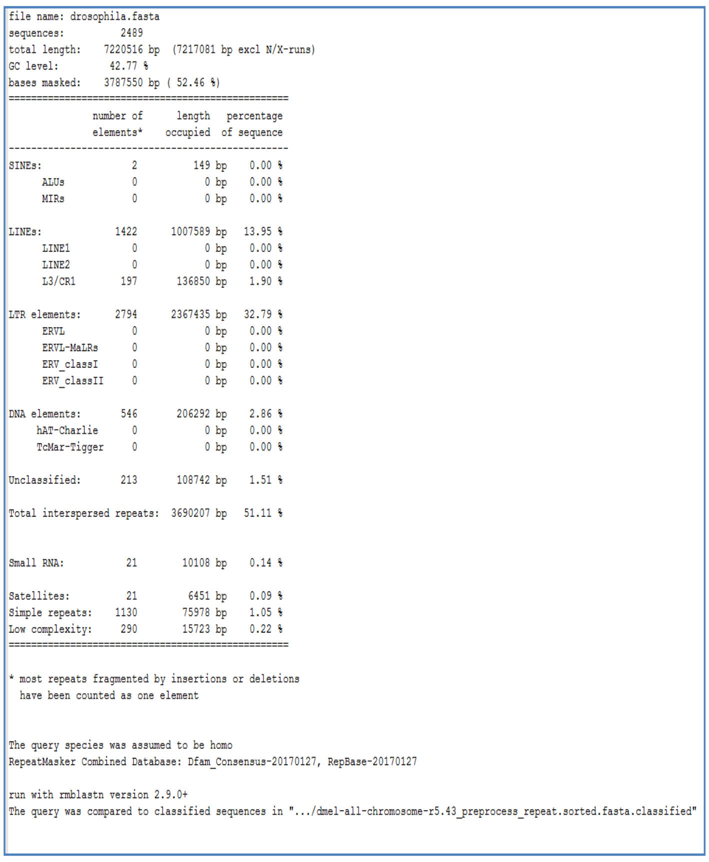
Fig. S8. A screenshot of the final repeat library with annotation information.



/home/liaoxingyu/tools/LongRepMarker_master/Results/Repeats_report.info - liaoxy@192.168.254.3 - 编辑器 - WinSCP
编码 ▾
□ 颜色 ▾
行 : 1/141979
列 : 1
字符 : 32 (0x20)
编码 : 936 (ANSI/OEM

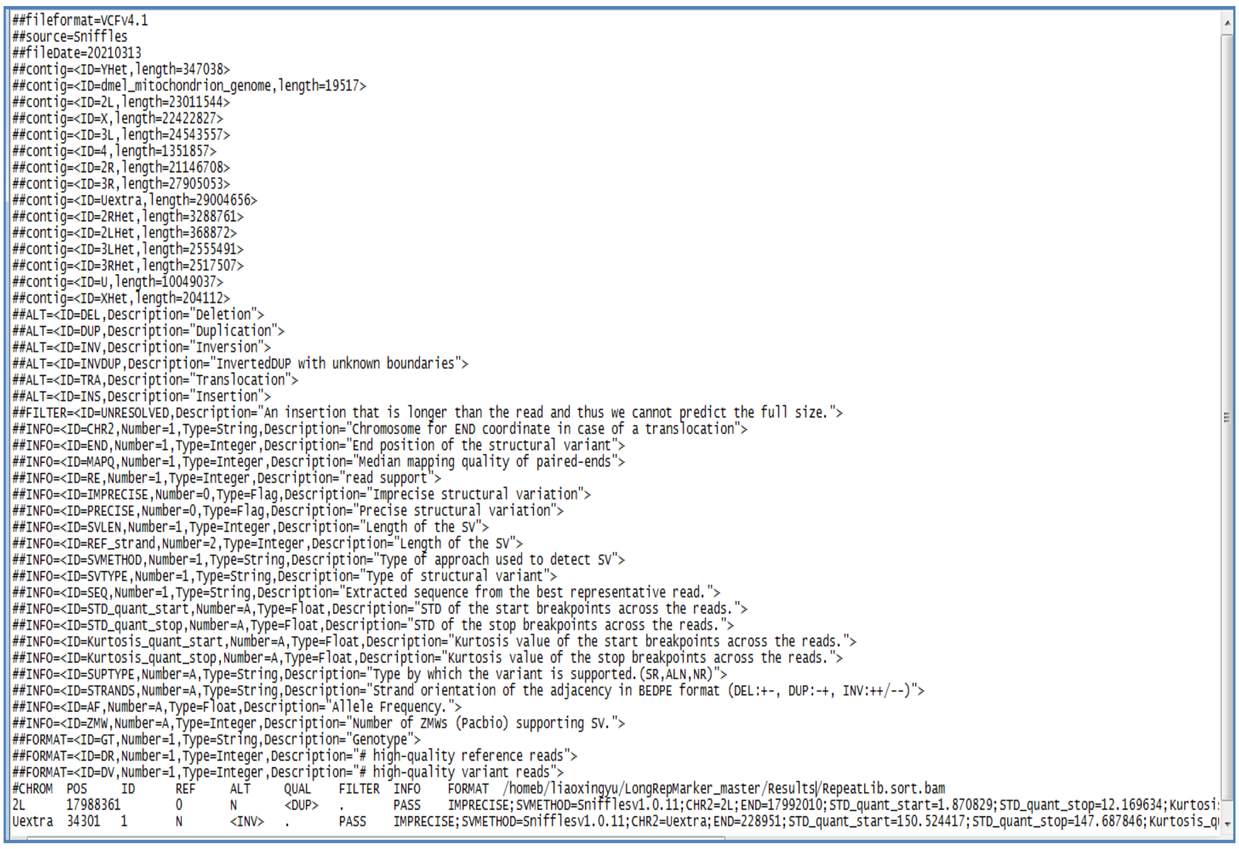
```
  [Repeat_id]      [r(repeat)_start]   [r_end]   [chromosome]   [R(Reference)_start]   [R_end]            [cigar]                          [identity(%)]
 NODE_10210_688     0                  664       Uextra         14676059               14676724           650M1D14M24S                     99.699
 NODE_10210_688     0                  664       Uextra         19717683               19718345           598M1I7M1I57M24S                 99.699
 NODE_10210_688     0                  688       Uextra         11111988               11112675           37M1D561M1I7M1I81M               99.419
 NODE_10210_688     0                  688       Uextra         11731436               11732123           37M1D561M1I7M1I81M               99.274
 NODE_10210_688     0                  688       Uextra         14470343               14471030           80M1I7M1I561M1D38M               99.419
 NODE_10210_688     0                  688       Uextra         18484471               18485159           688M                             100.0
 NODE_31319_89      0                  89        YHet           227186                 227275             89M                              100.0
 NODE_31319_89      0                  89        U              4100082                4100171            89M                              98.876
 NODE_31319_89      0                  89        Uextra         15117254               15117343           89M                              100.0
 NODE_15965_612     0                  612       Uextra         22689421               22690033           612M                             100.0
 NODE_15965_612     40                 612       Uextra         11395991               11396563           40S572M                          100.0
 NODE_13940_644     0                  614       Uextra         10774706               10775320           614M30S                          100.0
 NODE_13940_644     0                  641       Uextra         17157264               17157905           641M3S                           100.0
 NODE_13940_644     0                  644       Uextra         9886093                9886736            18M1I625M                        99.689
 NODE_13940_644     0                  644       Uextra         10439146               10439791           431M1D213M                       99.845
 NODE_13940_644     0                  644       Uextra         16641203               16641847           644M                             100.0
 NODE_13940_644     1                  644       Uextra         16456032               16456674           1S621M1I21M                      99.689
 NODE_19162_532     0                  530       Uextra         127                    653                96M2I321M1I104M1I5M2S            97.547
 NODE_19162_532     0                  530       3RHet          1799052                1799581            96M2I426M1D6M2S                  98.305
 NODE_19162_532     0                  532       3RHet          935534                 936067             8M1D524M                         99.812
 NODE_19162_532     0                  532       3RHet          943169                 943700             8M1D421M2I101M                   98.874
 NODE_19162_532     0                  532       Uextra         13843255               13843787           532M                             100.0
 NODE_19162_532     26                 532       Uextra         5884137                5884644            26S498M1D8M                      99.803
 NODE_24723_304     0                  304       U              2819526                2819830            304M                             100.0
 NODE_24723_304     0                  304       U              2831544                2831849            196M1D108M                       98.689
 NODE_24723_304     0                  304       Uextra         4857890                4858196            101M2D203M                       98.366
 NODE_24723_304     0                  304       U              6552910                6553216            196M2D108M                       98.366
 NODE_24723_304     0                  304       Uextra         16314530               16314834           304M                             100.0
 NODE_24723_304     6                  304       Uextra         15846912               15847212           6S95M2D203M                      98.333
 NODE_4923_797      0                  797       U              6709680                6710467            162M1D44M6I27M4I43M1I510M        96.742
 NODE_4923_797      0                  797       Uextra         7548698                7549495            797M                             100.0
 NODE_22651_395     0                  395       3R             27339712               27340107           395M                             100.0
 NODE_22651_395     1                  391       3R             27323360               27323750           1S390M4S                         97.692
 NODE_11265_676     0                  585       Uextra         17479635               17480221           562M1D23M91S                     99.659
 NODE_11265_676     0                  607       Uextra         8748670                8749283            376M7D176M1I54M69S               98.046
 NODE_11265_676     0                  646       Uextra         19782599               19783251           376M6D270M30S                    99.08
 NODE_11265_676     0                  676       2RHet          3180632                3181315            299M7D377M                       98.243
 NODE_11265_676     0                  676       Uextra         17863793               17864469           676M                             100.0
 NODE_11265_676     9                  676       Uextra         16600168               16600842           9S367M7D300M                     98.22
 NODE_14680_633     0                  632       U              4446633                4447266            453M1D179M1S                     98.42
 NODE_14680_633     0                  632       U              6376912                6377545            453M1D179M1S                     98.42
 NODE_14680_633     0                  633       2R             2318927                2319561            453M1D180M                       97.161
 NODE_14680_633     0                  633       2RHet          2924888                2925518            176M1D188M3I1M1I264M             96.53
 NODE_14680_633     0                  633       Uextra         16742507               16743139           4M1I14M1I433M1D180M              97.95
 NODE_14680_633     0                  633       Uextra         25038436               25039069           633M                             100.0
 NODE_20950_466     0                  466       Uextra         2741896                2742362            466M                             100.0
 NODE_20950_466     0                  466       Uextra         13118465               13118931           466M                             98.927
 NODE_20950_466     0                  466       Uextra         13632570               13633035           9M1D331M2I124M                   96.788
 NODE_20950_466     0                  466       Uextra         17898115               17898579           122M2I342M                       96.996
 NODE_20950_466     0                  466       Uextra         21306589               21307055           466M                             99.142
 NODE_20950_466     0                  466       Uextra         22829827               22830292           122M2I333M1D9M                   96.788
 NODE_16342_605     0                  605       Uextra         25883802               25884407           605M                             100.0
 NODE_16342_605     133                605       Uextra         1928413                1928885            133S472M                         99.788
 NODE_20292_491     0                  472       Uextra         10587997               10588463           250M1I37M3I31M2I148M19S          98.305
 NODE_20292_491     0                  491       3LHet          2527500                2527984            11M1I142M2I39M3I35M1I257M        97.556
 NODE_20292_491     0                  491       Uextra         11784031               11784521           478M1I12M                        98.982
 NODE_20292_491     0                  491       Uextra         25251294               25251775           96M1D78M2I20M1I91M3I31M4I152M1I12M 96.138
 NODE_20292_491     0                  491       Uextra         28081808               28082299           491M                             100.0
 NODE_20292_491     24                 491       Uextra         14697299               14697765           24S454M1I12M                     99.143
 NODE_1606_1668     0                  1649      2RHet          571460                 573108             447M1D562M2I638M19S              99.576
 NODE_1606_1668     0                  1668      2RHet          453956                 455624             1668M                            100.0
 NODE_6052_752      0                  752       U              7968925                7969676            731M1I20M                        99.335
 NODE_6052_752      0                  752       Uextra         8830934                8831686            752M                             100.0
 NODE_6052_752      0                  752       2L             22410281               22411059           133M19D394M5D1M2D224M            96.144
 NODE_6052_752      0                  752       2L             22413081               22413856           134M17D1M2D393M4D224M            96.903
 NODE_6052_752      0                  752       2L             22414478               22415235           473M1D55M4D224M                  98.943
 NODE_6052_752      19                 752       Uextra         9566557                9567313            19S114M19D395M4D224M             96.429
 NODE_10098_689     0                  660       Uextra         15382705               15383365           660M29S                          99.848
 NODE_10098_689     0                  686       U              9543701                9544391            207M1D5M2D317M1D157M3S           98.696
 NODE_10098_689     0                  689       XHet           201577                 202266             689M                             99.855
 NODE_10098_689     0                  689       U              3111341                3112030            689M                             99.129
 NODE_10098_689     0                  689       Uextra         17063542               17064231           689M                             100.0
```

Fig. S9. A screenshot of the report that detailed records the distribution of repetitive sequences in the genome.



```
file name: drosophila.fasta
sequences:         2489
total length:    7220516 bp  (7217081 bp excl N/X-runs)
GC level:         42.77 %
bases masked:    3787550 bp ( 52.46 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:                2          149 bp    0.00 %
      ALUs            0            0 bp    0.00 %
      MIRs            0            0 bp    0.00 %

LINEs:             1422      1007589 bp   13.95 %
      LINE1           0            0 bp    0.00 %
      LINE2           0            0 bp    0.00 %
      L3/CR1        197       136850 bp    1.90 %

LTR elements:      2794      2367435 bp   32.79 %
      ERVL            0            0 bp    0.00 %
      ERVL-MaLRs      0            0 bp    0.00 %
      ERV_classI      0            0 bp    0.00 %
      ERV_classII     0            0 bp    0.00 %

DNA elements:       546       206292 bp    2.86 %
     hAT-Charlie      0            0 bp    0.00 %
     TcMar-Tigger     0            0 bp    0.00 %

Unclassified:       213       108742 bp    1.51 %

Total interspersed repeats:  3690207 bp   51.11 %


Small RNA:           21        10108 bp    0.14 %

Satellites:          21         6451 bp    0.09 %
Simple repeats:    1130        75978 bp    1.05 %
Low complexity:     290        15723 bp    0.22 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element


The query species was assumed to be homo
RepeatMasker Combined Database: Dfam_Consensus-20170127, RepBase-20170127

run with rmblastn version 2.9.0+
The query was compared to classified sequences in ".../dmel-all-chromosome-r5.43_preprocess_repeat.sorted.fasta.classified"
```

Fig. S10. A screenshot of the statistical report which detailed records the number of repeats, the proportion and the covered bases of each type of repeats.



```
##fileformat=VCFv4.1
##source=Sniffles
##fileDate=20210313
##contig=<ID=YHet,length=347038>
##contig=<ID=dmel_mitochondrion_genome,length=19517>
##contig=<ID=2L,length=23011544>
##contig=<ID=X,length=22422827>
##contig=<ID=3L,length=24543557>
##contig=<ID=4,length=1351857>
##contig=<ID=2R,length=21146708>
##contig=<ID=3R,length=27905053>
##contig=<ID=Uextra,length=29004656>
##contig=<ID=2RHet,length=3288761>
##contig=<ID=2LHet,length=368872>
##contig=<ID=3LHet,length=2555491>
##contig=<ID=3RHet,length=2517507>
##contig=<ID=U,length=10049037>
##contig=<ID=XHet,length=204112>
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=INVDUP,Description="InvertedDUP with unknown boundaries">
##ALT=<ID=TRA,Description="Translocation">
##ALT=<ID=INS,Description="Insertion">
##FILTER=<ID=UNRESOLVED,Description="An insertion that is longer than the read and thus we cannot predict the full size.">
##INFO=<ID=CHR2,Number=1,Type=String,Description="Chromosome for END coordinate in case of a translocation">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the structural variant">
##INFO=<ID=MAPQ,Number=1,Type=Integer,Description="Median mapping quality of paired-ends">
##INFO=<ID=RE,Number=1,Type=Integer,Description="read support">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise structural variation">
##INFO=<ID=PRECISE,Number=0,Type=Flag,Description="Precise structural variation">
##INFO=<ID=SVLEN,Number=1,Type=Integer,Description="Length of the SV">
##INFO=<ID=REF_strand,Number=2,Type=Integer,Description="Length of the SV">
##INFO=<ID=SVMETHOD,Number=1,Type=String,Description="Type of approach used to detect SV">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=SEQ,Number=1,Type=String,Description="Extracted sequence from the best representative read.">
##INFO=<ID=STD_quant_start,Number=A,Type=Float,Description="STD of the start breakpoints across the reads.">
##INFO=<ID=STD_quant_stop,Number=A,Type=Float,Description="STD of the stop breakpoints across the reads.">
##INFO=<ID=Kurtosis_quant_start,Number=A,Type=Float,Description="Kurtosis value of the start breakpoints across the reads.">
##INFO=<ID=Kurtosis_quant_stop,Number=A,Type=Float,Description="Kurtosis value of the stop breakpoints across the reads.">
##INFO=<ID=SUPTYPE,Number=A,Type=String,Description="Type by which the variant is supported.(SR,ALN,NR)">
##INFO=<ID=STRANDS,Number=A,Type=String,Description="Strand orientation of the adjacency in BEDPE format (DEL:+-, DUP:-+, INV:++/--)">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency.">
##INFO=<ID=ZMW,Number=A,Type=Integer,Description="Number of ZMWs (Pacbio) supporting SV.">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DR,Number=1,Type=Integer,Description="# high-quality reference reads">
##FORMAT=<ID=DV,Number=1,Type=Integer,Description="# high-quality variant reads">
#CHROM  POS      ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  /homeb/liaoxingyu/LongRepMarker_master/Results/RepeatLib.sort.bam
2L      17988361        0       N       <DUP>   .       PASS    IMPRECISE;SVMETHOD=Snifflesv1.0.11;CHR2=2L;END=17992010;STD_quant_start=1.870829;STD_quant_stop=12.169634;Kurtosi:
Uextra  34301   1       N       <INV>   .       PASS    IMPRECISE;SVMETHOD=Snifflesv1.0.11;CHR2=Uextra;END=228951;STD_quant_start=150.524417;STD_quant_stop=147.687846;Kurtosis_q
```

Fig. S11. A screenshot of the structural variation statistical report with VCF format.

1.4.7 How to choose the *k-mer* size

The size of k-mer has a certain impact on the processing efficiency of LongRepMarker, because the smaller the size of k, the easier it is for *k-mers* to aggregate into unique *k-mer* (*k-mers* to their canonical representation with respect to reverse-complementation which called the unique *k-mer*), which makes the final unique k-mer set smaller, thus reducing the time and computational overhead of the subsequent alignment process. Theoretically, the influence of *k-mer* size on the accuracy of the test results is not significant, because the LongRepMarker is to find candidate repetitive sequences by looking for multiple alignment unique *k-mer* and their coverage regions on the reference genome or assembly results. Theoretically, the size of *k-mer* does not affect the acquisition of multiple alignment unique *k-mer* and their coverage regions on the reference genome or assemblies. However, in fact, due to the existence of sequencing error, the size of *k-mer* will have a certain impact on the accuracy of detection results, which is mainly manifested in the small size of k (such as less than 11bp). The main reason for this effect is that when the size of *k-mer* is short, it is easy to cause coupling alignment under the combined effect of sequencing error and alignment fault tolerance strategy, which leads to the ordinary unique *k-mer* which is not in the range of multiple alignment unique *k-mer* to be screened into the process of detection, resulting in the final detection results containing a large number of non repetitive elements. In order to solve this problem, we need to limit the minimum value of k. In practical application, the following formula is usually used to limit the size of k.

$$k \geq \lceil \log_4 G + 1 \rceil \tag{3}$$

Where k represents the *k-mer* size, G represents the genome size or the total length of assemblies. This formula refers to the readme document of RepeatScout tool (https://github.com/mmcco/RepeatScout/blob/master/README). The larger the size of *k-mer*, the greater the number of unique *k-mer* generated. Further more, the larger the size of *k-mer* is, the more disk space and memory are needed. At the same time, the time-consuming of subsequent alignment process will also increase sharply. Due to the limited fault tolerance of the alignment process, the longer the segment is, the lower the probability that it can be aligned to the different locations of the genome. As the size of *k-mer* increases, the accuracy of the detection results is promoted to a certain extent, but it will also cause the excessive consumption of computing resources and affect the integrity of the detection results. Therefore, the size of *k-mer* is not the bigger the better, but it is very difficult to control in practical application, so we can only give an upper limit range of k ($k \in$ [21,59], k is odd) which is generated based on the actual experience of testing. For example, all experimental results in this study are obtained under the condition of $k = 49$.

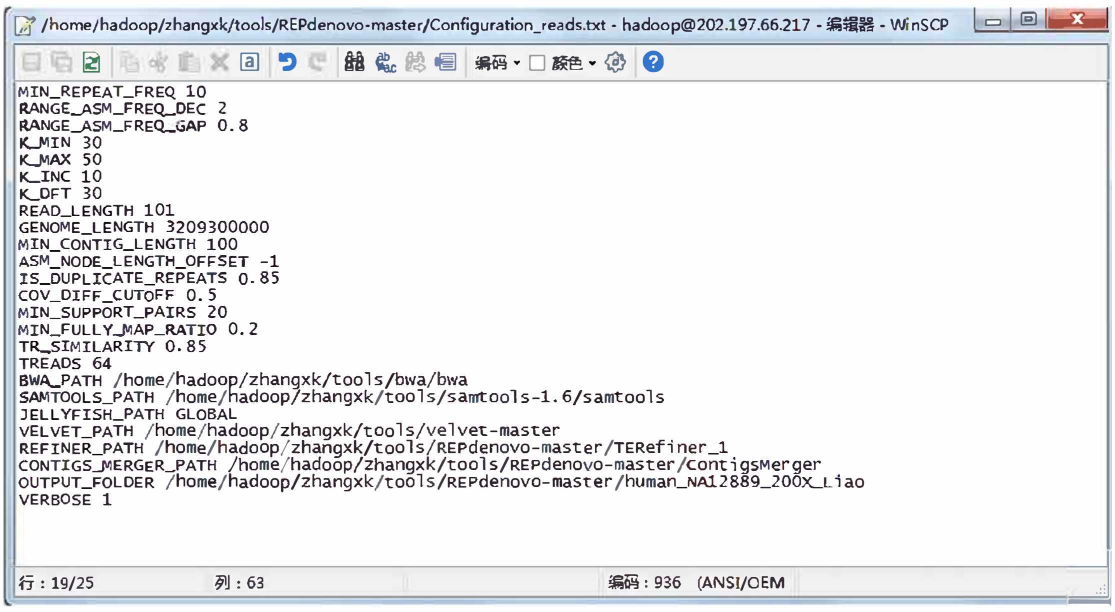
1.4.8 Reproduction of REPdenovo's detection results on the NA12889 dataset

We have redesigned all comparative experiments in this study. For example, a new detection mode based only on NGS short reads is added in the framework of LongRepMarker, and only the detection results in this mode are compared with RepARK and REPdenovo. In addition, in this round of revision, we re-tested the results of REPdenovo (Supplementary Section S3.5.3). The commands and parameters used in this round of testing are as follows:

```
python ./main.py -c Assembly -g configuration-file-name -r raw-reads-file-name

python ./main.py -c Scaffolding -g configuration-file-name -r raw-reads-file-name
```

Among the two commands, the first one is to splice the high-frequency k-mers into contigs by assembly strategy, while the second one is mainly to scaffolding the contigs generated in the first step, so as to generate longer detection fragments. Since the datasets used in this study is different from the datasets mentioned in the corresponding article of the REPdenovo tool, the corresponding test results will also be different. However, from the overall point of view, although there are some large fragments in REPdenovo's detection results, the total number of fragments is small and the detection results are sparse. This phenomenon can be seen from figures S19-S22 in the supplementary materials.

In order to prove that the commands and parameters were not missed in the process of calling REPdenovo in this study, we reproduced the results of REPdenovo based on the NA12889 dataset. The detailed processing steps and results are shown in Figs.S12-S17, and Tables S7-S8, respectively. Among them, Fig. S12 shows a screenshot of the configuration of the tools that REPdenovo depends. Fig. S13 shows a screenshot of the configuration of the paired-end reads. Fig. S14 shows a screenshot of the running commands and parameters. Fig. S15 shows a screenshot of the detection files generated by REPdenovo. Fig. S16 shows a screenshot of the value of the average coverage of reads, and Fig. S17 shows a screenshot of the final repeat library generated by REPdenovo.

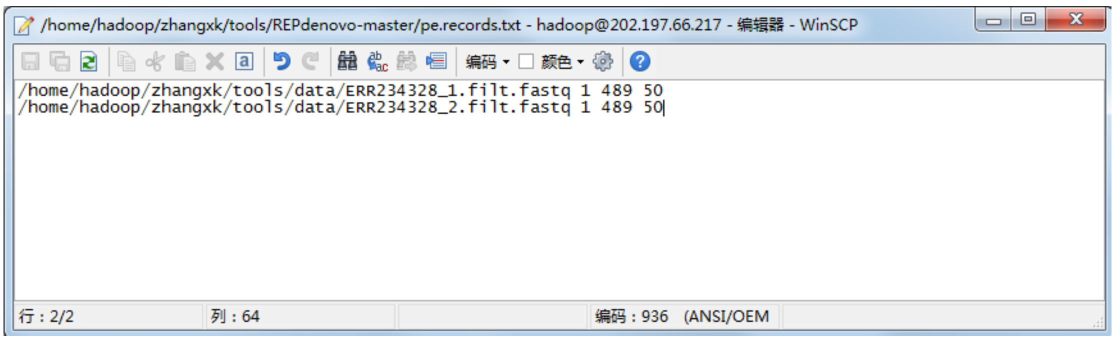


/home/hadoop/zhangxk/tools/REPdenovo-master/Configuration_reads.txt - hadoop@202.197.66.217 - 编辑器 - WinSCP

```
MIN_REPEAT_FREQ 10
RANGE_ASM_FREQ_DEC 2
RANGE_ASM_FREQ_GAP 0.8
K_MIN 30
K_MAX 50
K_INC 10
K_DFT 30
READ_LENGTH 101
GENOME_LENGTH 3209300000
MIN_CONTIG_LENGTH 100
ASM_NODE_LENGTH_OFFSET -1
IS_DUPLICATE_REPEATS 0.85
COV_DIFF_CUTOFF 0.5
MIN_SUPPORT_PAIRS 20
MIN_FULLY_MAP_RATIO 0.2
TR_SIMILARITY 0.85
TREADS 64
BWA_PATH /home/hadoop/zhangxk/tools/bwa/bwa
SAMTOOLS_PATH /home/hadoop/zhangxk/tools/samtools-1.6/samtools
JELLYFISH_PATH GLOBAL
VELVET_PATH /home/hadoop/zhangxk/tools/velvet-master
REFINER_PATH /home/hadoop/zhangxk/tools/REPdenovo-master/TERefiner_1
CONTIGS_MERGER_PATH /home/hadoop/zhangxk/tools/REPdenovo-master/ContigsMerger
OUTPUT_FOLDER /home/hadoop/zhangxk/tools/REPdenovo-master/human_NA12889_200X_Liao
VERBOSE 1
```

行：19/25 列：63 编码：936 (ANSI/OEM

Fig. S12. A screenshot of the configuration of the tools that REPdenovo depends.

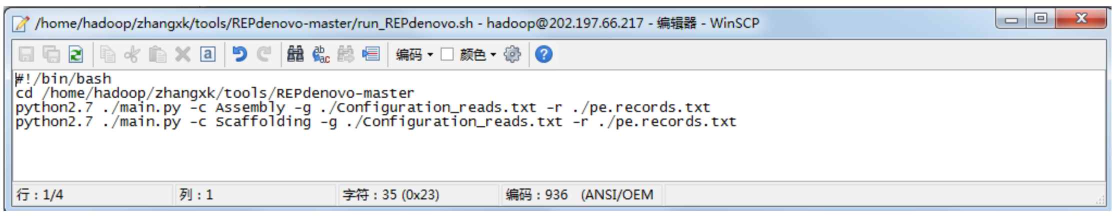


/home/hadoop/zhangxk/tools/REPdenovo-master/pe.records.txt - hadoop@202.197.66.217 - 编辑器 - WinSCP

```
/home/hadoop/zhangxk/tools/data/ERR234328_1.filt.fastq 1 489 50
/home/hadoop/zhangxk/tools/data/ERR234328_2.filt.fastq 1 489 50
```

行：2/2 列：64 编码：936 (ANSI/OEM

Fig. S13. A screenshot of the configuration of the paired-end reads.

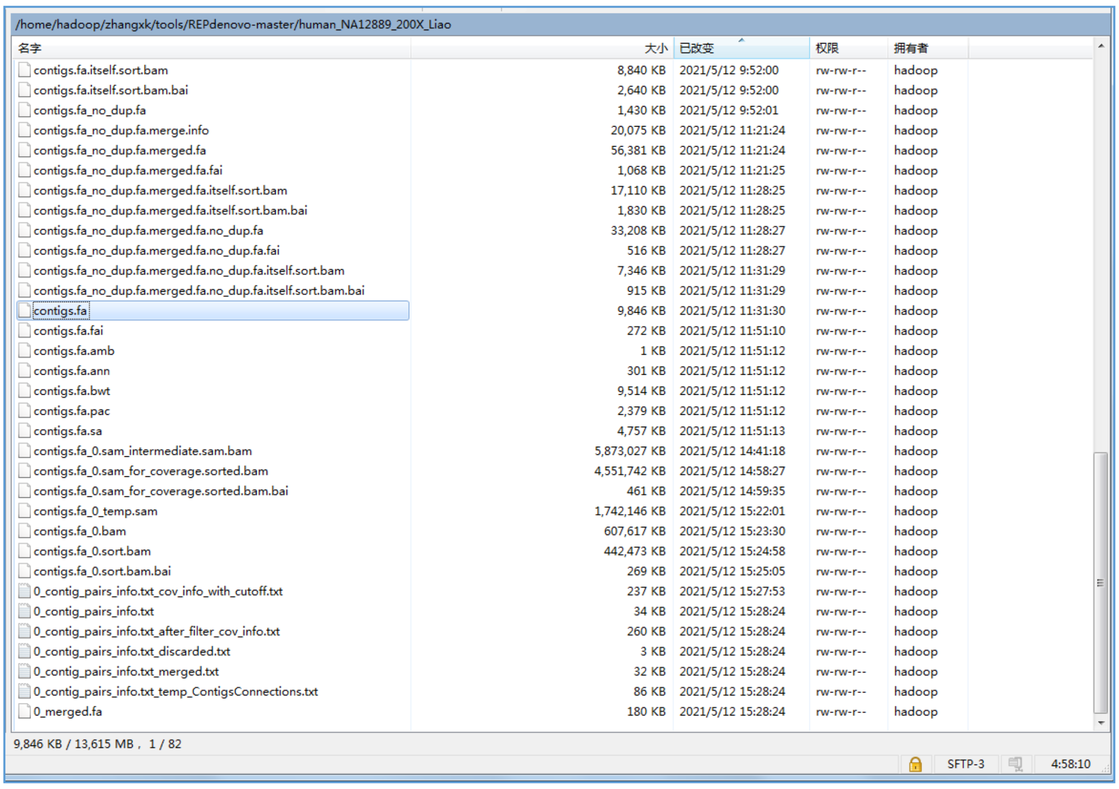


/home/hadoop/zhangxk/tools/REPdenovo-master/run_REPdenovo.sh - hadoop@202.197.66.217 - 编辑器 - WinSCP

```
#!/bin/bash
cd /home/hadoop/zhangxk/tools/REPdenovo-master
python2.7 ./main.py -c Assembly -g ./Configuration_reads.txt -r ./pe.records.txt
python2.7 ./main.py -c Scaffolding -g ./Configuration_reads.txt -r ./pe.records.txt
```

行：1/4 列：1 字符：35 (0x23) 编码：936 (ANSI/OEM

Fig. S14. A screenshot of the running commands and parameters.



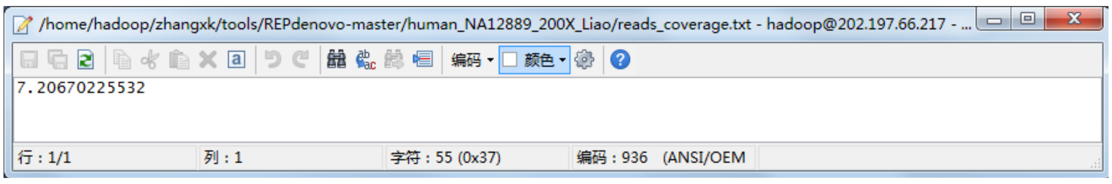
/home/hadoop/zhangxk/tools/REPdenovo-master/human_NA12889_200X_Liao

名字	大小	已改变	权限	拥有者
contigs.fa.itself.sort.bam	8,840 KB	2021/5/12 9:52:00	rw-rw-r--	hadoop
contigs.fa.itself.sort.bam.bai	2,640 KB	2021/5/12 9:52:00	rw-rw-r--	hadoop
contigs.fa_no_dup.fa	1,430 KB	2021/5/12 9:52:01	rw-rw-r--	hadoop
contigs.fa_no_dup.fa.merge.info	20,075 KB	2021/5/12 11:21:24	rw-rw-r--	hadoop
contigs.fa_no_dup.fa.merged.fa	56,381 KB	2021/5/12 11:21:24	rw-rw-r--	hadoop
contigs.fa_no_dup.fa.merged.fa.fai	1,068 KB	2021/5/12 11:21:25	rw-rw-r--	hadoop
contigs.fa_no_dup.fa.merged.fa.itself.sort.bam	17,110 KB	2021/5/12 11:28:25	rw-rw-r--	hadoop
contigs.fa_no_dup.fa.merged.fa.itself.sort.bam.bai	1,830 KB	2021/5/12 11:28:25	rw-rw-r--	hadoop
contigs.fa_no_dup.fa.merged.fa.no_dup.fa	33,208 KB	2021/5/12 11:28:27	rw-rw-r--	hadoop
contigs.fa_no_dup.fa.merged.fa.no_dup.fa.fai	516 KB	2021/5/12 11:28:27	rw-rw-r--	hadoop
contigs.fa_no_dup.fa.merged.fa.no_dup.fa.itself.sort.bam	7,346 KB	2021/5/12 11:31:29	rw-rw-r--	hadoop
contigs.fa_no_dup.fa.merged.fa.no_dup.fa.itself.sort.bam.bai	915 KB	2021/5/12 11:31:29	rw-rw-r--	hadoop
contigs.fa	9,846 KB	2021/5/12 11:31:30	rw-rw-r--	hadoop
contigs.fa.fai	272 KB	2021/5/12 11:51:10	rw-rw-r--	hadoop
contigs.fa.amb	1 KB	2021/5/12 11:51:12	rw-rw-r--	hadoop
contigs.fa.ann	301 KB	2021/5/12 11:51:12	rw-rw-r--	hadoop
contigs.fa.bwt	9,514 KB	2021/5/12 11:51:12	rw-rw-r--	hadoop
contigs.fa.pac	2,379 KB	2021/5/12 11:51:12	rw-rw-r--	hadoop
contigs.fa.sa	4,757 KB	2021/5/12 11:51:13	rw-rw-r--	hadoop
contigs.fa_0.sam_intermediate.sam.bam	5,873,027 KB	2021/5/12 14:41:18	rw-rw-r--	hadoop
contigs.fa_0.sam_for_coverage.sorted.bam	4,551,742 KB	2021/5/12 14:58:27	rw-rw-r--	hadoop
contigs.fa_0.sam_for_coverage.sorted.bam.bai	461 KB	2021/5/12 14:59:35	rw-rw-r--	hadoop
contigs.fa_0_temp.sam	1,742,146 KB	2021/5/12 15:22:01	rw-rw-r--	hadoop
contigs.fa_0.bam	607,617 KB	2021/5/12 15:23:30	rw-rw-r--	hadoop
contigs.fa_0.sort.bam	442,473 KB	2021/5/12 15:24:58	rw-rw-r--	hadoop
contigs.fa_0.sort.bam.bai	269 KB	2021/5/12 15:25:05	rw-rw-r--	hadoop
0_contig_pairs_info.txt_cov_info_with_cutoff.txt	237 KB	2021/5/12 15:27:53	rw-rw-r--	hadoop
0_contig_pairs_info.txt	34 KB	2021/5/12 15:28:24	rw-rw-r--	hadoop
0_contig_pairs_info.txt_after_filter_cov_info.txt	260 KB	2021/5/12 15:28:24	rw-rw-r--	hadoop
0_contig_pairs_info.txt_discarded.txt	3 KB	2021/5/12 15:28:24	rw-rw-r--	hadoop
0_contig_pairs_info.txt_merged.txt	32 KB	2021/5/12 15:28:24	rw-rw-r--	hadoop
0_contig_pairs_info.txt_temp_ContigsConnections.txt	86 KB	2021/5/12 15:28:24	rw-rw-r--	hadoop
0_merged.fa	180 KB	2021/5/12 15:28:24	rw-rw-r--	hadoop

9,846 KB / 13,615 MB，1 / 82

SFTP-3 4:58:10

Fig. S15. A screenshot of the detection files generated by REPdenovo.

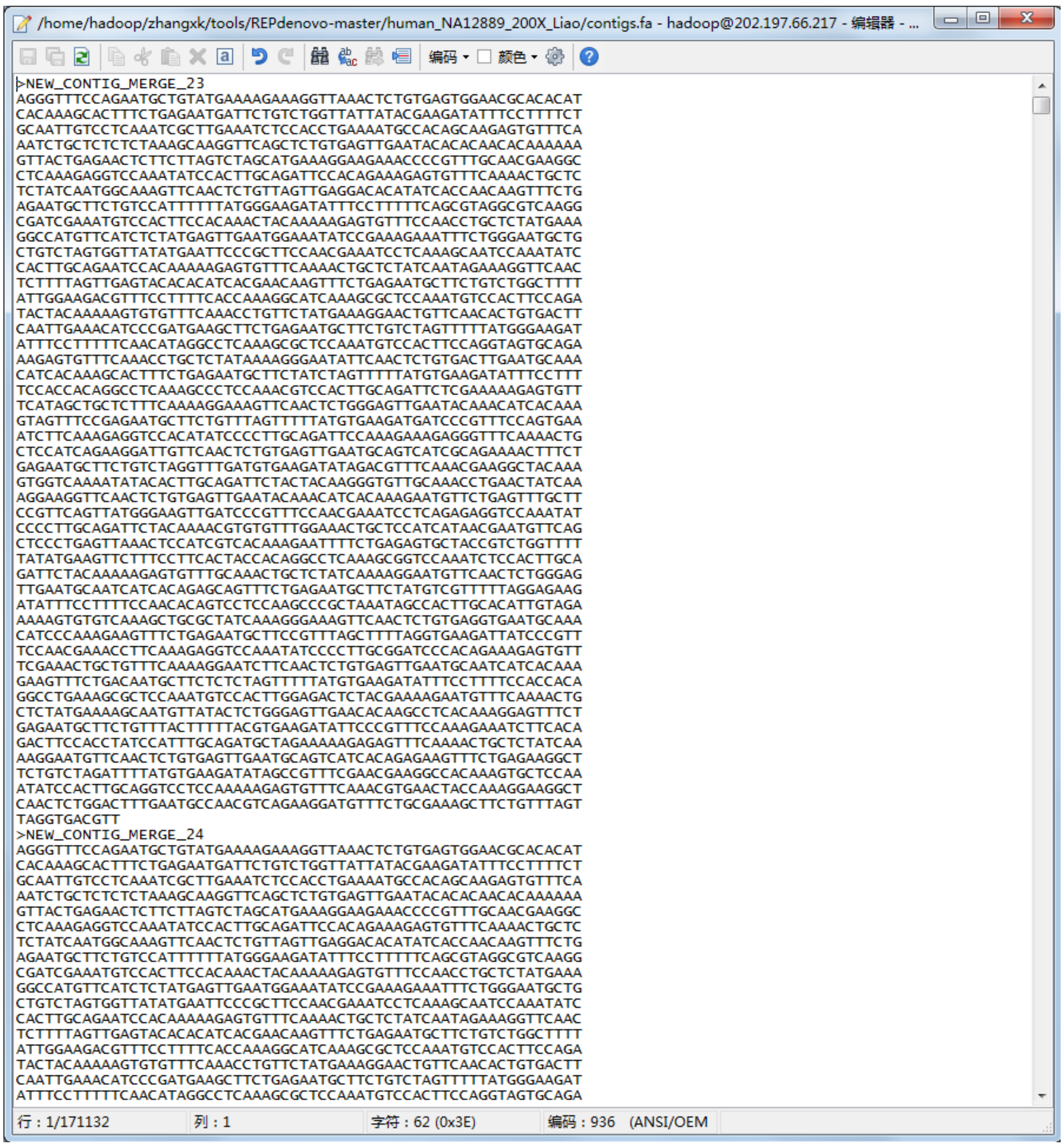


/home/hadoop/zhangxk/tools/REPdenovo-master/human_NA12889_200X_Liao/reads_coverage.txt - hadoop@202.197.66.217 - ...

```
7.20670225532
```

行：1/1 列：1 字符：55 (0x37) 编码：936 (ANSI/OEM

Fig. S16. A screenshot of the value of the average coverage of reads.



/home/hadoop/zhangxk/tools/REPdenovo-master/human_NA12889_200X_Liao/contigs.fa - hadoop@202.197.66.217 - 编辑器 - ...

```
>NEW_CONTIG_MERGE_23
AGGGTTTCCAGAATGCTGTATGAAAAGAAAGGTTAAACTCTGTGAGTGGAACGCACACAT
CACAAAGCACTTTCTGAGAATGATTCTGTCTGGTTATTATACGAAGATATTTCCTTTTCT
GCAATTGTCCTCAAATCGCTTGAAATCTCCACCTGAAAATGCCACAGCAAGAGTGTTTCA
AATCTGCTCTCTCTAAAGCAAGGTTCAGCTCTGTGAGTTGAATACACACAACACAAAAAA
GTTACTGAGAACTCTTCTTAGTCTAGCATGAAAGGAAGAAACCCCGTTTGCAACGAAGGC
CTCAAAGAGGTCCAAATATCCACTTGCAGATTCCACAGAAAGAGTGTTTCAAAACTGCTC
TCTATCAATGGCAAAGTTCAACTCTGTTAGTTGAGGACACATATCACCAACAAGTTTCTG
AGAATGCTTCTGTCCATTTTTTTATGGGAAGATATTTCCTTTTTCAGCGTAGGCGTCAAGG
CGATCGAAATGTCCACTTCCACAAACTACAAAAAGAGTGTTTCCAACCTGCTCTATGAAA
GGCCATGTTCATCTCTATGAGTTGAATGGAAATATCCGAAAGAAATTTCTGGGAATGCTG
CTGTCTAGTGGTTATATGAATTCCCGCTTCCAACGAAATCCTCAAAGCAATCCAAATATC
CACTTGCAGAATCCACAAAAAGAGTGTTTCAAAACTGCTCTATCAATAGAAAGGTTCAAC
TCTTTTAGTTGAGTACACACATCACGAACAAGTTTCTGAGAATGCTTCTGTCTGGCTTTT
ATTGGAAGACGTTTCCTTTTCACCAAAGGCATCAAAGCGCTCCAAATGTCCACTTCCAGA
TACTACAAAAAGTGTGTTTCAAACCTGTTCTATGAAAGGAACTGTTCAACACTGTGACTT
CAATTGAAACATCCCGATGAAGCTTCTGAGAATGCTTCTGTCTAGTTTTTATGGGAAGAT
ATTTCCTTTTTCAACATAGGCCTCAAAGCGCTCCAAATGTCCACTTCCAGGTAGTGCAGA
AAGAGTGTTTCAAACCTGCTCTATAAAAGGGAATATTCAACTCTGTGACTTGAATGCAAA
CATCACAAAGCACTTTCTGAGAATGCTTCTATCTAGTTTTTATGTGAAGATATTTCCTTT
TCCACCACAGGCCTCAAAGCCCTCCAAACGTCCACTTGCAGATTCTCGAAAAAGAGTGTT
TCATAGCTGCTCTTTCAAAAGGAAAGTTCAACTCTGGGAGTTGAATACAAACATCACAAA
GTAGTTTCCGAGAATGCTTCTGTTTAGTTTTTATGTGAAGATGATCCCGTTTCCAGTGAA
ATCTTCAAAGAGGTCCACATATCCCCTTGCAGATTCCAAAGAAAGAGGGTTTCAAAACTG
CTCCATCAGAAGGATTGTTCAACTCTGTGAGTTGAATGCAGTCATCGCAGAAAACTTTCT
GAGAATGCTTCTGTCTAGGTTTGATGTGAAGATATAGACGTTTCAAACGAAGGCTACAAA
GTGGTCAAAATATACACTTGCAGATTCTACTACAAGGGTGTTGCAAACCTGAACTATCAA
AGGAAGGTTCAACTCTGTGAGTTGAATACAAACATCACAAAGAATGTTCTGAGTTTGCTT
CCGTTCAGTTATGGGAAGTTGATCCCGTTTCCAACGAAATCCTCAGAGAGGTCCAAATAT
CCCCTTGCAGATTCTACAAAACGTGTGTTTGGAAACTGCTCCATCATAACGAATGTTCAG
CTCCCTGAGTTAAACTCCATCGTCACAAAGAATTTTCTGAGAGTGCTACCGTCTGGTTTT
TATATGAAGTTCTTTCCTTCACTACCACAGGCCTCAAAGCGGTCCAAATCTCCACTTGCA
GATTCTACAAAAAGAGTGTTTGCAAACTGCTCTATCAAAAGGAATGTTCAACTCTGGGAG
TTGAATGCAATCATCACAGAGCAGTTTCTGAGAATGCTTCTATGTCGTTTTTAGGAGAAG
ATATTTCCTTTTCCAACACAGTCCTCCAAGCCCGCTAAATAGCCACTTGCACATTGTAGA
AAAAGTGTGTCAAAGCTGCGCTATCAAAGGGAAAGTTCAACTCTGTGAGGTGAATGCAAA
CATCCCAAAGAAGTTTCTGAGAATGCTTCCGTTTAGCTTTTAGGTGAAGATTATCCCGTT
TCCAACGAAACCTTCAAAGAGGTCCAAATATCCCCTTGCGGATCCCACAGAAAGAGTGTT
TCGAAACTGCTGTTTCAAAAGGAATCTTCAACTCTGTGAGTTGAATGCAATCATCACAAA
GAAGTTTCTGACAATGCTTCTCTCTAGTTTTTATGTGAAGATATTTCCTTTTCCACCACA
GGCCTGAAAGCGCTCCAAATGTCCACTTGGAGACTCTACGAAAAGAATGTTTCAAAACTG
CTCTATGAAAAGCAATGTTATACTCTGGGAGTTGAACACAAGCCTCACAAAGGAGTTTCT
GAGAATGCTTCTGTTTACTTTTTACGTGAAGATATTCCCGTTTCCAAAGAAATCTTCACA
GACTTCCACCTATCCATTTGCAGATGCTAGAAAAAGAGAGTTTCAAAACTGCTCTATCAA
AAGGAATGTTCAACTCTGTGAGTTGAATGCAGTCATCACAGAGAAGTTTCTGAGAAGGCT
TCTGTCTAGATTTTATGTGAAGATATAGCCGTTTCGAACGAAGGCCACAAAGTGCTCCAA
ATATCCACTTGCAGGTCCTCCAAAAAGAGTGTTTCAAACGTGAACTACCAAAGGAAGGCT
CAACTCTGGACTTTGAATGCCAACGTCAGAAGGATGTTTCTGCGAAAGCTTCTGTTTAGT
TAGGTGACGTT
>NEW_CONTIG_MERGE_24
AGGGTTTCCAGAATGCTGTATGAAAAGAAAGGTTAAACTCTGTGAGTGGAACGCACACAT
CACAAAGCACTTTCTGAGAATGATTCTGTCTGGTTATTATACGAAGATATTTCCTTTTCT
GCAATTGTCCTCAAATCGCTTGAAATCTCCACCTGAAAATGCCACAGCAAGAGTGTTTCA
AATCTGCTCTCTCTAAAGCAAGGTTCAGCTCTGTGAGTTGAATACACACAACACAAAAAA
GTTACTGAGAACTCTTCTTAGTCTAGCATGAAAGGAAGAAACCCCGTTTGCAACGAAGGC
CTCAAAGAGGTCCAAATATCCACTTGCAGATTCCACAGAAAGAGTGTTTCAAAACTGCTC
TCTATCAATGGCAAAGTTCAACTCTGTTAGTTGAGGACACATATCACCAACAAGTTTCTG
AGAATGCTTCTGTCCATTTTTTTATGGGAAGATATTTCCTTTTTCAGCGTAGGCGTCAAGG
CGATCGAAATGTCCACTTCCACAAACTACAAAAAGAGTGTTTCCAACCTGCTCTATGAAA
GGCCATGTTCATCTCTATGAGTTGAATGGAAATATCCGAAAGAAATTTCTGGGAATGCTG
CTGTCTAGTGGTTATATGAATTCCCGCTTCCAACGAAATCCTCAAAGCAATCCAAATATC
CACTTGCAGAATCCACAAAAAGAGTGTTTCAAAACTGCTCTATCAATAGAAAGGTTCAAC
TCTTTTAGTTGAGTACACACATCACGAACAAGTTTCTGAGAATGCTTCTGTCTGGCTTTT
ATTGGAAGACGTTTCCTTTTCACCAAAGGCATCAAAGCGCTCCAAATGTCCACTTCCAGA
TACTACAAAAAGTGTGTTTCAAACCTGTTCTATGAAAGGAACTGTTCAACACTGTGACTT
CAATTGAAACATCCCGATGAAGCTTCTGAGAATGCTTCTGTCTAGTTTTTATGGGAAGAT
ATTTCCTTTTTCAACATAGGCCTCAAAGCGCTCCAAATGTCCACTTCCAGGTAGTGCAGA
```

行：1/171132 列：1 字符：62 (0x3E) 编码：936 (ANSI/OEM

Fig. S17. A screenshot of the final repeat library generated by REPdenovo.

Table S7. Sequence reads information of dataset NA12889.

Individual	Population	#of reads	Read length	Coverage
NA12889	CEU	228,944,748	101	7.2067

Table S8. Assembly quality of REPdenovo on dataset NA12889.

Individual	K_MIN	K_MAX	K_INT	K_DFT	N	N_h	Max fragment (bp)	N50 (bp)
NA12889	30	50	10	30	5890	80	23,178	3,090

$'N'$ represents the number of assembled contigs. $'N_h'$ represents the number of complete RepBase hits from the N repeats.

1.4.9 Which repeat group is used in the RepBase for the comparison of performance between LongRepMarker and other tools? What are those not covered ones?

The evaluations in this study are all based on RepeatMasker. The library used by RepeatMasker is a combination of RepBase and Dfam. The latest value of the proportion of the detection results of LongRepMarker covering the combination of RepBase and Dfam is 82.45%, just as shown in Table S19. The total number of sequences in combination of RepBase and Dfam of human that cannot be covered by the detection results of LongRepMarker is 154, accounting for 17.42% of the total sequences in the corresponding library. The classification of repetitive sequences in human RepBase library that cannot be covered by the detection results of LongRepMarker is shown in Table S9 below.

Table S9. The classification of repetitive sequences in human RepBase library that cannot be covered by the detection results of LongRepMarker.

Main class	Subclass	Number
ARTEFACT	-	7
Low complexity	-	2
DNA	Merlin	3
	Crypton-A?	2
	Crypton	5
	TcMar-Pogo	1
	PIF-Harbinger	2
	TcMar-Tigger	5
	hAT-Tip100?	1
	hAT-Ac	1
	hAT	4
	hAT-Blackjack	1
	TcMar-Mariner	1
	hAT-Charlie	1
	hAT-hAT19	1
	hAT-Tip100	5
	TcMar-Tc1	1
	TcMar?	1
	Other	13
LTR	ERVL	1
	ERV1?	1
	Gypsy	1
	Other	1
scRNA	-	1
snRNA	-	2
tRNA	-	7
RC?	Helitron?	1
SINE	tRNA	1
	tRNA-Deu	1
LINE	Dong-R4	1
	CR1	13
	L1	1
	Penelope	1
	L2	3
	RTE-BovB	1
	I-Jockey	1
DNA?	PiggyBac?	1
	Other	12
LTR?	-	4
Unknown	-	42
Total number of sequence is 154		

2 Methods

2.1 The specific commands and parameters for obtaining the overlap sequences

In this step, minimap2[88] is used for generating overlap sequences between and within chromosomes or contigs. The command line for running minimap2 is as follows: ′ minimap2 -x ava-pb ./contigs.fasta ./contigs.fasta > ./ovlp.paf ′.

2.2 Conversion of overlap sequences into unique *k-mers*

The counting of *k-mer* frequency in the sequencing data can be carried out using many of the currently available tools, such as Jellyfish[89], DSK[90] and KMC2[91]. In this study, we use DSK for *k-mer* frequency statistics. Note that before counting, *k-mer* size k needs to be determined. k should be kept small to prevent the overuse of computer memory. Counting k-mer frequency from either the reference genome sequence or sequenced reads data is one of the most fundamental analyses in bioinformatics. To estimate genomic characters, the *k-mer* size k should be determined under the logic that the space of *k-mer* (4^k) should be several times larger than the genome size (G), so that few *k-mers* derived from different genomic positions will merge together by chance, i.e. most *k-mers* in the genome will appear uniquely. In practice, we often require the *k-mer* space to be at least 5 times larger than the genome size and the larger the better. The rule for *k-mer* size selection is expressed as follows.

$$4^k > 5 * G \tag{4}$$

DSK (disk streaming of *k-mers*) is a new streaming algorithm for *k-mer* counting, which only requires a fixed, user-defined amount of memory and disk space. This approach realizes a memory, time and disk trade-off. The multi-set of all *k-mers* present in the reads is partitioned and partitions are saved to disk. Then, each partition is separately loaded in memory in a temporary hash table. The *k-mer* counts are returned by traversing each hash table. Low-abundance *k-mers* are optionally filtered. DSK is the first approach that is able to count all the 27-mers of a human genome dataset using only 4.0 GB of memory and moderate disk space (160 GB), in 17.9 hours. DSK can replace popular *k-mer* counting software (Jellyfish) on small-memory servers. In this step, DSK is used for generating unique *k-mers*, the specific operations and calculation complexity analysis are as follows. Assuming that there are n sequences, which respectively correspond to n fragments in an overlap file. Let c_i be the i-th fragment (i=1,2,...,n), and lc_i be the length of c_i. Given a fix length k of *k-mers* ($k << lc_i$), c_i can be represented as a list of (lc_i-k+1) *k-mers*. Therefore, the total number of *k-mers* (Num_k) that are transferred from all fragments can be expressed as follows.

$$Num_k = \sum_{i=1}^{n} \quad (lc_i - k + 1) \tag{5}$$

When the value of lc is large and the value of k is small, the total number of *k-mer* generated from these overlap sequences is very large. In order to further reduce the total number of *k-mers*, DSK converts all *k-mers* to their canonical representation with respect to reverse-complementation which called the unique *k-mers*. In other words, a *k-mer* and its reverse complement are considered to be the same object. For example, with k=3 and assuming the *k-mer* AAA and its reverse complement TTT are both present in the input dataset, DSK considers that one of them is the canonical *k-mer*, e.g. AAA. If AAA is present 2 times and TTT 3 times, then DSK outputs that the count of AAA is 5 and does not return the count of TTT at all. It is noticed that a canonical *k-mer* is not necessarily the lexicographically smallest one.

DSK uses a different ordering for faster performance. Specifically, it considers that A<C<T<G and returns the lexicographically smaller *k-mer* in this alphabet order. So, in the example above, AAA is indeed the canonical *k-mer*. For the GTA/TAC pair, the lexicographically smallest is GTA while the canonical *k-mer* is TAC (as DSK considers that T<G). The *k-mer* obtained by merging *k-mer* and its reverse-complementation is called the unique *k-mer*. Assuming that the length of c_i is lc_i, the average read coverage of c_i is C_i. Then, each *k-mer* appearing on c_i has an average of C_i reads corresponding to it, just as shown in Fig. S18. The relationship between the total number of reads (Num_r) and the total number of *k-mers* (Num_k) can be expressed as follows.

$$Num_r = C_i \times Num_k = C_i \times \sum_{i=1}^{n} \quad (lc_i - k + 1) \tag{6}$$

From Equation (6), we can see that the total number of reads appeared on the i-th fragment is C times the total number of unique *k-mers* appeared on the i-th fragment, where C is the average read coverage of sequencing. In practical applications, the average depth of the next generation sequencing data can reach dozens to hundreds of layers. Therefore, using the converted unique *k-mers* instead of the reads for mapping can greatly reduce the complexity of the alignment. The quantitative relationship among reads, *k-mers* and the unique *k-mers* is shown in Fig. S18.

In addition, a *k-mer* and its reverse complement are considered to be the same object in DSK, so that the actual number of converted unique *k-mers* is much smaller than the actual number of *k-mers* directly converted from the original sequences. Therefore, using unique *k-mer* instead of *k-mer* for mapping can further greatly reduce the complexity of the alignment. The relationship between the total number of

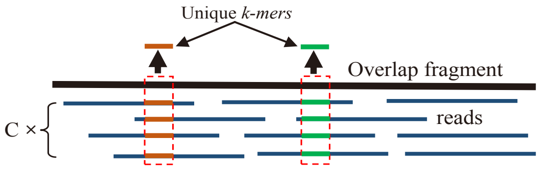


Fig. S18. The quantitative relationship between reads and *k-mers*, where the long black line represents chromosome or contig, short black lines represent reads and C indicates the average read coverage of the chromosome or contig.

alignment using read ($AlignNum_r$), the total number of alignment using *k-mer* ($AlignNum_k$) and the total number of alignment using the unique *k-mer* ($AlignNum_{uk}$) can be expressed as follows.

$$AlignNum_r = C_i \times AlignNum_k = C_i \times \sum_{i=1}^{n} (lc_i - k + 1) \gg AlignNum_{uk} \tag{7}$$

A practical example of the quantitative relationship among *k-mers* and the unique *k-mers* is shown in Table S10. From Table S10, we can see that a piece of DNA sequence which is composed as ′ ATTCGCTGATCATGTTCGA ′. Then, there are 16 *4-mer* generated from this sequence. After that, there are 13 unique *4-mer* generated from this sequence. From the quantitative relationship, we can know that there is a big difference between the *k-mer* and the unique *k-mer* in this short sequence, and this difference will be more obvious when the sequence increase and grows.

Table S10. A practical example of the quantitative relationship among *k-mers* and the unique *k-mers*.

Here is a piece of DNA sequence, its base arrangement is composed as follows: ATTCGCTGATCATGTTCGA
Then the generated *k-mer* set is: ATTC, TTCG, TCGC, CGCT, GCTG, CTGA, TGAT, GATC, ATCA, TCAT, CATG, ATGT, TGTT, GTTC, TTCG, TCGA
The generated unique *k-mer* set is: ATTC, TTCG, TCGC, CGCT, GCTG, CTGA, TGAT, ATCA TCAT, ATGT, TGTT, GTTC, TTCG

2.3 Generation of multi-alignment unique *k-mers* and their coverage regions on overlap sequences

LongRepMarker uses the multiple sequence alignment to find out the unique *k-mers* which can be aligned to different locations on overlap sequences and the regions on overlap sequences that can be covered by these multi-alignment unique *k-mers*. The process of generating the multi-alignment unique *k-mers* is described by *Algorithm 1*, and the process of generating the regions on overlap sequences that can be covered with these multi-alignment unique *k-mers* is described in *Algorithm 2*. The time complexity of these two algorithms is $O(n)$. The results of the multiple sequence alignment are stored in a bam file. When LongRepMarker gets the bam file. First of all, LongRepMarker filters the file and keeps the multiple alignment records and the ID of multi-alignment unique *k-mers* when it gets the bam file. Secondly, it converts the filtered bam file into a depth file via the samtools[92]. Finally, based on the information provided by the depth file and the ID records of multi-alignment unique *k-mers*, it extracts the regions on overlap sequences that can be covered with these multi-alignment unique *k-mers*, and forms several sequence fragments which are called sequence fragments with high probability of repetitive regions. The principle of generating fragments with high probability of repetitive regions is shown in Figs. S19(A) and S19(B).

Algorithm 1: Generation of multi-alignment unique *k-mers*

```
Input: Unique k-mers; overlap sequences
Output: The bam file; a set of serial numbers of multi-alignment unique k-mers
Initialize;
Let $T$<k-mer_id,k-mer_fre> be the set used to store the ID and frequency of each unique k-mer;
Let $M$<k-mer_id> be the set used to store the ID of each multi-alignment unique k-mer;
k-mer_id ← null; //Initialize of k-mer_id;
k-mer_fre ← 0; //Initialize of k-mer_fre;
Map all unique k-mers to overlap sequences to generate a bam file;
while not at end of the bam file do
    Get the $ID$ of the unique k-mer in the current line;
    if $ID \notin T$ then
        Create a new unit $R$ in set $T$;
        $R$.k-mer_id ← $ID$;
        $R$.k-mer_fre ← 1;
    else
        Get the unit $S$ marked with $ID$;
        $S$.k-mer_fre ← $S$.k-mer_fre+1;
    end
end
while not at end of the set $T$ do
    Get a record $C$ in set $T$;
    if $C$.k-mer_fre>1 then
        Create a new unit $G$ in set $M$;
        $G$.k-mer_id ← $C$.k-mer_id;
    else
        Continue;
    end
end
$T \leftarrow \emptyset$;
Return $M$ and bam file;
```

Algorithm 2: Generation of the regions on overlap sequences that can be covered by multi-alignment *k-mers*

```
Input: The bam file generated by Algorithm 1; $E$ ( a set of overlap sequence); $U$ (a set of serial numbers of
       the multi-alignment k-mers)
Output: The regions on overlap sequences that can be covered by multi-alignment k-mers
Initialize;
Let file $F$ be used to store the filtered bam file;
Let $D[i][E_i.length]$ be the character array used to store the $i$-th fragment in overlap sequences. Each unit from
    $D[i][0]$ to $D[i][E_i.length]$ is used to store every character of the $i$-th fragment in overlap sequences;
Let $S$ be the set used to store the regions on overlap sequences that can be covered by multi-alignment k-mers;
$S \leftarrow$ null; //Initialize $S$;
$D \leftarrow$ null; //Initialize $D$;
$F \leftarrow$ null; //Initialize $F$;
while not at end of the bam file do
    Get the $ID$ of the unique k-mer in the current line;
    if $ID \in U$ or line.StartWith('@') then
        Write the current line to the file $F$;
    else
        Continue; //Non-multiple alignment record;
    end
end
Convert the file $F$ to a depth file $P$ by samtools;
Group file $P$ by the first column;
while not at end of the file $P$ do
    Get a record $W$ in file $P$;
    $c_1 \leftarrow W.firstcolumn$ //The first column of depth file indicates the segment ID;
    $c_2 \leftarrow W.secondcolumn$ //The second column of depth file indicates the location on segment;
    $c_3 \leftarrow W.thirdcolumn$ //The third column of depth file indicates the depth of location on segment;
    if ($c_3 > 0$) then
        $D[c_1][c_2] = E[c_1].charAt(c_2 - 1)$ ;
        //Get the corresponding character;
    else
        $D[c_1][c_2] = 'N'$ ;
        //Uncovered location filling 'N' ;
    end
end
for each $i \in E.size()$ do
    $SeqString \leftarrow (D[i][E_i.length])$.toString();
    //Array to string;
    Create a new unit $G$ in set $S$;
    $G \leftarrow SeqString.SplitBy('N')$;
    //Splitting string by 'N';
end
$F \leftarrow \emptyset$; $D \leftarrow \emptyset$;
Return $S$;
```

2.4 The detailed principle of the combination of multiple threads parallel computing model and *k-mer* based multiple sequence alignment

LongRepMarker is designed based on the multiple threads parallel computing model and *k-mer* based multiple sequence alignment. It cuts a file consisting of all unique *k-mers* into multiple sub-files and aligns them separately with the overlap sequences in parallel (to facilitate parallel computing, the number of subfiles is set to the number of threads). The alignment results generated in the previous step (generation of multi-alignment unique *k-mers* and their coverage regions on original sequences) need to be used in this step. The principle of generating the regions on overlap sequences that can be covered by multi-alignment unique *k-mers* based on the parallel computing model is shown in Fig. S19(C). In Fig. S19(C), the different colored areas in different dashed boxes represent parallel units.

2.5 Classification of regions on overlap sequences that can be covered by the multi-alignment *k-mers*

The regions on original sequences (chromosomes or contigs) covered by the multi-alignment *k-mers* can be divided into two categories. The regions in the first category can be aligned to different locations ($\geq$2 locations) of the overlap sequences, which are highly likely to be repeats, so they are stored into the final repeat library directly. The regions in the second category cannot be aligned to the overlap sequences many times, but some sub-segments of them can be aligned to overlap sequences many times, which are probably caused by coupling matches due to sequencing errors (e.g., the two sequences are originally not repetitive sequences, due to sequencing errors that form some coupled alignments under error-tolerant conditions, resulting in multiple subsequences within them that can be aligned with each other. These two sequences should be removed) or the genetic variations (e.g., the two sequences are originally repetitive sequences, due to structural variations, multiple subsequences within them cannot be aligned with each other. These two sequences should be retained). The characteristic of coupling alignment due to the sequencing errors is that the alignment region is short and scattered, and it accounts for a relatively small proportion of the entire sequence fragment. On the contrary, the distribution of structural variation regions on the sequence fragment is relatively concentrated, and all have a certain length (e.g., greater than 50bp). Based on these obvious features, we can further filter these non-multiple aligned sequences. The coupling alignment due to sequencing errors, and conflict alignment due to SVs are shown in Fig. S20. The method used to distinguish two categories of alignment regions is described in detail in Section *S3.7*.

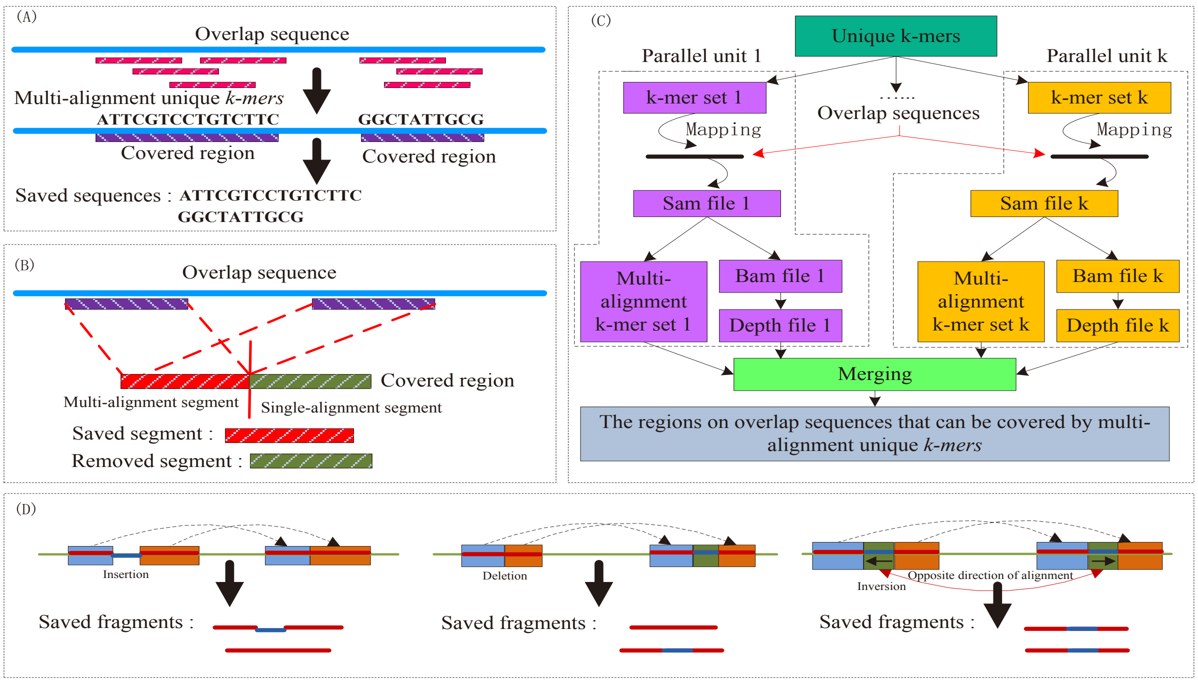


Fig. S19. The processing principle of some steps of LongRepMarker. The sub-graph(A) shows the principle of generating the regions on overlap sequences that can be covered by the multi-alignment unique *k-mers*; The sub-graph(B) shows the principle of processing the maximum multi-alignment sub-segments in the non multi-alignment fragment; The sub-graph (C) shows the principle of the parallel computing model in LongRepMarker; The sub-graph (D) shows the principle of identification of the genetic variations existing in the repetitive regions; Rectangles of red color in Sub-graph(A) represent the multi-alignment unique *k-mers*, and the rectangles of violet color represent their covered regions on overlap sequence; Rectangles of red color in in sub-graph(B) represent the multi-alignment segments in covered regions, and rectangles of green color represent the single-alignment segments in covered regions; In sub-graph(C) ,the different colored areas in different dashed boxes represent parallel units; The orange and blue rectangles in Sub-graph(D) represent the two subsequences in a pair of repetitive regions, and the green rectangles represent the regions of genetic variations.

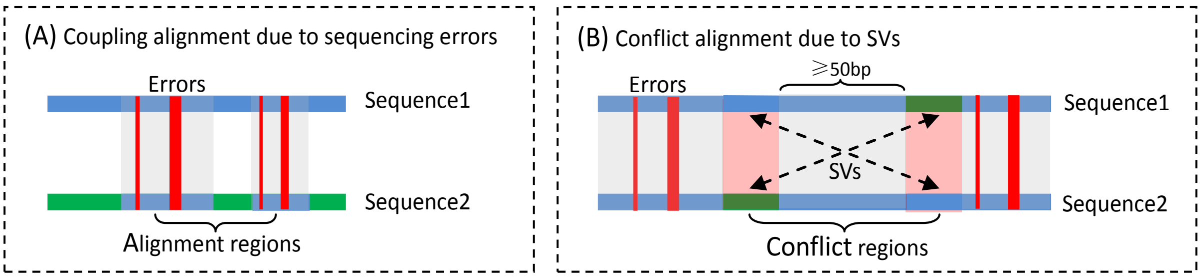


Fig. S20. Coupling alignment due to sequencing errors, and conflict alignment due to SVs. The light blue regions in sub-figure (A) indicate the coupling alignment regions caused by sequencing errors. The light blue regions in sub-figure (B) indicate the alignment regions, and the light red regions indicate the conflict regions caused by structural variations.

3 Results

3.1 Benchmarking methods and its configurations

In order to further illustrate the superior effectiveness of LongRepMarker, we compared it to the four famous methods in the field of repeat detection which are: RepeatScout, RepeatModel2, RepeatMasker, RepARK, REPdenovo and RepLong.

The running commands of RepeatScout are configured as ′./build_lmer_table -sequence $ref -freq $freq; ./RepeatScout -sequence $ref3 -output $repeats_out3 -freq $freq; ./filter-stage-1.prl $repeats_out > $filtered_out′. The running commands of RepeatModel2 are configured as ′ < RepeatModelerPath > / BuildDatabase -name elephant elephant.fa, nohup < RepeatModelerPath > /RepeatModeler -database elephant -pa 20 -LTRStruct > & run.out & ′. The running command of RepeatMasker is configured as: ′RepeatMasker -pa 128 -lib ./RepeatLib.fa -dir ./DetectionResults/ ./reference.fasta′.

The running command of RepARK is configured as ′RepARK.pl -p 24 -l ./left.fq -l ./right.fq -o output′. The running command of REPdenovo is configured as ′main.py -c Assembly -g ./sample_configuration.txt -r ./sample_input_paired_end_reads.txt, python ./main.py -c Scaffolding -g configuration-file-name -r raw-reads-file-name′. The running command of RepLong is configured as ′./replong.sh -f human_100k.fa -s 500M -t /2T/hum_100k′. When processing the dataset of Drosophila melanogaster, we use two different parameters to evaluate the performance of RepARK, the first one is the default parameter and the second parameter is t=84 which is given in the corresponding paper. Except for Drosophila dataset, RepARK uses default parameters when testing on other datasets. In this experiment, the parameters of REPdenovo are set to the default.

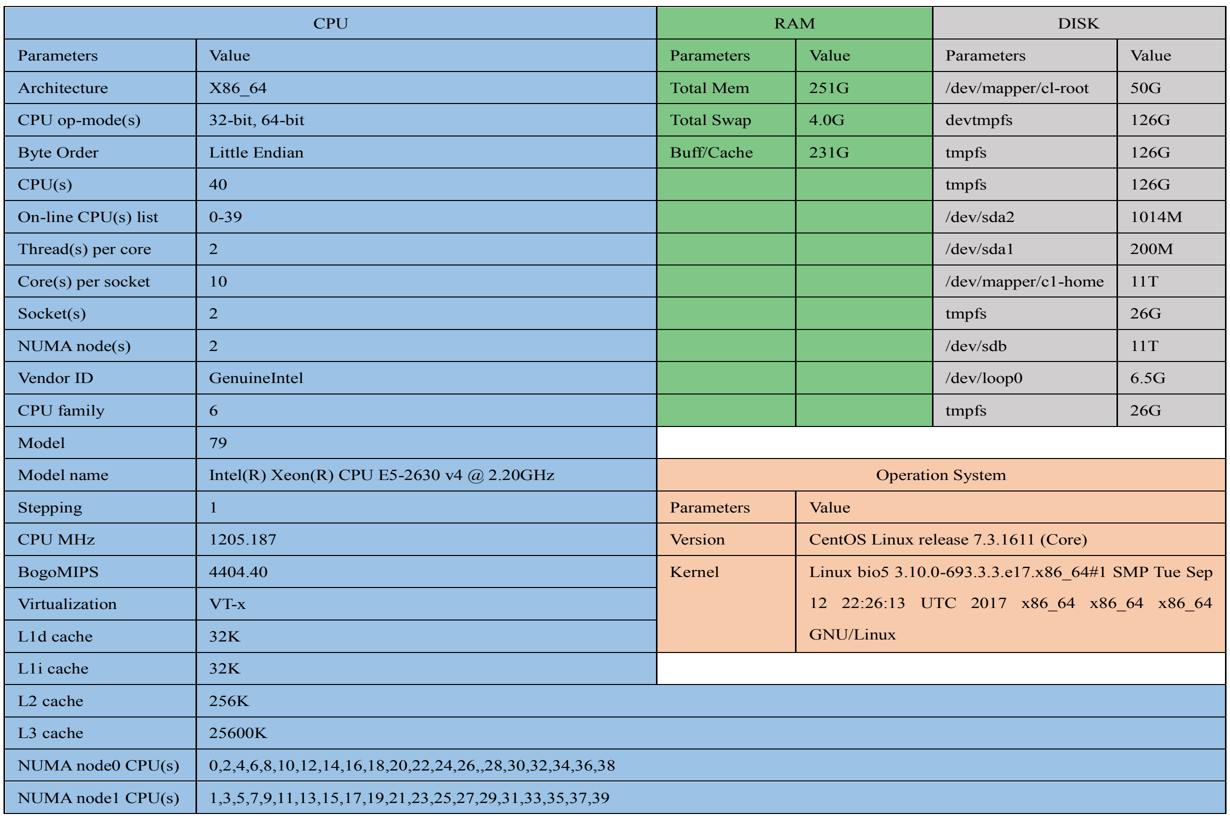
To provide unbiased evaluation, we used the assembly evaluation tool Quast, the multiple sequence alignment tool minimp2, and the repeat annotation tool RepeatMasker to comprehensively evaluate the repetitive regions generated from each of the compared tools in the experiment. The configuration of tools (Benchmarking and evaluation tools) is shown in Table S11.

3.2 Hardware Configuration

All benchmarkings were done on a computer with 24 cores and the memory of 512GB (Intel Xeon E5-2620 2.00 GHz), and the hardware configuration is shown in Fig. S21.

Table S11. Details of benchmarking tools

	Benchmarking tools		
Tools	Version	Requirements	Running commands and parameters configuration
ABySS	v2.2.4	gcc 4.6 or above;Boost;Open MPI	abyss-pe k=49 l=40 n=5 s=1000 name=test in='./left.fastq ./right.fastq'
SOAPdenovo2	2.04-r241	gcc 3.0 or above	./SOAPdenovo all -s config_file -K 25 -R -D 1 -d -o graph_prefix 1>ass.log 2>ass.err
SPAdes	v3.10.0	gcc 4.8.2;cmake2.8.12;zlib;libbz2	spades.py -t 64 –pe1-1 ./left.fastq –pe1-2 ./right.fastq -o ./output > /output/out.log
IDBA-UD	v1.3.0	gcc 4.7 or above	fq2fa –merge –filter ./fastq1 ./fastq2 ./read.fa idba_ud -r ./read.fa -o ./output
RepeatScout	v3.2.1	gcc 4.7 or above	./build_lmer_table -sequence $ref -freq $freq ./RepeatScout -sequence $ref3 -output $repeats_out3 -freq $freq ./filter-stage-1.prl $repeats_out > $filtered_out
RepeatModel2	v3.2.1	gcc 4.7 or above	<RepeatModelerPath> / BuildDatabase -name elephant elephant.fa nohup <RepeatModelerPath> /RepeatModeler -database elephant -pa 20 -LTRStruct > & run.out &
RepARK	v1	Perl5;Jellyfish2.1.4;Velvet1.2.08	RepARK.pl -p 24 -l ./left.fastq -l ./right.fastq -o output_folder
REPdenovo	v1.0.3	Python2.7;Jellyfish2.1.4;Velvet1.2.08	python ./main.py -c Assembly -g configuration-file-name -r raw-reads-file-name python ./main.py -c Scaffolding -g configuration-file-name -r raw-reads-file-name
MECAT2	v2	gcc 4.7 or above; HDF5	mecat2pw -j 0 -d ecoli_filtered.fastq -o ecoli_filtered.fastq.pm.can -w wrk_dir -t 16 mecat2cns -i 0 -t 16 ecoli_filtered.fastq.pm.can ecoli_filtered.fastq corrected_ecoli_filtered
RepLong	v0.0.1	faidx version 0.4.7.1; Canu 1.4; python(2.7 or 3.4 above) ; jdk1.8	./replong.sh -f <long reads fasta> -s <an estimate of the whole genome size> -t <place of temporary files> [-a <faidx path>] [-j <java path>] [-r minimum read length] [-o minimum overlap length] [-h maximum thread] [-e maximum memory] [-c true]
LongRepMarker	v0.0.1	gcc 4.7 or above; Python2.7; jdk1.8	java LongRepMarker [-r reference sequence] [-q1 1-th fastq file] [-q2 2-th fastq file] [-k k-mer size] [-X 10X linked reads] [-l SMS long reads] [-m the minimum length of repeat] [-t threads] [-Q effective size evaluation] [-M multi-alignment evaluation] [-v structure variation detection] [-o output]
LRSIM	v0.0.1	Math.Ramdom; Inline; perl5	./simulateLinkedReads.pl -r/-g <reference/haplotypes> -p <output prefix> [options]
minimap2	v2.2.17	GNU g++ 4.0	minimap2 -d ./ref_Align.mmi ./reference.fasta minimap2 -a -t 32 ./ref_Align.mmi ./regions.fasta > ./regions.sam
quast	v4.3	Python3.5	quast.py ./RepLib.fa -r ./Ref.fa -o ./QuastResults
RepeatMasker	v1.1.3	Perl5	RepeatMasker -parallel 30 -lib ./Lib.fa -html -gff -dir ./output ./Repbase.fa



CPU	
Parameters	Value
Architecture	X86_64
CPU op-mode(s)	32-bit, 64-bit
Byte Order	Little Endian
CPU(s)	40
On-line CPU(s) list	0-39
Thread(s) per core	2
Core(s) per socket	10
Socket(s)	2
NUMA node(s)	2
Vendor ID	GenuineIntel
CPU family	6
Model	79
Model name	Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz
Stepping	1
CPU MHz	1205.187
BogoMIPS	4404.40
Virtualization	VT-x
L1d cache	32K
L1i cache	32K
L2 cache	256K
L3 cache	25600K
NUMA node0 CPU(s)	0,2,4,6,8,10,12,14,16,18,20,22,24,26,,28,30,32,34,36,38
NUMA node1 CPU(s)	1,3,5,7,9,11,13,15,17,19,21,23,25,27,29,31,33,35,37,39

RAM		DISK	
Parameters	Value	Parameters	Value
Total Mem	251G	/dev/mapper/cl-root	50G
Total Swap	4.0G	devtmpfs	126G
Buff/Cache	231G	tmpfs	126G
		tmpfs	126G
		/dev/sda2	1014M
		/dev/sda1	200M
		/dev/mapper/c1-home	11T
		tmpfs	26G
		/dev/sdb	11T
		/dev/loop0	6.5G
		tmpfs	26G

Operation System	
Parameters	Value
Version	CentOS Linux release 7.3.1611 (Core)
Kernel	Linux bio5 3.10.0-693.3.3.e17.x86_64#1 SMP Tue Sep 12 22:26:13 UTC 2017 x86_64 x86_64 x86_64 GNU/Linux

Fig. S21. Hardware Configuration.

3.3 Evaluation metrics

In order to comprehensively evaluate the performance of the compared methods, we used 13 evaluation metrics in this experiment, which are Num, Max(kb), N50(kb), N75(kb), N90(kb), 0 times, 1 times, >1 times, Mapping Rate(%), Reference(%), Repbase(%), Time (hour) and Memory(MB). ′Num′ denotes the number of segments; ′Max(kb)′ denotes the length of the largest segment; ′N50(kb)′ is the length of the longest segment such that all the segments longer than this segment cover at least half (50%) of the total length of all segments; ′N75′ and ′N90′ are calculated in a similar way; ′0 times′ indicates the proportion of segments that cannot be aligned to the reference sequence in all segments; ′1 times′ indicates the proportion of segments that can be aligned to a unique location on the reference sequence in all segments; ′ >1 times’ indicates the proportion of segments that can be aligned to multiple locations on the reference sequence in all segments; ′Mapping Rate(%)′ indicates the proportion of segments that can be aligned to the reference sequence in all segments; ′Reference(%)′ indicates the proportion of regions marked as repetitive regions in the reference sequence that can be covered with the segments; ′Repbase(%)′ indicates the proportion of fragments in Repbase that can be covered with segments; ′Time (hour)′ indicates the time consumption of algorithms; ′Memory(MB)′ indicates the peak memory consumption of algorithms. Evaluation metrics are shown in Table S12.

Table S12. Evaluation metrics

Metrics	Meaning
Num	The number of segment
Max(kb)	The length of the largest segment
N50(kb)	The length of the longest segment such that all the segments longer than this segment cover at least 50% of the total length of all segments
N75(kb)	The length of the longest segment such that all the segments longer than this segment cover at least 75% of the total length of all segments
N90(kb)	The length of the longest segment such that all the segments longer than this segment cover at least 90% of the total length of all segments
0 times	The proportion of segments that cannot be aligned to the reference sequence in all segments
1 times	The proportion of segments that can be aligned to a unique location on the reference sequence in all segments
> 1 times	The proportion of segments that can be aligned to multiple locations on the reference sequence in all segments
Mapping Rate(%)	The proportion of segments that can be aligned to the reference sequence in all segments
Repbase(%)	The proportion of fragments in Repbase that can be covered with segments
Reference(%)	The proportion of regions marked as repetitive regions in the reference sequence that can be covered with the segments
Time (hour)	The time consumption of algorithms
Memory(MB)	The peak memory consumption of algorithms

3.4 Details of the experimental data

We evaluated the performance of five detection modes (Reference-assisted mode, *de novo* mode based on only NGS short reads, *de novo* mode based on the NGS short reads + barcode linked reads, *de novo* mode based on the NGS short reads + SMS long reads and *de novo* mode based on only SMS long reads) of LongRepMarker on 21 kinds of real sequencing data which include the reference genomes of 6 species, NGS sequencing data of 5 samples, NGS data of 3 samples and corresponding barcode linked data, NGS data of 3 samples and corresponding SMS data and SMS data of 4 samples. The details of these 21 kinds of data are shown in the Table S13.

Table S13. Details of the experimental data

Test items	Species	Dataset Name	Datasize (KB)	Source
Reference mode	Leafcutter Ant	GCA_000204515.1_Aech_3.9 _genomic_Ant.fna	293,052	https://www.ncbi.nlm.nih.gov/
	D.melanogaster	dmel-all-chromosome- r5.43.fasta	168,080	https://www.ncbi.nlm.nih.gov/
	Soybean	Glycine_max_Soybean.fna	968,211	https://www.ncbi.nlm.nih.gov/
	Gallus	Gallus_gallus.fna	1,053,454	https://www.ncbi.nlm.nih.gov/
	Mouse	GCA_000001635.8_GRCm38.p6 _genomic_Mouse.fna	2,787,341	https://www.ncbi.nlm.nih.gov/
	Human(hg38)	GCF_000001405.39_genomic _Human.fna	3,196,759	https://www.ncbi.nlm.nih.gov/
NGS only mode	Leafcutter Ant	ERR034186_1.fastq	17,580,863	https://www.ncbi.nlm.nih.gov/
		ERR034186_2.fastq	17,580,863	https://www.ncbi.nlm.nih.gov/
	D.melanogaster	SRR350908_1.fastq	5,767,698	https://www.ncbi.nlm.nih.gov/
		SRR350908_2.fastq	5,767,698	https://www.ncbi.nlm.nih.gov/
	Mouse	ERR2894257_1.fastq	26,655,537	https://www.ncbi.nlm.nih.gov/
		ERR2894257_2.fastq	26,655,537	https://www.ncbi.nlm.nih.gov/
	Human-chr14	frag_1.fastq	4,913,897	http://gage.cbcb.umd.edu/
		frag_2.fastq	4,913,897	http://gage.cbcb.umd.edu/
	HG003_24149_father	D2_S2_L001_R1_001.fastq	23,534,426	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
		D2_S2_L001_R2_001.fastq	23,534,426	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
NGS+linked mode	HG003_24149_father	D2_S2_L001_R1_001.fastq	23,544,498	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
		D2_S2_L001_R2_001.fastq	23,534,426	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
		10Xgenomics_ChromiumGenome	83,701,326,271	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
	HG004_NA24143_mother	D3_S3_L001_R1_001.fastq	20,115,293	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
		D3_S3_L001_R2_001.fastq	20,106,259	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
		10Xgenomics_ChromiumGenome	83,701,326,271	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
	HG002_NA24385_son	D1_S1_L001_R1_001.fastq	18,307,018	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
		D1_S1_L001_R2_001.fastq	18,315,730	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
		10Xgenomics_ChromiumGenome	172,283,140,716	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
NGS+SMS mode	HG003_24149_father	D2_S2_L001_R1_001.fastq	23,544,498	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
		D2_S2_L001_R2_001.fastq	23,534,426	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
		m64017_191110_150028.fastq	38,082,258	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
	HG004_NA24143_mother	D3_S3_L001_R1_001.fastq	20,115,293	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
		D3_S3_L001_R2_001.fastq	20,106,259	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
		m64017_191113_230503.fastq	43,819,465	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
	HG002_NA24385_son	D1_S1_L001_R1_001.fastq	18,307,018	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
		D1_S1_L001_R2_001.fastq	18,315,730	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
		m54316_180630_153952.Q20.fastq	5,285,851	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data
SMS only mode	D.melanogaster	dro_100k.fa	919,162	https://github.com/ruiguo-bio/replong
	Human	human_100k.fa	507,871	https://github.com/ruiguo-bio/replong
	D.melanogaster	dmel_filtered.fastq	30,885,716	https://github.com/ruiguo-bio/replong
	Human	human_polished.fastq	109,716,724	https://github.com/ruiguo-bio/replong

′Reference mode′ represents the Reference-assisted mode; ′NGS only mode′ represents the *de novo* mode based on only NGS short reads; ′NGS+linked mode′ represents the *de novo* mode based on the NGS short reads + barcode linked reads; ′NGS+SMS mode′ represents the *de novo* mode based on the NGS short reads + SMS long reads; ′SMS only mode′ represents the *de novo* mode based on only SMS long reads.

3.5 Detailed test results

3.5.1 Detection results of reference-assisted mode

We evaluated the performance of LongRepMarker in the reference-assisted mode on six eukaryote genomes (Table S13). The genome size of these six species are 3.196Gb (*H.sapiens(hg38)*), 2.752Gb (*Mouse*), 289Mb (*Leafcutter Ant*), 168Mb (*D.melanogaster*), 956Mb (*soybean*) and 1.040Gb (*Gallus*). The results are shown in Figs. S22-S23 and Tables S14-S26. We compared the reference-assisted mode of LongRepMarker with RepeatMasker (Table S14), RepeatScout and RepeatModeler2.

Since the RepeatMasker can only be used to mask the repeats in the genome, it cannot classify the masked repeats in detail, so we can only compare the performance of LongRepMarker with RepeatMasker by detecting the size and alignment rate of detected fragments, just as the shown in Table S14. As can be seen from Table S14, LongRepMarker is superior to RepeatMasker in terms of running time, memory consumption, fragment size and alignment rate. In order to further analyze the difference between the detection results of those two tools, we carried out two comparative experiments, and the results are shown in the Tables S61 and S62 of Section 3.8. Comparative experiments show that LongRepMaker can find some new repetitive sequence types which cannot be found by RepeatMasker.

Comparison of the detection results of LongRepMaker with that of RepeatScout and RepeatModeler2 is shown in Fig. S22 and Tables S15-S26. In Fig. S22, the X-axis represents the length distribution of the detected fragments and Y-axis represents the repetition frequency of the detected fragments in the genome, and the three images in each row respectively represent the frequency and length distribution of the repeated sequences detected by the LongRepMaker, RepeatScout, and RepeatModeler2 in a certain species. The coordinates of the Y-axis are divided into left and right displays, where the low frequency on the left is represented by purple, and the high frequency on the right is represented by green. As can be seen from Fig. S22, the repetition frequency and length distribution of the fragments detected by LongRepMarker have significant advantages over the latter two tools. Tables S15-S26 show the proportion and detailed classification of the detection results generated from the three tools on six species covering the corresponding RepBase library and reference genome. From the perspective of the coverage of the total base ratio, LongRepMarker has certain advantages compared with the latter two tools.

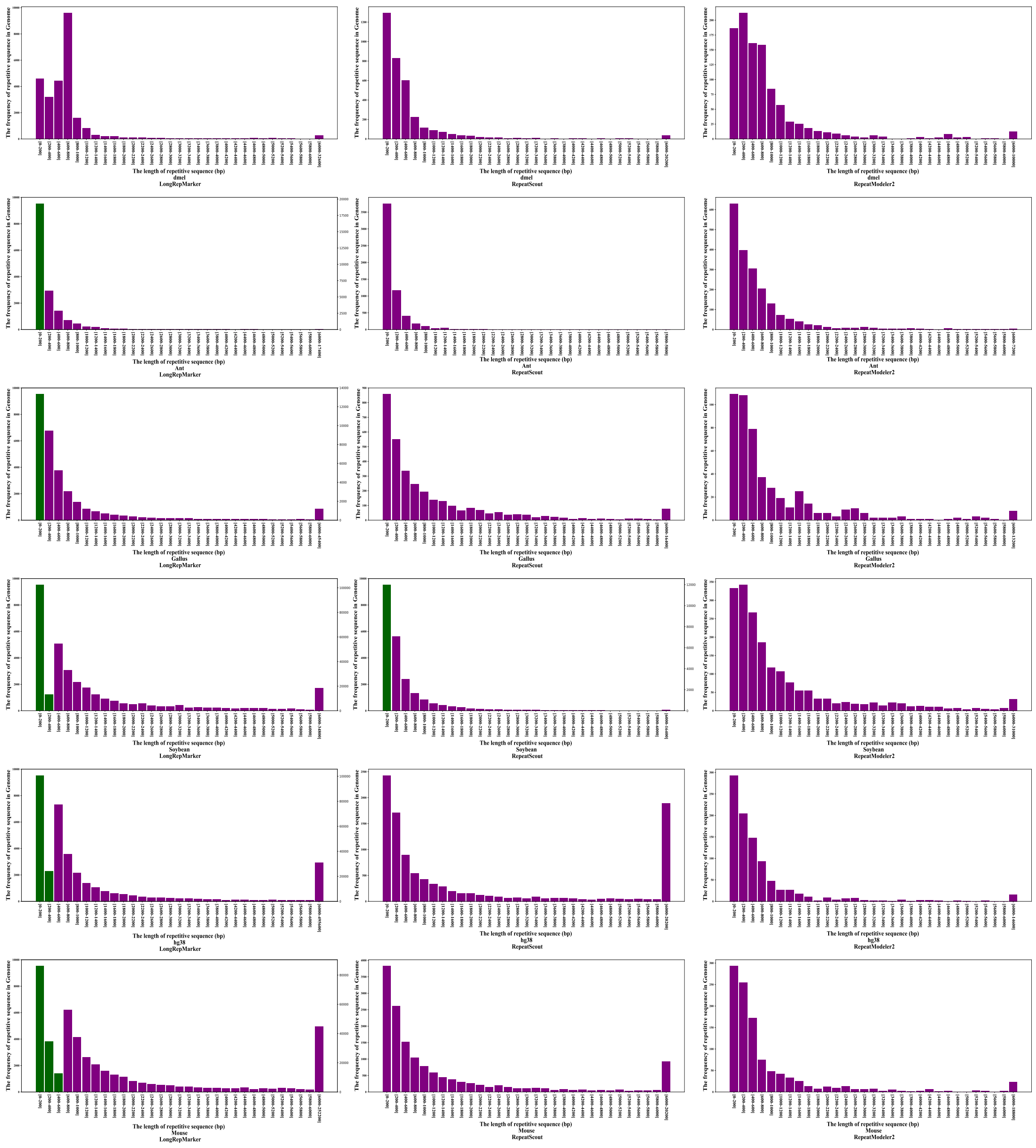


Fig. S22. Comparison of the repetition frequency and length distribution of the detected fragments generated from three tools. The X-axis represents the length distribution of the detected fragments and Y-axis represents the repetition frequency of the detected fragments in the genome, and the three images in each row respectively represent the frequency and length distribution of the repeated sequences detected by the three tools in a certain species. The coordinates of the Y-axis are divided into left and right displays, where the low frequency on the left is represented by purple, and the high frequency on the right is represented by green.

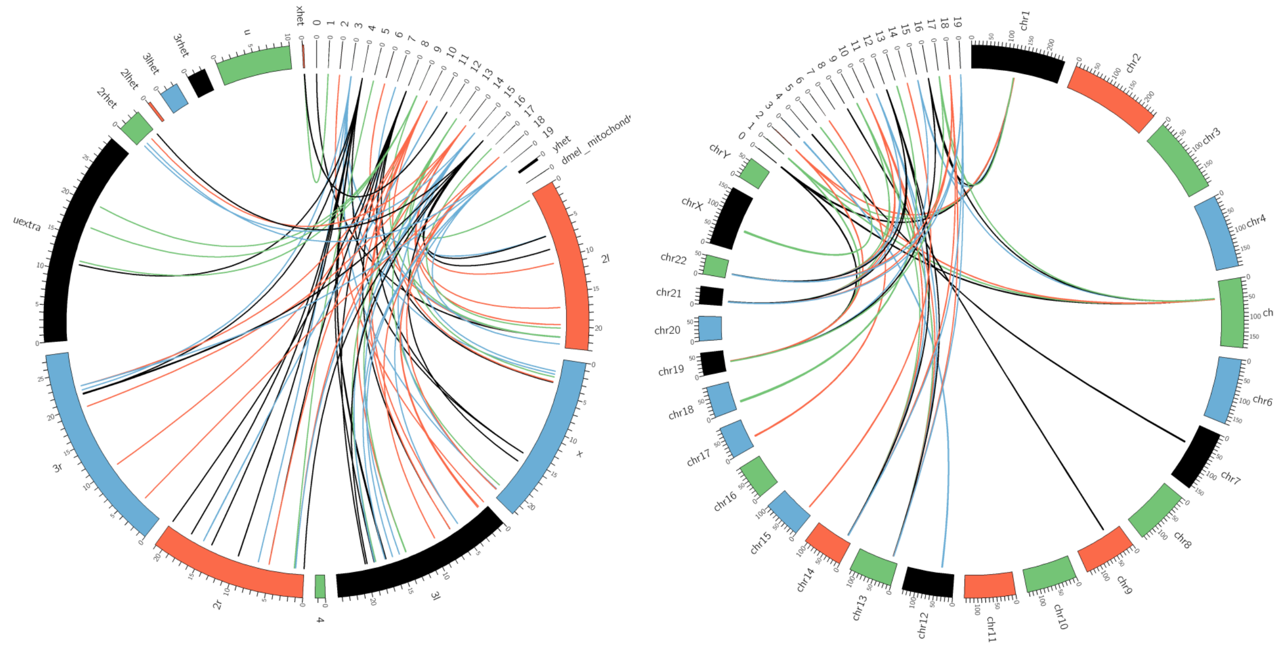


Fig. S23. The sub-figure on the left shows the alignment between the reference genome of Drosophila and 20 repetitive fragments randomly selected from the detection results of the Drosophila dataset which generated from the reference-assisted mode of LongRepMarker. The sub-figure on the right shows the alignment between the reference genome of Human(hg38) and 20 repetitive fragments randomly selected from the detection results of the Human dataset which generated from the reference-assisted mode of LongRepMarker.

Table S14. Comparison of the detection results generated by LongRepMarker in the reference-assisted mode and RepeatMasker.

		Quast (length ≥ 5000bp)					Minimap2			
Species	Tool	Time(min)/Peak Mem(GB)	Max (kb)	N50 (kb)	N75 (kb)	N90 (kb)	0 time	1 time	>1 time	Mapping Rate (%)
H.sapiens(hg38)	LongRepMarker	2863.539/46.688	1034.338	83.195	28.812	10.281	0.00%	11.75%	88.25%	100.0%
	RepeatMasker	12696.500/71.808	1499.996	7.228	6.133	5.616	0.00%	92.63%	7.37%	100.0%
Mouse	LongRepMarker	2979.584/42.868	339.188	16.526	7.112	6.061	0.00%	24.49%	75.51%	100.0%
	RepeatMasker	11734.183/65.234	78.144	6.409	6.092	5.391	0.01%	82.61%	17.39%	99.99%

The middle sub-table shows the size statistics of detection results. The right sub-table shows the alignment of repetitive fragments in reference genome. The main evaluation indicators are introduced in Table S12.

Table S15. Comparison of the proportion and detailed classification of detection results generated by three tools on Drosophila dataset covering the corresponding RepBase library.

LongRepMarker				RepeatScout			RepeatModeler2		
sequence: 2489				**sequence: 2489**			**sequence: 2489**		
total length: 7220516bp				**total length: 7220516bp**			**total length: 7220516bp**		
GC level: 42.77%				**GC level: 42.77%**			**GC level: 42.77%**		
bases masked: 3746452 bp (51.89%))				**bases masked: 3491131 bp (48.35%))**			**bases masked: 3336440 bp (46.21%))**		
Repeat Types	**Number of elements**	**Length occupied**	**Percentage of sequence**	**Number of elements**	**Length occupied**	**Percentage of sequence**	**Number of elements**	**Length occupied**	**Percentage of sequence**
SINEs:	1	73 bp	0.00%	1	74 bp	0.00%	0	0 bp	0.00%
–ALUs:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–MIRs:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
LINEs:	1317	1043230 bp	14.45%	1187	949761 bp	13.15%	1152	955570 bp	13.23%
–LINE1:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–LINE2:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–L3/CR1:	202	143230 bp	1.98%	147	106999 bp	1.48%	108	104755 bp	1.45%
LTR elements:	2515	2355715 bp	32.63%	2631	2194498 bp	30.39%	2254	2065761 bp	28.61%
–ERVL	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–ERVL-MaLRs	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–ERV_classI	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–ERV_classII	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
DNA transposon elements:	421	193452 bp	2.68%	524	177269 bp	2.46%	409	170180 bp	2.36%
–hAT-Charlie:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–TcMar-Tigger:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
Unclassified:	166	53666 bp	0.74%	179	59066 bp	0.82%	139	32370 bp	0.45%
Total interspersed repeats:		3646136 bp	50.50%		3380668 bp	46.82%		3223881 bp	44.65%
Small RNA:	29	13271 bp	0.18%	27	13792 bp	0.19%	15	6003 bp	0.08%
Satellites:	17	6719 bp	0.09%	15	6065 bp	0.08%	4	544 bp	0.03%
Simple repeats:	1108	74336 bp	1.03%	1172	76644 bp	1.06%	1224	80026 bp	1.11%
Low complexity:	291	15714 bp	0.22%	295	16084 bp	0.22%	318	17088 bp	0.24%

Table S16. Comparison of the proportion and detailed classification of detection results generated by three tools on Ant dataset covering the corresponding RepBase library.

LongRepMarker				RepeatScout			RepeatModeler2		
sequence: 254				**sequence: 254**			**sequence: 254**		
total length: 214457bp				**total length: 214457bp**			**total length: 214457bp**		
GC level: 45.07%				**GC level: 45.07%**			**GC level: 45.07%**		
bases masked: 168915 bp (78.76%))				**bases masked: 173383 bp (80.85%))**			**bases masked: 169070 bp (78.84%))**		
Repeat Types	**Number of elements**	**Length occupied**	**Percentage of sequence**	**Number of elements**	**Length occupied**	**Percentage of sequence**	**Number of elements**	**Length occupied**	**Percentage of sequence**
SINEs:	1	69 bp	0.03%	0	0 bp	0.00%	0	0 bp	0.00%
–ALUs:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–MIRs:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
LINEs:	19	8223 bp	3.83%	29	12137 bp	5.66%	16	13379 bp	6.24%
–LINE1:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–LINE2:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–L3/CR1:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
LTR elements:	71	48344 bp	22.54%	43	49839 bp	23.24%	38	48684 bp	22.70%
–ERVL	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–ERVL-MaLRs	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–ERV_classI	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–ERV_classII	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
DNA transposon elements:	95	71724 bp	33.44%	111	72618 bp	33.86%	116	69591 bp	32.45%
–hAT-Charlie:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–TcMar-Tigger:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
Unclassified:	39	9733 bp	4.54%	34	7607 bp	3.55%	13	7330 bp	3.42%
Total interspersed repeats:		138093 bp	64.39%		142201 bp	66.31%		138984 bp	64.81%
Small RNA:	6	566 bp	0.26%	7	746 bp	0.35%	0	0 bp	0.00%
Satellites:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
Simple repeats:	184	30485 bp	14.21%	184	30449 bp	14.20%	181	30285 bp	14.12%
Low complexity:	2	110 bp	0.05%	2	110 bp	0.05%	2	110 bp	0.05%

Table S17. Comparison of the proportion and detailed classification of detection results generated by three tools on Gallus dataset covering the corresponding RepBase library.

LongRepMarker				RepeatScout			RepeatModeler2		
sequence: 512				**sequence: 512**			**sequence: 512**		
total length: 362626bp				**total length: 362626bp**			**total length: 362626bp**		
GC level: 49.51%				**GC level: 49.51%**			**GC level: 49.51%**		
bases masked: 255267 bp (70.39%))				**bases masked: 246110 bp (67.87%))**			**bases masked: 225717 bp (62.25%))**		
Repeat Types	**Number of elements**	**Length occupied**	**Percentage of sequence**	**Number of elements**	**Length occupied**	**Percentage of sequence**	**Number of elements**	**Length occupied**	**Percentage of sequence**
SINEs:	9	784 bp	0.22%	15	1292 bp	0.36%	13	1935 bp	0.53%
–ALUs:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–MIRs:	0	0 bp	0.00%	0	0 bp	0.00%	5	839 bp	0.23%
LINEs:	163	119330 bp	32.91%	94	100795 bp	27.80%	84	88718 bp	24.47%
–LINE1:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–LINE2:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–L3/CR1:	163	119330 bp	32.91%	94	100795 bp	27.80%	84	88718 bp	24.47%
LTR elements:	68	78553 bp	21.94%	107	99403 bp	27.41%	109	85021 bp	23.45%
–ERVL	40	45157 bp	12.45%	55	52791 bp	14.56%	69	54457 bp	15.02%
–ERVL-MaLRs	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–ERV_classI	7	9997 bp	2.76%	14	11837 bp	3.26%	16	14306 bp	3.95%
–ERV_classII	20	23453 bp	6.47%	38	34775 bp	9.59%	24	16258 bp	4.48%
DNA transposon elements:	18	3974 bp	1.10%	21	7645 bp	2.11%	35	11014 bp	3.04%
–hAT-Charlie:	2	346 bp	0.10%	7	2249 bp	0.62%	5	4653 bp	1.28%
–TcMar-Tigger:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
Unclassified:	8	8430 bp	2.32%	2	397 bp	0.11%	5	652 bp	0.18%
Total interspersed	repeats:	212071 bp	58.48%		209532 bp	57.78%		187340 bp	51.66%
Small RNA:	39	6135 bp	1.69%	46	4770 bp	1.32%	2	380 bp	0.10%
Satellites:	5	5709 bp	1.57%	3	583 bp	0.16%	7	6199 bp	1.71%
Simple repeats:	197	31810 bp	8.77%	198	31808 bp	8.77%	200	32031 bp	8.83%
Low complexity:	3	147 bp	0.04%	3	147 bp	0.04%	3	147 bp	0.04%

Table S18. Comparison of the proportion and detailed classification of detection results generated by three tools on Soybean dataset covering the corresponding RepBase library.

LongRepMarker				RepeatScout			RepeatModeler2		
sequence: 758				**sequence: 758**			**sequence: 758**		
total length: 1646292bp				**total length: 1646292bp**			**total length: 1646292bp**		
GC level: 42.57%				**GC level: 42.57%**			**GC level: 42.57%**		
bases masked: 1536173 bp (93.31%))				**bases masked: 1535709 bp (93.28%))**			**bases masked: 1375693 bp (83.56%))**		
Repeat Types	**Number of elements**	**Length occupied**	**Percentage of sequence**	**Number of elements**	**Length occupied**	**Percentage of sequence**	**Number of elements**	**Length occupied**	**Percentage of sequence**
SINEs:	1	18 bp	0.00%	0	0 bp	0.00%	2	145 bp	0.01%
–ALUs:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–MIRs:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
LINEs:	45	80754 bp	4.91%	52	85283 bp	5.18%	67	72838 bp	4.42%
–LINE1:	44	77578 bp	4.71%	50	81968 bp	4.98%	65	69502 bp	4.22%
–LINE2:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–L3/CR1:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
LTR elements:	881	1238562 bp	75.23%	1030	1238480 bp	75.23%	815	1114450 bp	67.69%
–ERVL	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–ERVL-MaLRs	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–ERV_classI	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–ERV_classII	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
DNA transposon elements:	130	154184 bp	9.37%	145	146753 bp	8.91%	139	123780 bp	7.52%
–hAT-Charlie:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
–TcMar-Tigger:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
Unclassified:	28	25065 bp	1.52%	34	31451 bp	1.91%	20	23872 bp	1.45%
Total interspersed	repeats:	1498583 bp	91.03%		1501967 bp	91.23%		1335085 bp	81.10%
Small RNA:	22	6625 bp	0.40%	26	6988 bp	0.42%	2	5216 bp	0.32%
Satellites:	0	0 bp	0.00%	0	0 bp	0.00%	3	301 bp	0.02%
Simple repeats:	200	31493 bp	1.91%	215	32310 bp	1.96%	255	33830 bp	2.05%
Low complexity:	9	1018 bp	0.06%	11	824 bp	0.05%	21	1344 bp	0.08%

Table S19. Comparison of the proportion and detailed classification of detection results generated by three tools on Human(hg38) dataset covering the corresponding RepBase library.

LongRepMarker				RepeatScout			RepeatModeler2		
sequence: 1512				**sequence: 1512**			**sequence: 1512**		
total length: 1647075bp				**total length: 1647075bp**			**total length: 1647075bp**		
GC level: 45.30%				**GC level: 45.30%**			**GC level: 45.30%**		
bases masked: 1213841 bp (82.45%))				**bases masked: 1213841 bp (73.70%))**			**bases masked: 1213841 bp (63.33%))**		
Repeat Types	**Number of elements**	**Length occupied**	**Percentage of sequence**	**Number of elements**	**Length occupied**	**Percentage of sequence**	**Number of elements**	**Length occupied**	**Percentage of sequence**
SINEs:	709	255186 bp	15.49%	690	189225 bp	11.49%	87	26863 bp	1.63%
–ALUs:	690	251356 bp	15.26%	676	188001 bp	11.41%	74	25030 bp	1.52%
–MIRs:	17	3552 bp	0.22%	9	613 bp	0.04%	10	1566 bp	0.10%
LINEs:	1376	690975 bp	41.95%	624	298720 bp	18.14%	275	254509 bp	15.45%
–LINE1:	1337	682454 bp	41.43%	608	295084 bp	17.92%	244	242822 bp	14.74%
–LINE2:	11	1455 bp	0.09%	9	2517 bp	0.15%	9	5981 bp	0.36%
–L3/CR1:	5	708 bp	0.04%	4	805 bp	0.05%	18	4208 bp	0.26%
LTR elements:	566	327086 bp	19.86%	903	571647 bp	34.71%	1011	612530 bp	37.19%
–ERVL	98	39268 bp	2.38%	152	87126 bp	5.29%	188	119764 bp	7.27%
–ERVL-MaLRs	32	8118 bp	0.49%	47	13970 bp	0.85%	47	19783 bp	1.20%
–ERV_classI	370	220447 bp	13.38%	634	388908 bp	23.61%	709	402655 bp	24.45%
–ERV_classII	54	57466 bp	3.49%	56	79131 bp	4.80%	65	69756 bp	4.24%
DNA transposon elements:	110	25838 bp	1.57%	223	64465 bp	3.91%	344	106865 bp	6.49%
–hAT-Charlie:	41	8781 bp	0.53%	48	11006 bp	0.67%	102	28766 bp	1.75%
–TcMar-Tigger:	35	11048 bp	0.67%	84	34543 bp	2.10%	107	36801 bp	2.23%
Unclassified:	185	57213 bp	3.47%	141	54603 bp	3.32%	8	2266 bp	0.14%
Total interspersed	repeats:	1356298 bp	82.35%		1178660 bp	71.56%		1003033 bp	60.90%
Small RNA:	14	1276 bp	0.08%	51	11176 bp	0.68%	6	742 bp	0.05%
Satellites:	24	10205 bp	0.62%	31	12727 bp	0.77%	14	3414 bp	0.21%
Simple repeats:	216	31821 bp	1.93%	255	34087 bp	2.07%	279	34727 bp	2.11%
Low complexity:	11	483 bp	0.03%	18	851 bp	0.05%	27	1228 bp	0.07%

Table S20. Comparison of the proportion and detailed classification of detection results generated by three tools on Mouse dataset covering the corresponding RepBase library.

LongRepMarker				RepeatScout			RepeatModeler2		
sequence: 1561				**sequence: 1561**			**sequence: 1561**		
total length: 1680566bp				**total length: 1680566bp**			**total length: 1680566bp**		
GC level: 44.70%				**GC level: 44.70%**			**GC level: 44.70%**		
bases masked: 1044496 bp (62.15%))				**bases masked: 987478 bp (58.76%))**			**bases masked: 907532 bp (54.00%))**		
Repeat Types	**Number of elements**	**Length occupied**	**Percentage of sequence**	**Number of elements**	**Length occupied**	**Percentage of sequence**	**Number of elements**	**Length occupied**	**Percentage of sequence**
SINEs:	325	62219 bp	3.70%	292	60826 bp	3.62%	77	10662 bp	0.63%
–ALUs:	272	52570 bp	3.13%	218	50691 bp	3.02%	44	6817 bp	0.41%
–MIRs:	8	1464 bp	0.09%	5	439 bp	0.03%	10	1220 bp	0.07%
LINEs:	822	581349 bp	34.59%	477	417481 bp	24.84%	276	357395 bp	21.27%
–LINE1:	820	579118 bp	34.46%	472	416637 bp	24.79%	275	357213 bp	21.26%
–LINE2:	1	94 bp	0.01%	4	205 bp	0.01%	1	182 bp	0.01%
–L3/CR1:	0	0 bp	0.00%	0	0 bp	0.00%	0	0 bp	0.00%
LTR elements:	493	343322 bp	20.43%	659	458817 bp	27.30%	848	483400 bp	28.76%
–ERVL	85	56252 bp	3.35%	83	54386 bp	3.24%	76	61939 bp	3.69%
–ERVL-MaLRs	57	10909 bp	0.65%	71	16563 bp	0.99%	59	16407 bp	0.98%
–ERV_classI	78	65804 bp	3.92%	117	86999 bp	5.18%	135	82306 bp	4.90%
–ERV_classII	265	207985 bp	12.38%	383	291835 bp	17.37%	575	31922 bp	18.99%
DNA transposon elements:	57	7136 bp	0.42%	33	4009 bp	0.24%	36	9446 bp	0.56%
–hAT-Charlie:	32	3880 bp	0.23%	25	3117 bp	0.19%	24	5297 bp	0.32%
–TcMar-Tigger:	9	1410 bp	0.08%	1	107 bp	0.01%	9	3608 bp	0.21%
Unclassified:	53	20086 bp	1.20%	33	9440 bp	0.56%	18	6587 bp	0.39%
Total interspersed	repeats:	1014112 bp	60.34%		950573 bp	56.56%		867490 bp	51.62%
Small RNA:	29	3693 bp	0.22%	55	4815 bp	0.29%	2	323 bp	0.02%
Satellites:	8	4208 bp	0.25%	9	4033 bp	0.24%	4	544 bp	0.03%
Simple repeats:	314	36351 bp	2.16%	327	36959 bp	2.20%	333	37141 bp	2.21%
Low complexity:	35	1618 bp	0.10%	40	1873 bp	0.11%	45	2227 bp	0.13%

Table S21. Comparison of the proportion and detailed classification of detection results generated by three tools covering the repeat regions on the reference genome of Drosophila.

LongRepMarker				RepeatScout			RepeatModeler2		
sequence: 15 **total length: 168736537bp** **bases masked: 34098031 bp (20.21%))**				**sequence: 15** **total length: 168736537bp** **bases masked: 11038265 bp (6.54%))**			**sequence: 15** **total length: 168736537bp** **bases masked: 9653964 bp (5.72%))**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	14298	1746619bp	1.04%	3436	635949bp	0.38%	2061	544394bp	0.32%
–Academ-1:	6	3667bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–CMC-EnSpm:	9	3415bp	0.00%	42	4015bp	0.00%	1	87bp	0.00%
–CMC-Transib:	1448	153013bp	0.09%	950	109301bp	0.06%	206	62009bp	0.04%
–Crypton-C:	0	0bp	0.00%	0	0bp	0.00%	3	499bp	0.00%
–Crypton-S:	0	0bp	0.00%	7	11414bp	0.01%	0	0bp	0.00%
–MULE-MuDR:	12	7323bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MULE-NOF:	726	100063bp	0.06%	91	9623bp	0.01%	16	7460bp	0.00%
–Maverick:	14	6413bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MuLE-NOF:	6	239bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	206	22100bp	0.01%	132	21553bp	0.01%	34	4726bp	0.00%
–P:	6582	935918bp	0.55%	879	223592bp	0.13%	1091	229500bp	0.14%
–PIF-Harbinger:	16	894bp	0.00%	57	5029bp	0.00%	7	1306bp	0.00%
–PiggyBac:	19	955bp	0.00%	20	1242bp	0.00%	0	0bp	0.00%
–TcMar-Mariner:	0	0bp	0.00%	55	2516bp	0.00%	0	0bp	0.00%
–TcMar-Pogo:	717	54830bp	0.03%	45	22374bp	0.01%	84	28827bp	0.02%
–TcMar-Tc1:	2423	260265bp	0.15%	794	130478bp	0.08%	325	128066bp	0.08%
–hAT-Ac:	823	82579bp	0.05%	118	34518bp	0.02%	102	26511bp	0.02%
–hAT-Pegasus:	0	0bp	0.00%	0	0bp	0.00%	12	2875bp	0.00%
–hAT-Tag1:	61	14006bp	0.01%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Tip100:	25	14147bp	0.01%	0	0bp	0.00%	0	0bp	0.00%
–hAT-hATm:	13	9519bp	0.01%	45	5764bp	0.00%	0	0bp	0.00%
–hAT-hobo:	1192	133839bp	0.08%	201	56995bp	0.03%	180	52528bp	0.03%
LINE:	63010	7968013bp	4.72%	8653	2218122bp	1.31%	9548	2424426bp	1.44%
–CR1:	3231	604771bp	0.36%	510	135201bp	0.08%	536	168043bp	0.10%
–I:	1627	302608bp	0.18%	218	112053bp	0.07%	224	140677bp	0.08%
–I-Jockey:	17000	2648503bp	1.57%	3752	885268bp	0.52%	2788	729222bp	0.43%
–Jockey:	16801	2866297bp	1.70%	2530	666464bp	0.39%	3827	1058867bp	0.63%
–LOA:	1108	166353bp	0.10%	251	104209bp	0.06%	257	98076bp	0.06%
–OTHER:	127	94706bp	0.06%	30	17419bp	0.01%	0	0bp	0.00%
–R1:	19563	2618643bp	1.55%	1007	322219bp	0.19%	1495	447221bp	0.27%
–R1-LOA:	448	130728bp	0.08%	234	69698bp	0.04%	174	66079bp	0.04%
–R2:	3087	205770bp	0.12%	121	58916bp	0.03%	247	64798bp	0.04%
–R2-NeSL:	8	617bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–RTE-X:	10	17039bp	0.01%	0	0bp	0.00%	0	0bp	0.00%
LTR:	127541	15911903bp	9.43%	22050	6160703bp	3.65%	17199	5094134bp	3.02%
–Copia:	7021	961221bp	0.57%	1548	475276bp	0.28%	1172	296734bp	0.18%
–ERVK:	15	6458bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Gypsy:	83678	11517560bp	6.83%	17622	4765804bp	2.82%	12089	3532468bp	2.09%
–Gypsy-Cigr:	5	559bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	83	70018bp	0.04%	48	18971bp	0.01%	12	15980bp	0.01%
–Pao:	35945	3426976bp	2.03%	2780	915238bp	0.54%	3716	1163954bp	0.69%
–Viper:	794	34755bp	0.02%	52	5354bp	0.00%	210	84998bp	0.05%
Other:	7666	660761bp	0.39%	602	152415bp	0.09%	1162	244620bp	0.14%
–OTHER:	7666	660761bp	0.39%	602	152415bp	0.09%	1162	244620bp	0.14%
RC:	1457	306200bp	0.18%	1295	183764bp	0.11%	1091	78447bp	0.05%
–Helitron:	1457	306200bp	0.18%	1295	183764bp	0.11%	1091	78447bp	0.05%
RNA:	116	20947bp	0.01%	33	1782bp	0.00%	0	0bp	0.00%
–OTHER:	116	20947bp	0.01%	33	1782bp	0.00%	0	0bp	0.00%
SINE:	236	31505bp	0.02%	48	2925bp	0.00%	0	0bp	0.00%
–5S:	98	25877bp	0.02%	0	0bp	0.00%	0	0bp	0.00%
–ID:	35	929bp	0.00%	18	884bp	0.00%	0	0bp	0.00%
–U:	31	2135bp	0.00%	12	1273bp	0.00%	0	0bp	0.00%
–tRNA:	31	1468bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–tRNA-RTE:	41	1096bp	0.00%	18	768bp	0.00%	0	0bp	0.00%
Satellite:	25288	3682739bp	2.18%	1004	311451bp	0.18%	1921	447397bp	0.27%
–OTHER:	25288	3682739bp	2.18%	1004	311451bp	0.18%	1921	447397bp	0.27%
Simple:	4161	194894bp	0.12%	21	952bp	0.00%	47	1280bp	0.00%
–repeat:	4161	194894bp	0.12%	21	952bp	0.00%	47	1280bp	0.00%
Unknown:	42126	4185948bp	2.48%	8897	1152166bp	0.68%	1829	536752bp	0.32%
–OTHER:	42126	4185948bp	2.48%	8897	1152166bp	0.68%	1829	536752bp	0.32%
rRNA:	32958	1783949bp	1.06%	1311	501598bp	0.30%	1069	382183bp	0.23%
–OTHER:	32958	1783949bp	1.06%	1311	501598bp	0.30%	1069	382183bp	0.23%
snRNA:	34	1869bp	0.00%	13	697bp	0.00%	0	0bp	0.00%
–OTHER:	34	1869bp	0.00%	13	697bp	0.00%	0	0bp	0.00%
tRNA:	937	23307bp	0.01%	244	10805bp	0.01%	30	1870bp	0.00%
–OTHER:	937	23307bp	0.01%	244	10805bp	0.01%	30	1870bp	0.00%

Table S22. Comparison of the proportion and detailed classification of detection results generated by three tools covering the repeat regions on the reference genome of Ant.

LongRepMarker				RepeatScout				RepeatModeler2			
sequence: 4339				**sequence: 4339**				**sequence: 4339**			
total length: 295944863bp				**total length: 295944863bp**				**total length: 295944863bp**			
bases masked: 11663344 bp (3.94%))				**bases masked: 8107022 bp (2.74%))**				**bases masked: 6083534 bp (2.06%))**			
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**		**Num of elements**	**Length occupied**	**Percentage of sequence**	
DNA:	61552	4928388bp	1.67%	21001	3243339bp	1.10%		13753	2645687bp	0.89%	
–Academ-1:	5	661bp	0.00%	5	195bp	0.00%		2	855bp	0.00%	
–CMC-Chapaev-3:	188	28829bp	0.01%	290	30211bp	0.01%		147	19438bp	0.01%	
–CMC-EnSpm:	150	44955bp	0.02%	312	29517bp	0.01%		506	81773bp	0.03%	
–CMC-Transib:	13	862bp	0.00%	47	7799bp	0.00%		33	8941bp	0.00%	
–Crypton-V:	40	4320bp	0.00%	6	441bp	0.00%		32	8088bp	0.00%	
–IS3EU:	8	3241bp	0.00%	0	0bp	0.00%		4	3048bp	0.00%	
–Kolobok:	0	0bp	0.00%	15	890bp	0.00%		0	0bp	0.00%	
–Kolobok-Hydra:	904	76209bp	0.03%	252	58495bp	0.02%		196	45029bp	0.02%	
–Kolobok-T2:	8759	585078bp	0.20%	1287	235738bp	0.08%		1316	197060bp	0.07%	
–MULE-NOF:	18	5612bp	0.00%	133	11398bp	0.00%		278	21158bp	0.01%	
–Maverick:	15263	860598bp	0.29%	3905	647332bp	0.22%		1608	366768bp	0.12%	
–Merlin:	69	21185bp	0.01%	236	13812bp	0.00%		236	34922bp	0.01%	
–MuLE-NOF:	4	2252bp	0.00%	0	0bp	0.00%		0	0bp	0.00%	
–MuLE-NOF?:	6	593bp	0.00%	0	0bp	0.00%		0	0bp	0.00%	
–OTHER:	661	179910bp	0.06%	865	107644bp	0.04%		1064	179133bp	0.06%	
–P:	34	2854bp	0.00%	49	3117bp	0.00%		10	661bp	0.00%	
–PIF-Harbinger:	15	1743bp	0.00%	10	579bp	0.00%		30	8444bp	0.00%	
–PIF-ISL2EU:	0	0bp	0.00%	12	2164bp	0.00%		0	0bp	0.00%	
–PIF-Spy:	18	1991bp	0.00%	88	6472bp	0.00%		70	10680bp	0.00%	
–PiggyBac:	63	8101bp	0.00%	40	12797bp	0.00%		63	13186bp	0.00%	
–TcMar:	343	28338bp	0.01%	212	21515bp	0.01%		112	17965bp	0.01%	
–TcMar-Cweed:	31	4601bp	0.00%	13	3442bp	0.00%		20	5722bp	0.00%	
–TcMar-Fot1:	24	14562bp	0.00%	75	7535bp	0.00%		106	26638bp	0.01%	
–TcMar-Mariner:	22053	2085501bp	0.70%	8170	1347878bp	0.46%		4977	873158bp	0.30%	
–TcMar-Tc1:	11262	892895bp	0.30%	3641	543181bp	0.18%		2075	527674bp	0.18%	
–TcMar-Tc2:	0	0bp	0.00%	12	1208bp	0.00%		0	0bp	0.00%	
–TcMar-Tc4:	1184	63484bp	0.02%	562	80535bp	0.03%		351	82578bp	0.03%	
–TcMar-Tigger:	5	434bp	0.00%	0	0bp	0.00%		5	361bp	0.00%	
–hAT:	79	9856bp	0.00%	12	4035bp	0.00%		45	13936bp	0.00%	
–hAT-Ac:	47	19373bp	0.01%	78	6889bp	0.00%		132	22484bp	0.01%	
–hAT-Blackjack:	210	19502bp	0.01%	538	47604bp	0.02%		171	44496bp	0.02%	
–hAT-Charlie:	70	28741bp	0.01%	99	10070bp	0.00%		65	17356bp	0.01%	
–hAT-Pegasus:	12	7622bp	0.00%	0	0bp	0.00%		0	0bp	0.00%	
–hAT-Tip100:	14	6534bp	0.00%	20	1961bp	0.00%		25	6450bp	0.00%	
–hAT-Tol2:	0	0bp	0.00%	0	0bp	0.00%		6	1256bp	0.00%	
–hAT-hAT5:	0	0bp	0.00%	17	2873bp	0.00%		49	4665bp	0.00%	
–hAT-hATm:	0	0bp	0.00%	0	0bp	0.00%		19	2698bp	0.00%	
LINE:	9011	812146bp	0.27%	3027	588373bp	0.20%		1216	486742bp	0.16%	
–CR1:	20	19115bp	0.01%	18	6510bp	0.00%		0	0bp	0.00%	
–I:	192	19925bp	0.01%	248	69549bp	0.02%		32	58207bp	0.02%	
–I-Jockey:	9	4831bp	0.00%	37	6688bp	0.00%		0	0bp	0.00%	
–I-Nimb:	8	2342bp	0.00%	13	2880bp	0.00%		0	0bp	0.00%	
–Jockey:	12	1190bp	0.00%	99	13519bp	0.00%		17	10396bp	0.00%	
–L1-Tx1:	4	629bp	0.00%	0	0bp	0.00%		0	0bp	0.00%	
–L2:	91	6760bp	0.00%	133	16803bp	0.01%		75	42141bp	0.01%	
–LOA:	17	1255bp	0.00%	15	662bp	0.00%		16	4685bp	0.00%	
–OTHER:	13	3399bp	0.00%	78	21893bp	0.01%		0	0bp	0.00%	
–Penelope:	6790	664008bp	0.22%	1222	308625bp	0.10%		826	232478bp	0.08%	
–R1:	1721	78389bp	0.03%	770	109806bp	0.04%		190	88219bp	0.03%	
–R1-LOA:	7	929bp	0.00%	0	0bp	0.00%		0	0bp	0.00%	
–R2:	0	0bp	0.00%	18	772bp	0.00%		0	0bp	0.00%	
–R2-NeSL:	36	2863bp	0.00%	204	15207bp	0.01%		11	7645bp	0.00%	
–RTE-BovB:	0	0bp	0.00%	18	480bp	0.00%		16	6982bp	0.00%	
–RTE-X:	91	9335bp	0.00%	154	24814bp	0.01%		33	35992bp	0.01%	
LTR:	18197	1058147bp	0.36%	6199	1143348bp	0.39%		1531	646274bp	0.22%	
–Copia:	1236	154160bp	0.05%	819	103383bp	0.03%		82	36843bp	0.01%	
–DIRS:	20	2470bp	0.00%	38	2339bp	0.00%		0	0bp	0.00%	
–ERV1:	7	1010bp	0.00%	0	0bp	0.00%		0	0bp	0.00%	
–Gypsy:	10849	606715bp	0.21%	3330	627049bp	0.21%		1119	399530bp	0.14%	
–Gypsy-Cigr:	3	217bp	0.00%	25	6012bp	0.00%		10	15761bp	0.01%	
–Ngaro:	0	0bp	0.00%	0	0bp	0.00%		8	536bp	0.00%	
–OTHER:	26	3672bp	0.00%	6	794bp	0.00%		0	0bp	0.00%	
–Pao:	6056	296097bp	0.10%	1981	405227bp	0.14%		312	193604bp	0.07%	
RC:	875	61951bp	0.02%	755	96732bp	0.03%		242	54578bp	0.02%	
–Helitron:	875	61951bp	0.02%	755	96732bp	0.03%		242	54578bp	0.02%	
RC?:	34	2757bp	0.00%	41	7518bp	0.00%		57	10149bp	0.00%	
–Helitron:	34	2757bp	0.00%	41	7518bp	0.00%		57	10149bp	0.00%	
SINE:	36	4021bp	0.00%	91	10888bp	0.00%		8	593bp	0.00%	
–5S-Deu-L2:	16	1836bp	0.00%	31	1429bp	0.00%		0	0bp	0.00%	
–ID:	13	507bp	0.00%	0	0bp	0.00%		0	0bp	0.00%	
–U:	4	631bp	0.00%	0	0bp	0.00%		0	0bp	0.00%	
–tRNA-Deu:	0	0bp	0.00%	0	0bp	0.00%		8	593bp	0.00%	
–tRNA-Deu-RTE:	0	0bp	0.00%	32	2021bp	0.00%		0	0bp	0.00%	
–tRNA-RTE:	3	1047bp	0.00%	28	7438bp	0.00%		0	0bp	0.00%	
Satellite:	8	510bp	0.00%	24	1801bp	0.00%		84	26817bp	0.01%	
–OTHER:	8	510bp	0.00%	24	1801bp	0.00%		84	26817bp	0.01%	
Simple:	2789	118800bp	0.04%	134	5618bp	0.00%		233	9291bp	0.00%	
–repeat:	2789	118800bp	0.04%	134	5618bp	0.00%		233	9291bp	0.00%	
Unknown:	90822	5172819bp	1.75%	37373	3080906bp	1.04%		16856	2217117bp	0.75%	
–OTHER:	90822	5172819bp	1.75%	37373	3080906bp	1.04%		16856	2217117bp	0.75%	
rRNA:	10	651bp	0.00%	9	820bp	0.00%		0	0bp	0.00%	
–OTHER:	10	651bp	0.00%	9	820bp	0.00%		0	0bp	0.00%	
tRNA:	24	813bp	0.00%	41	1404bp	0.00%		0	0bp	0.00%	
–OTHER:	24	813bp	0.00%	41	1404bp	0.00%		0	0bp	0.00%	

Table S23. Comparison of the proportion and detailed classification of detection results generated by three tools covering the repeat regions on the reference genome of Gallus.

LongRepMarker				RepeatScout			RepeatModeler2		
sequence: 464				**sequence: 464**			**sequence: 464**		
total length: 1065365434bp				**total length: 1065365434bp**			**total length: 1065365434bp**		
bases masked: 42433541 bp (3.98%))				**bases masked: 20799675 bp (1.95%))**			**bases masked: 5536215 bp (0.52%))**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	3822	1511276bp	0.14%	5031	912863bp	0.09%	1140	181739bp	0.02%
–Academ-1:	23	2714bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–CMC-EnSpm:	78	25210bp	0.00%	72	132192bp	0.01%	0	0bp	0.00%
–Crypton-H:	33	2518bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton-V:	8	6263bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Ginger:	6	3423bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Kolobok:	8	13471bp	0.00%	0	0bp	0.00%	1	84bp	0.00%
–Kolobok-T2:	4	10886bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MULE-MuDR:	283	89603bp	0.01%	6	1500bp	0.00%	0	0bp	0.00%
–Maverick:	262	114224bp	0.01%	41	75620bp	0.01%	0	0bp	0.00%
–Merlin:	0	0bp	0.00%	0	0bp	0.00%	1	72bp	0.00%
–OTHER:	317	247749bp	0.02%	66	80479bp	0.01%	48	69962bp	0.01%
–P:	15	2610bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PIF-Harbinger:	302	107185bp	0.01%	0	0bp	0.00%	0	0bp	0.00%
–Sola-1:	502	22630bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Sola-3:	150	115060bp	0.01%	37	51791bp	0.00%	0	0bp	0.00%
–TcMar-Fot1:	12	1689bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-ISRm11:	2	1051bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Mariner:	808	403624bp	0.04%	3758	353035bp	0.03%	613	32204bp	0.00%
–TcMar-Tc1:	38	69659bp	0.01%	153	47006bp	0.00%	0	0bp	0.00%
–TcMar-Tc2:	26	16269bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Zisupton:	4	316bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Ac:	390	70181bp	0.01%	12	21691bp	0.00%	98	7270bp	0.00%
–hAT-Charlie:	482	199204bp	0.02%	856	170526bp	0.02%	309	29345bp	0.00%
–hAT-Pegasus:	34	3633bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Tag1:	32	25488bp	0.00%	0	0bp	0.00%	70	43049bp	0.00%
–hAT-Tip100:	3	1764bp	0.00%	30	26124bp	0.00%	0	0bp	0.00%
LINE:	56995	18851002bp	1.77%	21417	9128535bp	0.86%	5733	1644540bp	0.15%
–CR1:	55670	18534186bp	1.74%	20926	8781356bp	0.82%	5637	1581687bp	0.15%
–CRE:	14	5276bp	0.00%	27	7600bp	0.00%	0	0bp	0.00%
–Dualen:	8	1107bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–I:	5	3620bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–I-Jockey:	38	12045bp	0.00%	35	16396bp	0.00%	0	0bp	0.00%
–Jockey:	6	809bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–L1:	149	72778bp	0.01%	44	14629bp	0.00%	0	0bp	0.00%
–L1-Tx1:	3	1078bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–L2:	91	39329bp	0.00%	8	3299bp	0.00%	0	0bp	0.00%
–OTHER:	36	63317bp	0.01%	30	53672bp	0.01%	0	0bp	0.00%
–Penelope:	84	27397bp	0.00%	12	11438bp	0.00%	1	68bp	0.00%
–R1:	18	5410bp	0.00%	6	1848bp	0.00%	0	0bp	0.00%
–R1-LOA:	5	708bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–R2:	108	44374bp	0.00%	13	10243bp	0.00%	0	0bp	0.00%
–R2-NeSL:	0	0bp	0.00%	32	3639bp	0.00%	0	0bp	0.00%
–RTE-BovB:	621	190155bp	0.02%	246	148447bp	0.01%	68	33152bp	0.00%
–RTE-RTE:	4	3340bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–RTE-X:	20	806bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Rex-Babar:	21	7980bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Tad1:	94	46982bp	0.00%	38	108756bp	0.01%	27	29633bp	0.00%
LTR:	43275	11135452bp	1.05%	13471	6181751bp	0.58%	4671	1667393bp	0.16%
–Caulimovirus:	2	144bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Copia:	327	175566bp	0.02%	157	44498bp	0.00%	18	3003bp	0.00%
–DIRS:	44	32002bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERV:	106	296823bp	0.03%	66	17155bp	0.00%	0	0bp	0.00%
–ERV1:	10494	2734741bp	0.26%	2796	1434650bp	0.13%	596	354862bp	0.03%
–ERVK:	4556	1884644bp	0.18%	2169	1111518bp	0.10%	387	371455bp	0.03%
–ERVL:	26177	6443868bp	0.60%	7655	3383117bp	0.32%	3567	813946bp	0.08%
–Gypsy:	1098	559901bp	0.05%	377	280499bp	0.03%	96	81314bp	0.01%
–Ngaro:	118	82877bp	0.01%	115	88812bp	0.01%	0	0bp	0.00%
–OTHER:	102	49259bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Pao:	186	48475bp	0.00%	121	47928bp	0.00%	7	42854bp	0.00%
–Viper:	65	5407bp	0.00%	15	3421bp	0.00%	0	0bp	0.00%
RC:	306	156071bp	0.01%	0	0bp	0.00%	0	0bp	0.00%
–Helitron:	306	156071bp	0.01%	0	0bp	0.00%	0	0bp	0.00%
SINE:	1305	106966bp	0.01%	171	94937bp	0.01%	32	18473bp	0.00%
–5S:	1146	65107bp	0.01%	27	71939bp	0.01%	0	0bp	0.00%
–5S-Deu-L2:	47	5513bp	0.00%	9	1712bp	0.00%	1	51bp	0.00%
–5S-Sauria-RTE:	33	3924bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Alu:	4	672bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ID:	5	746bp	0.00%	31	9037bp	0.00%	17	780bp	0.00%
–MIR:	8	2034bp	0.00%	0	0bp	0.00%	3	259bp	0.00%
–U:	42	2643bp	0.00%	16	2405bp	0.00%	0	0bp	0.00%
–tRNA:	20	34709bp	0.00%	56	3977bp	0.00%	11	17383bp	0.00%
–tRNA-RTE:	0	0bp	0.00%	15	666bp	0.00%	0	0bp	0.00%
–tRNA-Sauria-L2:	0	0bp	0.00%	17	5201bp	0.00%	0	0bp	0.00%
SINE?:	4	23697bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	4	23697bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Satellite:	52126	5494345bp	0.52%	1073	1000942bp	0.09%	918	312522bp	0.03%
–OTHER:	4012	954380bp	0.09%	483	354123bp	0.03%	657	90115bp	0.01%
–W-chromosome:	39006	3152502bp	0.30%	150	83642bp	0.01%	118	58903bp	0.01%
–macro:	9035	1414666bp	0.13%	440	565025bp	0.05%	143	163504bp	0.02%
–telomeric:	73	3671bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Simple:	17471	1919486bp	0.18%	148	49253bp	0.00%	600	28340bp	0.00%
–repeat:	17471	1919486bp	0.18%	148	49253bp	0.00%	600	28340bp	0.00%
Unknown:	164468	14768863bp	1.39%	25911	5449676bp	0.51%	3434	1710324bp	0.16%
–OTHER:	164468	14768863bp	1.39%	25911	5449676bp	0.51%	3434	1710324bp	0.16%
rRNA:	120	44087bp	0.00%	104	4719bp	0.00%	0	0bp	0.00%
–OTHER:	120	44087bp	0.00%	104	4719bp	0.00%	0	0bp	0.00%
snRNA:	25	1699bp	0.00%	39	2858bp	0.00%	0	0bp	0.00%
–OTHER:	25	1699bp	0.00%	39	2858bp	0.00%	0	0bp	0.00%
tRNA:	245	32157bp	0.00%	209	19120bp	0.00%	0	0bp	0.00%
–OTHER:	245	32157bp	0.00%	209	19120bp	0.00%	0	0bp	0.00%

Table S24. Comparison of the proportion and detailed classification of detection results generated by three tools covering the repeat regions on the reference genome of Soybean.

LongRepMarker				RepeatScout			RepeatModeler2		
sequence: 1192				**sequence: 1192**			**sequence: 1192**		
total length: 979046046bp				**total length: 979046046bp**			**total length: 979046046bp**		
bases masked: 114369952 bp (11.68%))				**bases masked: 71264358 bp (7.28%))**			**bases masked: 34575219 bp (3.53%))**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	106809	14542224bp	1.49%	20051	10557566bp	1.08%	7539	3646588bp	0.37%
–Academ:	5	459bp	0.00%	4	548bp	0.00%	0	0bp	0.00%
–CMC-Chapaev-3:	0	0bp	0.00%	7	701bp	0.00%	0	0bp	0.00%
–CMC-EnSpm:	38904	5942067bp	0.61%	7066	4754427bp	0.49%	2202	1635440bp	0.17%
–Crypton-S:	0	0bp	0.00%	0	0bp	0.00%	50	11078bp	0.00%
–Dada:	0	0bp	0.00%	9	3312bp	0.00%	0	0bp	0.00%
–Ginger:	9	952bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Kolobok-Hydra:	4	9588bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Kolobok-T2:	0	0bp	0.00%	30	5822bp	0.00%	0	0bp	0.00%
–MULE-MuDR:	46960	4916688bp	0.50%	6121	2079584bp	0.21%	2484	536042bp	0.05%
–Maverick:	23	27518bp	0.00%	17	6160bp	0.00%	6	3296bp	0.00%
–MuLE-MuDR:	7112	2351881bp	0.24%	2161	1913232bp	0.20%	1230	598916bp	0.06%
–OTHER:	417	159727bp	0.02%	133	22098bp	0.00%	82	34876bp	0.00%
–P:	4	2481bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PIF-Harbinger:	2649	605943bp	0.06%	912	438339bp	0.04%	203	173191bp	0.02%
–PIF-ISL2EU:	4	1727bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PiggyBac:	8	4081bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PiggyBac-X:	20	25786bp	0.00%	39	8935bp	0.00%	0	0bp	0.00%
–Sola-2:	4	777bp	0.00%	30	7437bp	0.00%	0	0bp	0.00%
–TcMar-Ant1:	0	0bp	0.00%	9	643bp	0.00%	0	0bp	0.00%
–TcMar-Mariner:	4	4551bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Pogo:	639	34588bp	0.00%	72	54816bp	0.01%	30	50049bp	0.01%
–TcMar-Sagan:	4	615bp	0.00%	7	1307bp	0.00%	0	0bp	0.00%
–TcMar-Stowaway:	257	34581bp	0.00%	111	38163bp	0.00%	87	7277bp	0.00%
–TcMar-Tc1:	4	995bp	0.00%	20	53065bp	0.01%	0	0bp	0.00%
–TcMar-Tc2:	4	1260bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Zisupton:	0	0bp	0.00%	16	3683bp	0.00%	0	0bp	0.00%
–hAT:	0	0bp	0.00%	0	0bp	0.00%	3	1238bp	0.00%
–hAT-Ac:	4312	567858bp	0.06%	1442	539494bp	0.06%	526	227784bp	0.02%
–hAT-Charlie:	47	72861bp	0.01%	89	55172bp	0.01%	7	841bp	0.00%
–hAT-Tag1:	3393	609331bp	0.06%	1102	471859bp	0.05%	462	246197bp	0.03%
–hAT-Tip100:	2016	326734bp	0.03%	654	311843bp	0.03%	167	125997bp	0.01%
–hAT-hATm:	6	922bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LINE:	37045	6097857bp	0.62%	7344	2716876bp	0.28%	1131	484156bp	0.05%
–CR1:	8	1506bp	0.00%	4	292bp	0.00%	0	0bp	0.00%
–I:	28	36121bp	0.00%	10	643bp	0.00%	0	0bp	0.00%
–I-Jockey:	32	14094bp	0.00%	24	17633bp	0.00%	0	0bp	0.00%
–L1:	29043	4557691bp	0.47%	6382	2450341bp	0.25%	768	302687bp	0.03%
–L1-DRE:	3	980bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–L1-Tx1:	8	13550bp	0.00%	18	1742bp	0.00%	0	0bp	0.00%
–L2:	60	26784bp	0.00%	41	12442bp	0.00%	0	0bp	0.00%
–Penelope:	17	1300bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–R1:	4	891bp	0.00%	79	9697bp	0.00%	0	0bp	0.00%
–RTE-BovB:	7834	1442598bp	0.15%	743	219475bp	0.02%	363	181469bp	0.02%
–RTE-X:	8	2343bp	0.00%	23	1784bp	0.00%	0	0bp	0.00%
–Tad1:	0	0bp	0.00%	20	7517bp	0.00%	0	0bp	0.00%
LTR:	295825	61894913bp	6.32%	63182	37377343bp	3.82%	39408	26440652bp	2.70%
–Cassandra:	101	2593bp	0.00%	172	56698bp	0.01%	20	44852bp	0.00%
–Caulimovirus:	11692	1038067bp	0.11%	677	563737bp	0.06%	162	274336bp	0.03%
–Copia:	110672	24312006bp	2.48%	26274	14979052bp	1.53%	14652	7864320bp	0.80%
–DIRS:	0	0bp	0.00%	7	841bp	0.00%	0	0bp	0.00%
–ERV1:	60	85955bp	0.01%	115	108958bp	0.01%	3	3467bp	0.00%
–ERV4:	4	5441bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERVK:	27	3588bp	0.00%	9	447bp	0.00%	4	2719bp	0.00%
–ERVL:	4	1868bp	0.00%	6	366bp	0.00%	0	0bp	0.00%
–Gypsy:	172762	36628195bp	3.74%	35605	21647150bp	2.21%	24478	18348200bp	1.87%
–Ngaro:	8	1217bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	422	93749bp	0.01%	231	41367bp	0.00%	70	12408bp	0.00%
–Pao:	73	39864bp	0.00%	86	54808bp	0.01%	19	55790bp	0.01%
RC:	6109	1139535bp	0.12%	2412	1621569bp	0.17%	309	295295bp	0.03%
–Helitron:	6109	1139535bp	0.12%	2412	1621569bp	0.17%	309	295295bp	0.03%
RC?:	7	10011bp	0.00%	7	1218bp	0.00%	0	0bp	0.00%
–Helitron:	7	10011bp	0.00%	7	1218bp	0.00%	0	0bp	0.00%
Retroposon:	21	8073bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	21	8073bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
SINE:	1330	172682bp	0.02%	116	15489bp	0.00%	63	26309bp	0.00%
–ID:	29	1449bp	0.00%	30	1901bp	0.00%	2	116bp	0.00%
–OTHER:	14	2977bp	0.00%	0	0bp	0.00%	22	1975bp	0.00%
–tRNA:	16	5194bp	0.00%	0	0bp	0.00%	5	6402bp	0.00%
–tRNA-RTE:	1271	167864bp	0.02%	86	13588bp	0.00%	34	17816bp	0.00%
SINE?:	9	2128bp	0.00%	0	0bp	0.00%	9	1516bp	0.00%
–OTHER:	9	2128bp	0.00%	0	0bp	0.00%	9	1516bp	0.00%
Satellite:	1261	176627bp	0.02%	608	54962bp	0.01%	106	83562bp	0.01%
–OTHER:	1253	166304bp	0.02%	608	54962bp	0.01%	106	83562bp	0.01%
–centromeric:	8	10323bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Simple:	16713	731173bp	0.07%	331	18706bp	0.00%	69	2855bp	0.00%
–repeat:	16713	731173bp	0.07%	331	18706bp	0.00%	69	2855bp	0.00%
Unknown:	495250	35074975bp	3.58%	170755	20649747bp	2.11%	10664	2827024bp	0.29%
–OTHER:	495250	35074975bp	3.58%	170755	20649747bp	2.11%	10664	2827024bp	0.29%
rRNA:	28544	854797bp	0.09%	70	189019bp	0.02%	88	837906bp	0.09%
–OTHER:	28544	854797bp	0.09%	70	189019bp	0.02%	88	837906bp	0.09%
snRNA:	88	3651bp	0.00%	92	4331bp	0.00%	30	3285bp	0.00%
–OTHER:	88	3651bp	0.00%	92	4331bp	0.00%	30	3285bp	0.00%
tRNA:	644	33897bp	0.00%	662	40603bp	0.00%	0	0bp	0.00%
–OTHER:	644	33897bp	0.00%	662	40603bp	0.00%	0	0bp	0.00%

Table S25. Comparison of the proportion and detailed classification of detection results generated by three tools covering the repeat regions on the reference genome of Human(hg38).

LongRepMarker				RepeatScout			RepeatModeler2		
sequence: 455				**sequence: 455**			**sequence: 455**		
total length: 3209286105bp				**total length: 3209286105bp**			**total length: 3209286105bp**		
bases masked: 246167245 bp (7.67%))				**bases masked: 115637763 bp (3.60%))**			**bases masked: 8446785 bp (0.26%))**		
Repeat Types	Num of elements	Length occupied	Percentage of sequence	Num of elements	Length occupied	Percentage of sequence	Num of elements	Length occupied	Percentage of sequence
DNA:	42345	5122539bp	0.16%	28858	4622136bp	0.14%	4291	759693bp	0.02%
–Academ-1:	197	120305bp	0.00%	133	178620bp	0.01%	0	0bp	0.00%
–CMC-EnSpm:	2049	179617bp	0.01%	126	27771bp	0.00%	0	0bp	0.00%
–Crypton-A:	25	5146bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton-H:	16	1216bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton-S:	5	2162bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton-V:	6	438bp	0.00%	35	14606bp	0.00%	0	0bp	0.00%
–Dada:	50	14384bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Ginger:	138	13870bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–IS3EU:	18	4983bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Kolobok:	77	14315bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Kolobok-T2:	51	8646bp	0.00%	141	43242bp	0.00%	0	0bp	0.00%
–MULE-MuDR:	334	45856bp	0.00%	409	76115bp	0.00%	59	4252bp	0.00%
–MULE-NOF:	10	2764bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Maverick:	58	7480bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Merlin:	0	0bp	0.00%	1	41bp	0.00%	0	0bp	0.00%
–MuLE-MuDR:	16	802bp	0.00%	64	6936bp	0.00%	42	9822bp	0.00%
–MuLE-NOF:	9	950bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Novosib:	26	8595bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	97	40966bp	0.00%	38	16849bp	0.00%	4	156bp	0.00%
–P:	10	841bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PIF-Harbinger:	34	10053bp	0.00%	46	4239bp	0.00%	0	0bp	0.00%
–PIF-Spy:	4	226bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PiggyBac:	681	40472bp	0.00%	518	41542bp	0.00%	53	10810bp	0.00%
–PiggyBac-X:	3	180bp	0.00%	16	2088bp	0.00%	0	0bp	0.00%
–Sola:	3	143bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Sola-1:	12	394bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Sola-2:	18	978bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Sola-3:	29	10311bp	0.00%	32	3038bp	0.00%	0	0bp	0.00%
–TcMar-Fot1:	14	2050bp	0.00%	9	948bp	0.00%	0	0bp	0.00%
–TcMar-ISRm11:	24	6061bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Mariner:	4232	243778bp	0.01%	713	187883bp	0.01%	78	16514bp	0.00%
–TcMar-Tc1:	51	24889bp	0.00%	36	38040bp	0.00%	1	332bp	0.00%
–TcMar-Tc2:	74	9687bp	0.00%	172	40049bp	0.00%	10	1015bp	0.00%
–TcMar-Tigger:	25728	1744247bp	0.05%	13632	1818789bp	0.06%	2062	494046bp	0.02%
–Zisupton:	313	195919bp	0.01%	62	15979bp	0.00%	0	0bp	0.00%
–hAT:	35	21304bp	0.00%	37	1964bp	0.00%	3	415bp	0.00%
–hAT-Ac:	68	25503bp	0.00%	50	128670bp	0.00%	4	941bp	0.00%
–hAT-Blackjack:	168	26762bp	0.00%	452	40453bp	0.00%	109	9503bp	0.00%
–hAT-Charlie:	6922	1939181bp	0.06%	10910	1446177bp	0.05%	1659	184073bp	0.01%
–hAT-Pegasus:	8	353bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Tag1:	20	7131bp	0.00%	14	14793bp	0.00%	0	0bp	0.00%
–hAT-Tip100:	702	357294bp	0.01%	1212	485752bp	0.02%	207	27991bp	0.00%
–hAT-Tip100?:	4	1312bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-hATm:	6	2414bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
DNA?:	0	0bp	0.00%	15	2951bp	0.00%	0	0bp	0.00%
–hAT-Tip100:	0	0bp	0.00%	15	2951bp	0.00%	0	0bp	0.00%
LINE:	253482	116091245bp	3.62%	91505	40916219bp	1.27%	9309	4076749bp	0.13%
–CR1:	307	211040bp	0.01%	303	229534bp	0.01%	8	1211bp	0.00%
–CRE:	8	6604bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–I:	22	15578bp	0.00%	45	105043bp	0.00%	0	0bp	0.00%
–I-Jockey:	8	373bp	0.00%	33	9887bp	0.00%	0	0bp	0.00%
–I-Nimb:	0	0bp	0.00%	6	870bp	0.00%	0	0bp	0.00%
–Jockey:	8	749bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–L1:	249969	113923085bp	3.55%	87289	39160105bp	1.22%	9124	4042496bp	0.13%
–L1-Tx1:	78	27322bp	0.00%	109	26170bp	0.00%	0	0bp	0.00%
–L2:	2597	1413646bp	0.04%	3300	1277255bp	0.04%	149	18423bp	0.00%
–OTHER:	134	1938577bp	0.06%	95	175321bp	0.01%	0	0bp	0.00%
–Penelope:	18	4355bp	0.00%	6	10617bp	0.00%	0	0bp	0.00%
–R1:	75	49254bp	0.00%	19	10667bp	0.00%	0	0bp	0.00%
–R1-LOA:	4	1556bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–R2:	37	15809bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–R2-Hero:	0	0bp	0.00%	8	1394bp	0.00%	0	0bp	0.00%
–R2-NeSL:	0	0bp	0.00%	0	0bp	0.00%	12	8287bp	0.00%
–RTE-BovB:	137	33397bp	0.00%	204	81103bp	0.00%	13	6010bp	0.00%
–RTE-X:	68	25859bp	0.00%	88	30551bp	0.00%	3	322bp	0.00%
–Rex-Babar:	12	2544bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LTR:	233948	21644832bp	0.67%	142609	20851756bp	0.65%	12774	2398409bp	0.07%
–Caulimovirus:	15	3159bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Copia:	274	82430bp	0.00%	281	68939bp	0.00%	17	14913bp	0.00%
–DIRS:	83	84934bp	0.00%	30	8379bp	0.00%	0	0bp	0.00%
–ERV:	51	9043bp	0.00%	166	103236bp	0.00%	0	0bp	0.00%
–ERV1:	137547	9924078bp	0.31%	66494	10267624bp	0.32%	6493	1390674bp	0.04%
–ERV4:	12	13385bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERVK:	27750	3174763bp	0.10%	9972	3052567bp	0.10%	1316	502784bp	0.02%
–ERVL:	30329	4440975bp	0.14%	19263	3886403bp	0.12%	1839	287348bp	0.01%
–ERVL-MaLR:	36245	3745971bp	0.12%	45557	3789981bp	0.12%	3108	204839bp	0.01%
–Gypsy:	1427	430273bp	0.01%	445	143225bp	0.00%	0	0bp	0.00%
–Ngaro:	18	7415bp	0.00%	233	13345bp	0.00%	0	0bp	0.00%
–OTHER:	139	22011bp	0.00%	125	37855bp	0.00%	1	30bp	0.00%
–Pao:	50	9434bp	0.00%	31	5272bp	0.00%	0	0bp	0.00%
–Viper:	8	768bp	0.00%	12	1624bp	0.00%	0	0bp	0.00%
Other:	31	4060bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	27	1457bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–subtelomeric:	4	2603bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
RC:	232	66283bp	0.00%	45	15199bp	0.00%	0	0bp	0.00%
–Helitron:	232	66283bp	0.00%	45	15199bp	0.00%	0	0bp	0.00%
RC?:	0	0bp	0.00%	29	2460bp	0.00%	0	0bp	0.00%
–Helitron:	0	0bp	0.00%	29	2460bp	0.00%	0	0bp	0.00%
RNA:	59	2808bp	0.00%	77	4050bp	0.00%	12	1584bp	0.00%
–OTHER:	59	2808bp	0.00%	77	4050bp	0.00%	12	1584bp	0.00%
Retroposon:	3377	1270842bp	0.04%	526	308116bp	0.01%	97	3762bp	0.00%
–SVA:	3377	1270842bp	0.04%	526	308116bp	0.01%	97	3762bp	0.00%
SINE:	129729	75310083bp	2.35%	319830	51254750bp	1.60%	5728	530547bp	0.02%
–5S:	89	26210bp	0.00%	112	6575bp	0.00%	35	2327bp	0.00%
–5S-Deu-L2:	0	0bp	0.00%	6	657bp	0.00%	0	0bp	0.00%
–5S-Sauria-RTE:	0	0bp	0.00%	11	2942bp	0.00%	0	0bp	0.00%
–Alu:	126088	74205313bp	2.31%	314932	50079001bp	1.56%	5307	433360bp	0.01%
–ID:	0	0bp	0.00%	101	7308bp	0.00%	0	0bp	0.00%
–MIR:	3456	1439502bp	0.04%	4520	1625596bp	0.05%	371	94663bp	0.00%
–U:	62	14456bp	0.00%	82	25826bp	0.00%	15	552bp	0.00%
–tRNA:	0	0bp	0.00%	8	1428bp	0.00%	0	0bp	0.00%
–tRNA-7SL:	16	1411bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–tRNA-Core-L2:	4	184bp	0.00%	16	2392bp	0.00%	0	0bp	0.00%
–tRNA-Core-RTE:	0	0bp	0.00%	13	680bp	0.00%	0	0bp	0.00%
–tRNA-Meta:	10	1898bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–tRNA-RTE:	4	179bp	0.00%	29	7788bp	0.00%	0	0bp	0.00%
SINE?:	16	21177bp	0.00%	15	16871bp	0.00%	0	0bp	0.00%
–OTHER:	16	21177bp	0.00%	15	16871bp	0.00%	0	0bp	0.00%
Satellite:	1339490	51375051bp	1.60%	2623	921845bp	0.03%	350	102667bp	0.00%
–OTHER:	8157	1392631bp	0.04%	1061	454086bp	0.01%	143	52327bp	0.00%
–Y-chromosome:	321517	24008707bp	0.75%	93	30431bp	0.00%	56	6113bp	0.00%
–acromeric:	203	91137bp	0.00%	467	143680bp	0.00%	0	0bp	0.00%
–centromeric:	1008827	43052149bp	1.34%	588	253744bp	0.01%	151	44227bp	0.00%
–telomeric:	786	192714bp	0.01%	414	60934bp	0.00%	0	0bp	0.00%
Simple:	19483	1534323bp	0.05%	434	109413bp	0.00%	199	12847bp	0.00%
–repeat:	19483	1534323bp	0.05%	434	109413bp	0.00%	199	12847bp	0.00%
Unknown:	158348	38949784bp	1.21%	87921	19053503bp	0.59%	1103	567000bp	0.02%
–OTHER:	158348	38949784bp	1.21%	87921	19053503bp	0.59%	1103	567000bp	0.02%
rRNA:	256	98971bp	0.00%	460	62804bp	0.00%	0	0bp	0.00%
–OTHER:	256	98971bp	0.00%	460	62804bp	0.00%	0	0bp	0.00%
scRNA:	35	3213bp	0.00%	247	22162bp	0.00%	16	1016bp	0.00%
–OTHER:	35	3213bp	0.00%	247	22162bp	0.00%	16	1016bp	0.00%
snRNA:	140	46208bp	0.00%	329	24052bp	0.00%	39	2378bp	0.00%
–OTHER:	140	46208bp	0.00%	329	24052bp	0.00%	39	2378bp	0.00%
tRNA:	235	19643bp	0.00%	498	54895bp	0.00%	0	0bp	0.00%
–OTHER:	235	19643bp	0.00%	498	54895bp	0.00%	0	0bp	0.00%

Table S26. Comparison of the proportion and detailed classification of detection results generated by three tools covering the repeat regions on the reference genome of Mouse.

LongRepMarker				RepeatScout			RepeatModeler2		
sequence: 239				**sequence: 239**			**sequence: 239**		
total length: 2818974548bp				**total length: 2818974548bp**			**total length: 2818974548bp**		
bases masked: 260838424 bp (9.25%))				**bases masked: 127719817 bp (4.53%))**			**bases masked: 28407842 bp (1.01%))**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	12442	2962978bp	0.11%	8860	1567808bp	0.06%	1038	285958bp	0.01%
–Academ:	36	2694bp	0.00%	39	4198bp	0.00%	0	0bp	0.00%
–CMC-Chapaev:	18	1331bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–CMC-Chapaev-3:	0	0bp	0.00%	16	2900bp	0.00%	0	0bp	0.00%
–CMC-EnSpm:	1327	324764bp	0.01%	225	98766bp	0.00%	0	0bp	0.00%
–Crypton:	3	7046bp	0.00%	33	12389bp	0.00%	0	0bp	0.00%
–Crypton-A:	8	2275bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton-F:	4	556bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton-H:	31	2916bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton-S:	24	1936bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton-V:	93	21394bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Dada:	88	8493bp	0.00%	16	1812bp	0.00%	0	0bp	0.00%
–Ginger:	187	32591bp	0.00%	56	15635bp	0.00%	0	0bp	0.00%
–IS3EU:	88	17399bp	0.00%	36	8126bp	0.00%	0	0bp	0.00%
–Kolobok-Hydra:	6	10035bp	0.00%	29	4557bp	0.00%	0	0bp	0.00%
–Kolobok-T2:	322	40466bp	0.00%	12	3750bp	0.00%	0	0bp	0.00%
–MULE:	12	1097bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MULE-MuDR:	252	54222bp	0.00%	219	31970bp	0.00%	19	8432bp	0.00%
–MULE-NOF:	4	3892bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Maverick:	92	14315bp	0.00%	93	24063bp	0.00%	0	0bp	0.00%
–MuLE-MuDR:	48	5816bp	0.00%	35	4343bp	0.00%	0	0bp	0.00%
–MuLE-NOF:	8	249bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MuLE-NOF?:	5	434bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Novosib:	37	7666bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	475	89064bp	0.00%	87	117282bp	0.00%	0	0bp	0.00%
–P:	7	146bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PIF-Harbinger:	179	24980bp	0.00%	117	28806bp	0.00%	0	0bp	0.00%
–PIF-ISL2EU:	61	5336bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PiggyBac:	23	17800bp	0.00%	66	23944bp	0.00%	0	0bp	0.00%
–PiggyBac-X:	0	0bp	0.00%	11	1482bp	0.00%	0	0bp	0.00%
–Sola-1:	28	1870bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Sola-2:	13	8115bp	0.00%	14	3446bp	0.00%	0	0bp	0.00%
–Sola-3:	56	10163bp	0.00%	14	2876bp	0.00%	0	0bp	0.00%
–TcMar-Ant1:	0	0bp	0.00%	8	1080bp	0.00%	0	0bp	0.00%
–TcMar-IS885:	4	718bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-ISRm11:	23	4096bp	0.00%	12	4352bp	0.00%	0	0bp	0.00%
–TcMar-Mariner:	33	8551bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Sagan:	4	227bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Tc1:	188	44269bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Tc2:	21	6560bp	0.00%	24	4696bp	0.00%	0	0bp	0.00%
–TcMar-Tigger:	1762	293909bp	0.01%	265	52357bp	0.00%	33	2675bp	0.00%
–Zisupton:	164	35683bp	0.00%	42	9898bp	0.00%	0	0bp	0.00%
–hAT:	118	10626bp	0.00%	0	0bp	0.00%	5	345bp	0.00%
–hAT-Ac:	286	381717bp	0.01%	184	93383bp	0.00%	0	0bp	0.00%
–hAT-Blackjack:	54	5169bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Charlie:	5970	1415986bp	0.05%	7025	965773bp	0.03%	944	229709bp	0.01%
–hAT-Tag1:	16	2181bp	0.00%	72	18044bp	0.00%	37	44797bp	0.00%
–hAT-Tip100:	195	67198bp	0.00%	110	48745bp	0.00%	0	0bp	0.00%
–hAT-hATw:	69	6239bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LINE:	575223	130879619bp	4.64%	72870	61339201bp	2.18%	13799	10891520bp	0.39%
–CR1:	84	94797bp	0.00%	17	61029bp	0.00%	0	0bp	0.00%
–CRE-Ambal:	11	6172bp	0.00%	28	5980bp	0.00%	0	0bp	0.00%
–I:	39	66907bp	0.00%	154	136938bp	0.00%	0	0bp	0.00%
–I-Jockey:	132	47906bp	0.00%	74	14780bp	0.00%	0	0bp	0.00%
–Jockey:	17	2700bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–L1:	573389	130388803bp	4.63%	71301	60840527bp	2.16%	13764	10850626bp	0.38%
–L1-Tx1:	175	80576bp	0.00%	40	111835bp	0.00%	0	0bp	0.00%
–L2:	476	102990bp	0.00%	371	70192bp	0.00%	5	982bp	0.00%
–LOA:	3	253bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	15	46447bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Penelope:	48	14272bp	0.00%	0	0bp	0.00%	24	36831bp	0.00%
–R1:	79	28939bp	0.00%	31	8611bp	0.00%	0	0bp	0.00%
–R2:	62	48509bp	0.00%	21	10650bp	0.00%	0	0bp	0.00%
–R2-NeSL:	8	2836bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–RTE-BovB:	607	155230bp	0.01%	829	249314bp	0.01%	6	3081bp	0.00%
–RTE-RTE:	8	2433bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–RTE-X:	66	12053bp	0.00%	4	350bp	0.00%	0	0bp	0.00%
–Rex-Babar:	4	938bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LTR:	522898	79586508bp	2.82%	128730	40705662bp	1.44%	29905	11851534bp	0.42%
–Caulimovirus:	82	13899bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Copia:	991	277753bp	0.01%	363	139928bp	0.00%	24	8425bp	0.00%
–DIRS:	86	23958bp	0.00%	82	25787bp	0.00%	0	0bp	0.00%
–ERV:	749	1443253bp	0.05%	204	2275532bp	0.08%	0	0bp	0.00%
–ERV1:	122353	16628052bp	0.59%	18435	7550772bp	0.27%	3877	1590035bp	0.06%
–ERV4:	0	0bp	0.00%	45	6877bp	0.00%	0	0bp	0.00%
–ERVK:	282465	43357961bp	1.54%	77413	21616510bp	0.77%	17748	5091469bp	0.18%
–ERVL:	25836	8286233bp	0.29%	7345	3311531bp	0.12%	2718	2516274bp	0.09%
–ERVL-MaLR:	86571	12498839bp	0.44%	23827	6837893bp	0.24%	5487	2643985bp	0.09%
–Gypsy:	3149	591138bp	0.02%	901	354197bp	0.01%	46	34264bp	0.00%
–Lenti:	19	842bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Ngaro:	19	2384bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	119	26437bp	0.00%	22	5674bp	0.00%	5	3608bp	0.00%
–Pao:	429	114264bp	0.00%	93	14168bp	0.00%	0	0bp	0.00%
–Viper:	30	6470bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Other:	4696	698096bp	0.02%	2221	159797bp	0.01%	367	27887bp	0.00%
–OTHER:	4696	698096bp	0.02%	2221	159797bp	0.01%	367	27887bp	0.00%
RC:	385	46054bp	0.00%	107	43070bp	0.00%	36	1808bp	0.00%
–Helitron:	385	46054bp	0.00%	107	43070bp	0.00%	36	1808bp	0.00%
RNA:	0	0bp	0.00%	25	881bp	0.00%	8	897bp	0.00%
–OTHER:	0	0bp	0.00%	25	881bp	0.00%	8	897bp	0.00%
Retroposon:	10	1577bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–L1:	6	563bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–SVA:	4	1014bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
SINE:	99855	33312607bp	1.18%	61225	19846695bp	0.70%	10918	3649881bp	0.13%
–5S:	208	23801bp	0.00%	60	3994bp	0.00%	45	3159bp	0.00%
–5S-Deu-L2:	3	1522bp	0.00%	9	3436bp	0.00%	0	0bp	0.00%
–7SL:	98	31742bp	0.00%	27	1464bp	0.00%	0	0bp	0.00%
–Alu:	61154	26362743bp	0.94%	29656	13942791bp	0.49%	7988	1918217bp	0.07%
–B2:	29260	6335683bp	0.22%	24861	5386533bp	0.19%	2363	1469013bp	0.05%
–B4:	7230	1633364bp	0.06%	4652	1059477bp	0.04%	311	250062bp	0.01%
–ID:	1400	567613bp	0.02%	1269	294968bp	0.01%	165	6833bp	0.00%
–MIR:	382	155299bp	0.01%	452	150017bp	0.01%	24	1771bp	0.00%
–U:	80	7672bp	0.00%	203	6863bp	0.00%	22	2506bp	0.00%
–tRNA:	18	11026bp	0.00%	36	1620bp	0.00%	0	0bp	0.00%
–tRNA-7SL:	18	2263bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–tRNA-Core-L2:	4	866bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Satellite:	14413	2699863bp	0.10%	1751	1045108bp	0.04%	542	178380bp	0.01%
–5S:	4	1078bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	14409	2698785bp	0.10%	1751	1045108bp	0.04%	542	178380bp	0.01%
Simple:	77612	3656460bp	0.13%	410	58513bp	0.00%	782	41962bp	0.00%
–repeat:	77612	3656460bp	0.13%	410	58513bp	0.00%	782	41962bp	0.00%
Unknown:	967058	60613644bp	2.15%	106671	13870586bp	0.49%	4032	2542822bp	0.09%
–OTHER:	964526	60438386bp	2.14%	106671	13870586bp	0.49%	4032	2542822bp	0.09%
–Y-chromosome:	2532	227820bp	0.01%	0	0bp	0.00%	0	0bp	0.00%
rRNA:	204	11801bp	0.00%	157	9129bp	0.00%	0	0bp	0.00%
–OTHER:	204	11801bp	0.00%	157	9129bp	0.00%	0	0bp	0.00%
scRNA:	13	5519bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	13	5519bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
snRNA:	1244	41044bp	0.00%	593	41433bp	0.00%	9	671bp	0.00%
–OTHER:	1244	41044bp	0.00%	593	41433bp	0.00%	9	671bp	0.00%
tRNA:	275	19033bp	0.00%	369	33998bp	0.00%	0	0bp	0.00%
–OTHER:	275	19033bp	0.00%	369	33998bp	0.00%	0	0bp	0.00%

3.5.2 Assembly effect comparison of several tools before and after using barcode linked reads

In order to verify that long sequencing fragments can effectively resolve the problem of repetitive regions that cannot be solved during the assembly process of short sequencing fragments, we used four well-known NGS based assemblers (SOAPdenovo2, Abyss, IDBA-UD and SPAdes) to perform assembly tests on three real datasets of *Drosophila*, *Saccharomyces* and *human-chr-14*. The test results are shown in Tables S27, S28 and S29. The left sub-table shows the assemblies of various tools when barcode linked reads are not used, and the right sub-table shows the assemblies of each tools when barcode linked reads are used. ′Max′ indicates the largest contig. ′MA′ indicates the number of misassembly events. ′GF(%)′ indicates the genome fraction. From Tables S27, S28 and S29, we can find that barcode linked reads have played an important role in resolving the unresolved repetitive regions encountered during the assembly of short paired-end reads, reducing assembly errors, and improving the integrity of assembly results.

Table S27. Assemblies of several tools on Saccharomyces dataset before and after using barcode linked reads.

	Scaffolds before using barcode linked-reads							**Scaffolds after using barcode linked-reads**						
Assembler	Max(kb)	N50 (kb)	NG50 (kb)	MA	NA50 (kb)	NGA50 (kb)	GF(%)	Max(kb)	N50 (kb)	NG50 (kb)	MA	NA50 (kb)	NGA50 (kb)	GF(%)
SOAP2	69.393	20.631	19.153	**1**	20.631	19.153	93.196	78.579	22.412	21.412	5	22.412	21.020	93.103
Abyss	**417.251**	**126.494**	**123.308**	20	**117.356**	**114.685**	95.208	417.305	130.535	126.499	20	119.325	117.360	95.210
IDBA-UD	52.339	6.221	5.960	21	6.190	5.893	87.551	52.339	6.240	5.981	19	6.209	5.895	87.546
SPAdes	380.265	93.183	93.077	21	89.740	86.943	**95.924**	**723.706**	**420.315**	**418.964**	**11**	**371.687**	**371.687**	**97.007**

The left sub-table shows the assemblies of various tools when barcode linked reads are not used, and the right sub-table shows the assemblies of each tools when barcode linked reads are used. ′Max′ indicates the largest contig. ′MA′ indicates the number of misassembly events. ′GF(%)′ indicates the genome fraction.

Table S28. Assemblies of several tools on Human-chr14 dataset before and after using barcode linked reads.

	Scaffolds before using barcode linked-reads							**Scaffolds after using barcode linked-reads**						
Assembler	Max(kb)	N50 (kb)	NG50 (kb)	MA	NA50 (kb)	NGA50 (kb)	GF(%)	Max(kb)	N50 (kb)	NG50 (kb)	MA	NA50 (kb)	NGA50 (kb)	GF(%)
SOAP2	106.303	3.765	3.226	3100	2.580	2.084	74.229	108.294	14.515	10.320	1018	12.482	8.995	78.536
Abyss	97.603	11.612	8.320	1162	10.119	7.344	77.998	108.294	14.515	10.320	1018	12.482	8.995	78.536
IDBA-UD	**146.481**	**20.684**	**15.450**	224	**20.236**	**15.092**	**79.743**	301.267	50.736	37.876	**36**	50.483	37.739	**80.428**
SPAdes	74.796	8.176	5.834	136	8.100	5.758	78.373	**880.789**	**203.996**	**147.576**	687	**112.745**	**82.897**	79.363

The left sub-table shows the assemblies of various tools when barcode linked reads are not used, and the right sub-table shows the assemblies of each tools when barcode linked reads are used. ′Max′ indicates the largest contig. ′MA′ indicates the number of misassembly events. ′GF(%)′ indicates the genome fraction.

Table S29. Assemblies of several tools on Drosophila melanogaster dataset before and after using barcode linked reads.

	Scaffolds before using barcode linked-reads							**Scaffolds after using barcode linked-reads**						
Assembler	Max(kb)	N50 (kb)	NG50 (kb)	MA	NA50 (kb)	NGA50 (kb)	GF(%)	Max(kb)	N50 (kb)	NG50 (kb)	MA	NA50 (kb)	NGA50 (kb)	GF(%)
SOAP2	13.074	1.474	0.562	219	0.944	NA	36.512	498.151	76.351	36.886	23	76.351	36.886	68.513
Abyss	15.802	0.857	NA	**43**	**0.835**	NA	26.017	906.991	136.028	66.342	249	122.650	58.530	68.497
IDBA-UD	66.115	6.776	3.301	1795	6.441	3.061	65.958	243.301	24.091	12.335	**19**	23.565	11.929	70.865
SPAdes	109.693	13.065	6.452	1119	12.422	5.998	**67.459**	**1423.779**	**276.610**	**141.947**	1672	**194.398**	**100.383**	**71.118**

The left sub-table shows the assemblies of various tools when barcode linked reads are not used, and the right sub-table shows the assemblies of each tools when barcode linked reads are used. ′Max′ indicates the largest contig. ′MA′ indicates the number of misassembly events. ′GF(%)′ indicates the genome fraction.

3.5.3 Detection results of *de novo* mode based on only NGS short reads

The detection results of LongRepMarker in *de novo* mode based on only NGS short reads are shown in Fig. S24 and Tables S30-S39. Five NGS datasets (Drosophila, Ant, Mouse, Human-chr14 and HG003_NA24149_father (WGS)) are used in this test, and the performance of LongRepMarker is compared with the two similar tools (RepARK and REPdenovo). From Fig. S24, we can find that the distribution range of length and repetition frequency of the repetitive sequences found by LongRepMarker is larger than that of those two similar tools, which also means that the detection results of LongRepMarker are more comprehensive and complete than that of the two simiar tools. For example, the detected repetitive fragment length of LongRepMarker on dataset Mouse ranged from 1bp to 23.6kb, while that of RepARK and REPdenovo ranged from 1bp to 16.4kp and from 1bp to 6.1kp, respectively. Tables S30-S39 show the proportion and detailed classification of the detection results generated from three tools on these five NGS datasets covering the corresponding RepBase library and reference genome.

From the perspective of the coverage of the total base ratio, LongRepMarker has certain advantages compared with the latter two tools. For example, the detection results of LongRepMarker on Mouse dataset covered 69.48% of the bases in the corresponding RepBase library, while the corresponding ratios of RepARK and REPdenovo are 51.62% and 22.70%, respectively. Some practice example show the completeness and coverage of the repetitive sequences detected by LongRepMarker, RepARK and REPdenovo in the same region of the mouse genome, just as shown in Fig. S25-S28.

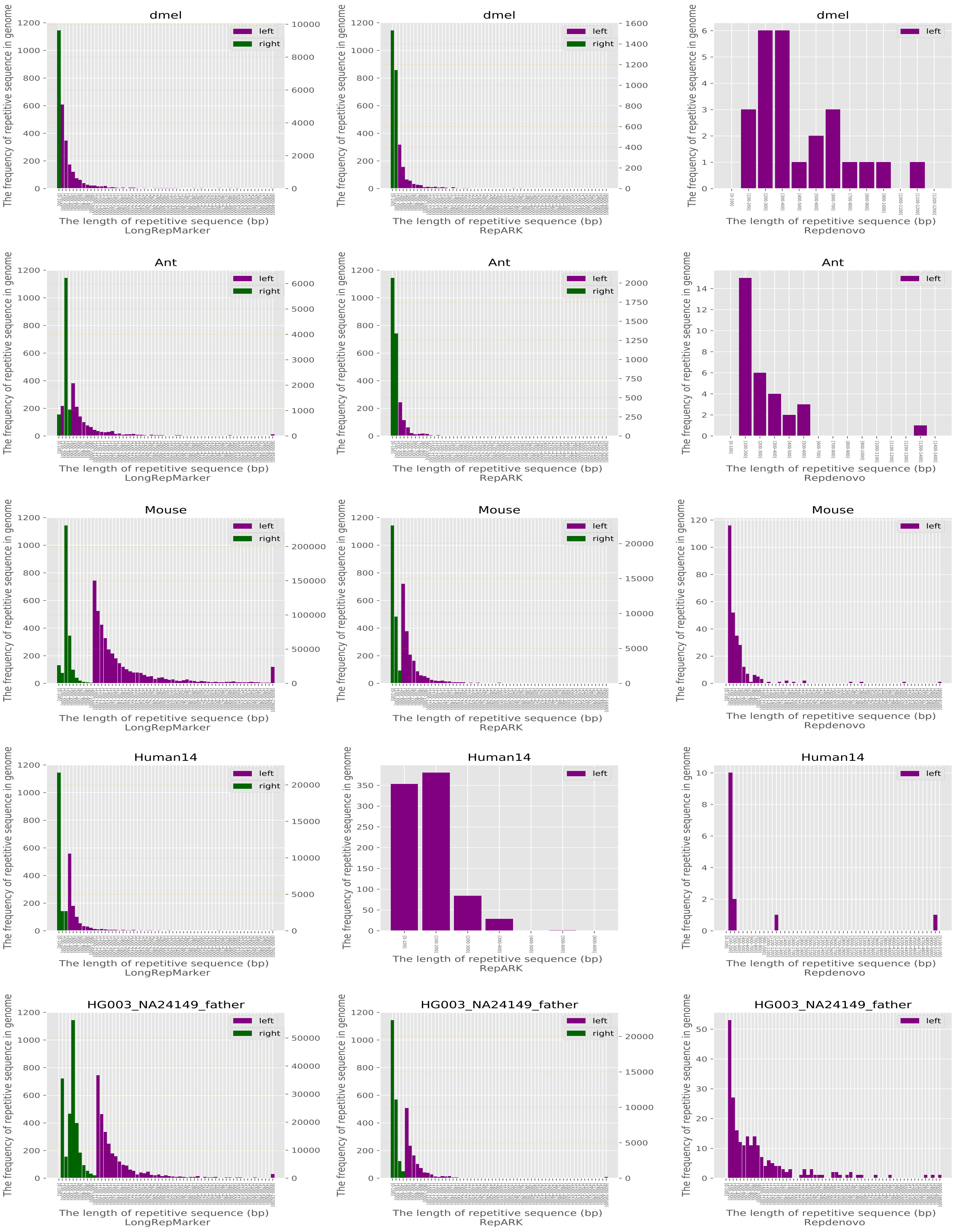


Fig. S24. Comparison of the repetition frequency and length distribution of the detected fragments generated from three tools. The X-axis represents the length distribution of the detected fragments and Y-axis represents the repetition frequency of the detected fragments in the genome, and the three images in each row respectively represent the frequency and length distribution of the repeated sequences detected by the three tools in a certain species. The coordinates of the Y-axis are divided into left and right displays, where the low frequency on the left is represented by purple, and the high frequency on the right is represented by green.

Table S30. Comparison of the proportion and detailed classification of detection results generated by three tools on Drosophila dataset covering the corresponding RepBase library.

LongRepMarker				RepARK			REPdenovo		
sequence: 2489 **total length: 7220516bp** **bases masked: 3051295 bp (42.26%)**				**sequence: 2489** **total length: 7220516bp** **bases masked: 2820283 bp (39.06%)**			**sequence: 2489** **total length: 7220516bp** **bases masked: 199042 bp (2.76%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA transposon elements:	243	84091bp	1.16%	419	67480bp	0.93%	12	1758bp	0.02%
–TcMar-Tigger:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Charlie:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LINEs:	1203	893842bp	12.38%	2304	790389bp	10.95%	189	74869bp	1.04%
–L3/CR1:	178	90613bp	1.25%	345	77798bp	1.08%	0	0bp	0.00%
–LINE1:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–LINE2:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LTR elements:	2328	1948634bp	26.99%	3444	1834947bp	25.41%	2	317bp	0.00%
–ERVL:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERVL-MaLRs:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERV-classI:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERV-classII:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Low complexity:	333	18139bp	0.25%	350	19015bp	0.26%	456	24145bp	0.33%
SINEs:	1	135bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ALUs:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MIRs:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Satellites:	20	4926bp	0.07%	36	4122bp	0.06%	6	1791bp	0.02%
Simple repeats:	1253	81143bp	1.12%	1288	82461bp	1.14%	1494	92425bp	1.28%
Small RNA:	15	13770bp	0.19%	20	13760bp	0.19%	5	593bp	0.01%
Total interspersed repeats:		2935431bp	40.65%		2702177bp	37.42%		80127bp	1.11%
Unclassified:	63	8729bp	0.12%	70	9361bp	0.13%	7	3183bp	0.04%

Table S31. Comparison of the proportion and detailed classification of detection results generated by three tools on Ant dataset covering the corresponding RepBase library.

LongRepMarker				RepARK			REPdenovo		
sequence: 254 **total length: 214457bp** **bases masked: 181755 bp (84.75%)**				**sequence: 254** **total length: 214457bp** **bases masked: 142209 bp (66.31%)**			**sequence: 254** **total length: 214457bp** **bases masked: 46235 bp (21.56%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA transposon elements:	108	73712bp	34.37%	261	62768bp	29.27%	21	14153bp	6.60%
–TcMar-Tigger:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Charlie:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LINEs:	24	14161bp	6.60%	59	9312bp	4.34%	0	0bp	0.00%
–L3/CR1:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–LINE1:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–LINE2:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LTR elements:	40	44578bp	20.79%	91	24272bp	11.32%	0	0bp	0.00%
–ERVL:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERVL-MaLRs:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERV-classI:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERV-classII:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Low complexity:	1	72bp	0.03%	2	137bp	0.06%	8	351bp	0.16%
SINEs:	0	0bp	0.00%	1	45bp	0.02%	0	0bp	0.00%
–ALUs:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MIRs:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Satellites:	0	0bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Simple repeats:	185	30802bp	14.36%	194	30860bp	14.39%	208	31731bp	14.80%
Small RNA:	15	13646bp	6.36%	15	13826bp	6.45%	0	0bp	0.00%
Total interspersed repeats:		138052bp	64.37%		97695bp	45.55%		14153bp	6.60%
Unclassified:	11	5601bp	2.61%	17	1298bp	0.61%	0	0bp	0.00%

Table S32. Comparison of the proportion and detailed classification of detection results generated by three tools on Mouse dataset covering the corresponding RepBase library.

LongRepMarker				RepARK			REPdenovo		
sequence: 1561 **total length: 1680566bp** **bases masked: 1167584 bp (69.48%)**				**sequence: 1561** **total length: 1680566bp** **bases masked: 867535 bp (51.62%)**			**sequence: 1561** **total length: 1680566bp** **bases masked: 381565 bp (22.70%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA transposon elements:	395	69181bp	4.12%	40	8027bp	0.48%	0	0bp	0.00%
–TcMar-Tigger:	75	13659bp	0.81%	2	222bp	0.01%	0	0bp	0.00%
–hAT-Charlie:	145	28233bp	1.68%	27	6071bp	0.36%	0	0bp	0.00%
LINEs:	646	454590bp	27.05%	371	334128bp	19.88%	241	299948bp	17.85%
–L3/CR1:	21	2981bp	0.18%	0	0bp	0.00%	0	0bp	0.00%
–LINE1:	591	447010bp	26.60%	367	333804bp	19.86%	241	299948bp	17.85%
–LINE2:	22	3006bp	0.18%	4	324bp	0.02%	0	0bp	0.00%
LTR elements:	981	532020bp	31.66%	1795	450620bp	26.81%	118	31176bp	1.86%
–ERVL:	183	73304bp	4.36%	265	35675bp	2.12%	32	11265bp	0.67%
–ERVL-MaLRs:	156	31941bp	1.90%	102	23074bp	1.37%	43	10176bp	0.61%
–ERV-classI:	209	107341bp	6.39%	338	85642bp	5.10%	0	0bp	0.00%
–ERV-classII:	399	313254bp	18.64%	1086	305486bp	18.18%	43	9735bp	0.58%
Low complexity:	26	1069bp	0.06%	51	2566bp	0.15%	97	5116bp	0.30%
SINEs:	273	46075bp	2.74%	110	13784bp	0.82%	29	2553bp	0.15%
–ALUs:	164	31988bp	1.90%	56	7858bp	0.47%	24	1826bp	0.11%
–MIRs:	5	943bp	0.06%	4	589bp	0.04%	0	0bp	0.00%
Satellites:	10	3642bp	0.22%	22	4181bp	0.25%	2	734bp	0.04%
Simple repeats:	286	34991bp	2.08%	354	37830bp	2.25%	426	42038bp	2.50%
Small RNA:	41	14171bp	0.84%	46	12537bp	0.75%	0	0bp	0.00%
Total interspersed repeats:		1126788bp	67.05%		815147bp	48.50%		333677bp	19.86%
Unclassified:	172	24922bp	1.48%	63	8588bp	0.51%	0	0bp	0.00%

Table S33. Comparison of the proportion and detailed classification of detection results generated by three tools on Human-chr14 dataset covering the corresponding RepBase library.

LongRepMarker				RepARK			REPdenovo		
sequence: 1512 **total length: 1647075bp** **bases masked: 452080 bp (27.45%)**				**sequence: 1512** **total length: 1647075bp** **bases masked: 229636 bp (13.94%)**			**sequence: 1512** **total length: 1647075bp** **bases masked: 183245 bp (11.13%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA transposon elements:	81	13426bp	0.82%	32	2828bp	0.17%	0	0bp	0.00%
–TcMar-Tigger:	40	8134bp	0.49%	23	2045bp	0.12%	0	0bp	0.00%
–hAT-Charlie:	23	3836bp	0.23%	5	461bp	0.03%	0	0bp	0.00%
LINEs:	345	191120bp	11.60%	544	118330bp	7.18%	126	116694bp	7.08%
–L3/CR1:	6	243bp	0.01%	1	75bp	0.00%	0	0bp	0.00%
–LINE1:	309	178143bp	10.82%	540	117752bp	7.15%	126	116694bp	7.08%
–LINE2:	10	2268bp	0.14%	3	503bp	0.03%	0	0bp	0.00%
LTR elements:	539	155596bp	9.45%	260	35630bp	2.16%	15	1427bp	0.09%
–ERVL:	119	36766bp	2.23%	30	3894bp	0.24%	0	0bp	0.00%
–ERVL-MaLRs:	70	12274bp	0.75%	49	8109bp	0.49%	15	1427bp	0.09%
–ERV-classI:	310	91588bp	5.56%	144	19917bp	1.21%	0	0bp	0.00%
–ERV-classII:	28	12635bp	0.77%	37	3710bp	0.23%	0	0bp	0.00%
Low complexity:	60	3018bp	0.18%	85	4170bp	0.25%	82	4030bp	0.24%
SINEs:	183	39215bp	2.38%	74	19956bp	1.21%	71	18201bp	1.11%
–ALUs:	173	38321bp	2.33%	70	19316bp	1.17%	71	18201bp	1.11%
–MIRs:	10	894bp	0.05%	4	640bp	0.04%	0	0bp	0.00%
Satellites:	14	3334bp	0.20%	19	1908bp	0.12%	6	524bp	0.03%
Simple repeats:	393	39165bp	2.38%	408	40077bp	2.43%	419	40550bp	2.46%
Small RNA:	0	0bp	0.00%	2	224bp	0.01%	0	0bp	0.00%
Total interspersed repeats:		407126bp	24.72%		183285bp	11.13%		138141bp	8.39%
Unclassified:	21	7769bp	0.47%	51	6541bp	0.40%	10	1819bp	0.11%

Table S34. Comparison of the proportion and detailed classification of detection results generated by three tools on HG003_NA24149_father dataset covering the corresponding RepBase library.

	LongRepMarker			RepARK			REPdenovo		
sequence: 1512				**sequence: 1512**			**sequence: 1512**		
total length: 1647075bp				**total length: 1647075bp**			**total length: 1647075bp**		
bases masked: 1210784 bp (73.51%)				**bases masked: 870210 bp (52.83%)**			**bases masked: 199536 bp (12.11%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA transposon elements:	448	121106bp	7.35%	378	60164bp	3.65%	0	0bp	0.00%
–TcMar-Tigger:	126	37522bp	2.28%	129	23339bp	1.42%	0	0bp	0.00%
–hAT-Charlie:	143	39145bp	2.38%	121	17883bp	1.09%	0	0bp	0.00%
LINEs:	653	291629bp	17.71%	504	213163bp	12.94%	129	123854bp	7.52%
–L3/CR1:	26	4388bp	0.27%	5	1076bp	0.07%	0	0bp	0.00%
–LINE1:	586	278984bp	16.94%	481	209312bp	12.71%	129	123854bp	7.52%
–LINE2:	21	4123bp	0.25%	12	1992bp	0.12%	0	0bp	0.00%
LTR elements:	1082	620894bp	37.70%	2467	483315bp	29.34%	14	1927bp	0.12%
–ERVL:	219	92790bp	5.63%	336	68843bp	4.18%	0	0bp	0.00%
–ERVL-MaLRs:	102	23018bp	1.40%	93	26417bp	1.60%	14	1927bp	0.12%
–ERV-classI:	649	418886bp	25.43%	1713	320240bp	19.44%	0	0bp	0.00%
–ERV-classII:	63	76152bp	4.62%	309	65213bp	3.96%	0	0bp	0.00%
Low complexity:	16	652bp	0.04%	40	2039bp	0.12%	81	3926bp	0.24%
SINEs:	503	112306bp	6.82%	198	37531bp	2.28%	72	19991bp	1.21%
–ALUs:	469	108143bp	6.57%	162	34020bp	2.07%	72	19991bp	1.21%
–MIRs:	25	3428bp	0.21%	17	1893bp	0.11%	0	0bp	0.00%
Satellites:	39	9924bp	0.60%	102	16427bp	1.00%	11	1683bp	0.10%
Simple repeats:	234	32433bp	1.97%	318	36600bp	2.22%	414	40423bp	2.45%
Small RNA:	23	13124bp	0.80%	44	12839bp	0.78%	0	0bp	0.00%
Total interspersed repeats:		1165100bp	70.74%		806029bp	48.94%		153504bp	9.32%
Unclassified:	80	19165bp	1.16%	109	11856bp	0.72%	34	7732bp	0.47%

Table S35. Comparison of the proportion and detailed classification of detection results generated by three tools covering the repeat regions on the reference genome of Drosophila.

	LongRepMarker			RepARK			REPdenovo		
sequence: 15				**sequence: 15**			**sequence: 15**		
total length: 168736537bp				**total length: 168736537bp**			**total length: 168736537bp**		
bases masked: 13152663 bp (7.79%)				**bases masked: 10328668 bp (6.12%)**			**bases masked: 445253 bp (0.26%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	8667	472889bp	0.28%	4295	413650bp	0.25%	33	6463bp	0.00%
–CMC-Transib:	131	42384bp	0.03%	485	38315bp	0.02%	0	0bp	0.00%
–MULE-NOF:	100	20167bp	0.01%	100	26640bp	0.02%	0	0bp	0.00%
–P:	7820	270518bp	0.16%	2330	166430bp	0.10%	33	6463bp	0.00%
–TcMar-Pogo:	132	21336bp	0.01%	158	27829bp	0.02%	0	0bp	0.00%
–TcMar-Tc1:	310	63549bp	0.04%	876	92146bp	0.05%	0	0bp	0.00%
–hAT-Ac:	78	17967bp	0.01%	103	16133bp	0.01%	0	0bp	0.00%
–hAT-hATm:	14	1781bp	0.00%	8	344bp	0.00%	0	0bp	0.00%
–hAT-hobo:	82	35271bp	0.02%	235	46543bp	0.03%	0	0bp	0.00%
LINE:	23646	2335075bp	1.38%	20242	2317797bp	1.37%	391	126677bp	0.08%
–CR1:	506	113230bp	0.07%	969	114922bp	0.07%	0	0bp	0.00%
–I:	279	137164bp	0.08%	740	136293bp	0.08%	0	0bp	0.00%
–I-Jockey:	3432	926877bp	0.55%	5504	795385bp	0.47%	0	0bp	0.00%
–Jockey:	2587	298888bp	0.18%	3126	455050bp	0.27%	85	29279bp	0.02%
–LOA:	311	49228bp	0.03%	661	47273bp	0.03%	0	0bp	0.00%
–OTHER:	34	15250bp	0.01%	0	0bp	0.00%	0	0bp	0.00%
–R1:	16228	786758bp	0.47%	8665	747875bp	0.44%	306	97398bp	0.06%
–R1-LOA:	246	32082bp	0.02%	575	38344bp	0.02%	0	0bp	0.00%
–R2:	23	17294bp	0.01%	2	3609bp	0.00%	0	0bp	0.00%
LTR:	17205	4938812bp	2.93%	37735	6177355bp	3.66%	107	22321bp	0.01%
–Copia:	920	279327bp	0.17%	1987	322203bp	0.19%	0	0bp	0.00%
–Gypsy:	11210	3737284bp	2.21%	28143	4337017bp	2.57%	0	0bp	0.00%
–Ngaro:	2	197bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	19	3058bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Pao:	5054	918950bp	0.54%	7605	1518146bp	0.90%	107	22321bp	0.01%
Other:	41610	531208bp	0.31%	616	51603bp	0.03%	1077	176415bp	0.10%
–OTHER:	41610	531208bp	0.31%	616	51603bp	0.03%	1077	176415bp	0.10%
RC:	166	34553bp	0.02%	173	17915bp	0.01%	0	0bp	0.00%
–Helitron:	166	34553bp	0.02%	173	17915bp	0.01%	0	0bp	0.00%
RNA:	250	15785bp	0.01%	39	9470bp	0.01%	0	0bp	0.00%
–OTHER:	250	15785bp	0.01%	39	9470bp	0.01%	0	0bp	0.00%
SINE:	34	11232bp	0.01%	0	0bp	0.00%	0	0bp	0.00%
–5S:	34	11232bp	0.01%	0	0bp	0.00%	0	0bp	0.00%
Satellite:	115731	2478996bp	1.47%	5050	325076bp	0.19%	211	56629bp	0.03%
–OTHER:	115731	2478996bp	1.47%	5050	325076bp	0.19%	211	56629bp	0.03%
Simple:	315254	561147bp	0.33%	12909	96206bp	0.06%	0	0bp	0.00%
–repeat:	315254	561147bp	0.33%	12909	96206bp	0.06%	0	0bp	0.00%
Unknown:	101943	2502217bp	1.48%	11038	866605bp	0.51%	83	36312bp	0.02%
–OTHER:	101943	2502217bp	1.48%	11038	866605bp	0.51%	83	36312bp	0.02%
rRNA:	960	306780bp	0.18%	307	77755bp	0.05%	332	94037bp	0.06%
–OTHER:	960	306780bp	0.18%	307	77755bp	0.05%	332	94037bp	0.06%
snRNA:	0	0bp	0.00%	12	803bp	0.00%	0	0bp	0.00%
–OTHER:	0	0bp	0.00%	12	803bp	0.00%	0	0bp	0.00%

Table S36. Comparison of the proportion and detailed classification of detection results generated by three tools covering the repeat regions on the reference genome of Ant.

LongRepMarker				RepARK			Repdenovo		
sequence: 4339				**sequence: 4339**			**sequence: 4339**		
total length: 295944863bp				**total length: 295944863bp**			**total length: 295944863bp**		
bases masked: 8200834 bp (2.77%)				**bases masked: 4151102 bp (1.40%)**			**bases masked: 113526 bp (0.04%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	24099	3286029bp	1.11%	18041	1978917bp	0.67%	258	70280bp	0.02%
–CMC-Chapaev-3:	27	11562bp	0.00%	56	9445bp	0.00%	0	0bp	0.00%
–CMC-EnSpm:	153	28552bp	0.01%	14	1046bp	0.00%	0	0bp	0.00%
–CMC-Transib:	102	11545bp	0.00%	41	3717bp	0.00%	0	0bp	0.00%
–Crypton-V:	48	4622bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Kolobok-Hydra:	88	45412bp	0.02%	99	34781bp	0.01%	0	0bp	0.00%
–Kolobok-T2:	1140	182520bp	0.06%	3567	195621bp	0.07%	10	631bp	0.00%
–MULE-MuDR:	16	2775bp	0.00%	0	0bp	0.00%	21	4762bp	0.00%
–MULE-NOF:	139	21557bp	0.01%	9	1667bp	0.00%	0	0bp	0.00%
–Maverick:	1108	405696bp	0.14%	4155	338328bp	0.11%	0	0bp	0.00%
–Merlin:	69	13038bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MuLE-MuDR:	2	301bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MuLE-NOF:	10	1188bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	504	150450bp	0.05%	168	19164bp	0.01%	0	0bp	0.00%
–P:	48	19003bp	0.01%	39	2140bp	0.00%	0	0bp	0.00%
–PIF-Harbinger:	12	1969bp	0.00%	5	290bp	0.00%	0	0bp	0.00%
–PIF-ISL2EU:	2	1243bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PIF-Spy:	19	4617bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PiggyBac:	20	5393bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar:	99	12648bp	0.00%	93	8560bp	0.00%	0	0bp	0.00%
–TcMar-Cweed:	38	7817bp	0.00%	13	3062bp	0.00%	0	0bp	0.00%
–TcMar-Fot1:	26	3562bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-ISRm11:	2	918bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Mariner:	15982	1566149bp	0.53%	5410	869293bp	0.29%	189	52602bp	0.02%
–TcMar-Tc1:	3549	596746bp	0.20%	4101	457886bp	0.15%	38	12285bp	0.00%
–TcMar-Tc4:	443	96868bp	0.03%	131	19788bp	0.01%	0	0bp	0.00%
–TcMar-Tigger:	0	0bp	0.00%	7	3290bp	0.00%	0	0bp	0.00%
–hAT:	76	18749bp	0.01%	17	2038bp	0.00%	0	0bp	0.00%
–hAT-Ac:	54	14405bp	0.00%	9	437bp	0.00%	0	0bp	0.00%
–hAT-Blackjack:	212	44039bp	0.01%	107	10127bp	0.00%	0	0bp	0.00%
–hAT-Charlie:	72	24877bp	0.01%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Tip100:	16	2953bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-hAT19:	23	2556bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LINE:	1892	520030bp	0.18%	2572	213280bp	0.07%	0	0bp	0.00%
–CR1:	29	9954bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–I:	130	65937bp	0.02%	76	5544bp	0.00%	0	0bp	0.00%
–I-Jockey:	8	3676bp	0.00%	26	1663bp	0.00%	0	0bp	0.00%
–Jockey:	35	15371bp	0.01%	80	7256bp	0.00%	0	0bp	0.00%
–L1:	13	1385bp	0.00%	6	429bp	0.00%	0	0bp	0.00%
–L2:	126	19399bp	0.01%	54	3154bp	0.00%	0	0bp	0.00%
–OTHER:	0	0bp	0.00%	30	2535bp	0.00%	0	0bp	0.00%
–Penelope:	1179	313785bp	0.11%	1543	130675bp	0.04%	0	0bp	0.00%
–R1:	258	59355bp	0.02%	669	55749bp	0.02%	0	0bp	0.00%
–R2-NeSL:	69	15271bp	0.01%	69	5162bp	0.00%	0	0bp	0.00%
–RTE-BovB:	2	235bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–RTE-X:	34	10360bp	0.00%	19	1291bp	0.00%	0	0bp	0.00%
–Tad1:	9	6599bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LTR:	1887	586849bp	0.20%	3931	366397bp	0.12%	0	0bp	0.00%
–Copia:	132	62359bp	0.02%	532	68777bp	0.02%	0	0bp	0.00%
–DIRS:	5	1300bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERV1:	2	446bp	0.00%	3	386bp	0.00%	0	0bp	0.00%
–ERVK:	3	30bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Gypsy:	1076	319483bp	0.11%	1627	174773bp	0.06%	0	0bp	0.00%
–OTHER:	10	9221bp	0.00%	16	4071bp	0.00%	0	0bp	0.00%
–Pao:	659	194332bp	0.07%	1753	118390bp	0.04%	0	0bp	0.00%
RC:	232	74081bp	0.03%	55	9784bp	0.00%	0	0bp	0.00%
–Helitron:	232	74081bp	0.03%	55	9784bp	0.00%	0	0bp	0.00%
RC?:	14	8650bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Helitron:	14	8650bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
SINE:	5	1124bp	0.00%	21	1716bp	0.00%	0	0bp	0.00%
–5S-Deu-L2:	5	1124bp	0.00%	14	1003bp	0.00%	0	0bp	0.00%
–U:	0	0bp	0.00%	7	713bp	0.00%	0	0bp	0.00%
Satellite:	2	211bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	2	211bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Simple:	12485	140921bp	0.05%	860	37841bp	0.01%	24	301bp	0.00%
–repeat:	12485	140921bp	0.05%	860	37841bp	0.01%	24	301bp	0.00%
Unknown:	33083	3626291bp	1.23%	23915	1572343bp	0.53%	438	42945bp	0.01%
–OTHER:	33083	3626291bp	1.23%	23915	1572343bp	0.53%	438	42945bp	0.01%
rRNA:	29	11625bp	0.00%	29	11727bp	0.00%	0	0bp	0.00%
–OTHER:	29	11625bp	0.00%	29	11727bp	0.00%	0	0bp	0.00%
tRNA:	10	1098bp	0.00%	12	673bp	0.00%	0	0bp	0.00%
–OTHER:	10	1098bp	0.00%	12	673bp	0.00%	0	0bp	0.00%

Table S37. Comparison of the proportion and detailed classification of detection results generated by three tools covering the repeat regions on the reference genome of Mouse.

LongRepMarker				RepARK			Repdenovo		
sequence: 239 **total length: 2818974548bp** **bases masked: 241946118 bp (8.58%)**				**sequence: 239** **total length: 2818974548bp** **bases masked: 43289265 bp (1.54%)**			**sequence: 239** **total length: 2818974548bp** **bases masked: 4627038 bp (0.16%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
ARTEFACT:	12	1751bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	12	1751bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
DNA:	34942	4028857bp	0.14%	2077	448811bp	0.02%	0	0bp	0.00%
–Academ:	6	768bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Academ-1:	14	4288bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Academ-H:	2	266bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–CMC-Chapaev:	8	1172bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–CMC-Chapaev-3:	8	1037bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–CMC-EnSpm:	12997	1039938bp	0.04%	76	10056bp	0.00%	0	0bp	0.00%
–CMC-Transib:	52	3109bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton:	8	1193bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton-A:	2	264bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton-F:	54	2072bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton-H:	84	6710bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton-S:	57	5428bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton-V:	322	31443bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Dada:	172	13527bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Ginger:	3533	268998bp	0.01%	0	0bp	0.00%	0	0bp	0.00%
–IS:	2	236bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–IS3EU:	174	27978bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Kolobok:	2	237bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Kolobok-Hydra:	79	6450bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Kolobok-T2:	183	11957bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MULE-MuDR:	797	110927bp	0.00%	50	6498bp	0.00%	0	0bp	0.00%
–MULE-NOF:	43	1495bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Maverick:	167	41376bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Merlin:	10	1328bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MuLE-F:	7	430bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MuLE-MuDR:	22	6205bp	0.00%	22	3596bp	0.00%	0	0bp	0.00%
–MuLE-NOF:	12	590bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MuLE-NOF?:	10	405bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Novosib:	660	69877bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	1593	133807bp	0.00%	15	468bp	0.00%	0	0bp	0.00%
–P:	31	2928bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–P-Fungi:	15	624bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PIF-Harbinger:	699	62618bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PIF-ISL2EU:	27	8105bp	0.00%	16	5937bp	0.00%	0	0bp	0.00%
–PIF-Spy:	18	995bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PiggyBac:	48	22828bp	0.00%	14	10924bp	0.00%	0	0bp	0.00%
–PiggyBac-X:	10	1471bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Sola-1:	321	32376bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Sola-2:	25	1172bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Sola-3:	214	25782bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar:	7	881bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Ant1:	16	3167bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Fot1:	58	2958bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-ISRm11:	30	6128bp	0.00%	12	903bp	0.00%	0	0bp	0.00%
–TcMar-Mariner:	62	8818bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Pogo:	4	586bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Sagan:	3	459bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Stowaway:	6	3330bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Tc1:	696	61841bp	0.00%	30	4344bp	0.00%	0	0bp	0.00%
–TcMar-Tc2:	78	8056bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Tc4:	4	706bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Tigger:	2122	275507bp	0.01%	117	23364bp	0.00%	0	0bp	0.00%
–Zisupton:	778	80877bp	0.00%	17	4063bp	0.00%	0	0bp	0.00%
–hAT:	34	7935bp	0.00%	8	1549bp	0.00%	0	0bp	0.00%
–hAT-Ac:	733	66700bp	0.00%	83	6542bp	0.00%	0	0bp	0.00%
–hAT-Blackjack:	133	11853bp	0.00%	17	2900bp	0.00%	0	0bp	0.00%
–hAT-Charlie:	6983	1496302bp	0.05%	1517	351264bp	0.01%	0	0bp	0.00%
–hAT-Pegasus:	44	3084bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Tag1:	90	9760bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Tip100:	435	74834bp	0.00%	83	16403bp	0.00%	0	0bp	0.00%
–hAT-hAT19:	4	531bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-hATm:	101	6526bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-hATw:	31	674bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-hobo:	2	255bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
DNA?:	6	728bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PiggyBac:	4	479bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT:	2	249bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LINE:	1739539	121108544bp	4.30%	148672	10063402bp	0.36%	10924	4148162bp	0.15%
–CR1:	231	59988bp	0.00%	9	563bp	0.00%	0	0bp	0.00%
–CR1-Zenon:	2	376bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–CRE-Ambal:	35	1485bp	0.00%	13	693bp	0.00%	0	0bp	0.00%
–CRE-Odin:	2	244bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Dong-R4:	8	857bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Dualen:	2	1562bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–I:	37	5332bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–I-Jockey:	375	29273bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Jockey:	2	270bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–L1:	1732439	120039364bp	4.26%	148373	9997561bp	0.35%	10924	4148162bp	0.15%
–L1-Tx1:	217	90021bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–L2:	5026	652714bp	0.02%	139	24633bp	0.00%	0	0bp	0.00%
–OTHER:	60	11132bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Penelope:	189	24121bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–R1:	147	26313bp	0.00%	73	24237bp	0.00%	0	0bp	0.00%
–R1-LOA:	8	1972bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–R2:	115	23063bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–R2-NeSL:	0	0bp	0.00%	24	12405bp	0.00%	0	0bp	0.00%
–RTE-BovB:	442	115547bp	0.00%	20	1638bp	0.00%	0	0bp	0.00%
–RTE-RTE:	4	1210bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–RTE-X:	130	21757bp	0.00%	11	412bp	0.00%	0	0bp	0.00%
–Rex-Babar:	52	4211bp	0.00%	10	1260bp	0.00%	0	0bp	0.00%
–Tad1:	16	1889bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LINE?:	4	484bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Penelope:	4	484bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LTR:	240965	32960648bp	1.17%	176053	14027673bp	0.50%	995	299622bp	0.01%
–Caulimovirus:	79	5495bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Copia:	1012	125717bp	0.00%	133	25564bp	0.00%	0	0bp	0.00%
–DIRS:	99	24377bp	0.00%	22	8037bp	0.00%	0	0bp	0.00%
–ERV:	19	6057bp	0.00%	53	25048bp	0.00%	0	0bp	0.00%
–ERV-Lenti:	13	501bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERV1:	13699	4272539bp	0.15%	27191	2458001bp	0.09%	0	0bp	0.00%
–ERV4:	18	2675bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERVK:	128277	17468583bp	0.62%	113800	8105149bp	0.29%	608	170732bp	0.01%
–ERVL:	19834	2729119bp	0.10%	15597	1562477bp	0.06%	102	53928bp	0.00%
–ERVL-MaLR:	74009	7983613bp	0.28%	18974	1796104bp	0.06%	285	74962bp	0.00%
–Gypsy:	3307	335974bp	0.01%	213	53713bp	0.00%	0	0bp	0.00%
–Ngaro:	82	9689bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	128	19268bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Pao:	389	66375bp	0.00%	70	9991bp	0.00%	0	0bp	0.00%
Other:	2912	418467bp	0.01%	533	69425bp	0.00%	0	0bp	0.00%
–OTHER:	2912	418467bp	0.01%	533	69425bp	0.00%	0	0bp	0.00%
RC:	1056	97924bp	0.00%	4	184bp	0.00%	0	0bp	0.00%
–Helitron:	1054	97549bp	0.00%	4	184bp	0.00%	0	0bp	0.00%
–Helitron-2:	2	375bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
RNA:	15	2131bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	15	2131bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Retroposon:	16	1823bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–SVA:	16	1823bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
SINE:	182926	23961883bp	0.85%	18093	2829668bp	0.10%	173	18049bp	0.00%
–5S:	80	26799bp	0.00%	92	7421bp	0.00%	0	0bp	0.00%
–5S-Deu-L2:	4	481bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–5S-Sauria-RTE:	4	488bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–7SL:	262	55265bp	0.00%	21	5203bp	0.00%	0	0bp	0.00%
–Alu:	111026	14163827bp	0.50%	8150	1373906bp	0.05%	80	8502bp	0.00%
–B2:	48480	6570506bp	0.23%	6707	883087bp	0.03%	93	9547bp	0.00%
–B4:	18023	2465836bp	0.09%	2283	407923bp	0.01%	0	0bp	0.00%
–ID:	3694	597285bp	0.02%	626	120052bp	0.00%	0	0bp	0.00%
–MIR:	1098	142144bp	0.01%	99	19356bp	0.00%	0	0bp	0.00%
–OTHER:	2	463bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–U:	6	2881bp	0.00%	19	1339bp	0.00%	0	0bp	0.00%
–tRNA:	130	30450bp	0.00%	37	2139bp	0.00%	0	0bp	0.00%
–tRNA-7SL:	105	10241bp	0.00%	59	13961bp	0.00%	0	0bp	0.00%
–tRNA-Core-L2:	8	754bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–tRNA-Core-RTE:	4	540bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Satellite:	1256869	1223315bp	0.04%	6069	477050bp	0.02%	9300	97755bp	0.00%
–5S:	14	574bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	1256836	1220988bp	0.04%	6063	476973bp	0.02%	9300	97755bp	0.00%
–centromeric:	11	1086bp	0.00%	6	137bp	0.00%	0	0bp	0.00%
–macro:	8	960bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Simple:	344562	8142216bp	0.29%	9189	410223bp	0.01%	986	73427bp	0.00%
–repeat:	344562	8142216bp	0.29%	9189	410223bp	0.01%	986	73427bp	0.00%
Unknown:	484681	54020241bp	1.92%	204866	15377767bp	0.55%	0	0bp	0.00%
–OTHER:	484678	54019731bp	1.92%	204866	15377767bp	0.55%	0	0bp	0.00%
–centromeric:	3	510bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
rRNA:	102	21283bp	0.00%	98	16523bp	0.00%	0	0bp	0.00%
–OTHER:	102	21283bp	0.00%	98	16523bp	0.00%	0	0bp	0.00%
scRNA:	8	902bp	0.00%	34	4583bp	0.00%	0	0bp	0.00%
–OTHER:	8	902bp	0.00%	34	4583bp	0.00%	0	0bp	0.00%
snRNA:	133	21764bp	0.00%	95	9137bp	0.00%	0	0bp	0.00%
–OTHER:	133	21764bp	0.00%	95	9137bp	0.00%	0	0bp	0.00%
tRNA:	41	8256bp	0.00%	38	4394bp	0.00%	0	0bp	0.00%
–OTHER:	41	8256bp	0.00%	38	4394bp	0.00%	0	0bp	0.00%

Table S38. Comparison of the proportion and detailed classification of detection results generated by three tools covering the repeat regions on the reference genome of Human-chr14.

LongRepMarker				RepARK			Repdenovo		
sequence: 1 **total length: 107349540bp** **bases masked: 5730369 bp (5.34%)**				**sequence: 1** **total length: 107349540bp** **bases masked: 549065 bp (0.51%)**			**sequence: 1** **total length: 107349540bp** **bases masked: 191540 bp (0.18%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	828	74218bp	0.07%	100	6861bp	0.01%	0	0bp	0.00%
–CMC-EnSpm:	490	11478bp	0.01%	10	260bp	0.00%	0	0bp	0.00%
–Ginger:	16	2557bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MULE-MuDR:	8	490bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MuLE-MuDR:	8	739bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Novosib:	6	1039bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	3	371bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PIF-Harbinger:	10	1559bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PiggyBac:	0	0bp	0.00%	4	302bp	0.00%	0	0bp	0.00%
–Sola-1:	6	2719bp	0.00%	4	578bp	0.00%	0	0bp	0.00%
–Sola-3:	10	2172bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Tc1:	10	1966bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Tc2:	6	1243bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Tigger:	148	29829bp	0.03%	53	2760bp	0.00%	0	0bp	0.00%
–Zisupton:	19	620bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Ac:	27	348bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Blackjack:	2	213bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Charlie:	55	17063bp	0.02%	25	2415bp	0.00%	0	0bp	0.00%
–hAT-Tag1:	2	240bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Tip100:	2	247bp	0.00%	4	546bp	0.00%	0	0bp	0.00%
LINE:	613297	2758108bp	2.57%	3483	238481bp	0.22%	171	183424bp	0.17%
–CR1:	4	240bp	0.00%	4	206bp	0.00%	0	0bp	0.00%
–L1:	612996	2712943bp	2.53%	3447	235647bp	0.22%	171	183424bp	0.17%
–L2:	267	33331bp	0.03%	32	2628bp	0.00%	0	0bp	0.00%
–OTHER:	8	8930bp	0.01%	0	0bp	0.00%	0	0bp	0.00%
–Penelope:	16	724bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–R2-NeSL:	6	1940bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LTR:	1733	257330bp	0.24%	1070	81337bp	0.08%	15	1014bp	0.00%
–Copia:	36	1919bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERV:	4	606bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERV1:	978	129105bp	0.12%	642	44995bp	0.04%	0	0bp	0.00%
–ERVK:	191	34933bp	0.03%	127	8215bp	0.01%	0	0bp	0.00%
–ERVL:	210	35160bp	0.03%	100	6899bp	0.01%	0	0bp	0.00%
–ERVL-MaLR:	233	47102bp	0.04%	199	21198bp	0.02%	15	1014bp	0.00%
–Gypsy:	75	8015bp	0.01%	2	122bp	0.00%	0	0bp	0.00%
–OTHER:	2	235bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Pao:	4	470bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
RC:	54	2155bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Helitron:	54	2155bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Retroposon:	959	44352bp	0.04%	212	17536bp	0.02%	11	1624bp	0.00%
–SVA:	959	44352bp	0.04%	212	17536bp	0.02%	11	1624bp	0.00%
SINE:	128817	1964602bp	1.83%	931	54659bp	0.05%	35	4688bp	0.00%
–Alu:	128739	1935008bp	1.80%	907	52591bp	0.05%	35	4688bp	0.00%
–MIR:	78	29742bp	0.03%	24	2068bp	0.00%	0	0bp	0.00%
Satellite:	734	66331bp	0.06%	172	9959bp	0.01%	8	593bp	0.00%
–OTHER:	269	24198bp	0.02%	30	1374bp	0.00%	0	0bp	0.00%
–Y-chromosome:	14	2200bp	0.00%	15	921bp	0.00%	0	0bp	0.00%
–centromeric:	443	37500bp	0.03%	107	5776bp	0.01%	8	593bp	0.00%
–telomeric:	8	2688bp	0.00%	20	2162bp	0.00%	0	0bp	0.00%
Simple:	21446	249378bp	0.23%	345	11739bp	0.01%	0	0bp	0.00%
–repeat:	21446	249378bp	0.23%	345	11739bp	0.01%	0	0bp	0.00%
Unknown:	7179	409303bp	0.38%	2182	128714bp	0.12%	2	197bp	0.00%
–OTHER:	7179	409303bp	0.38%	2182	128714bp	0.12%	2	197bp	0.00%
rRNA:	0	0bp	0.00%	2	116bp	0.00%	0	0bp	0.00%
–OTHER:	0	0bp	0.00%	2	116bp	0.00%	0	0bp	0.00%
snRNA:	2	284bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	2	284bp	0.00%	0	0bp	0.00%	0	0bp	0.00%

Table S39. Comparison of the proportion and detailed classification of detection results generated by three tools covering the repeat regions on the reference genome of Human(hg38).

LongRepMarker				RepARK			Repdenovo		
sequence: 455				**sequence: 455**			**sequence: 455**		
total length: 3209286105bp				**total length: 3209286105bp**			**total length: 3209286105bp**		
bases masked: 147019720 bp (4.58%)				**bases masked: 50576113 bp (1.58%)**			**bases masked: 7539026 bp (0.23%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	22247	3083222bp	0.10%	12587	1576732bp	0.05%	0	0bp	0.00%
–Academ-1:	63	167248bp	0.01%	44	3395bp	0.00%	0	0bp	0.00%
–CMC-Chapaev:	4	878bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–CMC-EnSpm:	8427	422704bp	0.01%	190	24218bp	0.00%	0	0bp	0.00%
–CMC-Transib:	17	2035bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Crypton-H:	246	17570bp	0.00%	7	206bp	0.00%	0	0bp	0.00%
–Crypton-V:	2	631bp	0.00%	21	4403bp	0.00%	0	0bp	0.00%
–Dada:	21	1323bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Ginger:	971	80111bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–IS3EU:	31	1150bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Kolobok:	6	894bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Kolobok-Hydra:	4	769bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Kolobok-T2:	16	1288bp	0.00%	22	3454bp	0.00%	0	0bp	0.00%
–MULE-MuDR:	405	52301bp	0.00%	591	45645bp	0.00%	0	0bp	0.00%
–MULE-NOF:	6	915bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Maverick:	41	3734bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Merlin:	2	471bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MuLE-MuDR:	4	804bp	0.00%	17	19219bp	0.00%	0	0bp	0.00%
–Novosib:	471	64506bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	191	25705bp	0.00%	29	14236bp	0.00%	0	0bp	0.00%
–P:	19	2086bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–P-Fungi:	2	417bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PIF-Harbinger:	131	15044bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–PiggyBac:	69	22615bp	0.00%	99	7503bp	0.00%	0	0bp	0.00%
–PiggyBac-X:	0	0bp	0.00%	10	640bp	0.00%	0	0bp	0.00%
–Sola-1:	92	6418bp	0.00%	12	2508bp	0.00%	0	0bp	0.00%
–Sola-2:	37	2992bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Sola-3:	116	23940bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Fot1:	264	11590bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–TcMar-Mariner:	484	113110bp	0.00%	379	101174bp	0.00%	0	0bp	0.00%
–TcMar-Tc1:	228	25853bp	0.00%	55	5227bp	0.00%	0	0bp	0.00%
–TcMar-Tc2:	135	28407bp	0.00%	106	10915bp	0.00%	0	0bp	0.00%
–TcMar-Tigger:	3545	783065bp	0.02%	5800	567017bp	0.02%	0	0bp	0.00%
–Zisupton:	382	62163bp	0.00%	34	8472bp	0.00%	0	0bp	0.00%
–hAT:	55	4977bp	0.00%	15	808bp	0.00%	0	0bp	0.00%
–hAT-Ac:	134	22225bp	0.00%	60	5422bp	0.00%	0	0bp	0.00%
–hAT-Blackjack:	184	30609bp	0.00%	82	9624bp	0.00%	0	0bp	0.00%
–hAT-Charlie:	4346	909988bp	0.03%	4546	670099bp	0.02%	0	0bp	0.00%
–hAT-Pegasus:	10	1173bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Tag1:	47	9790bp	0.00%	12	2556bp	0.00%	0	0bp	0.00%
–hAT-Tip100:	1025	178235bp	0.01%	456	70325bp	0.00%	0	0bp	0.00%
–hAT-hAT19:	2	387bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-hAT5:	10	1933bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-hATm:	2	476bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
DNA?:	6	832bp	0.00%	7	700bp	0.00%	0	0bp	0.00%
–hAT:	6	832bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–hAT-Tip100:	0	0bp	0.00%	7	700bp	0.00%	0	0bp	0.00%
LINE:	2385096	77705196bp	2.42%	77613	7434368bp	0.23%	1214	1024575bp	0.03%
–CR1:	361	134160bp	0.00%	128	40599bp	0.00%	0	0bp	0.00%
–Dong-R4:	6	823bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–I:	14	2769bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–I-Jockey:	80	12016bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–L1:	2379706	76629820bp	2.39%	74885	6895264bp	0.21%	1214	1024575bp	0.03%
–L1-Tx1:	49	8210bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–L2:	4412	814174bp	0.03%	2421	462834bp	0.01%	0	0bp	0.00%
–OTHER:	40	44303bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Penelope:	116	13680bp	0.00%	27	514bp	0.00%	0	0bp	0.00%
–Proto1:	2	401bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–R1:	15	433bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–R2:	33	7271bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–R2-NeSL:	16	2133bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–RTE-BovB:	16	4881bp	0.00%	44	15589bp	0.00%	0	0bp	0.00%
–RTE-X:	214	43075bp	0.00%	108	19630bp	0.00%	0	0bp	0.00%
–Rex-Babar:	16	504bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
LTR:	27965	7748892bp	0.24%	57872	5715985bp	0.18%	54	6791bp	0.00%
–Caulimovirus:	4	466bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–Copia:	539	83436bp	0.00%	175	34784bp	0.00%	0	0bp	0.00%
–DIRS:	23	3721bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–ERV:	40	16576bp	0.00%	18	1548bp	0.00%	0	0bp	0.00%
–ERV1:	10834	3244357bp	0.10%	27501	2476441bp	0.08%	0	0bp	0.00%
–ERVK:	1641	789578bp	0.02%	7918	640141bp	0.02%	0	0bp	0.00%
–ERVL:	5216	1386875bp	0.04%	8382	960204bp	0.03%	0	0bp	0.00%
–ERVL-MaLR:	8360	1950768bp	0.06%	13357	1487485bp	0.05%	54	6791bp	0.00%
–Gypsy:	967	208751bp	0.01%	363	90801bp	0.00%	0	0bp	0.00%
–Ngaro:	37	7094bp	0.00%	12	2310bp	0.00%	0	0bp	0.00%
–OTHER:	248	61774bp	0.00%	117	28456bp	0.00%	0	0bp	0.00%
–Pao:	56	6016bp	0.00%	29	1807bp	0.00%	0	0bp	0.00%
RC:	364	69262bp	0.00%	63	4886bp	0.00%	0	0bp	0.00%
–Helitron:	362	68888bp	0.00%	63	4886bp	0.00%	0	0bp	0.00%
–Helitron-2:	2	374bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
RC?:	19	7146bp	0.00%	13	1405bp	0.00%	0	0bp	0.00%
–Helitron:	19	7146bp	0.00%	13	1405bp	0.00%	0	0bp	0.00%
RNA:	2	532bp	0.00%	20	1236bp	0.00%	0	0bp	0.00%
–OTHER:	2	532bp	0.00%	20	1236bp	0.00%	0	0bp	0.00%
Retroposon:	1945	257205bp	0.01%	7116	266328bp	0.01%	114	22023bp	0.00%
–SVA:	1945	257205bp	0.01%	7116	266328bp	0.01%	114	22023bp	0.00%
SINE:	96269	15556436bp	0.48%	31279	3191422bp	0.10%	590	85945bp	0.00%
–5S:	11	2554bp	0.00%	54	17505bp	0.00%	0	0bp	0.00%
–7SL:	0	0bp	0.00%	41	8007bp	0.00%	0	0bp	0.00%
–Alu:	90345	14369393bp	0.45%	27884	2572731bp	0.08%	590	85945bp	0.00%
–B4:	0	0bp	0.00%	61	5348bp	0.00%	0	0bp	0.00%
–ID:	6	4328bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–L2:	4	791bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–MIR:	5849	1175266bp	0.04%	3182	564540bp	0.02%	0	0bp	0.00%
–U:	26	13119bp	0.00%	38	9164bp	0.00%	0	0bp	0.00%
–tRNA:	6	844bp	0.00%	9	14398bp	0.00%	0	0bp	0.00%
–tRNA-7SL:	2	428bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–tRNA-Core-L2:	4	443bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–tRNA-Core-RTE:	2	469bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–tRNA-RTE:	14	2458bp	0.00%	10	1250bp	0.00%	0	0bp	0.00%
SINE?:	8	3526bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
–OTHER:	8	3526bp	0.00%	0	0bp	0.00%	0	0bp	0.00%
Satellite:	390501	25815874bp	0.80%	398914	15445369bp	0.48%	76609	6253274bp	0.19%
–OTHER:	98480	1533391bp	0.05%	26951	686531bp	0.02%	505	94679bp	0.00%
–Y-chromosome:	58559	5560028bp	0.17%	106630	4401990bp	0.14%	44835	4272503bp	0.13%
–acromeric:	331	88222bp	0.00%	2520	108836bp	0.00%	0	0bp	0.00%
–centromeric:	232934	20314996bp	0.63%	262296	11108024bp	0.35%	31269	3399724bp	0.11%
–telomeric:	197	39613bp	0.00%	517	44235bp	0.00%	0	0bp	0.00%
Simple:	673702	4512722bp	0.14%	49891	954263bp	0.03%	1219	133991bp	0.00%
–repeat:	673702	4512722bp	0.14%	49891	954263bp	0.03%	1219	133991bp	0.00%
Unknown:	132734	15520152bp	0.48%	201476	16626979bp	0.52%	47	23608bp	0.00%
–OTHER:	132734	15520152bp	0.48%	201476	16626979bp	0.52%	47	23608bp	0.00%
rRNA:	25	111517bp	0.00%	198	63322bp	0.00%	0	0bp	0.00%
–OTHER:	25	111517bp	0.00%	198	63322bp	0.00%	0	0bp	0.00%
scRNA:	24	9941bp	0.00%	65	14038bp	0.00%	0	0bp	0.00%
–OTHER:	24	9941bp	0.00%	65	14038bp	0.00%	0	0bp	0.00%
snRNA:	57	22681bp	0.00%	145	17586bp	0.00%	0	0bp	0.00%
–OTHER:	57	22681bp	0.00%	145	17586bp	0.00%	0	0bp	0.00%
tRNA:	70	22076bp	0.00%	19	4088bp	0.00%	0	0bp	0.00%
–OTHER:	70	22076bp	0.00%	19	4088bp	0.00%	0	0bp	0.00%

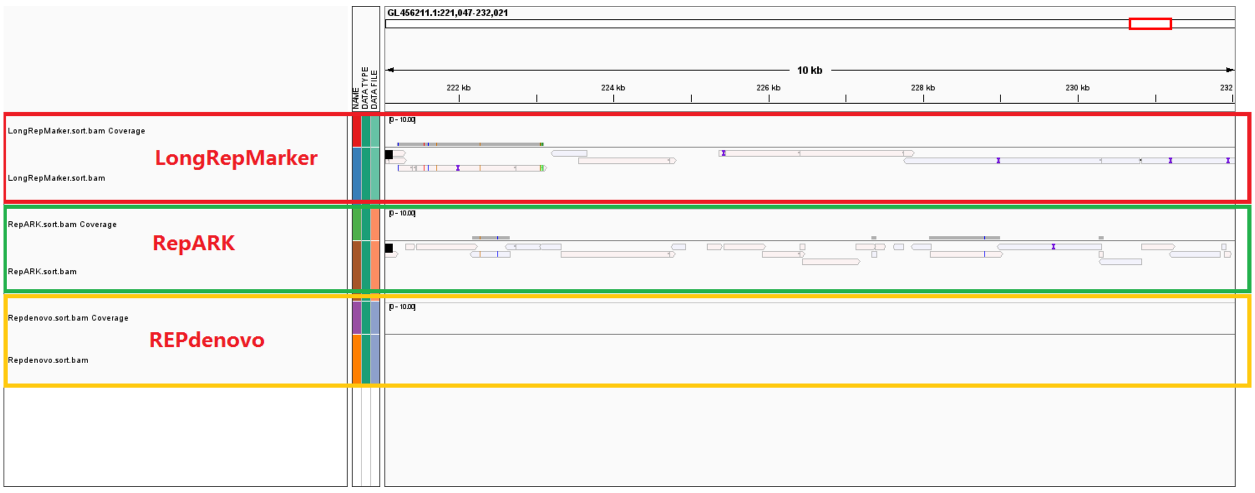


Fig. S25. An example showing the integrity and coverage of the repetitive sequences detected by LongRepMarker, RepARK and REPdenovo in the same region of the mouse genome. The pink arrow indicates the repetitive sequence covering the region (The pink arrow indicates the fragment that can be aligned multiple times which is defined in IGV[93]), and the direction of the arrow indicates the alignment direction. As can be seen from the figure, compared to RepARK, the repetitive sequence detected by LongRepMarker is more complete. In addition, we also noticed that REPdenovo did not detect any repetitive sequences in this region.

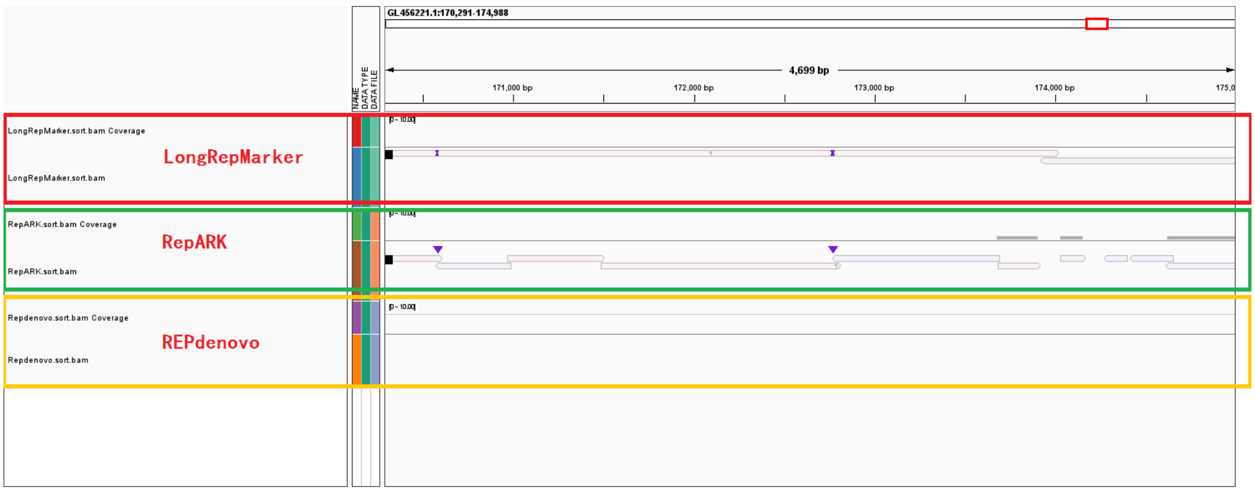


Fig. S26. Another example shows that the detection result of LongRepMarker is more complete than that of RepARK and REPdenovo. The pink arrow indicates the repetitive sequence covering the region (The pink arrow indicates the fragment that can be aligned multiple times which is defined in IGV), and the direction of the arrow indicates the alignment direction. It can be seen from the figure that the repetitive sequence detected by LongRepMarker is more complete than RepARK, and REPdenovo still has not detected any repeats in this region of the mouse genome.

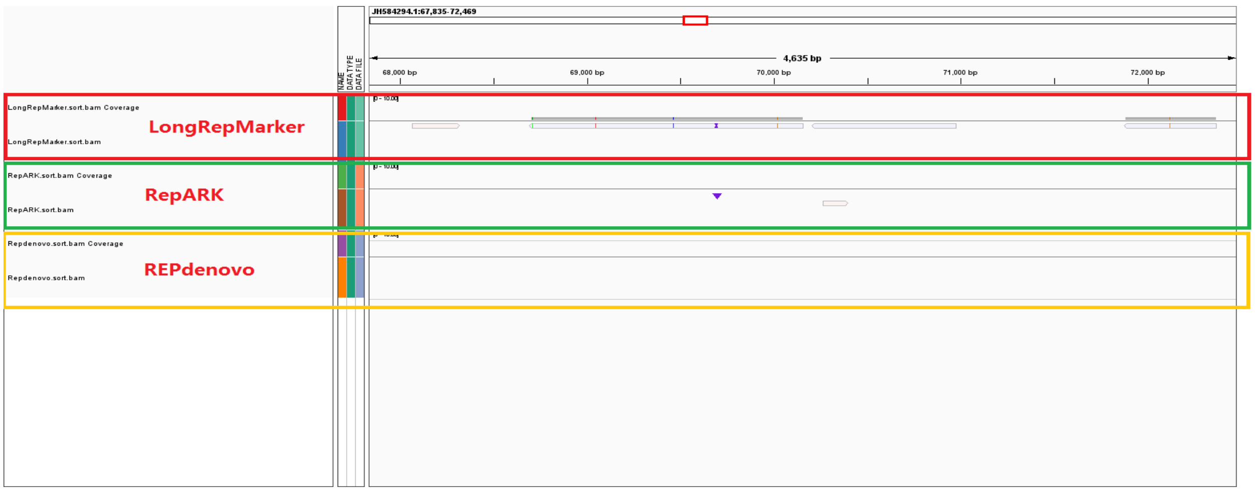


Fig. S27. An example showing that LongRepMarker can find repetitive sequences that the other two similar tools cannot recognize. The pink arrow indicates the repetitive sequence covering the region (The pink arrow indicates the fragment that can be aligned multiple times which is defined in IGV), and the direction of the arrow indicates the alignment direction. As can be seen from the figure, REPdenovo did not find any repetitive sequences in this region of the mouse genome, RepARk only found a short repetitive sequence, and LongRepMarker found large-scale repetitive sequences in this region (IGV identifies the fragments recognized by LongRepMarker as the fragments that can be aligned to multiple different locations on the mouse reference genome).

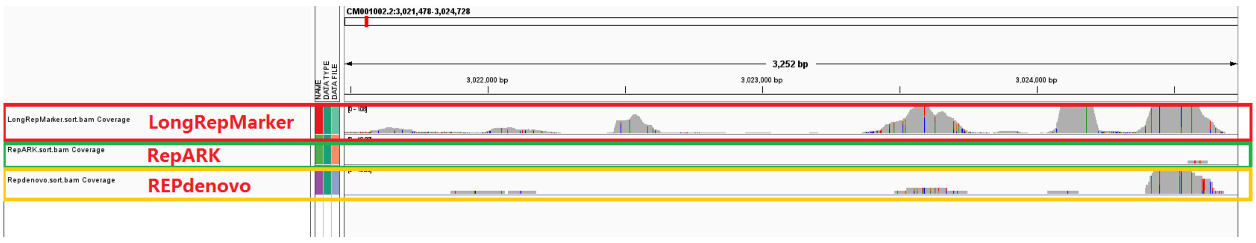


Fig. S28. A comparative example showing the coverage of repetitive sequences detected by LongRepMarker, RepARK and REPdenovo in the same region of the mouse genome. It can be seen from the figure that the high coverage regions detected by REPdenovo and RepARK can be well covered by the high coverage regions detected by LongRepMarker, and LongRepMarker can also find some high coverage regions that REPdenovo and RepARK cannot detect.

3.5.4 Detection results of *de novo* mode based on the NGS short reads + barcode linked reads / SMS long reads

In order to illustrate the effectiveness of the *de novo* mode based on the NGS short reads + barcode linked reads and the *de novo* mode based on the NGS short reads + SMS long reads of LongRepMarker, we tested these two kinds of *de novo* detection modes using three sets of real data respectively (Table S13). The NGS short paired-end reads, barcode linked reads and SMS long reads used in this experiment are downloaded from the NCBI website (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data). The detection results are shown in Fig.S29, Tables S40- S45. Since there is no tool has been proposed for repetitive sequence detection based on the mixed sequencing data, it will lose fairness if we compare the two modes of LongRepMarker with the detection methods based on single source sequencing data. Therefore, we did not compare LongRepMarker with other methods in this experiment.

The main purpose of this experiment is as follows: 1) LongRepMarker provides a detection mode based on mixed sequencing data, which can meet the needs of users to a greater extent; 2) This detection mode can make full use of the advantages of mixed sequencing data and make the detection results more superior than those obtained by using the single source sequencing data. For example, we can take HG003_NA24149_father dataset as an example to compare and analyze the test results of single source sequencing data and mixed sequencing data. The test results on single-source data cover the number and base length of DNA transposon elements in the RepBase library as 448 and 121106 bp, respectively, while the corresponding test results on mixed data are 529 and 157331 bp, respectively.

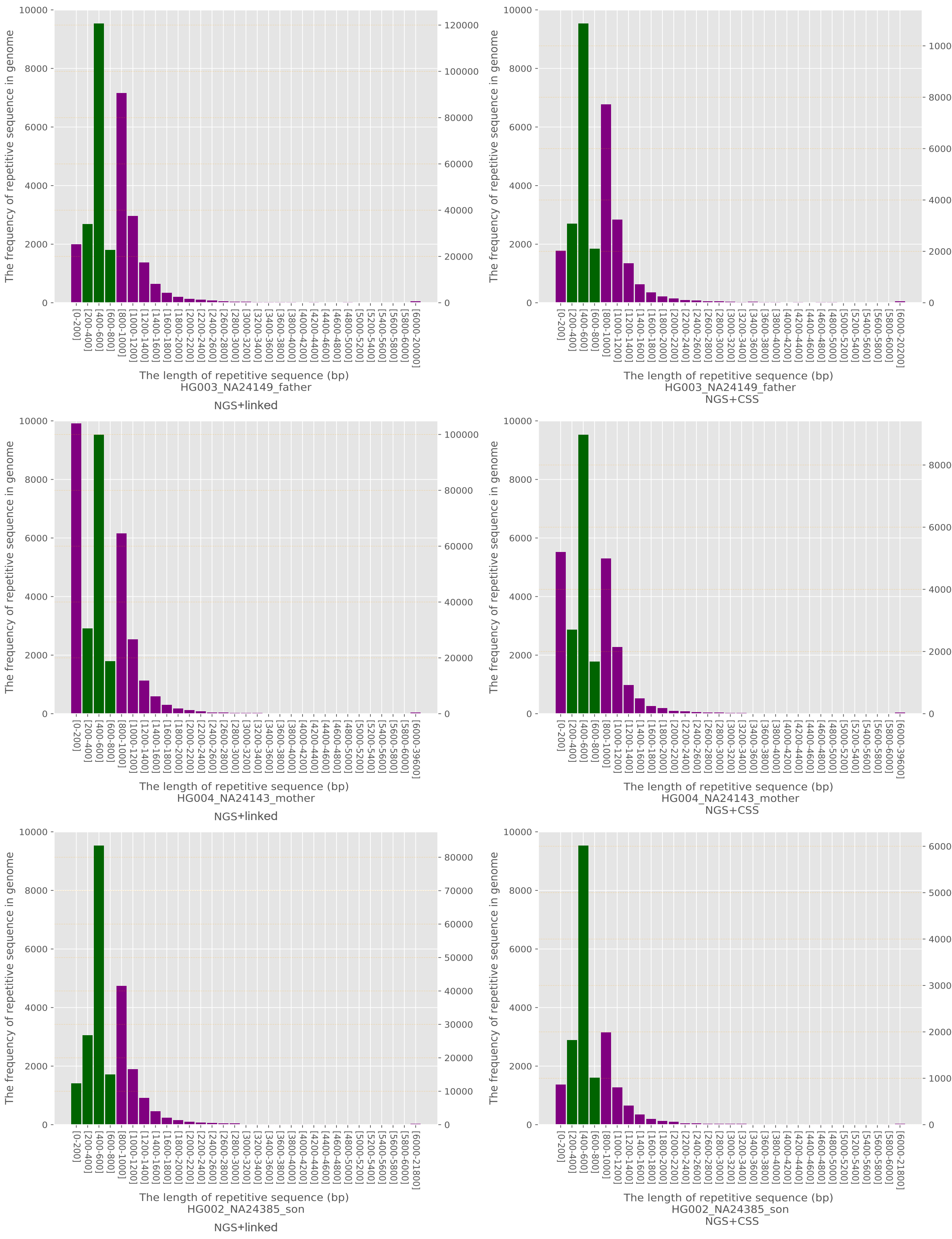


Fig. S29. Comparison of the repetition frequency and length distribution of the detected fragments generated from *de novo* mode based on the NGS short reads + barcode linked reads / SMS long reads. The X-axis represents the length distribution of the detected fragments and Y-axis represents the repetition frequency of the detected fragments in the genome, and the three images in each row respectively represent the frequency and length distribution of the repeated sequences detected by the three tools in a certain species. The coordinates of the Y-axis are divided into left and right displays, where the low frequency on the left is represented by purple, and the high frequency on the right is represented by green.

Table S40. The proportion and detailed classification of detection results in NGS+linked and NGS+CCS modes based on HG004_NA24143_father dataset covering the RepBase library of human.

	NGS+linked			NGS+CCS		
sequence: 1512				**sequence: 1512**		
total length: 1647075bp				**total length: 1647075bp**		
bases masked: 1356355 bp (82.35%)				**bases masked: 1329292 bp (80.71%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA transposon elements:	529	157331bp	9.55%	530	150076bp	9.11%
–TcMar-Tigger:	151	47828bp	2.90%	146	45008bp	2.73%
–hAT-Charlie:	155	51530bp	3.13%	153	49307bp	2.99%
LINEs:	755	342006bp	20.76%	716	317271bp	19.26%
–L3/CR1:	44	8178bp	0.50%	37	6059bp	0.37%
–LINE1:	656	322493bp	19.58%	632	299940bp	18.21%
–LINE2:	25	4521bp	0.27%	20	4231bp	0.26%
LTR elements:	1202	665236bp	40.39%	1146	657465bp	39.92%
–ERVL:	238	99978bp	6.07%	246	108103bp	6.56%
–ERVL-MaLRs:	125	30774bp	1.87%	104	25865bp	1.57%
–ERV_classI:	677	449556bp	27.29%	655	439668bp	26.69%
–ERV_classII:	81	68280bp	4.15%	72	70895bp	4.30%
Low complexity:	9	335bp	0.02%	9	335bp	0.02%
SINEs:	587	145272bp	8.82%	592	139383bp	8.46%
–ALUs:	529	136837bp	8.31%	534	131652bp	7.99%
–MIRs:	45	6844bp	0.42%	46	6414bp	0.39%
Satellites:	43	15892bp	0.96%	45	14089bp	0.86%
Simple repeats:	214	31621bp	1.92%	218	31991bp	1.94%
Small RNA:	24	2091bp	0.13%	30	2558bp	0.16%
Total interspersed repeats:		1334166bp	81.00%		1300226bp	78.94%
Unclassified:	135	24321bp	1.48%	142	36031bp	2.19%

Table S41. The proportion and detailed classification of detection results in NGS+linked and NGS+CCS modes based on HG004_NA24143_mother dataset covering the RepBase library of human.

	NGS+linked			NGS+CCS		
sequence: 1512				**sequence: 1512**		
total length: 1647075bp				**total length: 1647075bp**		
bases masked: 1329615 bp (80.73%)				**bases masked: 1293171 bp (78.51%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA transposon elements:	528	150342bp	9.13%	517	145580bp	8.84%
–TcMar-Tigger:	138	44170bp	2.68%	142	43229bp	2.62%
–hAT-Charlie:	153	49401bp	3.00%	157	42646bp	2.59%
LINEs:	775	337998bp	20.52%	701	325868bp	19.78%
–L3/CR1:	44	8516bp	0.52%	36	7402bp	0.45%
–LINE1:	671	317034bp	19.25%	606	306724bp	18.62%
–LINE2:	28	4816bp	0.29%	31	5798bp	0.35%
LTR elements:	1096	651711bp	39.57%	1085	631862bp	38.36%
–ERVL:	235	100300bp	6.09%	215	100056bp	6.07%
–ERVL-MaLRs:	105	27606bp	1.68%	96	26084bp	1.58%
–ERV_classI:	612	434666bp	26.39%	623	420143bp	25.51%
–ERV_classII:	68	74021bp	4.49%	77	71288bp	4.33%
Low complexity:	9	408bp	0.02%	9	400bp	0.02%
SINEs:	510	119810bp	7.27%	550	123213bp	7.48%
–ALUs:	449	112463bp	6.83%	490	116172bp	7.05%
–MIRs:	49	6213bp	0.38%	48	5929bp	0.36%
Satellites:	42	15290bp	0.93%	41	11923bp	0.72%
Simple repeats:	223	31906bp	1.94%	228	32195bp	1.95%
Small RNA:	30	5886bp	0.36%	25	9492bp	0.58%
Total interspersed repeats:		1291796bp	78.43%		1253440bp	76.10%
Unclassified:	134	31935bp	1.94%	139	26917bp	1.63%

Table S42. The proportion and detailed classification of detection results in NGS+linked and NGS+CCS modes based on HG002_NA24385_son dataset covering the RepBase library of human.

	NGS+linked			NGS+CCS		
sequence: 1512				**sequence: 1512**		
total length: 1647075bp				**total length: 1647075bp**		
bases masked: 1286981 bp (78.14%)				**bases masked: 1160200 bp (70.44%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA transposon elements:	534	139372bp	8.46%	456	110782bp	6.73%
–TcMar-Tigger:	143	42626bp	2.59%	112	37508bp	2.28%
–hAT-Charlie:	162	42019bp	2.55%	140	33184bp	2.01%
LINEs:	750	338807bp	20.57%	683	305328bp	18.54%
–L3/CR1:	30	6261bp	0.38%	13	2297bp	0.14%
–LINE1:	664	320914bp	19.48%	621	294719bp	17.89%
–LINE2:	23	4945bp	0.30%	26	4256bp	0.26%
LTR elements:	1120	627620bp	38.11%	1062	559412bp	33.96%
–ERVL:	201	101698bp	6.17%	191	84792bp	5.15%
–ERVL-MaLRs:	104	26163bp	1.59%	89	23499bp	1.43%
–ERV_classI:	672	420113bp	25.51%	653	384072bp	23.32%
–ERV_classII:	70	65574bp	3.98%	79	58565bp	3.56%
Low complexity:	12	465bp	0.03%	17	729bp	0.04%
SINEs:	530	107562bp	6.53%	534	117772bp	7.15%
–ALUs:	462	99461bp	6.04%	495	113330bp	6.88%
–MIRs:	56	6915bp	0.42%	32	3606bp	0.22%
Satellites:	41	14681bp	0.89%	41	15668bp	0.95%
Simple repeats:	229	32352bp	1.96%	249	33113bp	2.01%
Small RNA:	31	13146bp	0.80%	18	1299bp	0.08%
Total interspersed repeats:		1245300bp	75.61%		1122100bp	68.13%
Unclassified:	149	31939bp	1.94%	128	28806bp	1.75%

Table S43. The proportion and detailed classification of detection results in NGS+linked and NGS+CCS modes based on HG004_NA24143_father dataset covering the repetitive regions on the reference genome of human(hg38).

NGS+linked				NGS+CCS		
sequence: 455				**sequence: 455**		
total length: 3209286105bp				**total length: 3209286105bp**		
bases masked: 122984485 bp (3.83%)				**bases masked: 151962671 bp (4.74%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	13058	4401621bp	0.14%	32067	4803457bp	0.15%
–Academ-1:	18	59581bp	0.00%	55	157892bp	0.00%
–Academ-2:	1	455bp	0.00%	2	455bp	0.00%
–CMC-Chapaev:	3	1295bp	0.00%	8	939bp	0.00%
–CMC-EnSpm:	2430	339527bp	0.01%	8047	436802bp	0.01%
–CMC-Transib:	8	1964bp	0.00%	17	2035bp	0.00%
–Crypton:	3	1497bp	0.00%	6	1319bp	0.00%
–Crypton-H:	70	24174bp	0.00%	218	19281bp	0.00%
–Crypton-S:	2	834bp	0.00%	0	0bp	0.00%
–Crypton-V:	2	807bp	0.00%	6	1225bp	0.00%
–Dada:	19	6559bp	0.00%	60	6294bp	0.00%
–Ginger:	138	56000bp	0.00%	1031	94154bp	0.00%
–IS3EU:	6	486bp	0.00%	25	676bp	0.00%
–Kolobok:	5	1866bp	0.00%	10	1866bp	0.00%
–Kolobok-Hydra:	6	4440bp	0.00%	12	4440bp	0.00%
–Kolobok-T2:	19	5603bp	0.00%	34	4817bp	0.00%
–MULE-MuDR:	184	44392bp	0.00%	458	52349bp	0.00%
–MULE-NOF:	2	794bp	0.00%	4	501bp	0.00%
–Maverick:	46	11449bp	0.00%	76	9121bp	0.00%
–Merlin:	5	2263bp	0.00%	8	1792bp	0.00%
–MuLE-MuDR:	13	5206bp	0.00%	25	4439bp	0.00%
–MuLE-NOF:	1	851bp	0.00%	3	851bp	0.00%
–Novosib:	160	66797bp	0.00%	572	84340bp	0.00%
–OTHER:	142	48412bp	0.00%	295	42791bp	0.00%
–P:	13	4713bp	0.00%	30	5594bp	0.00%
–PIF-Harbinger:	88	31544bp	0.00%	213	26660bp	0.00%
–PiggyBac:	48	18644bp	0.00%	75	15797bp	0.00%
–Sola-1:	20	7998bp	0.00%	142	10959bp	0.00%
–Sola-2:	10	2012bp	0.00%	36	3262bp	0.00%
–Sola-3:	80	44548bp	0.00%	142	38834bp	0.00%
–TcMar:	1	408bp	0.00%	2	408bp	0.00%
–TcMar-Fot1:	59	6850bp	0.00%	264	11884bp	0.00%
–TcMar-ISRm11:	3	1430bp	0.00%	2	430bp	0.00%
–TcMar-Mariner:	311	117060bp	0.00%	708	135803bp	0.00%
–TcMar-Tc1:	128	48112bp	0.00%	337	52745bp	0.00%
–TcMar-Tc2:	175	72875bp	0.00%	284	59334bp	0.00%
–TcMar-Tigger:	3211	1274912bp	0.04%	6891	1415967bp	0.04%
–Zisupton:	65	29270bp	0.00%	328	27876bp	0.00%
–hAT:	49	14109bp	0.00%	101	13179bp	0.00%
–hAT-Ac:	105	37847bp	0.00%	204	37667bp	0.00%
–hAT-Blackjack:	210	71492bp	0.00%	342	60015bp	0.00%
–hAT-Charlie:	4196	1577047bp	0.05%	8969	1620371bp	0.05%
–hAT-Tag1:	48	19677bp	0.00%	92	16228bp	0.00%
–hAT-Tip100:	941	341704bp	0.01%	1921	340967bp	0.01%
–hAT-hAT1:	2	547bp	0.00%	0	0bp	0.00%
–hAT-hAT19:	2	950bp	0.00%	4	950bp	0.00%
–hAT-hAT5:	4	1852bp	0.00%	0	0bp	0.00%
–hAT-hATm:	1	476bp	0.00%	2	476bp	0.00%
–hAT-hobo:	5	1477bp	0.00%	6	999bp	0.00%
DNA?:	14	4480bp	0.00%	39	5487bp	0.00%
–PiggyBac:	4	998bp	0.00%	8	998bp	0.00%
–hAT:	6	2155bp	0.00%	10	1749bp	0.00%
–hAT-Tip100:	4	1327bp	0.00%	21	2740bp	0.00%
LINE:	141460	65356369bp	2.04%	468065	80098614bp	2.50%
–CR1:	357	148412bp	0.00%	865	185669bp	0.01%
–CRE:	1	495bp	0.00%	2	495bp	0.00%
–Dong-R4:	7	2404bp	0.00%	14	2404bp	0.00%
–I:	10	4075bp	0.00%	10	2050bp	0.00%
–I-Jockey:	16	6183bp	0.00%	31	5503bp	0.00%
–L1:	136324	63459384bp	1.98%	457411	78215864bp	2.44%
–L1-Tx1:	28	12876bp	0.00%	89	14787bp	0.00%
–L2:	4347	1571987bp	0.05%	8847	1503413bp	0.05%
–OTHER:	20	11111bp	0.00%	26	4934bp	0.00%
–Penelope:	44	15543bp	0.00%	135	17842bp	0.00%
–Proto1:	1	401bp	0.00%	2	401bp	0.00%
–R1:	21	7819bp	0.00%	74	17263bp	0.00%
–R2:	15	6859bp	0.00%	12	2748bp	0.00%
–R2-Hero:	1	401bp	0.00%	2	401bp	0.00%
–R2-NeSL:	8	2752bp	0.00%	29	4392bp	0.00%
–RTE-BovB:	56	29704bp	0.00%	106	57834bp	0.00%
–RTE-RTE:	1	391bp	0.00%	2	391bp	0.00%
–RTE-X:	188	72701bp	0.00%	370	73206bp	0.00%
–Rex-Babar:	9	1430bp	0.00%	20	1113bp	0.00%
–Tad1:	6	3288bp	0.00%	18	6538bp	0.00%
LTR:	19820	8562763bp	0.27%	45777	10284240bp	0.32%
–Caulimovirus:	3	1087bp	0.00%	6	998bp	0.00%
–Copia:	243	80077bp	0.00%	654	91005bp	0.00%
–DIRS:	5	1788bp	0.00%	13	2859bp	0.00%
–ERV:	10	4528bp	0.00%	18	4250bp	0.00%
–ERV1:	6685	3080698bp	0.10%	16275	3803333bp	0.12%
–ERV4:	1	386bp	0.00%	2	386bp	0.00%
–ERVK:	599	417919bp	0.01%	1733	744385bp	0.02%
–ERVL:	4117	1772072bp	0.06%	9471	2062304bp	0.06%
–ERVL-MaLR:	7343	2881961bp	0.09%	15475	3164255bp	0.10%
–Gypsy:	526	215484bp	0.01%	1406	300580bp	0.01%
–Ngaro:	44	16175bp	0.00%	168	23666bp	0.00%
–OTHER:	186	80434bp	0.00%	424	95892bp	0.00%
–Pao:	58	15618bp	0.00%	132	15181bp	0.00%
RC:	185	101904bp	0.00%	418	95050bp	0.00%
–Helitron:	184	101474bp	0.00%	416	94696bp	0.00%
–Helitron-2:	1	430bp	0.00%	2	354bp	0.00%
RC?:	14	7800bp	0.00%	23	7874bp	0.00%
–Helitron:	14	7800bp	0.00%	23	7874bp	0.00%
RNA:	6	2250bp	0.00%	9	1742bp	0.00%
–OTHER:	6	2250bp	0.00%	9	1742bp	0.00%
Retroposon:	204	81803bp	0.00%	1280	192958bp	0.01%
–SVA:	204	81803bp	0.00%	1280	192958bp	0.01%
SINE:	36757	16369642bp	0.51%	140349	21762573bp	0.68%
–5S:	7	2953bp	0.00%	14	2712bp	0.00%
–5S-Deu-L2:	6	3103bp	0.00%	10	2460bp	0.00%
–5S-Sauria-RTE:	2	1078bp	0.00%	4	785bp	0.00%
–7SL:	0	0bp	0.00%	2	415bp	0.00%
–Alu:	30496	14159380bp	0.44%	127439	19579671bp	0.61%
–ID:	2	846bp	0.00%	0	0bp	0.00%
–L2:	12	3249bp	0.00%	45	3877bp	0.00%
–MIR:	6192	2176568bp	0.07%	12768	2162197bp	0.07%
–U:	9	12021bp	0.00%	25	20051bp	0.00%
–tRNA:	12	4791bp	0.00%	20	7250bp	0.00%
–tRNA-7SL:	2	840bp	0.00%	4	840bp	0.00%
–tRNA-Core-L2:	2	443bp	0.00%	4	443bp	0.00%
–tRNA-Core-RTE:	2	780bp	0.00%	0	0bp	0.00%
–tRNA-Deu:	1	428bp	0.00%	0	0bp	0.00%
–tRNA-Deu-L2:	2	463bp	0.00%	0	0bp	0.00%
–tRNA-RTE:	10	3487bp	0.00%	14	2724bp	0.00%
SINE?:	4	3758bp	0.00%	0	0bp	0.00%
–OTHER:	4	3758bp	0.00%	0	0bp	0.00%
Satellite:	8510	3270195bp	0.10%	65619	9826360bp	0.31%
–OTHER:	5342	1199393bp	0.04%	23322	1544398bp	0.05%
–Y-chromosome:	959	1056067bp	0.03%	13276	4036474bp	0.13%
–acromeric:	34	44645bp	0.00%	126	73679bp	0.00%
–centromeric:	2066	948068bp	0.03%	28638	4457340bp	0.14%
–telomeric:	109	50047bp	0.00%	257	59463bp	0.00%
Simple:	26349	3667427bp	0.11%	375841	4398437bp	0.14%
–repeat:	26349	3667427bp	0.11%	375841	4398437bp	0.14%
Unknown:	59683	22588323bp	0.70%	167981	23913807bp	0.75%
–OTHER:	59683	22588323bp	0.70%	167981	23913807bp	0.75%
rRNA:	15	6123bp	0.00%	27	6354bp	0.00%
–OTHER:	15	6123bp	0.00%	27	6354bp	0.00%
scRNA:	16	5469bp	0.00%	16	2561bp	0.00%
–OTHER:	16	5469bp	0.00%	16	2561bp	0.00%
snRNA:	38	17729bp	0.00%	126	21510bp	0.00%
–OTHER:	38	17729bp	0.00%	126	21510bp	0.00%
tRNA:	37	23243bp	0.00%	72	19646bp	0.00%
–OTHER:	37	23243bp	0.00%	72	19646bp	0.00%

Table S44. The proportion and detailed classification of detection results in NGS+linked and NGS+CCS modes based on HG004_NA24143_mother dataset covering the repetitive regions on the reference genome of human(hg38).

NGS+linked				NGS+CCS		
sequence: 455				**sequence: 455**		
total length: 3209286105bp				**total length: 3209286105bp**		
bases masked: 103986182 bp (3.24%)				**bases masked: 138766579 bp (4.32%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	11430	3712078bp	0.12%	28587	4175426bp	0.13%
–Academ:	1	472bp	0.00%	2	472bp	0.00%
–Academ-1:	14	58634bp	0.00%	28	95924bp	0.00%
–CMC-Chapaev:	3	897bp	0.00%	2	406bp	0.00%
–CMC-EnSpm:	2540	326869bp	0.01%	8791	455183bp	0.01%
–Crypton:	4	1105bp	0.00%	6	684bp	0.00%
–Crypton-A:	8	3385bp	0.00%	17	4210bp	0.00%
–Crypton-H:	37	12813bp	0.00%	333	20073bp	0.00%
–Crypton-S:	9	1616bp	0.00%	15	1673bp	0.00%
–Crypton-V:	2	1053bp	0.00%	2	631bp	0.00%
–Dada:	16	4395bp	0.00%	35	6317bp	0.00%
–Ginger:	167	57925bp	0.00%	1235	92194bp	0.00%
–IS3EU:	7	990bp	0.00%	25	997bp	0.00%
–Kolobok:	4	990bp	0.00%	59	2762bp	0.00%
–Kolobok-Hydra:	3	1398bp	0.00%	6	1398bp	0.00%
–Kolobok-T2:	10	1430bp	0.00%	33	1764bp	0.00%
–MULE-MuDR:	173	41795bp	0.00%	369	45549bp	0.00%
–MULE-NOF:	0	0bp	0.00%	5	147bp	0.00%
–Maverick:	66	13335bp	0.00%	121	10689bp	0.00%
–Merlin:	6	2040bp	0.00%	6	1017bp	0.00%
–MuLE-MuDR:	16	8696bp	0.00%	63	20136bp	0.00%
–Novosib:	136	55993bp	0.00%	492	58875bp	0.00%
–OTHER:	119	40096bp	0.00%	301	44633bp	0.00%
–P:	15	3851bp	0.00%	37	3088bp	0.00%
–PIF-Harbinger:	59	20642bp	0.00%	107	19318bp	0.00%
–PIF-Spy:	7	1986bp	0.00%	13	1646bp	0.00%
–PiggyBac:	47	22035bp	0.00%	87	20492bp	0.00%
–PiggyBac-X:	1	417bp	0.00%	4	955bp	0.00%
–Sola-1:	24	10943bp	0.00%	56	10219bp	0.00%
–Sola-2:	5	1033bp	0.00%	10	1196bp	0.00%
–Sola-3:	51	28191bp	0.00%	98	24937bp	0.00%
–TcMar-Cweed:	1	431bp	0.00%	2	431bp	0.00%
–TcMar-Fot1:	73	15130bp	0.00%	222	15984bp	0.00%
–TcMar-Mariner:	238	97854bp	0.00%	485	94346bp	0.00%
–TcMar-Pogo:	1	427bp	0.00%	2	427bp	0.00%
–TcMar-Tc1:	81	25803bp	0.00%	253	29708bp	0.00%
–TcMar-Tc2:	113	37963bp	0.00%	222	37695bp	0.00%
–TcMar-Tigger:	2669	1103809bp	0.03%	5605	1380636bp	0.04%
–Zisupton:	39	21791bp	0.00%	233	32422bp	0.00%
–hAT:	41	14985bp	0.00%	90	9651bp	0.00%
–hAT-Ac:	55	18716bp	0.00%	134	16037bp	0.00%
–hAT-Blackjack:	196	62473bp	0.00%	359	58080bp	0.00%
–hAT-Charlie:	3550	1320017bp	0.04%	7061	1305964bp	0.04%
–hAT-Pegasus:	0	0bp	0.00%	2	315bp	0.00%
–hAT-Tag1:	41	15347bp	0.00%	56	10512bp	0.00%
–hAT-Tip100:	772	262038bp	0.01%	1495	256550bp	0.01%
–hAT-hAT5:	4	1860bp	0.00%	0	0bp	0.00%
–hAT-hATx:	1	427bp	0.00%	0	0bp	0.00%
–hAT-hobo:	5	802bp	0.00%	8	802bp	0.00%
DNA?:	14	4294bp	0.00%	21	3054bp	0.00%
–PiggyBac:	8	1688bp	0.00%	6	625bp	0.00%
–hAT:	4	1722bp	0.00%	13	2006bp	0.00%
–hAT-Tip100:	2	884bp	0.00%	2	423bp	0.00%
LINE:	124678	56785004bp	1.77%	632523	74183290bp	2.31%
–CR1:	321	125569bp	0.00%	526	112630bp	0.00%
–CRE-Ambal:	5	3148bp	0.00%	12	4410bp	0.00%
–Dong-R4:	10	4486bp	0.00%	18	3984bp	0.00%
–I:	3	2031bp	0.00%	8	2213bp	0.00%
–I-Jockey:	15	5449bp	0.00%	21	3510bp	0.00%
–Jockey:	5	2002bp	0.00%	10	2002bp	0.00%
–L1:	120704	55335616bp	1.72%	625070	72767785bp	2.27%
–L1-Tx1:	19	6525bp	0.00%	23	4638bp	0.00%
–L2:	3327	1195153bp	0.04%	6313	1157059bp	0.04%
–OTHER:	15	5696bp	0.00%	19	3362bp	0.00%
–Penelope:	21	5689bp	0.00%	42	5145bp	0.00%
–R1:	1	417bp	0.00%	27	10816bp	0.00%
–R1-LOA:	3	1241bp	0.00%	2	588bp	0.00%
–R2:	7	3415bp	0.00%	12	2801bp	0.00%
–R2-Hero:	3	710bp	0.00%	6	710bp	0.00%
–R2-NeSL:	2	207bp	0.00%	4	207bp	0.00%
–RTE-BovB:	25	11496bp	0.00%	41	8149bp	0.00%
–RTE-RTE:	1	390bp	0.00%	2	390bp	0.00%
–RTE-X:	187	75097bp	0.00%	358	96124bp	0.00%
–Rex-Babar:	1	389bp	0.00%	2	389bp	0.00%
–Tad1:	3	802bp	0.00%	7	1030bp	0.00%
LINE?:	1	446bp	0.00%	0	0bp	0.00%
–Penelope:	1	446bp	0.00%	0	0bp	0.00%
LTR:	16569	7084889bp	0.22%	37288	8698665bp	0.27%
–Caulimovirus:	3	437bp	0.00%	6	437bp	0.00%
–Copia:	224	66495bp	0.00%	552	63460bp	0.00%
–DIRS:	3	1263bp	0.00%	10	2107bp	0.00%
–ERV:	7	4989bp	0.00%	17	5906bp	0.00%
–ERV-Lenti:	1	560bp	0.00%	0	0bp	0.00%
–ERV1:	5574	2541733bp	0.08%	13854	3375313bp	0.11%
–ERV4:	1	409bp	0.00%	0	0bp	0.00%
–ERVK:	574	405326bp	0.01%	1338	670674bp	0.02%
–ERVL:	3554	1519841bp	0.05%	7471	1718916bp	0.05%
–ERVL-MaLR:	5883	2278539bp	0.07%	12364	2546184bp	0.08%
–Gypsy:	486	186334bp	0.01%	1108	249564bp	0.01%
–Ngaro:	44	15630bp	0.00%	186	20442bp	0.00%
–OTHER:	165	60005bp	0.00%	285	56800bp	0.00%
–Pao:	50	10819bp	0.00%	97	7900bp	0.00%
RC:	159	69151bp	0.00%	403	88167bp	0.00%
–Helitron:	159	69151bp	0.00%	403	88167bp	0.00%
RC?:	3	1583bp	0.00%	6	1371bp	0.00%
–Helitron:	3	1583bp	0.00%	6	1371bp	0.00%
Retroposon:	141	57991bp	0.00%	820	121527bp	0.00%
–SVA:	141	57991bp	0.00%	820	121527bp	0.00%
SINE:	30054	13133447bp	0.41%	106196	17610049bp	0.55%
–5S:	5	1714bp	0.00%	28	3188bp	0.00%
–5S-Deu-L2:	4	1954bp	0.00%	12	2144bp	0.00%
–5S-Sauria-RTE:	1	424bp	0.00%	0	0bp	0.00%
–7SL:	3	5788bp	0.00%	82	11270bp	0.00%
–Alu:	24901	11331590bp	0.35%	96239	15788453bp	0.49%
–B4:	1	457bp	0.00%	2	457bp	0.00%
–ID:	5	1801bp	0.00%	12	2067bp	0.00%
–L2:	7	2129bp	0.00%	18	1447bp	0.00%
–MIR:	5095	1775930bp	0.06%	9733	1804428bp	0.06%
–U:	6	4206bp	0.00%	19	9074bp	0.00%
–tRNA:	10	4338bp	0.00%	25	4853bp	0.00%
–tRNA-Core-RTE:	1	401bp	0.00%	2	401bp	0.00%
–tRNA-Deu:	1	446bp	0.00%	0	0bp	0.00%
–tRNA-RTE:	13	6432bp	0.00%	22	5829bp	0.00%
–tRNA-V:	1	381bp	0.00%	2	381bp	0.00%
SINE?:	1	376bp	0.00%	2	376bp	0.00%
–OTHER:	1	376bp	0.00%	2	376bp	0.00%
Satellite:	17270	3026886bp	0.09%	113853	13902539bp	0.43%
–OTHER:	13561	1061734bp	0.03%	28998	1364080bp	0.04%
–Y-chromosome:	875	837966bp	0.03%	21370	4763785bp	0.15%
–acromeric:	65	51957bp	0.00%	135	49213bp	0.00%
–centromeric:	2686	1067644bp	0.03%	63128	8687223bp	0.27%
–macro:	0	0bp	0.00%	8	1148bp	0.00%
–telomeric:	83	33340bp	0.00%	214	42649bp	0.00%
Simple:	25653	3270523bp	0.10%	379042	3956075bp	0.12%
–repeat:	25653	3270523bp	0.10%	379042	3956075bp	0.12%
Unknown:	48558	17971491bp	0.56%	133876	18590129bp	0.58%
–OTHER:	48558	17971491bp	0.56%	133876	18590129bp	0.58%
rRNA:	15	48620bp	0.00%	33	51027bp	0.00%
–OTHER:	15	48620bp	0.00%	33	51027bp	0.00%
scRNA:	20	10894bp	0.00%	46	14988bp	0.00%
–OTHER:	20	10894bp	0.00%	46	14988bp	0.00%
snRNA:	26	12907bp	0.00%	124	20343bp	0.00%
–OTHER:	26	12907bp	0.00%	124	20343bp	0.00%
tRNA:	26	24732bp	0.00%	54	25999bp	0.00%
–OTHER:	26	24732bp	0.00%	54	25999bp	0.00%

Table S45. The proportion and detailed classification of detection results in NGS+linked and NGS+CCS modes based on HG002_NA24385_son dataset covering the repetitive regions on the reference genome of human(hg38).

NGS+linked				NGS+CCS		
sequence: 455				**sequence: 455**		
total length: 3209286105bp				**total length: 3209286105bp**		
bases masked: 135221591 bp (4.21%)				**bases masked: 94132889 bp (2.93%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	27230	3869265bp	0.12%	20223	2881447bp	0.09%
–Academ-1:	30	111186bp	0.00%	78	286562bp	0.01%
–CMC-Chapaev:	10	703bp	0.00%	8	256bp	0.00%
–CMC-EnSpm:	7858	404955bp	0.01%	7713	406961bp	0.01%
–CMC-Transib:	27	2381bp	0.00%	23	1940bp	0.00%
–Crypton:	2	441bp	0.00%	2	441bp	0.00%
–Crypton-A:	6	913bp	0.00%	0	0bp	0.00%
–Crypton-C:	2	410bp	0.00%	0	0bp	0.00%
–Crypton-H:	313	19107bp	0.00%	311	18713bp	0.00%
–Crypton-S:	3	648bp	0.00%	3	648bp	0.00%
–Crypton-V:	8	990bp	0.00%	6	534bp	0.00%
–Dada:	8	969bp	0.00%	12	1483bp	0.00%
–Ginger:	615	64011bp	0.00%	560	56691bp	0.00%
–IS3EU:	25	676bp	0.00%	25	676bp	0.00%
–Kolobok-Hydra:	6	1077bp	0.00%	6	1077bp	0.00%
–Kolobok-T2:	21	2772bp	0.00%	13	1246bp	0.00%
–MULE-F:	0	0bp	0.00%	4	494bp	0.00%
–MULE-MuDR:	414	48306bp	0.00%	377	41204bp	0.00%
–MULE-NOF:	5	147bp	0.00%	5	147bp	0.00%
–Maverick:	83	8894bp	0.00%	71	6381bp	0.00%
–Merlin:	6	1378bp	0.00%	15	2153bp	0.00%
–MuLE-MuDR:	14	2994bp	0.00%	36	5416bp	0.00%
–MuLE-NOF?:	4	187bp	0.00%	4	187bp	0.00%
–Novosib:	810	80087bp	0.00%	871	74122bp	0.00%
–OTHER:	321	40205bp	0.00%	199	20973bp	0.00%
–P:	29	6695bp	0.00%	29	7157bp	0.00%
–PIF-Harbinger:	160	24218bp	0.00%	174	18455bp	0.00%
–PIF-Spy:	6	1345bp	0.00%	2	693bp	0.00%
–PiggyBac:	91	16627bp	0.00%	41	7990bp	0.00%
–PiggyBac-X:	8	1504bp	0.00%	8	1504bp	0.00%
–Sola-1:	28	4385bp	0.00%	27	3722bp	0.00%
–Sola-2:	31	2617bp	0.00%	23	925bp	0.00%
–Sola-3:	96	27459bp	0.00%	123	26075bp	0.00%
–TcMar:	10	1965bp	0.00%	2	387bp	0.00%
–TcMar-Fot1:	216	12765bp	0.00%	210	10990bp	0.00%
–TcMar-Mariner:	489	174261bp	0.01%	236	62599bp	0.00%
–TcMar-Pogo:	4	487bp	0.00%	10	2245bp	0.00%
–TcMar-Sagan:	0	0bp	0.00%	4	426bp	0.00%
–TcMar-Tc1:	187	30395bp	0.00%	165	23992bp	0.00%
–TcMar-Tc2:	167	38437bp	0.00%	98	26127bp	0.00%
–TcMar-Tigger:	5670	1059000bp	0.03%	3271	699262bp	0.02%
–Zisupton:	218	38702bp	0.00%	180	47816bp	0.00%
–hAT:	69	9792bp	0.00%	42	3213bp	0.00%
–hAT-Ac:	435	36821bp	0.00%	361	27142bp	0.00%
–hAT-Blackjack:	449	67699bp	0.00%	176	31575bp	0.00%
–hAT-Charlie:	6919	1284494bp	0.04%	3812	812111bp	0.03%
–hAT-Pegasus:	4	784bp	0.00%	2	393bp	0.00%
–hAT-Tag1:	96	16085bp	0.00%	27	4267bp	0.00%
–hAT-Tip100:	1235	229878bp	0.01%	842	146206bp	0.00%
–hAT-hAT19:	2	481bp	0.00%	0	0bp	0.00%
–hAT-hAT5:	4	851bp	0.00%	0	0bp	0.00%
–hAT-hATm:	14	1692bp	0.00%	14	1359bp	0.00%
–hAT-hobo:	2	436bp	0.00%	2	433bp	0.00%
DNA?:	40	9199bp	0.00%	13	4134bp	0.00%
–PiggyBac:	4	878bp	0.00%	0	0bp	0.00%
–hAT:	19	6465bp	0.00%	13	4134bp	0.00%
–hAT-Tip100:	17	1856bp	0.00%	0	0bp	0.00%
LINE:	924114	71797790bp	2.24%	287583	51970980bp	1.62%
–CR1:	530	127976bp	0.00%	329	106367bp	0.00%
–Dong-R4:	16	2894bp	0.00%	8	1202bp	0.00%
–I:	10	2143bp	0.00%	16	4872bp	0.00%
–I-Jockey:	95	12685bp	0.00%	87	9785bp	0.00%
–L1:	917469	70581215bp	2.20%	283294	51129816bp	1.59%
–L1-Tx1:	25	5551bp	0.00%	25	5019bp	0.00%
–L2:	5460	979090bp	0.03%	3432	654365bp	0.02%
–OTHER:	54	5723bp	0.00%	12	2655bp	0.00%
–Penelope:	68	7939bp	0.00%	70	6391bp	0.00%
–R1:	36	6228bp	0.00%	32	5496bp	0.00%
–R2:	14	2776bp	0.00%	35	10775bp	0.00%
–R2-NeSL:	4	207bp	0.00%	4	207bp	0.00%
–RTE-BovB:	57	10871bp	0.00%	44	7702bp	0.00%
–RTE-RTE:	4	780bp	0.00%	4	780bp	0.00%
–RTE-X:	253	65945bp	0.00%	174	42455bp	0.00%
–Rex-Babar:	17	494bp	0.00%	17	494bp	0.00%
–Tad1:	2	481bp	0.00%	0	0bp	0.00%
LTR:	34626	8043123bp	0.25%	22423	5570154bp	0.17%
–Caulimovirus:	38	9889bp	0.00%	16	4203bp	0.00%
–Copia:	532	74888bp	0.00%	337	51874bp	0.00%
–DIRS:	40	13144bp	0.00%	28	10640bp	0.00%
–ERV:	31	10494bp	0.00%	6	1211bp	0.00%
–ERV-Foamy:	6	609bp	0.00%	6	609bp	0.00%
–ERV1:	13056	3223269bp	0.10%	8316	2258265bp	0.07%
–ERVK:	1421	673655bp	0.02%	1054	582062bp	0.02%
–ERVL:	6449	1461089bp	0.05%	3839	889364bp	0.03%
–ERVL-MaLR:	11591	2300367bp	0.07%	7658	1568594bp	0.05%
–Gypsy:	1062	223175bp	0.01%	885	163522bp	0.01%
–Ngaro:	126	19841bp	0.00%	105	12036bp	0.00%
–OTHER:	204	45428bp	0.00%	120	29481bp	0.00%
–Pao:	70	6582bp	0.00%	53	3350bp	0.00%
RC:	289	60235bp	0.00%	173	40962bp	0.00%
–Helitron:	284	59033bp	0.00%	173	40962bp	0.00%
–Helitron-2:	5	1202bp	0.00%	0	0bp	0.00%
RC?:	27	7104bp	0.00%	17	5331bp	0.00%
–Helitron:	27	7104bp	0.00%	17	5331bp	0.00%
Retroposon:	1160	194251bp	0.01%	1263	235984bp	0.01%
–OTHER:	2	471bp	0.00%	0	0bp	0.00%
–SVA:	1158	193780bp	0.01%	1263	235984bp	0.01%
SINE:	99069	15412336bp	0.48%	72475	11746999bp	0.37%
–5S:	3	714bp	0.00%	0	0bp	0.00%
–5S-Deu-L2:	51	2918bp	0.00%	4	461bp	0.00%
–Alu:	90342	13835482bp	0.43%	68239	10796740bp	0.34%
–B4:	4	810bp	0.00%	2	431bp	0.00%
–ID:	6	1218bp	0.00%	10	1185bp	0.00%
–L2:	8	1116bp	0.00%	0	0bp	0.00%
–MIR:	8617	1553217bp	0.05%	4167	926547bp	0.03%
–U:	18	17429bp	0.00%	37	25511bp	0.00%
–tRNA:	6	1287bp	0.00%	4	860bp	0.00%
–tRNA-7SL:	4	834bp	0.00%	4	834bp	0.00%
–tRNA-Deu:	4	884bp	0.00%	2	465bp	0.00%
–tRNA-RTE:	6	1288bp	0.00%	6	1288bp	0.00%
SINE?:	14	1587bp	0.00%	0	0bp	0.00%
–OTHER:	14	1587bp	0.00%	0	0bp	0.00%
Satellite:	163786	16058560bp	0.50%	39325	7649229bp	0.24%
–OTHER:	45438	1329674bp	0.04%	15689	1132905bp	0.04%
–Y-chromosome:	26788	5530350bp	0.17%	7826	3882069bp	0.12%
–acromeric:	111	68159bp	0.00%	127	61652bp	0.00%
–centromeric:	91234	10922743bp	0.34%	15484	2777357bp	0.09%
–telomeric:	215	37712bp	0.00%	199	50810bp	0.00%
Simple:	459844	4216142bp	0.13%	352037	4062131bp	0.13%
–repeat:	459844	4216142bp	0.13%	352037	4062131bp	0.13%
Unknown:	132726	18003473bp	0.56%	89618	12107533bp	0.38%
–OTHER:	132726	18003473bp	0.56%	89618	12107533bp	0.38%
rRNA:	39	109343bp	0.00%	18	3097bp	0.00%
–OTHER:	39	109343bp	0.00%	18	3097bp	0.00%
scRNA:	28	11837bp	0.00%	21	8012bp	0.00%
–OTHER:	28	11837bp	0.00%	21	8012bp	0.00%
snRNA:	83	21098bp	0.00%	51	11892bp	0.00%
–OTHER:	83	21098bp	0.00%	51	11892bp	0.00%
tRNA:	88	30840bp	0.00%	58	25330bp	0.00%
–OTHER:	88	30840bp	0.00%	58	25330bp	0.00%

3.5.5 *de novo* detection mode based on only the SMS long reads

In order to better comply with market demand and further expand the application scope of this system, we have developed a new detection mode based on only the SMS long reads under the LongRepMarker framework. Compared with the existing detection methods based on the SMS long reads, this mode has the advantages of longer fragments, lower memory consumption, higher speed and higher detection accuracy. The input of this mode is only SMS long reads and the output is the detection results which contain the final repeat library and some reports.

RepLong is a novel *de novo* repeat elements identification method based on PacBio long reads. RepLong can handle lower coverage data and serve as a complementary solution to the existing methods to promote the repeat identification performance on long read sequencing data. In order to verify the detection performance of the *de novo* detection mode based on only the SMS long reads, we carried out a performance comparison between LongRepMarker and RepLong on 4 sets of pacbio real datasets. The detection results are shown in Fig. S30, and Tables S46-S53. From the results displayed in Fig. S30, we can find that the max fragment of detected results generated from LongRepMarker based on the drosophila_100k dataset is 31.400kb, while the corresponding value of RepLong is 14.800kb, and the proportion of detected fragments of LongRepMarker covering the RepBase library is 37.29%, as compared to 19.56% for RepLong. The data selected in the experiment comes from the RepLong website (Table S13), where the coverage of the first two datasets is low, and the coverage of the latter two datasets is relatively high. In order to compare with RepLong under the low and high coverage conditions, we also chose the same datasets for testing.

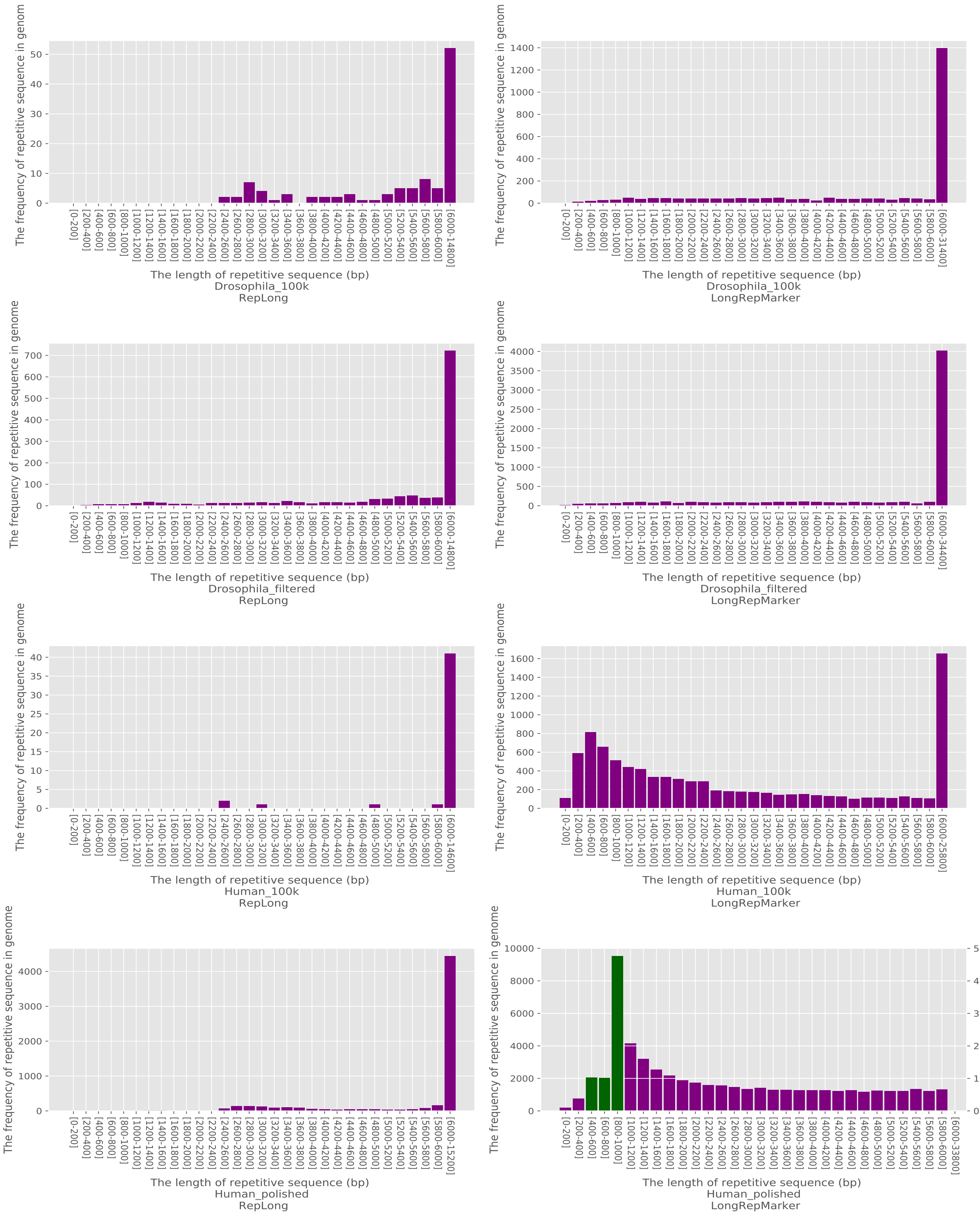


Fig. S30. Comparison of the repetition frequency and length distribution of the detected fragments generated from two tools. The X-axis represents the length distribution of the detected fragments and Y-axis represents the repetition frequency of the detected fragments in the genome, and the three images in each row respectively represent the frequency and length distribution of the repeated sequences detected by the three tools in a certain species. The coordinates of the Y-axis are divided into left and right displays, where the low frequency on the left is represented by purple, and the high frequency on the right is represented by green.

Table S46. Comparison of the proportion and detailed classification of detection results generated by two tools on drosophila_100k dataset covering the corresponding RepBase library.

RepLong				LongRepMarker		
sequence: 2489				**sequence: 2489**		
total length: 7220516bp				**total length: 7220516bp**		
bases masked: 1411983 bp (19.56%)				**bases masked: 2692788 bp (37.29%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA transposon elements:	28	12922bp	0.18%	94	34437bp	0.48%
–TcMar-Tigger:	0	0bp	0.00%	0	0bp	0.00%
–hAT-Charlie:	0	0bp	0.00%	0	0bp	0.00%
LINEs:	374	432450bp	5.99%	1427	932778bp	12.92%
–L3/CR1:	0	0bp	0.00%	128	75157bp	1.04%
–LINE1:	0	0bp	0.00%	1	87bp	0.00%
–LINE2:	0	0bp	0.00%	0	0bp	0.00%
LTR elements:	671	836759bp	11.59%	2043	1596662bp	22.11%
–ERVL:	0	0bp	0.00%	0	0bp	0.00%
–ERVL-MaLRs:	0	0bp	0.00%	0	0bp	0.00%
–ERV_classI:	0	0bp	0.00%	0	0bp	0.00%
–ERV_classII:	0	0bp	0.00%	0	0bp	0.00%
Low complexity:	379	20236bp	0.28%	329	17571bp	0.24%
SINEs:	0	0bp	0.00%	0	0bp	0.00%
–ALUs:	0	0bp	0.00%	0	0bp	0.00%
–MIRs:	0	0bp	0.00%	0	0bp	0.00%
Satellites:	8	4440bp	0.06%	18	5168bp	0.07%
Simple repeats:	1380	86969bp	1.20%	1218	78538bp	1.09%
Small RNA:	41	17926bp	0.25%	31	8910bp	0.12%
Total interspersed repeats:		1282432bp	17.76%		2598998bp	35.99%
Unclassified:	2	301bp	0.00%	159	35121bp	0.49%

Table S47. Comparison of the proportion and detailed classification of detection results generated by two tools on dmel_filtered dataset covering the corresponding RepBase library.

RepLong				LongRepMarker		
sequence: 2489				**sequence: 2489**		
total length: 7220516bp				**total length: 7220516bp**		
bases masked: 3369904 bp (46.67%)				**bases masked: 3172784 bp (43.94%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA transposon elements:	115	47278bp	0.65%	273	90505bp	1.25%
–TcMar-Tigger:	0	0bp	0.00%	0	0bp	0.00%
–hAT-Charlie:	0	0bp	0.00%	0	0bp	0.00%
LINEs:	1118	975622bp	13.51%	1532	830630bp	11.50%
–L3/CR1:	72	59325bp	0.82%	327	143147bp	1.98%
–LINE1:	0	0bp	0.00%	3	124bp	0.00%
–LINE2:	0	0bp	0.00%	0	0bp	0.00%
LTR elements:	2567	2232023bp	30.91%	2941	2092428bp	28.98%
–ERVL:	0	0bp	0.00%	0	0bp	0.00%
–ERVL-MaLRs:	0	0bp	0.00%	0	0bp	0.00%
–ERV_classI:	0	0bp	0.00%	0	0bp	0.00%
–ERV_classII:	0	0bp	0.00%	0	0bp	0.00%
Low complexity:	300	16307bp	0.23%	305	16291bp	0.23%
SINEs:	0	0bp	0.00%	1	74bp	0.00%
–ALUs:	0	0bp	0.00%	0	0bp	0.00%
–MIRs:	0	0bp	0.00%	0	0bp	0.00%
Satellites:	13	5512bp	0.08%	66	29926bp	0.41%
Simple repeats:	1152	75079bp	1.04%	1114	72963bp	1.01%
Small RNA:	22	8931bp	0.12%	40	11182bp	0.15%
Total interspersed repeats:		3273168bp	45.33%		3097456bp	42.90%
Unclassified:	67	18245bp	0.25%	321	83819bp	1.16%

Table S48. Comparison of the proportion and detailed classification of detection results generated by two tools on human_100k dataset covering the corresponding RepBase library.

RepLong				LongRepMarker		
sequence: 1512				**sequence: 1512**		
total length: 1647075bp				**total length: 1647075bp**		
bases masked: 344169 bp (20.90%)				**bases masked: 1206368 bp (73.24%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA transposon elements:	0	0bp	0.00%	63	29695bp	1.80%
–TcMar-Tigger:	0	0bp	0.00%	37	17618bp	1.07%
–hAT-Charlie:	0	0bp	0.00%	16	5773bp	0.35%
LINEs:	459	237488bp	14.42%	808	385211bp	23.39%
–L3/CR1:	0	0bp	0.00%	4	1460bp	0.09%
–LINE1:	459	237488bp	14.42%	790	378905bp	23.00%
–LINE2:	0	0bp	0.00%	4	1598bp	0.10%
LTR elements:	6	1779bp	0.11%	548	437452bp	26.56%
–ERVL:	6	1779bp	0.11%	100	66448bp	4.03%
–ERVL-MaLRs:	0	0bp	0.00%	28	8868bp	0.54%
–ERV_classI:	0	0bp	0.00%	355	291096bp	17.67%
–ERV_classII:	0	0bp	0.00%	48	64626bp	3.92%
Low complexity:	67	3106bp	0.19%	18	803bp	0.05%
SINEs:	166	48418bp	2.94%	916	274345bp	16.66%
–ALUs:	166	48418bp	2.94%	916	274345bp	16.66%
–MIRs:	0	0bp	0.00%	0	0bp	0.00%
Satellites:	7	1183bp	0.07%	11	1999bp	0.12%
Simple repeats:	383	39124bp	2.38%	253	33300bp	2.02%
Small RNA:	0	0bp	0.00%	4	268bp	0.02%
Total interspersed repeats:		301494bp	18.30%		1197844bp	72.73%
Unclassified:	53	13809bp	0.84%	212	71141bp	4.32%

Table S49. Comparison of the proportion and detailed classification of detection results generated by two tools on human_polished dataset covering the corresponding RepBase library.

RepLong				LongRepMarker		
sequence: 1512				**sequence: 1512**		
total length: 1647075bp				**total length: 1647075bp**		
bases masked: 1252608 bp (76.05%)				**bases masked: 1583587 bp (96.15%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA transposon elements:	36	16849bp	1.02%	106	42145bp	2.56%
–TcMar-Tigger:	16	8966bp	0.54%	36	7857bp	0.48%
–hAT-Charlie:	1	201bp	0.01%	41	23220bp	1.41%
LINEs:	1626	810377bp	49.20%	1152	647678bp	39.32%
–L3/CR1:	2	511bp	0.03%	19	6677bp	0.41%
–LINE1:	1607	805742bp	48.92%	1073	625152bp	37.96%
–LINE2:	7	1372bp	0.08%	31	5587bp	0.34%
LTR elements:	177	116258bp	7.06%	289	359583bp	21.83%
–ERVL:	78	52446bp	3.18%	68	46108bp	2.80%
–ERVL-MaLRs:	0	0bp	0.00%	52	13378bp	0.81%
–ERV_classI:	95	62670bp	3.80%	134	235553bp	14.30%
–ERV_classII:	4	1142bp	0.07%	23	58157bp	3.53%
Low complexity:	15	710bp	0.04%	1	37bp	0.00%
SINEs:	600	221950bp	13.48%	1117	449218bp	27.27%
–ALUs:	600	221950bp	13.48%	1069	440362bp	26.74%
–MIRs:	0	0bp	0.00%	47	8709bp	0.53%
Satellites:	7	1106bp	0.07%	24	9705bp	0.59%
Simple repeats:	240	33199bp	2.02%	185	30359bp	1.84%
Small RNA:	0	0bp	0.00%	6	1295bp	0.08%
Total interspersed repeats:		1247284bp	75.73%		1618750bp	98.28%
Unclassified:	237	81850bp	4.97%	278	120126bp	7.29%

Table S50. The proportion and detailed classification of detection results generated by two tools based on drosophila_100k dataset covering the repetitive regions on the reference genome of drosophila.

RepLong				LongRepMarker		
sequence: 15				**sequence: 15**		
total length: 168736537bp				**total length: 168736537bp**		
bases masked: 2969428 bp (1.76%)				**bases masked: 20146354 bp (11.94%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	222	17645bp	0.01%	4227	786564bp	0.47%
–Academ-1:	0	0bp	0.00%	14	7962bp	0.00%
–CMC-Chapaev-3:	0	0bp	0.00%	13	22560bp	0.01%
–CMC-EnSpm:	0	0bp	0.00%	56	71262bp	0.04%
–CMC-Transib:	0	0bp	0.00%	25	22921bp	0.01%
–Crypton-V:	0	0bp	0.00%	3	7309bp	0.00%
–Dada:	0	0bp	0.00%	4	7326bp	0.00%
–Kolobok-T2:	0	0bp	0.00%	6	9499bp	0.01%
–MULE-MuDR:	0	0bp	0.00%	20	27111bp	0.02%
–MULE-NOF:	0	0bp	0.00%	636	44060bp	0.03%
–Maverick:	0	0bp	0.00%	12	7064bp	0.00%
–OTHER:	0	0bp	0.00%	15	14919bp	0.01%
–P:	57	7947bp	0.00%	2449	240060bp	0.14%
–PIF-Harbinger:	0	0bp	0.00%	15	19600bp	0.01%
–PiggyBac:	0	0bp	0.00%	10	10674bp	0.01%
–TcMar-Fot1:	0	0bp	0.00%	9	17511bp	0.01%
–TcMar-Pogo:	0	0bp	0.00%	36	37399bp	0.02%
–TcMar-Tc1:	165	9698bp	0.01%	695	105851bp	0.06%
–TcMar-Tc2:	0	0bp	0.00%	3	2318bp	0.00%
–TcMar-m44:	0	0bp	0.00%	5	5647bp	0.00%
–Zisupton:	0	0bp	0.00%	5	9434bp	0.01%
–hAT:	0	0bp	0.00%	16	11486bp	0.01%
–hAT-Ac:	0	0bp	0.00%	36	31151bp	0.02%
–hAT-Charlie:	0	0bp	0.00%	42	48068bp	0.03%
–hAT-hobo:	0	0bp	0.00%	102	30987bp	0.02%
LINE:	5742	781210bp	0.46%	77663	5312380bp	3.15%
–CR1:	0	0bp	0.00%	237	192553bp	0.11%
–I:	284	47810bp	0.03%	155	53095bp	0.03%
–I-Jockey:	135	36300bp	0.02%	9953	1396401bp	0.83%
–Jockey:	2298	341290bp	0.20%	14315	1421176bp	0.84%
–L1:	0	0bp	0.00%	71	73781bp	0.04%
–L1-Tx1:	0	0bp	0.00%	40	36227bp	0.02%
–L2:	0	0bp	0.00%	24	15983bp	0.01%
–LOA:	0	0bp	0.00%	75	35902bp	0.02%
–OTHER:	0	0bp	0.00%	1872	308892bp	0.18%
–Penelope:	0	0bp	0.00%	5	1606bp	0.00%
–R1:	2748	326794bp	0.19%	40636	2525940bp	1.50%
–R1-LOA:	0	0bp	0.00%	206	23130bp	0.01%
–R2:	277	41471bp	0.02%	9351	819532bp	0.49%
–R2-Hero:	0	0bp	0.00%	7	9664bp	0.01%
–R2-NeSL:	0	0bp	0.00%	4	4126bp	0.00%
–RTE-X:	0	0bp	0.00%	4	2344bp	0.00%
–Tad1:	0	0bp	0.00%	708	6107bp	0.00%
LTR:	8897	1777857bp	1.05%	77647	6192042bp	3.67%
–Caulimovirus:	0	0bp	0.00%	11	7741bp	0.00%
–Copia:	1112	139787bp	0.08%	10402	799842bp	0.47%
–DIRS:	0	0bp	0.00%	13	14109bp	0.01%
–ERV:	0	0bp	0.00%	21	25055bp	0.01%
–ERV1:	0	0bp	0.00%	32	56228bp	0.03%
–ERV4:	0	0bp	0.00%	16	25400bp	0.02%
–ERVK:	0	0bp	0.00%	3	2490bp	0.00%
–ERVL:	0	0bp	0.00%	9	6159bp	0.00%
–ERVL-MaLR:	0	0bp	0.00%	4	7504bp	0.00%
–Gypsy:	4405	1128076bp	0.67%	18745	3294613bp	1.95%
–Ngaro:	0	0bp	0.00%	9	10875bp	0.01%
–OTHER:	0	0bp	0.00%	187	106548bp	0.06%
–Pao:	3380	509994bp	0.30%	48195	2144191bp	1.27%
Other:	0	0bp	0.00%	6878	444394bp	0.26%
–OTHER:	0	0bp	0.00%	6878	444394bp	0.26%
RC:	0	0bp	0.00%	1053	512929bp	0.30%
–Helitron:	0	0bp	0.00%	1053	512929bp	0.30%
SINE:	0	0bp	0.00%	24	37399bp	0.02%
–Alu:	0	0bp	0.00%	19	14653bp	0.01%
–U:	0	0bp	0.00%	5	22746bp	0.01%
SINE?:	0	0bp	0.00%	10	11564bp	0.01%
–OTHER:	0	0bp	0.00%	10	11564bp	0.01%
Satellite:	1352	120893bp	0.07%	82645	3889289bp	2.30%
–OTHER:	1352	120893bp	0.07%	82645	3889289bp	2.30%
Simple:	0	0bp	0.00%	135931	483006bp	0.29%
–repeat:	0	0bp	0.00%	135931	483006bp	0.29%
Unknown:	148	60256bp	0.04%	35009	4510469bp	2.67%
–OTHER:	148	60256bp	0.04%	35009	4510469bp	2.67%
rRNA:	2166	277854bp	0.16%	38614	1562614bp	0.93%
–OTHER:	2166	277854bp	0.16%	38614	1562614bp	0.93%
tRNA:	0	0bp	0.00%	65	79615bp	0.05%
–OTHER:	0	0bp	0.00%	65	79615bp	0.05%

Table S51. The proportion and detailed classification of detection results generated by two tools based on human_100k dataset covering the repetitive regions on the reference genome of human(hg38).

RepLong				LongRepMarker		
sequence: 455				**sequence: 455**		
total length: 3209286105bp				**total length: 3209286105bp**		
bases masked: 567067 bp (0.02%)				**bases masked: 31170736 bp (0.97%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	0	0bp	0.00%	396	551207bp	0.02%
–Academ-1:	0	0bp	0.00%	4	4226bp	0.00%
–CMC-EnSpm:	0	0bp	0.00%	3	1074bp	0.00%
–MuLE-MuDR:	0	0bp	0.00%	1	3138bp	0.00%
–PIF-Harbinger:	0	0bp	0.00%	2	5611bp	0.00%
–PiggyBac:	0	0bp	0.00%	45	4360bp	0.00%
–TcMar-Mariner:	0	0bp	0.00%	36	40639bp	0.00%
–TcMar-Tigger:	0	0bp	0.00%	214	361402bp	0.01%
–hAT-Ac:	0	0bp	0.00%	2	7020bp	0.00%
–hAT-Blackjack:	0	0bp	0.00%	1	812bp	0.00%
–hAT-Charlie:	0	0bp	0.00%	88	122925bp	0.00%
LINE:	36	381497bp	0.01%	4472	14812233bp	0.46%
–CR1:	0	0bp	0.00%	4	13257bp	0.00%
–L1:	36	381497bp	0.01%	4447	14671368bp	0.46%
–L2:	0	0bp	0.00%	11	80124bp	0.00%
–OTHER:	0	0bp	0.00%	10	55004bp	0.00%
LTR:	1	2525bp	0.00%	1628	2403527bp	0.07%
–Copia:	0	0bp	0.00%	7	1539bp	0.00%
–DIRS:	0	0bp	0.00%	1	650bp	0.00%
–ERV:	0	0bp	0.00%	4	20864bp	0.00%
–ERV1:	0	0bp	0.00%	706	860621bp	0.03%
–ERVK:	0	0bp	0.00%	82	218467bp	0.01%
–ERVL:	1	2525bp	0.00%	537	1093366bp	0.03%
–ERVL-MaLR:	0	0bp	0.00%	284	206370bp	0.01%
–Gypsy:	0	0bp	0.00%	7	3073bp	0.00%
RNA:	0	0bp	0.00%	4	564bp	0.00%
–OTHER:	0	0bp	0.00%	4	564bp	0.00%
Retroposon:	1	2849bp	0.00%	16	45324bp	0.00%
–SVA:	1	2849bp	0.00%	16	45324bp	0.00%
SINE:	8	89458bp	0.00%	3999	10785656bp	0.34%
–Alu:	8	89458bp	0.00%	3995	10781150bp	0.34%
–MIR:	0	0bp	0.00%	4	4506bp	0.00%
Satellite:	8	56176bp	0.00%	478	1567189bp	0.05%
–OTHER:	2	6431bp	0.00%	256	133037bp	0.00%
–Y-chromosome:	6	49745bp	0.00%	205	1414782bp	0.04%
–centromeric:	0	0bp	0.00%	17	24306bp	0.00%
Simple:	0	0bp	0.00%	212	255167bp	0.01%
–repeat:	0	0bp	0.00%	212	255167bp	0.01%
Unknown:	3	34562bp	0.00%	549	2334726bp	0.07%
–OTHER:	3	34562bp	0.00%	549	2334726bp	0.07%
tRNA:	0	0bp	0.00%	3	13351bp	0.00%
–OTHER:	0	0bp	0.00%	3	13351bp	0.00%

Table S52. The proportion and detailed classification of detection results generated by two tools based on dmel_filtered dataset covering the repetitive regions on the reference genome of drosophila.

	RepLong			LongRepMarker		
sequence: 15				**sequence: 15**		
total length: 168736537bp				**total length: 168736537bp**		
bases masked: 11808450 bp (7.00%)				**bases masked: 45193046 bp (26.78%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	417	223449bp	0.13%	11162	3357943bp	1.99%
–Academ-1:	0	0bp	0.00%	18	32652bp	0.02%
–Academ-H:	0	0bp	0.00%	7	4414bp	0.00%
–CMC-Chapaev:	0	0bp	0.00%	2	5540bp	0.00%
–CMC-EnSpm:	0	0bp	0.00%	313	239334bp	0.14%
–CMC-Transib:	8	5191bp	0.00%	89	132453bp	0.08%
–Crypton-F:	0	0bp	0.00%	12	21341bp	0.01%
–Crypton-H:	0	0bp	0.00%	10	11836bp	0.01%
–Crypton-S:	0	0bp	0.00%	16	9749bp	0.01%
–Crypton-V:	0	0bp	0.00%	1	3852bp	0.00%
–Dada:	0	0bp	0.00%	22	36871bp	0.02%
–Kolobok:	0	0bp	0.00%	4	964bp	0.00%
–Kolobok-T2:	0	0bp	0.00%	61	59757bp	0.04%
–MULE-MuDR:	0	0bp	0.00%	113	172365bp	0.10%
–MULE-NOF:	63	39498bp	0.02%	170	108095bp	0.06%
–Maverick:	0	0bp	0.00%	56	84675bp	0.05%
–MuLE-NOF:	0	0bp	0.00%	4	7662bp	0.00%
–OTHER:	0	0bp	0.00%	5127	169184bp	0.10%
–P:	203	119252bp	0.07%	2448	576561bp	0.34%
–PIF-Harbinger:	0	0bp	0.00%	102	126502bp	0.07%
–PIF-ISL2EU:	0	0bp	0.00%	10	25136bp	0.01%
–PIF-Spy:	0	0bp	0.00%	18	31163bp	0.02%
–PiggyBac:	0	0bp	0.00%	24	36090bp	0.02%
–PiggyBac-X:	0	0bp	0.00%	11	10672bp	0.01%
–Sola-1:	0	0bp	0.00%	2	1035bp	0.00%
–Sola-2:	0	0bp	0.00%	1	12632bp	0.01%
–TcMar-Ant1:	0	0bp	0.00%	5	20700bp	0.01%
–TcMar-Cweed:	0	0bp	0.00%	21	24396bp	0.01%
–TcMar-Fot1:	0	0bp	0.00%	27	30426bp	0.02%
–TcMar-ISRm11:	0	0bp	0.00%	36	34197bp	0.02%
–TcMar-Mariner:	0	0bp	0.00%	62	86226bp	0.05%
–TcMar-Pogo:	0	0bp	0.00%	844	72555bp	0.04%
–TcMar-Stowaway:	0	0bp	0.00%	1	1959bp	0.00%
–TcMar-Tc1:	55	28985bp	0.02%	1004	471642bp	0.28%
–TcMar-Tc2:	0	0bp	0.00%	3	41303bp	0.02%
–TcMar-Tc4:	0	0bp	0.00%	8	13442bp	0.01%
–Zator:	0	0bp	0.00%	2	11353bp	0.01%
–Zisupton:	0	0bp	0.00%	33	32305bp	0.02%
–hAT:	0	0bp	0.00%	33	28156bp	0.02%
–hAT-Ac:	0	0bp	0.00%	130	225640bp	0.13%
–hAT-Blackjack:	0	0bp	0.00%	6	17312bp	0.01%
–hAT-Charlie:	0	0bp	0.00%	30	41926bp	0.02%
–hAT-Pegasus:	0	0bp	0.00%	9	7331bp	0.00%
–hAT-Tag1:	0	0bp	0.00%	9	10894bp	0.01%
–hAT-Tip100:	0	0bp	0.00%	70	101127bp	0.06%
–hAT-hAT19:	0	0bp	0.00%	9	16365bp	0.01%
–hAT-hAT5:	0	0bp	0.00%	7	11084bp	0.01%
–hAT-hATw:	0	0bp	0.00%	11	14610bp	0.01%
–hAT-hATx:	0	0bp	0.00%	15	12867bp	0.01%
–hAT-hobo:	88	30523bp	0.02%	146	140805bp	0.08%
LINE:	16977	4189163bp	2.48%	103919	9287397bp	5.50%
–CR1:	30	41952bp	0.02%	595	648428bp	0.38%
–CR1-Zenon:	0	0bp	0.00%	17	23964bp	0.01%
–CRE:	0	0bp	0.00%	4	8204bp	0.00%
–I:	371	126854bp	0.08%	602	288452bp	0.17%
–I-Jockey:	1633	986651bp	0.58%	12323	3000052bp	1.78%
–Jockey:	2991	1411229bp	0.84%	17379	2157429bp	1.28%
–L1:	0	0bp	0.00%	284	341060bp	0.20%
–L1-Tx1:	0	0bp	0.00%	98	141618bp	0.08%
–L2:	0	0bp	0.00%	80	92941bp	0.06%
–LOA:	8	35475bp	0.02%	194	37588bp	0.02%
–OTHER:	653	512921bp	0.30%	3980	578051bp	0.34%
–Penelope:	0	0bp	0.00%	57	61763bp	0.04%
–Proto2:	0	0bp	0.00%	14	28853bp	0.02%
–R1:	6576	1988893bp	1.18%	55979	3055649bp	1.81%
–R1-LOA:	365	122428bp	0.07%	30	29116bp	0.02%
–R2:	4350	808229bp	0.48%	12198	950896bp	0.56%
–R2-NeSL:	0	0bp	0.00%	7	19021bp	0.01%
–RTE-BovB:	0	0bp	0.00%	31	46286bp	0.03%
–RTE-RTE:	0	0bp	0.00%	9	15719bp	0.01%
–Tad1:	0	0bp	0.00%	38	54789bp	0.03%
LTR:	16863	7162246bp	4.24%	105682	13827910bp	8.19%
–Caulimovirus:	0	0bp	0.00%	18	61114bp	0.04%
–Copia:	1690	515479bp	0.31%	10466	1755110bp	1.04%
–DIRS:	0	0bp	0.00%	43	78207bp	0.05%
–ERV-Foamy:	0	0bp	0.00%	16	24473bp	0.01%
–ERV1:	0	0bp	0.00%	102	204122bp	0.12%
–ERV4:	0	0bp	0.00%	24	36215bp	0.02%
–ERVK:	0	0bp	0.00%	70	88760bp	0.05%
–ERVL:	0	0bp	0.00%	10	13218bp	0.01%
–Gypsy:	11817	5056654bp	3.00%	42655	9056526bp	5.37%
–Ngaro:	0	0bp	0.00%	15	20648bp	0.01%
–OTHER:	178	401090bp	0.24%	1942	446355bp	0.26%
–Pao:	3178	1731467bp	1.03%	50321	2948938bp	1.75%
Other:	1529	9445bp	0.01%	5245	381735bp	0.23%
–OTHER:	1529	9445bp	0.01%	5245	381735bp	0.23%
RC:	6	13723bp	0.01%	2413	2506443bp	1.49%
–Helitron:	6	13723bp	0.01%	2413	2506443bp	1.49%
Retroposon:	0	0bp	0.00%	36	42026bp	0.02%
–L1:	0	0bp	0.00%	13	10051bp	0.01%
–SVA:	0	0bp	0.00%	23	31975bp	0.02%
SINE:	0	0bp	0.00%	11	9265bp	0.01%
–ID:	0	0bp	0.00%	11	9265bp	0.01%
SINE?:	0	0bp	0.00%	31	41984bp	0.02%
–OTHER:	0	0bp	0.00%	31	41984bp	0.02%
Satellite:	1787	538934bp	0.32%	133184	4745508bp	2.81%
–OTHER:	1787	538934bp	0.32%	133184	4745508bp	2.81%
Simple:	14942	85425bp	0.05%	263786	562282bp	0.33%
–repeat:	14942	85425bp	0.05%	263786	562282bp	0.33%
Unknown:	5000	640969bp	0.38%	89301	15112684bp	8.96%
–OTHER:	5000	640969bp	0.38%	89301	15112684bp	8.96%
rRNA:	25627	1658931bp	0.98%	49815	1857060bp	1.10%
–OTHER:	25627	1658931bp	0.98%	49815	1857060bp	1.10%
snRNA:	0	0bp	0.00%	14	13618bp	0.01%
–OTHER:	0	0bp	0.00%	14	13618bp	0.01%
tRNA:	0	0bp	0.00%	251	281363bp	0.17%
–OTHER:	0	0bp	0.00%	251	281363bp	0.17%

Table S53. The proportion and detailed classification of detection results generated by two tools based on human_polished dataset covering the repetitive regions on the reference genome of human(hg38).

RepLong				LongRepMarker		
sequence: 455				**sequence: 455**		
total length: 3209286105bp				**total length: 3209286105bp**		
bases masked: 51105613 bp (1.59%)				**bases masked: 698754064 bp (21.77%)**		
Repeat Types	**Num of elements**	**Length occupied**	**Percentage of sequence**	**Num of elements**	**Length occupied**	**Percentage of sequence**
DNA:	33	294043bp	0.01%	13963	14730125bp	0.46%
–Academ-1:	0	0bp	0.00%	255	388681bp	0.01%
–Academ-H:	0	0bp	0.00%	2	5572bp	0.00%
–CMC-EnSpm:	0	0bp	0.00%	273	82737bp	0.00%
–CMC-Transib:	0	0bp	0.00%	18	2358bp	0.00%
–Crypton:	0	0bp	0.00%	12	21456bp	0.00%
–Crypton-A:	0	0bp	0.00%	2	622bp	0.00%
–Crypton-I:	0	0bp	0.00%	2	2557bp	0.00%
–Crypton-S:	0	0bp	0.00%	2	767bp	0.00%
–Crypton-V:	0	0bp	0.00%	4	3831bp	0.00%
–Dada:	0	0bp	0.00%	2	1529bp	0.00%
–Ginger:	0	0bp	0.00%	28	7771bp	0.00%
–Kolobok-T2:	0	0bp	0.00%	2	3182bp	0.00%
–MULE-MuDR:	0	0bp	0.00%	29	29452bp	0.00%
–Maverick:	0	0bp	0.00%	12	20019bp	0.00%
–Merlin:	0	0bp	0.00%	12	8369bp	0.00%
–MuLE-MuDR:	1	11882bp	0.00%	49	77343bp	0.00%
–Novosib:	0	0bp	0.00%	9	456bp	0.00%
–OTHER:	0	0bp	0.00%	205	253903bp	0.01%
–P:	0	0bp	0.00%	2	1097bp	0.00%
–PIF-Harbinger:	0	0bp	0.00%	14	21798bp	0.00%
–PiggyBac:	0	0bp	0.00%	55	50936bp	0.00%
–Sola-1:	0	0bp	0.00%	2	2557bp	0.00%
–Sola-2:	0	0bp	0.00%	4	3465bp	0.00%
–Sola-3:	0	0bp	0.00%	6	3564bp	0.00%
–TcMar:	0	0bp	0.00%	6	3705bp	0.00%
–TcMar-Fot1:	0	0bp	0.00%	6	3479bp	0.00%
–TcMar-ISRm11:	0	0bp	0.00%	2	583bp	0.00%
–TcMar-Mariner:	0	0bp	0.00%	363	676972bp	0.02%
–TcMar-Sagan:	0	0bp	0.00%	2	526bp	0.00%
–TcMar-Tc1:	0	0bp	0.00%	22	42807bp	0.00%
–TcMar-Tc2:	0	0bp	0.00%	96	138622bp	0.00%
–TcMar-Tigger:	17	164448bp	0.01%	3987	7057574bp	0.22%
–Zator:	0	0bp	0.00%	2	3176bp	0.00%
–Zisupton:	0	0bp	0.00%	3925	146143bp	0.00%
–hAT:	0	0bp	0.00%	44	21983bp	0.00%
–hAT-Ac:	2	22601bp	0.00%	48	99506bp	0.00%
–hAT-Blackjack:	1	7554bp	0.00%	155	170949bp	0.01%
–hAT-Charlie:	10	68709bp	0.00%	3557	4535663bp	0.14%
–hAT-Pegasus:	0	0bp	0.00%	2	2020bp	0.00%
–hAT-Tag1:	0	0bp	0.00%	30	25258bp	0.00%
–hAT-Tip100:	2	18849bp	0.00%	709	838475bp	0.03%
–hAT-hAT1:	0	0bp	0.00%	2	1583bp	0.00%
–hAT-hATm:	0	0bp	0.00%	4	2279bp	0.00%
DNA?:	0	0bp	0.00%	10	8332bp	0.00%
–PiggyBac:	0	0bp	0.00%	2	1065bp	0.00%
–hAT:	0	0bp	0.00%	6	3417bp	0.00%
–hAT-Tip100:	0	0bp	0.00%	2	3850bp	0.00%
LINE:	4030	32158646bp	1.00%	129957	306015073bp	9.54%
–CR1:	1	9166bp	0.00%	477	818172bp	0.03%
–CRE:	0	0bp	0.00%	8	12564bp	0.00%
–Dong-R4:	0	0bp	0.00%	14	55288bp	0.00%
–I:	0	0bp	0.00%	8	3595bp	0.00%
–I-Jockey:	0	0bp	0.00%	24	19361bp	0.00%
–Jockey:	0	0bp	0.00%	2	9723bp	0.00%
–L1:	3990	31825325bp	0.99%	125281	298299458bp	9.29%
–L1-Tx1:	0	0bp	0.00%	24	17164bp	0.00%
–L2:	20	158438bp	0.00%	3466	5158961bp	0.16%
–OTHER:	19	212718bp	0.01%	408	1895218bp	0.06%
–Penelope:	0	0bp	0.00%	6	4547bp	0.00%
–R1:	0	0bp	0.00%	5	1795bp	0.00%
–R1-LOA:	0	0bp	0.00%	2	586bp	0.00%
–R2:	0	0bp	0.00%	4	10887bp	0.00%
–R2-NeSL:	0	0bp	0.00%	2	6307bp	0.00%
–RTE-BovB:	0	0bp	0.00%	46	65023bp	0.00%
–RTE-RTE:	0	0bp	0.00%	2	583bp	0.00%
–RTE-X:	0	0bp	0.00%	164	183832bp	0.01%
–Rex-Babar:	0	0bp	0.00%	4	8888bp	0.00%
–Tad1:	0	0bp	0.00%	10	6929bp	0.00%
LTR:	322	1819104bp	0.06%	22677	34495883bp	1.07%
–Caulimovirus:	0	0bp	0.00%	6	9350bp	0.00%
–Copia:	0	0bp	0.00%	143	121612bp	0.00%
–DIRS:	0	0bp	0.00%	12	7553bp	0.00%
–ERV:	0	0bp	0.00%	68	268741bp	0.01%
–ERV1:	169	663109bp	0.02%	6629	10678063bp	0.33%
–ERVK:	5	11034bp	0.00%	1114	1996168bp	0.06%
–ERVL:	144	1125666bp	0.04%	6511	12322032bp	0.38%
–ERVL-MaLR:	3	15671bp	0.00%	7388	8684526bp	0.27%
–Gypsy:	0	0bp	0.00%	587	440325bp	0.01%
–Ngaro:	0	0bp	0.00%	14	16541bp	0.00%
–OTHER:	1	3624bp	0.00%	183	125577bp	0.00%
–Pao:	0	0bp	0.00%	22	14958bp	0.00%
RC:	0	0bp	0.00%	45	30615bp	0.00%
–Helitron:	0	0bp	0.00%	45	30615bp	0.00%
RNA:	0	0bp	0.00%	20	19129bp	0.00%
–OTHER:	0	0bp	0.00%	20	19129bp	0.00%
Retroposon:	49	250625bp	0.01%	1454	1762391bp	0.05%
–SVA:	49	250625bp	0.01%	1454	1762391bp	0.05%
SINE:	1239	9701289bp	0.30%	122000	263525569bp	8.21%
–5S:	0	0bp	0.00%	6	2915bp	0.00%
–5S-Deu-L2:	0	0bp	0.00%	14	13637bp	0.00%
–Alu:	1239	9701289bp	0.30%	115426	255871934bp	7.97%
–L2:	0	0bp	0.00%	8	3683bp	0.00%
–MIR:	0	0bp	0.00%	6518	7846668bp	0.24%
–U:	0	0bp	0.00%	2	860bp	0.00%
–tRNA:	0	0bp	0.00%	4	2594bp	0.00%
–tRNA-7SL:	0	0bp	0.00%	2	508bp	0.00%
–tRNA-Core-L2:	0	0bp	0.00%	4	1422bp	0.00%
–tRNA-Deu:	0	0bp	0.00%	6	1953bp	0.00%
–tRNA-RTE:	0	0bp	0.00%	10	4467bp	0.00%
Satellite:	1705	4475526bp	0.14%	260908	58245890bp	1.81%
–OTHER:	319	209323bp	0.01%	34753	1735367bp	0.05%
–Y-chromosome:	1381	4258059bp	0.13%	222817	55872645bp	1.74%
–acromeric:	0	0bp	0.00%	193	184934bp	0.01%
–centromeric:	5	8144bp	0.00%	3098	1698524bp	0.05%
–telomeric:	0	0bp	0.00%	47	47932bp	0.00%
Simple:	858	678220bp	0.02%	73682	1795242bp	0.06%
–repeat:	858	678220bp	0.02%	73682	1795242bp	0.06%
Unknown:	424	3443330bp	0.11%	54472	66289814bp	2.07%
–OTHER:	424	3443330bp	0.11%	54472	66289814bp	2.07%
rRNA:	0	0bp	0.00%	155	179538bp	0.01%
–OTHER:	0	0bp	0.00%	155	179538bp	0.01%
scRNA:	0	0bp	0.00%	8	6150bp	0.00%
–OTHER:	0	0bp	0.00%	8	6150bp	0.00%
snRNA:	0	0bp	0.00%	25	19397bp	0.00%
–OTHER:	0	0bp	0.00%	25	19397bp	0.00%
tRNA:	0	0bp	0.00%	15	8361bp	0.00%
–OTHER:	0	0bp	0.00%	15	8361bp	0.00%

3.6 Running time and peak memory consumption statistics

LongRepMarker has problems with long running time and large memory requirements on all NGS short read datasets. To confirm this problem, we tested the running time and peak memory consumptions of four tools (LongRepMarker, RepARK, REPdenovo and RepLong) on seven datasets (Human-chr14, Leafcutter ant, D.melano, Mouse, HG004_NA24143_father, Dro_100k and Human_100k), just as shown in Fig S31. Through the analysis of the running time and peak memory consumptions of each step in the LongRepMarker processing flow, we found that the sequence assembly consumes the most running time and memory. Therefore, as long as the running time and memory consumptions of this step can be controlled by adjusting the parameters W and t, the tool can still run normally under the condition of limited resources. SPAdes uses 512 Mb per thread for buffers, which results in higher memory consumption (The default value of parameters W and t in SPAdes are set to 250GB and 16 respectively). If you set memory limit manually, SPAdes will use smaller buffers and thus less RAM. The parameter W set memory limit in Gb. SPAdes terminates if it reaches this limit. Actual amount of consumed RAM will be below this limit. Make sure this value is correct for the given machine. SPAdes uses the limit value to automatically determine the sizes of various buffers, etc. The parameter t is used to set the number of threads using in SPAdes assembly, and the default value of it is 16. The larger the number of threads is, the faster the SPAdes assembly speed, and the memory consumption will also increase.

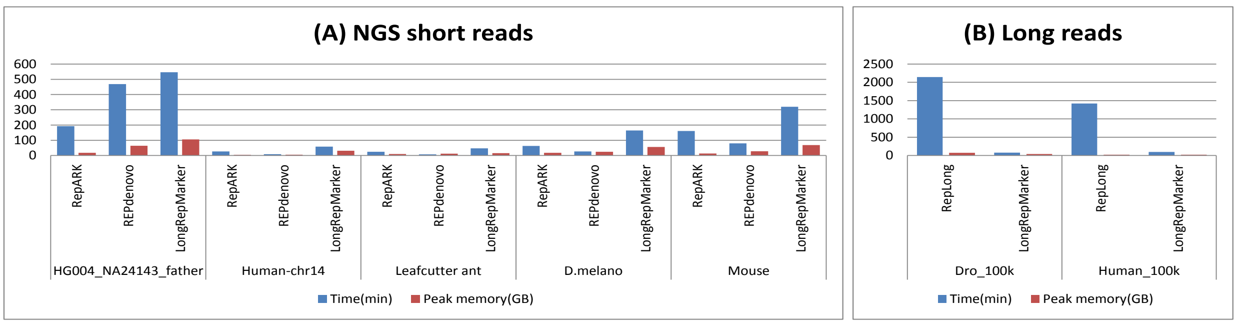


Fig. S31. Comparison of the running time and peak memory consumption of four tools on seven datasets.

3.7 Performance evaluation of structural variation detection in *de novo* mode

LongRepMarker is designed based on technologies of the *de novo* sequence assembly and multiple sequence alignment to identify repetitive regions in a genome (Fig. S32). From the perspective of implementation principles, it can identify the genomic structural variations contained in the repetitive sequences. The genomic variation regions between repeating segments generate a chimeras, which can negatively affect the alignment of entire segment. Chimeras consist of two or more repetitive regions and some genomic variation regions, which cannot be aligned to overlap sequences many times. However, the genomic variation regions between repeating segments are also the important component of repeating regions, and they are an important manifestation of repetitive regions polymorphism. In addition, the study and analysis of genomic structural variations that occur within the repetitive regions can provide a new perspective for understanding life processes and analyzing life mechanisms. Based on the above reasons, we have completely preserved the genomic variations that occur inside the repetitive regions along with the repetitive fragments.

In this step, LongRepMarker mainly analyzes the four kinds of mutations that occur in the repetitive regions, which are insertion, deletion, inversion and translocation. Among them, insertion refers to an extra sequence segment in a repetitive region that does not belong to this region; Deletion refers to a missing sequence segment in the repetitive region that should belong to this region; Inversion refers to a pair of repeats in which there is a sub segment with opposite alignment direction. If the positions of these two sub fragments are swapped, the pair of repeats match perfectly; Translocation refers to a pair of repeats in which there is a sub segment from each other at different locations. If the positions of these two fragments are swapped, the pair of repeats match perfectly. The alignment results generated from section 2.5 are the basic for the analysis in this section. The detection methods of four kinds of mutations are introduced in detail in the following sections.

In the insertion and deletion discovering, if there is a pair of sequence fragments A and B, all the subsequences of fragment B belong to the fragment A, but there are some subsequences in fragment A that dose not belong to fragment B, just as shown in Fig.S19(D). Those subsequences distributed in the middle of fragment A can be identified as insertions in fragment A or deletions in fragment B. To further determine whether these subsequences are insertions or deletions, LongRepMarker needs the support of the reference genome. By mapping the detected repetitive fragments to the reference genome, the matching regions and the differences between them can be obtained. For example, if there is a segment in fragment A that dose not exist in the reference genome, then this segment should be an insertion variation of the sequenced sample. However, whether the insertion variation belongs to the category of structural variation further depends on its size. If its size is greater than 50bp, it is an insertion in the concept of structural variation. If its size is less than 50bp, it can only be regarded as a micro insertion variation. Similarly, if there is a segment in reference genome that dose not exist in fragment B, then this segment should be a deletion variation of the sequenced sample. Among the alignment results produced by the alignment tool, there is a column of data called cigar, which details the alignment between the fragments and the reference sequences. For example, when the cigar value of a record is '87M109I547D', it means that the sub-segment in the region [0,86] on the segment is completely matched to the reference genome, the sub-segments in the region [87,195] can not be found on the reference genome, and the following region marked as [196,742] does not present on the segment, but it can be found on the reference genome. Therefore, by analyzing the cigar string of the alignment results, we can get the mutations that occurred in the detection fragments. For example, we also perform a simple analysis of the structural variations contained in the detected repetitive sequences. For example, a structural variation (deletion) detected by LongRepMarker in the dataset of Drosophila melanogaster is shown in Fig. S33. This figure is obtained by mapping the detection results of LongRepMarker to the reference genome with the evaluation tool minimap2. In Fig. S33, symbol 'gi' represents a chromosome in the Drosophila

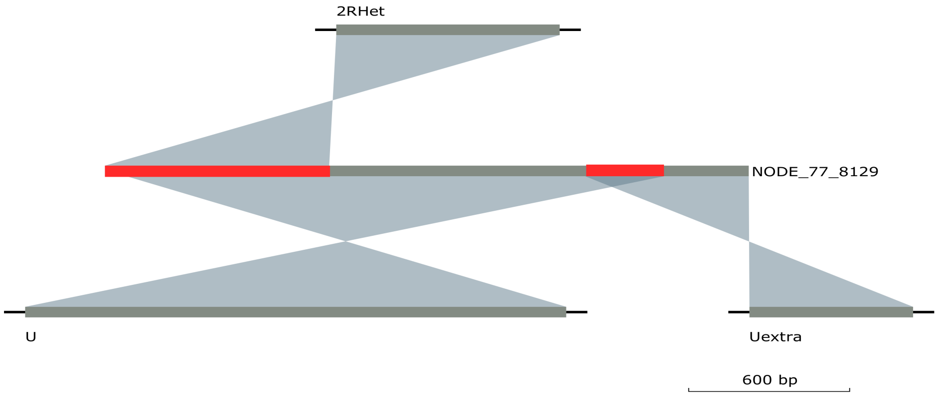


Fig. S32. An example of the structure of repetitive regions between multiple chromosomes found on Drosophila melanogaster dataset. $'NODE\_77\_8129'$ indicates the repetitive fragment which detected by LongRepMarker. $'2RHet'$, $'u'$ and $'Uextra'$ represent three different chromosomes in the Drosophila melanogaster genome, respectively. Twisted shading between fragments indicates reverse alignment.

melanogaster genome. The shaded areas indicate the alignment, and the gaps indicate the deletion that occurs in the detected fragment. Fig. S33 shows that LongRepMarker can accurately identify the structural variations (such as insertion, deletion, inversion and translocations) contained in the repeat regions. The detailed INDEL (Insertion and deletion) variants found in LongRepMarker's detection results on seven sequencing samples (Ant, human-chr14, drosophila, mouse, HG003_NA24149_father, HG002_NA24385_Son and HG004_NA24143_mother) are summarized in Tables S54-S60.

In inversion discovering, if in a pair of fragments A and B there are some sub segments with the opposite alignment direction, and if the positions of these sub fragments are swapped, the pair of fragments A and B match perfectly. Those sub segments with the opposite alignment direction may be inversions in this pair of repeats. During the detection process, LongRepMarker needs to determine the position, length and the alignment direction of subsegments in fragments A , and their corresponding alignment segments on the reference genome. By analyzing the position, length and alignment direction of sub-fragments, the inversion in the fragment pairs between query sequences and the reference genome can be accurately determined. The alignment results generated from section 2.5 are the basic for the inversion discovering. In translocation discovering, consider that a pair of fragments A and B have some subsegments C and D belong respectively. For example, segment C should have belonged to B, but now it is embedded in A, segment D should have belonged to A, but now it is embedded in B. If the positions of these two fragments C and D are swapped, the pair of repeats match perfectly, these sub segments may be translocation in this pair of repeats. The alignment results generated from section 2.5 are also the basic for the translocation discovering. The INDEL (insertion and deletion) mutations within repetitive sequences are usually easy to be found, but it is very difficult for LongRepMarker to find and determine the inversion and translocation mutations that occur within the repetitive sequences. In order to ensure the accuracy and reliability of detection, LongRepMarker needs the assistance of two tools, ngmlr (https://github.com/philres/ngmlr) and Sniffles (https://github.com/fritzsedlazeck/Sniffles), when detecting inversion and inversion mutations that occur within the repetitive sequences. Since detections of inversion and inversion mutations in LongRepMarker are done with the support of other tools, so the results of these two detections are not shown here.

The structural variation detection report of LongRepMarker contains the information of SNPs, INDELs, inversions and translocations. Finally, LongRepMarker generates a VCF format structural variation statistical report in the detection results, as shown in Figure S11. VCF (Variant Calling Format) is a tab-delimited text file that is used to describe single nucleotide variants (SNVs) as well as insertions, deletions, and other sequence variations.

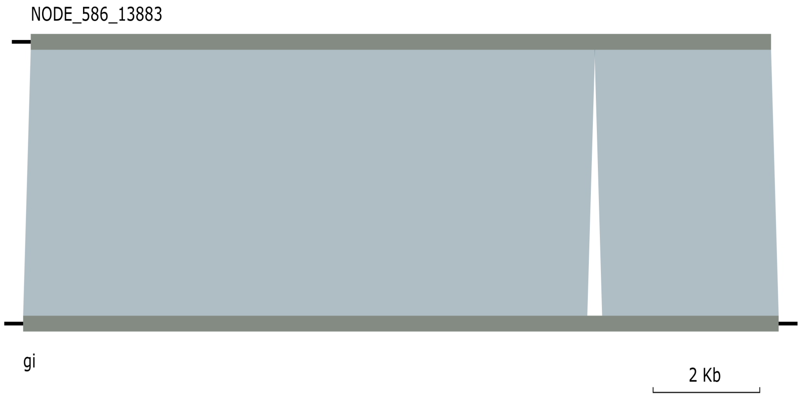


Fig. S33. An example of the structural variantion (deletion) found by LongRepMarker on human-chr14 dataset. $'NODE\_586\_13883'$ indicates the repetitive fragment which detected by LongRepMarker. $'gi'$ represents a chromosome in the reference genome of human-chr14. The shaded areas indicate the alignment, and the gaps indicate the deletion that occurred in the detected fragment.

Table S54. Partial INDEL variation statistics of detection results generated by the *de novo* mode of LongRepMarker on the drosophila dataset.

Repeat fragment id	Location on Fragment	Reference id	Location on Ref.	Variation/Length.
NODE_80_7585	1526	2R	8678507	Deletion/462bp
NODE_80_7585	1526	3L	24007213	Deletion/449bp
NODE_847_2791	1172	2L	22773069	Deletion/381bp
NODE_1628_1445	584	3R	17435374	Deletion/1003bp
NODE_1628_1445	580	X	21925103	Deletion/1189bp
NODE_1628_1445	871	3R	17435652	Deletion/1151bp
NODE_282_5678	3834	2RHet	1127969	Insertion/460bp
NODE_1020_2324	1849	U	2915525	Deletion/402bp
NODE_2508_1056	575	2L	8015354	Deletion/417bp
NODE_2508_1056	473	2L	11566727	Deletion/417bp
NODE_2508_1056	473	X	14715888	Deletion/417bp
NODE_2508_1056	473	3R	15938353	Deletion/417bp
NODE_2508_1056	575	2L	20160701	Deletion/417bp
NODE_1706_1380	1048	3R	17435828	Deletion/832bp
NODE_12850_624	315	3R	19665344	Deletion/385bp
NODE_12850_624	315	2R	21033391	Deletion/379bp
NODE_12850_624	312	X	10165933	Deletion/385bp
NODE_12850_624	312	X	16120612	Deletion/385bp
NODE_3268_912	539	3RHet	808287	Deletion/391bp
NODE_3268_912	539	3RHet	1058889	Deletion/381bp
NODE_3268_912	371	3LHet	643037	Deletion/378bp
NODE_3268_912	371	2L	22574496	Deletion/379bp
NODE_1188_1997	1226	X	18852406	Deletion/421bp
NODE_3383_894	435	Uextra	677898	Deletion/406bp
NODE_3383_894	435	3RHet	2013216	Deletion/420bp
NODE_3383_894	435	U	2054869	Deletion/420bp
NODE_3383_894	505	2RHet	323045	Deletion/420bp
NODE_143_7329	1926	3LHet	1240631	Insertion/426bp
NODE_725_3289	2288	3LHet	341984	Deletion/433bp
NODE_1211_1957	1591	3RHet	1796974	Deletion/371bp
NODE_1437_1661	977	3L	6406260	Deletion/404bp
NODE_1437_1661	977	3L	6423567	Deletion/404bp
NODE_1437_1661	683	3R	18066515	Deletion/404bp
NODE_1437_1661	977	3R	25715242	Deletion/404bp
NODE_2419_1076	585	3R	833296	Deletion/370bp
NODE_2419_1076	585	U	1305472	Deletion/370bp
NODE_2419_1076	503	2R	5615377	Deletion/370bp
NODE_2419_1076	585	3L	10059225	Deletion/370bp
NODE_2419_1076	585	3L	11262940	Deletion/370bp
NODE_1197_1980	1161	3R	18277464	Deletion/770bp
NODE_580_4119	3012	3LHet	1171134	Deletion/464bp
NODE_580_4119	3012	2R	16673984	Deletion/464bp
NODE_580_4119	3012	2L	17276389	Deletion/464bp
NODE_580_4119	1108	2R	219417	Deletion/464bp
NODE_580_4119	1108	3LHet	1012372	Deletion/464bp
NODE_1078_2215	570	3LHet	1583478	Deletion/435bp
NODE_3425_889	651	2L	12784671	Deletion/388bp
NODE_3425_889	656	X	6068489	Deletion/388bp
NODE_3425_889	250	2L	15724376	Deletion/388bp
NODE_3425_889	250	3L	22809548	Deletion/388bp
NODE_1480_1612	769	X	2718952	Deletion/411bp
NODE_158_7230	6754	2R	1796459	Deletion/404bp
NODE_158_7230	6745	X	3654240	Deletion/407bp
NODE_158_7230	6769	3L	10060023	Deletion/427bp
NODE_158_7230	631	2R	5614720	Deletion/428bp
NODE_158_7230	627	3L	24211428	Deletion/417bp
NODE_5872_739	356	3L	18766517	Deletion/489bp
NODE_5872_739	356	X	18852111	Deletion/396bp
NODE_14483_586	313	3L	18766795	Deletion/403bp
NODE_404_4981	4734	3LHet	539937	Deletion/437bp
NODE_264_5990	4175	X	727684	Insertion/495bp
NODE_2862_978	429	2LHet	241152	Deletion/413bp
NODE_2862_978	429	3RHet	1277786	Deletion/410bp
NODE_2862_978	554	2RHet	1432111	Deletion/414bp
NODE_2862_978	554	X	21331292	Deletion/414bp
NODE_2862_978	429	3L	22757987	Deletion/414bp
NODE_207_6588	775	3RHet	2274615	Insertion/434bp
NODE_6012_736	444	U	7172533	Deletion/445bp
NODE_94_7495	2068	3LHet	1240774	Insertion/426bp
NODE_2224_1121	702	U	1700433	Deletion/475bp
NODE_2224_1121	702	U	2320101	Deletion/462bp
NODE_2224_1121	702	2L	22544356	Deletion/475bp
NODE_937_2514	724	2RHet	20653	Deletion/438bp
NODE_4295_805	580	2RHet	3845	Deletion/375bp
NODE_4295_805	580	2L	22574496	Deletion/379bp
NODE_4295_805	232	3RHet	808296	Deletion/391bp
NODE_11987_641	419	2R	14000545	Deletion/407bp
NODE_11987_641	224	3L	23350254	Deletion/407bp
NODE_2749_1008	576	2R	8601140	Deletion/495bp
NODE_2749_1008	439	3R	194290	Deletion/495bp
NODE_2749_1008	439	2R	8678610	Deletion/495bp
NODE_2749_1008	445	3L	24007322	Deletion/476bp
NODE_343_5228	4106	U	1678935	Deletion/483bp
NODE_633_3773	3430	X	11643279	Deletion/376bp
NODE_633_3773	439	3R	27716384	Deletion/378bp
NODE_2548_1049	435	X	21763683	Deletion/382bp
NODE_2548_1049	435	X	21771317	Deletion/382bp
NODE_1366_1737	648	X	135413	Deletion/700bp
NODE_4134_816	434	2R	17703607	Deletion/441bp
NODE_554_4287	3317	U	280932	Deletion/429bp
NODE_554_4287	3428	X	22027109	Deletion/418bp
NODE_1694_1387	992	2RHet	2937156	Deletion/497bp
NODE_1694_1387	994	2R	14250884	Deletion/497bp
NODE_1694_1387	992	3L	19842896	Deletion/497bp
NODE_3166_925	664	2R	2221729	Deletion/426bp
NODE_3166_925	664	2R	66164	Deletion/426bp
NODE_3166_925	664	3LHet	2122945	Deletion/426bp
NODE_3166_925	266	3R	5133228	Deletion/426bp
NODE_3166_925	266	3R	22387052	Deletion/426bp
NODE_4856_772	422	3RHet	1074533	Deletion/476bp
NODE_4856_772	422	3RHet	1084644	Deletion/480bp
NODE_4856_772	355	X	21467764	Deletion/481bp
NODE_2974_956	374	2RHet	542460	Deletion/450bp
NODE_2974_956	375	3R	3929424	Deletion/451bp
NODE_2974_956	375	2L	20217022	Deletion/451bp
NODE_3047_944	311	X	21764651	Deletion/373bp
NODE_3047_944	311	X	21772285	Deletion/373bp
NODE_3047_944	335	U	6200108	Deletion/373bp
NODE_3540_875	559	3RHet	777415	Deletion/416bp
NODE_310_5442	1386	3LHet	1763738	Deletion/377bp
NODE_68_7806	948	X	18852100	Deletion/396bp
NODE_68_7806	950	3L	18766508	Deletion/489bp
NODE_383_5083	992	3LHet	1769072	Insertion/377bp
NODE_257_6121	721	3RHet	2274595	Insertion/403bp
NODE_219_6457	5246	3LHet	342219	Deletion/433bp
NODE_280_5702	2184	U	2106367	Deletion/602bp
NODE_573_4157	359	3LHet	798364	Deletion/479bp
NODE_289_5643	1997	2LHet	354357	Deletion/418bp
NODE_4732_778	477	X	6326319	Deletion/447bp
NODE_4732_778	477	2L	10140457	Deletion/447bp
NODE_4732_778	477	3L	12914679	Deletion/447bp
NODE_4732_778	307	3L	13570493	Deletion/447bp
NODE_4732_778	307	3R	23617817	Deletion/447bp
NODE_893_2653	2002	U	665644	Deletion/493bp
NODE_893_2653	2002	3L	13572103	Deletion/493bp
NODE_893_2653	707	X	6324714	Deletion/493bp
NODE_893_2653	718	2L	10138863	Deletion/493bp
NODE_893_2653	723	3R	1164826	Deletion/493bp
NODE_2053_1171	770	U	1147284	Deletion/426bp
NODE_2053_1171	770	3LHet	2121337	Deletion/420bp
NODE_2053_1171	770	X	4724354	Deletion/420bp
NODE_2053_1171	413	3R	22388665	Deletion/420bp
NODE_2053_1171	776	2RHet	323040	Deletion/420bp
NODE_516_4564	2500	3LHet	1841671	Deletion/422bp
NODE_2306_1103	594	U	280603	Deletion/402bp
NODE_256_6124	2152	3RHet	1056790	Deletion/384bp
NODE_1723_1370	1106	2R	20245182	Deletion/385bp
NODE_1723_1370	1106	2R	21033391	Deletion/379bp
NODE_1723_1370	1106	2L	21445294	Deletion/385bp
NODE_1723_1370	266	X	3391720	Deletion/385bp
NODE_1723_1370	266	3L	18526335	Deletion/385bp
NODE_4398_798	530	4	9723	Deletion/398bp
NODE_4398_798	272	3RHet	1657727	Deletion/398bp
NODE_1140_2083	1183	3RHet	2223261	Deletion/461bp
NODE_2241_1117	570	3LHet	443303	Deletion/405bp
NODE_2241_1117	570	3LHet	450742	Deletion/406bp
NODE_2241_1117	570	2RHet	894580	Deletion/405bp
NODE_2241_1117	543	2R	1174302	Deletion/400bp
NODE_2241_1117	570	3RHet	1355338	Deletion/405bp
NODE_215_6483	3651	2R	2264760	Insertion/407bp
NODE_112_7421	1893	U	1427390	Deletion/479bp
NODE_2026_1181	308	3L	8015858	Deletion/453bp
NODE_1094_2191	1805	3RHet	1796983	Deletion/371bp
NODE_424_4840	707	U	1199872	Deletion/529bp
NODE_424_4840	4259	X	21678035	Deletion/529bp
NODE_424_4840	4260	3L	23912538	Deletion/529bp
NODE_1387_1720	810	3L	18766795	Deletion/403bp
NODE_1271_1852	805	2R	2257462	Deletion/407bp

Table S55. Partial INDEL variation statistics of detection results generated by the *de novo* mode of LongRepMarker on the mouse dataset.

Repeat fragment id	Location on fragment	Reference id	Location on Ref.	Variation/Length.
NODE_1612_10359	7177	CM001014.2	2839743	Deletion/513bp
NODE_1612_10359	3281	CM001014.2	3296769	Deletion/509bp
NODE_1612_10359	7179	CM001014.2	3780992	Deletion/469bp
NODE_212_43740	35758	GL456210.1	113792	Deletion/498bp
NODE_408_29557	28233	KZ289081.1	147140	Insertion/498bp
NODE_1363_12001	10085	CM000997.2	60526366	Insertion/453bp
NODE_1363_12001	10083	CM000997.2	60765716	Insertion/453bp
NODE_6694_4510	4027	CM000995.2	176220701	Deletion/483bp
NODE_6694_4510	493	CM000995.2	176933206	Deletion/483bp
NODE_6694_4510	490	CM000995.2	177403768	Deletion/483bp
NODE_6694_4510	4026	CM000995.2	177753322	Deletion/483bp
NODE_1065_14600	12938	CM000994.2	85599594	Insertion/492bp
NODE_630_21955	6645	GL456068.1	29708	Deletion/782bp
NODE_211_43800	37118	CM000997.2	147654002	Deletion/453bp
NODE_7896_3664	2176	CM001005.2	22961142	Deletion/494bp
NODE_293_36608	2991	JH584328.1	417366	Deletion/520bp
NODE_293_36608	2991	JH584266.1	417477	Deletion/520bp
NODE_820_17868	5685	KQ030486.1	22319	Deletion/454bp
NODE_820_17868	12222	GL456077.1	69596	Deletion/454bp
NODE_820_17868	5679	CM000997.2	60686584	Deletion/453bp
NODE_820_17868	5685	CM000997.2	61171930	Deletion/454bp
NODE_1394_11820	10654	CM001001.2	71145707	Deletion/563bp
NODE_5613_5344	2708	CM001010.2	76040608	Deletion/499bp
NODE_357_32532	31953	CM001006.2	120281172	Deletion/498bp
NODE_0_252066	165145	GL456022.2	1250599	Deletion/534bp
NODE_225_42514	5810	CM001000.2	7330707	Deletion/471bp
NODE_10447_2486	760	CM001002.2	119221553	Deletion/460bp
NODE_5601_5351	4460	CM000995.2	177475324	Deletion/504bp
NODE_289_36862	9264	GL456022.2	1641517	Insertion/551bp
NODE_1458_11294	1597	CM000995.2	175212312	Insertion/483bp
NODE_1458_11294	9222	CM000995.2	175863876	Insertion/483bp
NODE_780_18602	1792	JH590470.1	779149	Insertion/514bp
NODE_780_18602	1794	CM001010.2	36824539	Insertion/514bp
NODE_1417_11624	5277	CM001000.2	7330707	Deletion/471bp
NODE_122_60184	22575	GL456022.2	1017588	Deletion/528bp
NODE_122_60184	22575	CM001010.2	34768809	Deletion/528bp
NODE_464_27081	26779	CM001012.2	9640161	Deletion/483bp
NODE_3948_6574	3828	JH584324.1	2589865	Deletion/471bp
NODE_3948_6574	3827	CM000994.2	8193887	Deletion/471bp
NODE_3948_6574	3827	CM001001.2	90385973	Deletion/471bp
NODE_3948_6574	3828	CM000997.2	131225987	Deletion/471bp
NODE_2419_7854	914	CM001003.2	81684217	Deletion/473bp
NODE_2419_7854	914	CM001003.2	81843319	Deletion/473bp
NODE_1750_9646	4275	CM001013.2	170833612	Insertion/796bp
NODE_3385_7046	1492	CM001001.2	21304634	Deletion/460bp
NODE_313_35149	20394	CM001005.2	23069048	Insertion/572bp
NODE_770_18860	10299	CM000997.2	147391141	Insertion/587bp
NODE_649_21443	20663	CM001013.2	5102273	Deletion/558bp
NODE_302_35903	4616	CM001014.2	17307496	Deletion/608bp
NODE_4884_6041	1320	JH584293.1	28699	Deletion/467bp
NODE_4884_6041	4901	CM000997.2	42146528	Deletion/467bp
NODE_4884_6041	4774	CM000997.2	42643739	Deletion/467bp
NODE_9504_2846	1128	CM001005.2	21016173	Deletion/499bp
NODE_9286_2926	508	CM001013.2	124364290	Deletion/521bp
NODE_9286_2926	508	CM001013.2	125562337	Deletion/521bp
NODE_9286_2926	510	CM001013.2	125299304	Deletion/521bp
NODE_3074_7130	2405	CM001002.2	88576268	Deletion/466bp
NODE_1162_13541	4229	KZ289068.1	113232	Deletion/588bp
NODE_1162_13541	4229	CM000997.2	146196060	Deletion/588bp
NODE_1162_13541	4210	CM000997.2	146718666	Deletion/587bp
NODE_1162_13541	4262	GL456053.2	123786	Deletion/587bp
NODE_10012_2638	1323	CM001007.2	3545620	Deletion/585bp
NODE_277_37824	2618	CM001000.2	14730687	Deletion/572bp
NODE_1790_9471	1929	GL456350.1	180728	Insertion/467bp
NODE_1790_9471	7117	CM000997.2	41935405	Insertion/467bp
NODE_1790_9471	7141	CM000997.2	42287518	Insertion/467bp
NODE_5808_5209	3213	KQ030486.1	22317	Deletion/454bp
NODE_5808_5209	2013	GL456077.1	69568	Deletion/454bp
NODE_5808_5209	3208	CM000997.2	60845802	Deletion/454bp
NODE_5808_5209	3216	CM000997.2	60686589	Deletion/453bp
NODE_2075_8718	5396	GL456019.1	1119971	Insertion/499bp
NODE_674_20781	6910	CM001001.2	3016402	Deletion/493bp
NODE_1264_12692	2994	CM000995.2	175212298	Insertion/483bp
NODE_1264_12692	9230	CM000995.2	175863872	Insertion/483bp
NODE_7667_3820	373	CM000994.2	173972552	Deletion/467bp
NODE_595_23059	18878	CM000994.2	85599679	Insertion/498bp
NODE_1626_10282	7033	CM001005.2	19976628	Deletion/504bp
NODE_82_73090	6420	JH584315.1	23920	Deletion/637bp
NODE_121_60755	6195	CM001010.2	35619726	Deletion/1035bp
NODE_121_60755	48972	CM001010.2	35662503	Deletion/532bp
NODE_11236_2248	1245	CM000995.2	176810116	Deletion/672bp
NODE_10343_2521	1438	CM001007.2	3545485	Deletion/719bp
NODE_2_240432	27098	JH584328.1	1319159	Insertion/534bp
NODE_2_240432	27098	JH584266.1	1319398	Insertion/534bp
NODE_1645_10153	5260	CM001000.2	17371414	Deletion/479bp
NODE_2699_7390	2743	CM001010.2	76040633	Deletion/499bp
NODE_496_25812	8978	CM000995.2	177274588	Deletion/476bp
NODE_594_23103	19132	CM000995.2	175011309	Insertion/476bp
NODE_594_23103	5604	CM000995.2	177579255	Insertion/476bp
NODE_294_36585	4767	CM001001.2	20267480	Insertion/582bp
NODE_294_36585	4843	CM001001.2	20394754	Insertion/581bp
NODE_344_33196	25292	GL456017.2	177521	Insertion/782bp
NODE_255_39288	28663	GL456019.1	362268	Insertion/455bp
NODE_255_39288	28661	CM001007.2	52708239	Insertion/455bp
NODE_205_44936	26635	CM001005.2	23850327	Deletion/572bp
NODE_417_29104	23351	CM001001.2	20120501	Deletion/470bp
NODE_782_18572	7363	JH584266.1	637119	Deletion/452bp
NODE_7972_3610	2458	GL456350.1	182638	Insertion/498bp
NODE_7972_3610	656	CM000997.2	41933300	Insertion/498bp
NODE_7972_3610	656	CM000997.2	42144581	Insertion/498bp
NODE_7972_3610	3112	CM000997.2	42147037	Deletion/467bp
NODE_7972_3610	656	CM000997.2	42285531	Insertion/498bp
NODE_5522_5411	5193	CM000997.2	42146552	Deletion/467bp
NODE_5522_5411	5182	CM000997.2	42643879	Deletion/467bp
NODE_5522_5411	261	JH584293.1	28700	Deletion/467bp
NODE_12230_1994	924	CM001014.2	41248569	Deletion/474bp
NODE_327_34436	747	KZ289080.1	162122	Deletion/656bp
NODE_862_17118	5518	CM001000.2	7330707	Deletion/471bp
NODE_14912_1563	445	JH584318.1	140602	Deletion/450bp
NODE_1054_14728	14359	CM001005.2	20584858	Deletion/730bp
NODE_425_28748	698	GL456022.2	1641157	Insertion/604bp
NODE_55_89804	22517	CM000997.2	60526381	Insertion/454bp
NODE_55_89804	22508	CM000997.2	60765735	Insertion/454bp
NODE_6041_5023	307	CM001010.2	76040607	Deletion/499bp
NODE_851_17276	5247	CM001005.2	23069036	Insertion/552bp
NODE_10725_2392	879	CM001014.2	41248560	Deletion/483bp
NODE_675_20780	6930	CM001001.2	3016403	Deletion/493bp
NODE_240_40967	11633	CM001000.2	7330740	Deletion/471bp
NODE_91_68995	40124	JH584328.1	1087366	Insertion/528bp
NODE_91_68995	40119	JH584266.1	1087613	Insertion/528bp
NODE_110_63963	63610	CM001004.2	82998283	Deletion/594bp
NODE_110_63963	64235	KZ289075.1	63909	Deletion/594bp
NODE_1548_10679	606	CM001007.2	4296669	Deletion/581bp
NODE_1548_10679	10093	CM001007.2	6146167	Deletion/586bp
NODE_1548_10679	1022	CM001007.2	6918779	Insertion/496bp
NODE_258_39166	28661	CM001007.2	53414908	Deletion/455bp
NODE_6115_4953	779	GL456350.1	182617	Insertion/498bp
NODE_6115_4953	3767	CM000997.2	41933322	Insertion/498bp
NODE_6115_4953	3911	CM000997.2	42285555	Insertion/498bp
NODE_6115_4953	3744	CM000997.2	42641775	Insertion/498bp
NODE_665_21072	18818	CM000994.2	85599639	Insertion/498bp
NODE_5347_5543	579	CM001011.2	6346958	Deletion/500bp
NODE_5347_5543	575	CM000995.2	25719536	Deletion/474bp
NODE_5347_5543	578	CM001000.2	42515603	Deletion/510bp
NODE_5347_5543	571	CM001006.2	77636047	Deletion/506bp
NODE_2685_7408	2696	CM001010.2	76040604	Deletion/499bp
NODE_2147_8550	7888	GL456019.1	580641	Deletion/515bp
NODE_2776_7305	2710	CM001010.2	76040619	Deletion/499bp
NODE_544_24263	4318	CM000995.2	175212309	Insertion/483bp
NODE_544_24263	19537	CM000995.2	175864564	Insertion/483bp
NODE_777_18682	1805	CM001006.2	65803978	Deletion/504bp
NODE_777_18682	16725	CM001006.2	66121810	Deletion/588bp

Table S56. Partial INDEL variation statistics of detection results generated by the *de novo* mode of LongRepMarker on the HG003_NA24149_father dataset.

Repeat fragment id	Location on fragment	Reference id	Location on Ref.	Variation/Length.
NODE_77_Length	341	chr3	195589799	Insertion/50bp
NODE_16_Length	2127	chr22	18205220	Deletion/189bp
NODE_29_Length	1521	chr4	49710771	Deletion/202bp
NODE_29_Length	373	chrUn_KI270333v1	702	Deletion/235bp
NODE_29_Length	1591	chrUn_KI270333v1	1920	Deletion/124bp
NODE_29_Length	860	chrUn_KI270333v1	653	Insertion/80bp
NODE_287_Length	323	chrY	10927895	Deletion/70bp
NODE_339_Length	277	chr5	175498939	Deletion/50bp
NODE_840_Length	404	chr5	49601788	Deletion/60bp
NODE_64_Length	675	chr4	49101938	Deletion/70bp
NODE_64_Length	696	chr4	49107712	Deletion/84bp
NODE_64_Length	676	chr4	49109130	Deletion/70bp
NODE_126_Length	272	chr22	11946235	Deletion/124bp
NODE_126_Length	261	chr3	75760593	Deletion/62bp
NODE_88_Length	596	chrY	10748856	Deletion/55bp
NODE_88_Length	167	chr16	34097217	Deletion/53bp
NODE_53_Length	854	chrUn_KI270333v1	983	Deletion/159bp
NODE_53_Length	355	chr4	49710796	Deletion/159bp
NODE_53_Length	431	chrUn_KI270333v1	760	Deletion/112bp
NODE_53_Length	900	chrUn_KI270333v1	1229	Deletion/83bp
NODE_53_Length	983	chr4	49711180	Deletion/84bp
NODE_53_Length	1104	chr4	49711301	Deletion/84bp
NODE_53_Length	357	chr4	49710019	Deletion/83bp
NODE_53_Length	823	chr4	49709863	Deletion/158bp
NODE_53_Length	1314	chr4	49710354	Deletion/203bp
NODE_401_Length	369	chrY	10755401	Deletion/50bp
NODE_99_Length	821	chrY	10846827	Deletion/84bp
NODE_356_Length	208	chrY	10926537	Insertion/54bp
NODE_747_Length	223	chrUn_KI270757v1	14572	Deletion/57bp
NODE_684_Length	279	chr5	49657241	Deletion/65bp
NODE_684_Length	286	chr5	49660385	Deletion/70bp
NODE_467_Length	345	chr5	49602549	Deletion/71bp
NODE_907_Length	372	chr20	31063965	Deletion/50bp
NODE_52_Length	502	chr18	110704	Deletion/68bp
NODE_52_Length	720	chr18	110922	Deletion/68bp
NODE_686_Length	347	chr9	41229817	Deletion/51bp
NODE_686_Length	271	chrUn_KI270442v1	391775	Deletion/50bp
NODE_86_Length	172	chr4	49120897	Insertion/100bp
NODE_758_Length	155	chr10	41860219	Deletion/55bp
NODE_368_Length	232	chr4	49711416	Deletion/84bp
NODE_368_Length	289	chr4	49710768	Deletion/202bp
NODE_368_Length	307	chrUn_KI270333v1	716	Deletion/152bp
NODE_368_Length	598	chrUn_KI270333v1	1007	Deletion/76bp
NODE_368_Length	170	chrUn_KI270333v1	1150	Deletion/167bp
NODE_368_Length	255	chr4	49709585	Deletion/160bp
NODE_368_Length	467	chr4	49709797	Deletion/74bp
NODE_778_Length	334	chrUn_KI270591v1	5290	Insertion/50bp
NODE_734_Length	268	chr12	126327075	Deletion/201bp
NODE_504_Length	173	chrY	10963389	Deletion/124bp
NODE_315_Length	286	chr4	49102087	Deletion/55bp
NODE_314_Length	298	chr17	21856473	Deletion/130bp
NODE_987_Length	275	chr4	49709544	Deletion/83bp
NODE_85_Length	434	chr1	25631827	Deletion/251bp
NODE_85_Length	328	chr3	95325563	Insertion/57bp
NODE_823_Length	249	chr5	49601674	Deletion/141bp
NODE_295_Length	347	chr5	49658415	Deletion/65bp
NODE_295_Length	353	chr5	49602289	Deletion/90bp
NODE_17_Length	1665	chrUn_GL000220v1	128360	Deletion/112bp
NODE_17_Length	1667	chr21	8229758	Deletion/112bp
NODE_17_Length	1871	chr21	8416985	Insertion/196bp
NODE_254_Length	471	chr4	49635424	Deletion/50bp
NODE_179_Length	456	chr1	144328486	Deletion/227bp
NODE_179_Length	765	chr1	144328795	Deletion/88bp
NODE_179_Length	463	chr1_KI270765v1_alt	148271	Deletion/101bp
NODE_179_Length	868	chr1_KI270765v1_alt	148676	Deletion/62bp
NODE_179_Length	229	chr1	143502774	Deletion/62bp
NODE_179_Length	573	chr1	143503118	Deletion/111bp
NODE_179_Length	782	chr1	143503327	Deletion/74bp
NODE_179_Length	462	chr1	144637772	Deletion/227bp
NODE_232_Length	260	chr9	61992880	Deletion/58bp
NODE_45_Length	998	chr4	51471879	Deletion/343bp
NODE_45_Length	1000	chr4	51523251	Deletion/343bp
NODE_45_Length	917	chr4	50822193	Insertion/167bp
NODE_926_Length	302	chrY	10769340	Deletion/74bp
NODE_926_Length	415	chrY	11039571	Deletion/75bp
NODE_754_Length	187	chr7	150003517	Deletion/54bp
NODE_10_Length	120	chr4	49709969	Insertion/126bp
NODE_10_Length	740	chr4	49710589	Insertion/84bp
NODE_10_Length	835	chrUn_KI270467v1	1519	Insertion/167bp
NODE_10_Length	1177	chrUn_KI270467v1	1861	Insertion/114bp
NODE_10_Length	1547	chrUn_KI270467v1	2231	Insertion/84bp
NODE_10_Length	2455	chrUn_KI270467v1	3139	Deletion/122bp
NODE_646_Length	313	chr1	4144694	Deletion/50bp
NODE_11_Length	1679	chr20	31061873	Insertion/95bp
NODE_418_Length	340	chr5	49601737	Deletion/59bp
NODE_21_Length	569	chr10	133687907	Deletion/139bp
NODE_21_Length	1819	chr10	133689157	Deletion/67bp
NODE_21_Length	2044	chr10	133689382	Deletion/68bp
NODE_21_Length	569	chr4	190178019	Deletion/139bp
NODE_21_Length	1819	chr4	190179269	Deletion/67bp
NODE_21_Length	2025	chr4	190179475	Deletion/68bp
NODE_21_Length	355	chr18	108442	Insertion/67bp
NODE_21_Length	1378	chr18	109465	Deletion/206bp
NODE_971_Length	489	chrY	10850620	Insertion/50bp
NODE_24_Length	234	chr21	8203053	Deletion/146bp
NODE_24_Length	234	chr21	8386100	Deletion/146bp
NODE_24_Length	218	chr21	8430485	Insertion/95bp
NODE_293_Length	394	chr4	49709516	Deletion/84bp
NODE_293_Length	130	chrUn_KI270333v1	1070	Deletion/77bp
NODE_293_Length	308	chr4	49711411	Deletion/126bp
NODE_14_Length	1667	chr17	26603196	Insertion/335bp
NODE_14_Length	1771	chr17	26620595	Insertion/336bp
NODE_609_Length	358	chrY	10808498	Deletion/74bp
NODE_269_Length	462	chr10	41860068	Deletion/50bp
NODE_269_Length	272	chr4	49098615	Deletion/85bp
NODE_508_Length	418	chr5	49658637	Insertion/52bp
NODE_219_Length	500	chr4	49710269	Deletion/113bp
NODE_219_Length	341	chrUn_KI270333v1	325	Deletion/84bp
NODE_819_Length	255	chr10	41895444	Deletion/53bp
NODE_264_Length	414	chrY	56829353	Deletion/80bp

Table S57. Partial INDEL variation statistics of detection results generated by the *de novo* mode of LongRepMarker on the human-chr14 dataset.

Repeat fragment id	Location on fragment	Reference id	Location on Ref.	Variation/Length.
NODE_132_Length	319	gi	19553531	Deletion/111bp
NODE_132_Length	506	gi	20019960	Deletion/111bp
NODE_314_Length	306	gi	65994298	Insertion/54bp
NODE_314_Length	106	gi	22818416	Insertion/54bp
NODE_314_Length	151	gi	52675979	Insertion/54bp
NODE_77_Length	849	gi	105326375	Deletion/154bp
NODE_77_Length	1120	gi	105326646	Deletion/180bp
NODE_23_Length	488	gi	106159686	Deletion/185bp
NODE_168_Length	394	gi	70663237	Insertion/55bp
NODE_317_Length	92	gi	98143931	Insertion/55bp
NODE_281_Length	331	gi	49529917	Insertion/56bp
NODE_131_Length	404	gi	38734571	Insertion/58bp
NODE_6484_Length	44	gi	102846348	Deletion/51bp
NODE_6484_Length	44	gi	102846450	Deletion/51bp
NODE_7966_Length	34	gi	102845867	Deletion/51bp
NODE_126_Length	707	gi	106084159	Deletion/56bp
NODE_7820_Length	34	gi	104791218	Deletion/86bp
NODE_354_Length	371	gi	97235304	Deletion/78bp
NODE_9819_Length	32	gi	52996244	Deletion/50bp
NODE_9819_Length	32	gi	52996369	Deletion/50bp
NODE_575_Length	148	gi	82144846	Deletion/101bp
NODE_286_Length	356	gi	22819081	Insertion/55bp
NODE_9085_Length	28	gi	101711567	Deletion/90bp
NODE_123_Length	642	gi	94986010	Insertion/51bp
NODE_147_Length	363	gi	95117490	Insertion/150bp

Table S58. Partial INDEL variation statistics of detection results generated by the *de novo* mode of LongRepMarker on the HG002_NA24385_Son dataset.

Repeat fragment id	Location on fragment	Reference id	Location on Ref.	Variation/Length.
NODE_286_Length	365	chr4	49633966	Deletion/60bp
NODE_286_Length	180	chr4	49112996	Deletion/75bp
NODE_286_Length	201	chr4	49111178	Deletion/60bp
NODE_694_Length	442	chr4	49154943	Deletion/70bp
NODE_262_Length	433	chr10	41879151	Deletion/60bp
NODE_147_Length	358	chrUn_KI270333v1	1639	Insertion/84bp
NODE_147_Length	469	chr4	49710761	Deletion/83bp
NODE_147_Length	293	chrUn_KI270337v1	308	Deletion/160bp
NODE_147_Length	587	chrUn_KI270337v1	602	Deletion/84bp
NODE_147_Length	42	chr4	49711116	Deletion/249bp
NODE_759_Length	272	chr4	49139241	Deletion/60bp
NODE_759_Length	261	chr4	49141694	Deletion/60bp
NODE_759_Length	300	chr4	49125891	Deletion/70bp
NODE_759_Length	291	chr4	49148761	Deletion/90bp
NODE_350_Length	259	chrUn_KI270337v1	304	Deletion/244bp
NODE_630_Length	212	chrY	11326315	Insertion/69bp
NODE_891_Length	251	chr7	51672065	Deletion/125bp
NODE_466_Length	222	chr10	132821261	Deletion/50bp
NODE_527_Length	298	chr9	86053399	Deletion/67bp
NODE_481_Length	125	chr6	43294135	Deletion/66bp
NODE_19_Length	1173	chr4	50562717	Insertion/169bp
NODE_19_Length	1173	chr4	50882017	Insertion/169bp
NODE_157_Length	363	chr14	24546480	Deletion/117bp
NODE_970_Length	199	chrUn_KI270333v1	1237	Deletion/83bp
NODE_624_Length	309	chr5	49658586	Deletion/55bp
NODE_12_Length	2195	chr22	18205216	Deletion/189bp
NODE_12_Length	1211	chr21	10271377	Insertion/354bp
NODE_131_Length	442	chrY	10940944	Insertion/76bp
NODE_131_Length	378	chr20	28889936	Deletion/55bp
NODE_131_Length	345	chrY	56830684	Deletion/79bp
NODE_490_Length	235	chrUn_KI270438v1	104554	Deletion/65bp
NODE_173_Length	326	chrUn_KI270438v1	46973	Insertion/50bp
NODE_127_Length	270	chrUn_KI270336v1	333	Deletion/210bp
NODE_717_Length	177	chr4	49710820	Deletion/113bp
NODE_717_Length	198	chrUn_KI270333v1	477	Deletion/82bp
NODE_14_Length	1496	chrUn_GL000216v2	156218	Insertion/205bp
NODE_14_Length	772	chrUn_GL000216v2	152829	Insertion/68bp
NODE_14_Length	1491	chrUn_GL000216v2	153548	Insertion/206bp
NODE_14_Length	795	chrY	11326153	Insertion/67bp
NODE_14_Length	995	chrY	11326353	Insertion/69bp
NODE_14_Length	1496	chrY	11326854	Insertion/206bp
NODE_902_Length	318	chr4	49137001	Deletion/80bp
NODE_902_Length	328	chr4	49122646	Deletion/80bp
NODE_902_Length	328	chr4	49129942	Deletion/70bp
NODE_902_Length	342	chr4	49148445	Deletion/50bp
NODE_124_Length	253	chr4	49709500	Deletion/74bp
NODE_418_Length	184	chr8	85654334	Deletion/110bp
NODE_418_Length	253	chr8	85737507	Deletion/110bp
NODE_418_Length	184	chr8	85744190	Deletion/110bp
NODE_418_Length	184	chr8	85790026	Deletion/110bp
NODE_418_Length	184	chr8	85802219	Deletion/110bp
NODE_418_Length	184	chr8	85814413	Deletion/110bp
NODE_45_Length	488	chr21	8247409	Insertion/95bp
NODE_45_Length	488	chr21	8430433	Insertion/95bp
NODE_102_Length	250	chr4	49710010	Deletion/83bp
NODE_102_Length	228	chrUn_KI270337v1	246	Insertion/126bp
NODE_261_Length	223	chr4	49137778	Deletion/50bp
NODE_733_Length	146	chr4	49124434	Deletion/50bp
NODE_268_Length	444	chr16	88054227	Deletion/51bp
NODE_439_Length	214	chr4	49710491	Deletion/84bp
NODE_439_Length	322	chrUn_KI270333v1	2300	Deletion/84bp
NODE_185_Length	370	chr4	49711112	Deletion/167bp
NODE_185_Length	263	chrUn_KI270333v1	1330	Deletion/125bp
NODE_185_Length	307	chr4	49710755	Deletion/83bp
NODE_44_Length	353	chr11	67619502	Deletion/121bp
NODE_44_Length	544	chr12	131228468	Deletion/236bp
NODE_44_Length	784	chr19	11941038	Deletion/160bp
NODE_44_Length	198	chr4	76042714	Insertion/102bp
NODE_765_Length	272	chrUn_KI270333v1	1517	Deletion/84bp
NODE_42_Length	739	chr20	31074197	Insertion/145bp
NODE_540_Length	319	chr5	49602310	Deletion/55bp
NODE_47_Length	406	chr22	45555989	Deletion/322bp

Table S59. Partial INDEL variation statistics of detection results generated by the *de novo* mode of LongRepMarker on the ant dataset.

Repeat fragment id	Location on fragment	Reference id	Location on Ref.	Variation/Length.
NODE_14_Length	3988	GL888247.1	235428	Insertion/143bp
NODE_14_Length	5551	GL888247.1	236991	Insertion/87bp
NODE_14_Length	5910	GL888247.1	237350	Insertion/175bp
NODE_14_Length	6700	GL888247.1	238140	Insertion/120bp
NODE_14_Length	1735	GL888385.1	39189	Insertion/291bp
NODE_14_Length	2075	GL888385.1	39529	Insertion/120bp
NODE_14_Length	5658	GL888385.1	43112	Deletion/134bp
NODE_14_Length	8227	GL888385.1	45681	Insertion/111bp
NODE_18_Length	445	GL888180.1	85614	Insertion/283bp
NODE_55_Length	626	GL888374.1	138919	Deletion/113bp
NODE_55_Length	1724	GL888374.1	140017	Insertion/75bp
NODE_55_Length	891	GL888828.1	2511334	Insertion/329bp
NODE_55_Length	1852	GL888828.1	2512295	Insertion/75bp
NODE_11_Length	2808	GL888090.1	212105	Insertion/212bp
NODE_35_Length	8294	GL888302.1	14449	Insertion/178bp
NODE_6291_Length	39	GL888360.1	22031	Deletion/124bp
NODE_39_Length	1500	GL888643.1	103932	Deletion/141bp
NODE_39_Length	2013	GL888643.1	104445	Insertion/88bp
NODE_38_Length	249	GL888359.1	100821	Deletion/271bp
NODE_8500_Length	27	GL888567.1	653	Deletion/114bp
NODE_44_Length	4440	GL888543.1	2838	Deletion/118bp
NODE_44_Length	3904	GL888283.1	37678	Insertion/71bp
NODE_1590_Length	26	GL887634.1	252976	Deletion/60bp
NODE_56_Length	8858	GL887834.1	165540	Deletion/412bp
NODE_56_Length	4382	GL887834.1	153182	Deletion/86bp
NODE_56_Length	4878	GL887834.1	153678	Insertion/211bp
NODE_56_Length	5814	GL887834.1	154614	Insertion/158bp
NODE_26_Length	2184	GL888775.1	95977	Deletion/354bp
NODE_57_Length	1316	GL887554.1	9378	Deletion/182bp
NODE_57_Length	1464	GL887793.1	97633	Deletion/266bp
NODE_57_Length	747	GL888539.1	112569	Deletion/235bp
NODE_57_Length	506	GL887917.1	515366	Insertion/91bp
NODE_57_Length	750	GL887917.1	515610	Deletion/176bp
NODE_57_Length	1096	GL887917.1	515956	Insertion/59bp
NODE_57_Length	610	GL888645.1	194067	Insertion/119bp
NODE_49_Length	899	GL888447.1	11975	Insertion/72bp
NODE_1694_Length	16	GL888608.1	731051	Deletion/63bp
NODE_50_Length	2367	GL888594.1	9495	Insertion/205bp
NODE_42_Length	824	GL887828.1	5706	Deletion/139bp
NODE_42_Length	538	GL888254.1	139671	Deletion/212bp
NODE_42_Length	581	GL888191.1	200522	Insertion/380bp
NODE_43_Length	408	GL887862.1	16535	Deletion/241bp
NODE_8588_Length	35	GL887601.1	533	Deletion/115bp
NODE_46_Length	242	GL888482.1	28738	Deletion/93bp
NODE_46_Length	2699	GL888409.1	172460	Insertion/192bp
NODE_24_Length	267	GL888528.1	4369	Deletion/142bp
NODE_24_Length	2233	GL888528.1	6335	Deletion/65bp

Table S60. Partial INDEL variation statistics of detection results generated by the *de novo* mode of LongRepMarker on the HG004_NA24143_mother dataset.

Repeat fragment id	Location on fragment	Reference id	Location on Ref.	Variation/Length.
NODE_555_Length	319	chr4	49144271	Deletion/75bp
NODE_50_Length	775	chrUn_KI270333v1	1786	Deletion/84bp
NODE_50_Length	938	chrUn_KI270333v1	1949	Deletion/201bp
NODE_50_Length	598	chr4	49709833	Deletion/158bp
NODE_50_Length	319	chr4	49710134	Deletion/83bp
NODE_50_Length	665	chr4	49711049	Deletion/83bp
NODE_50_Length	1234	chr4	49711618	Deletion/76bp
NODE_50_Length	701	chrUn_KI270333v1	1015	Deletion/76bp
NODE_50_Length	840	chrUn_KI270337v1	787	Insertion/253bp
NODE_188_Length	398	chr4	9574460	Insertion/69bp
NODE_161_Length	559	chr16	34097247	Deletion/55bp
NODE_12_Length	1507	chr17	26620589	Insertion/336bp
NODE_354_Length	301	chr1	144103736	Deletion/173bp
NODE_326_Length	100	chr21	5328592	Deletion/70bp
NODE_798_Length	59	chr8	43241326	Deletion/83bp
NODE_798_Length	294	chr22	11215073	Deletion/132bp
NODE_290_Length	338	chrUn_KI270756v1	873	Deletion/71bp
NODE_290_Length	256	chr16	34097939	Deletion/51bp
NODE_290_Length	396	chr16	34098079	Deletion/75bp
NODE_639_Length	43	chr4	49709799	Deletion/75bp
NODE_17_Length	906	chr4	190179261	Deletion/67bp
NODE_17_Length	1128	chr4	190179483	Deletion/67bp
NODE_17_Length	906	chr10	133689149	Deletion/67bp
NODE_17_Length	1128	chr10	133689371	Deletion/67bp
NODE_17_Length	1439	chr18	108453	Insertion/67bp
NODE_864_Length	221	chrY	11026376	Deletion/52bp
NODE_367_Length	310	chrUn_KI270333v1	328	Deletion/117bp
NODE_367_Length	335	chr4	49710793	Deletion/75bp
NODE_156_Length	566	chr22	35601933	Deletion/175bp
NODE_680_Length	275	chrUn_KI270467v1	2230	Deletion/122bp
NODE_116_Length	771	chrY	56834194	Deletion/58bp
NODE_495_Length	310	chr17	21971431	Deletion/140bp
NODE_495_Length	310	chr17	21974588	Deletion/140bp
NODE_106_Length	305	chrY	56856795	Insertion/60bp
NODE_672_Length	210	chr9	134599487	Deletion/55bp
NODE_325_Length	106	chrUn_KI270467v1	880	Deletion/125bp
NODE_325_Length	583	chrUn_KI270466v1	921	Deletion/84bp
NODE_585_Length	103	chr9	137327857	Deletion/93bp
NODE_60_Length	1037	chr4	49710350	Deletion/74bp
NODE_60_Length	262	chrUn_KI270333v1	244	Deletion/77bp
NODE_60_Length	799	chrUn_KI270333v1	781	Deletion/116bp
NODE_60_Length	1052	chrUn_KI270333v1	1034	Deletion/75bp
NODE_13_Length	989	chr4	49710329	Deletion/74bp
NODE_13_Length	304	chrUn_KI270467v1	946	Insertion/84bp
NODE_13_Length	1126	chrUn_KI270467v1	1768	Insertion/72bp
NODE_13_Length	1454	chrUn_KI270467v1	2096	Insertion/84bp
NODE_13_Length	2701	chrUn_KI270467v1	3343	Deletion/124bp
NODE_13_Length	3222	chrUn_KI270466v1	2043	Insertion/209bp
NODE_13_Length	3544	chrUn_KI270466v1	2365	Insertion/83bp
NODE_87_Length	288	chr5	49658457	Deletion/70bp
NODE_87_Length	915	chr17	21970045	Deletion/105bp
NODE_924_Length	339	chr4	49098632	Deletion/111bp
NODE_616_Length	413	chr1	25768178	Deletion/149bp
NODE_580_Length	344	chr4	49099911	Deletion/59bp
NODE_710_Length	358	chrUn_KI270337v1	600	Deletion/125bp
NODE_410_Length	134	chrY	10632157	Insertion/50bp
NODE_410_Length	306	chrY	10632329	Deletion/81bp
NODE_391_Length	248	chr4	49710486	Deletion/125bp
NODE_391_Length	249	chr4	49710773	Deletion/201bp
NODE_391_Length	307	chrUn_KI270337v1	604	Deletion/124bp
NODE_391_Length	291	chrUn_KI270333v1	1232	Deletion/209bp
NODE_374_Length	455	chrUn_KI270337v1	533	Deletion/84bp
NODE_374_Length	353	chr4	49711162	Deletion/132bp
NODE_735_Length	451	chr4	49123333	Deletion/51bp
NODE_27_Length	509	chrUn_KI270438v1	103623	Deletion/75bp
NODE_839_Length	118	chr1	790308	Deletion/69bp
NODE_81_Length	417	chr9	67594596	Deletion/288bp
NODE_81_Length	417	chr9	61532860	Deletion/215bp
NODE_81_Length	425	chr9	65463304	Deletion/215bp
NODE_81_Length	625	chr4	3568981	Deletion/172bp
NODE_81_Length	642	chr9	61993836	Deletion/288bp
NODE_81_Length	635	chr9	41782387	Deletion/217bp
NODE_202_Length	168	chr16	34064755	Deletion/64bp
NODE_127_Length	157	chr4	49710321	Deletion/74bp
NODE_33_Length	219	chr21	8247460	Insertion/95bp
NODE_33_Length	168	chr21	8430433	Insertion/95bp
NODE_238_Length	622	chr17	26939834	Deletion/51bp
NODE_704_Length	369	chr4	49710855	Deletion/76bp
NODE_704_Length	293	chr4	49710493	Deletion/84bp
NODE_63_Length	937	chr5	49602538	Deletion/75bp
NODE_526_Length	200	chrY	56826320	Deletion/50bp
NODE_287_Length	390	chrUn_KI270333v1	960	Deletion/118bp
NODE_287_Length	150	chr4	49709516	Deletion/84bp
NODE_287_Length	341	chr4	49710487	Deletion/83bp
NODE_287_Length	224	chr4	49711235	Deletion/50bp
NODE_107_Length	445	chr5	49601917	Deletion/55bp
NODE_198_Length	268	chr4	49709626	Deletion/66bp
NODE_549_Length	135	chr4	49113665	Deletion/54bp
NODE_549_Length	124	chr4	49125466	Deletion/64bp
NODE_225_Length	299	chr5	49666503	Insertion/55bp
NODE_225_Length	167	chr5	49659485	Insertion/55bp
NODE_752_Length	187	chr4	49709987	Deletion/83bp
NODE_752_Length	257	chrUn_KI270336v1	215	Insertion/84bp
NODE_113_Length	381	chr1	20023491	Deletion/298bp
NODE_275_Length	444	chr8	43238155	Deletion/90bp
NODE_122_Length	639	chr2	89814801	Deletion/75bp
NODE_834_Length	127	chr4	49100353	Insertion/60bp
NODE_705_Length	272	chrUn_KI270337v1	598	Deletion/84bp
NODE_705_Length	297	chrUn_KI270333v1	1511	Deletion/84bp
NODE_68_Length	701	chr5	49666905	Deletion/55bp
NODE_68_Length	519	chr5	49602440	Insertion/54bp
NODE_342_Length	205	chr22	16347120	Deletion/90bp
NODE_541_Length	318	chr2	89839843	Deletion/55bp
NODE_728_Length	317	chr17	21960016	Deletion/64bp
NODE_433_Length	305	chr5	49659267	Deletion/70bp
NODE_14_Length	930	chr2_KI270772v1_alt	3765	Deletion/222bp
NODE_14_Length	930	chr2_KI270772v1_alt	10428	Deletion/199bp
NODE_14_Length	1796	chr2_KI270894v1_alt	203525	Deletion/199bp
NODE_14_Length	1796	chr2_KI270894v1_alt	210165	Deletion/222bp
NODE_14_Length	1796	chr2	90391878	Deletion/199bp
NODE_14_Length	1796	chr2	90398518	Deletion/222bp
NODE_216_Length	253	chr4	49709500	Deletion/74bp
NODE_857_Length	450	chr15	91749847	Deletion/72bp
NODE_29_Length	1073	chr1	143213352	Deletion/468bp
NODE_507_Length	294	chr16	88054080	Deletion/51bp
NODE_507_Length	294	chr16	88053468	Deletion/51bp
NODE_891_Length	205	chr10	41908998	Insertion/55bp
NODE_891_Length	212	chr10	41915397	Deletion/54bp
NODE_891_Length	216	chr10	41895560	Insertion/65bp
NODE_799_Length	201	chr20	31157353	Deletion/55bp
NODE_20_Length	693	chr4	49710323	Deletion/76bp
NODE_20_Length	395	chrUn_KI270333v1	769	Deletion/112bp
NODE_20_Length	1210	chrUn_KI270333v1	1300	Deletion/83bp
NODE_20_Length	2098	chrUn_KI270333v1	2188	Deletion/84bp
NODE_20_Length	818	chr4	49709868	Deletion/242bp
NODE_20_Length	1787	chr4	49710837	Deletion/160bp
NODE_20_Length	616	chrUn_KI270333v1	545	Insertion/160bp
NODE_20_Length	1960	chrUn_KI270333v1	1889	Insertion/160bp
NODE_20_Length	1226	chr4	49709940	Insertion/85bp
NODE_323_Length	150	chr2	130070169	Insertion/127bp
NODE_323_Length	437	chr2	130662072	Insertion/79bp
NODE_323_Length	328	chr2	131268767	Insertion/139bp

3.8 Comparison of detection performance between LongRepMark and RepeatMasker

LongRepMarker can discover some new repetition types that RepeatMasker cannot find. In order to prove this conclusion, we conducted two specific experiments: 1) classify the detection results of the two tools by RepeatModeler2, and then compare the classification results, and 2) repetitive fragments in LongRepMarker's detection results covered by the detection results of RepeatMasker are removed, and then the remaining fragments are classified in detail by RepeatModeler2. Those two specific experiments are carried on the three species of Drosophila, Ant and Human-chr14. In order to fully demonstrate the high specificity of repeat sequences detected by LongRepMarker, the working mode of it is set to *de novo*, and its input is sequencing reads. The input of RepeatMasker can only be the reference genome of species, because it only supports the reference sequence as input.

The detailed steps of experiment 1 are as follows: 1) input the reference sequence into RepeatMasker to get the masked sequences; 2) extract the masked sequences as the repetitive sequences; 3) input the sequencing reads into LongRepMarker to obtain the repeated sequences; 4) classify the repetitive sequences generated from the two tools by RepeatModeler2; 5) compare the classification results. The detailed steps of experiment 2 are as follows: 1) input the reference sequence into RepeatMasker to get the masked sequences; 2) extract the masked sequences as the repetitive sequences; 3) input the sequencing reads into LongRepMarker to obtain the repeated sequences; 4) remove the repetitive fragments in LongRepMarker's detection results that can be covered by the detection results of RepeatMasker; 5) classify the remaining repetitive sequences in LongRepMarker's detection results; 6) Analyze the classification results. Some results of experiments 1 and 2 are shown in Table S61 and Table S62, respectively.

Table S61. Results of experiment 1.

	LongRepMarker			**RepeatMasker**		
species	Main class	Sub-class	amount	Main class	Sub-class	amount
Drosophila	LTR	Ngaro	1	LTR	Ngaro	0
	LTR	Gypsy	560	LTR	Gypsy	171
	LTR	Pao	130	LTR	Pao	5
	Other	None	130	Other	None	0
	LINE	Jockey	94	LINE	Jockey	52
	LINE	I-Jockey	152	LINE	I-Jockey	62
	LINE	R1-LOA	11	LINE	R1-LOA	0
	LINE	R2	2	LINE	R2	0
	Satellite	None	764	Satellite	None	4
	DNA	P	101	DNA	P	1
	DNA	Tc-Mar-Tc1	17	DNA	Tc-Mar-Tc1	5
	DNA	hAT-hATm	2	DNA	hAT-hATm	0
	DNA	IS	1	DNA	IS	0
	DNA	MULE-NOF	4	DNA	MULE-NOF	0
	DNA	hAT-hobo	4	DNA	hAT-hobo	0
Ant	RC?	Helitron	2	RC?	Helitron	0
	LTR	Gypsy	145	LTR	Gypsy	17
	LTR	Pao	69	LTR	Pao	17
	LTR	DIRS	1	LTR	DIRS	0
	LTR	ERVK	1	LTR	ERVK	0
	LTR	Other	1	LTR	Other	0
	LINE	Penelope	172	LINE	Penelope	5
	LINE	I	6	LINE	I	0
	LINE	R2-NeSL	11	LINE	R2-NeSL	0
	LINE	Tad1	1	LINE	Tad1	0
	DNA	Maverick	136	DNA	Maverick	4
	DNA	Kolobok-T2	97	DNA	Kolobok-T2	26
	DNA	TcMar-Tc1	227	DNA	TcMar-Tc1	136
	DNA	Kolobok-Hydra	5	DNA	Kolobok-Hydra	0
	DNA	MULE-NOF	14	DNA	MULE-NOF	0
	DNA	Crypton-V	6	DNA	Crypton-V	0
	DNA	hAT-hAT19	3	DNA	hAT-hAT19	0
	DNA	TcMar-ISRm11	1	DNA	TcMar-ISRm11	0
	DNA	CMC-Transib	7	DNA	CMC-Transib	0
	DNA	MuLE-MuDR	1	DNA	MuLE-MuDR	0
	DNA	MuLE-NOF	2	DNA	MuLE-NOF	0
	DNA	PIF-ISL2EU	1	DNA	PIF-ISL2EU	0
	DNA	PIF-Spy	5	DNA	PIF-Spy	0
	DNA	CMC-Chapaev-3	5	DNA	CMC-Chapaev-3	0
	DNA	PiggyBac	3	DNA	PiggyBac	0
	DNA	TcMar-Cweed	1	DNA	TcMar-Cweed	0
	DNA	IS	1	DNA	IS	0
Human-chr14	LINE	L1	17221	LINE	L1	16970
	Satellite	None	70	Satellite	None	15
	Satellite	centromeric	100	Satellite	centromeric	1
	DNA	Ginger	7	DNA	Ginger	0
	DNA	Novosib	3	DNA	Novosib	0
	DNA	Zisupton	3	DNA	Zisupton	0
	DNA	Sola-1	2	DNA	Sola-1	0
	DNA	Sola-3	5	DNA	Sola-3	0

Table S62. Results of experiment 2.

Drosophila			**Ant**			**Human-chr14**		
Main class	Sub-class	amount	Main class	Sub-class	amount	Main class	Sub-class	amount
LTR	Pao	136	LTR	Gypsy	152	LTR	ERVL-MaLR	2
LTR	Gypsy	613	LTR	Pao	64	LTR	ERV1	4
LTR	Copia	40	-	-	-	LTR	Gypsy	6
LTR	Other	2	-	-	-	LTR	Pao	1
-	-	-	-	-	-	LTR	Copia	4
-	-	-	-	-	-	LTR	ERVK	1
-	-	-	-	-	-	LTR	ERVL	1
LINE	R1	169	LINE	Penelope	175	LINE	L1	7
LINE	R1-LOA	13	-	-	-	LINE	R2-NeSL	1
LINE	Jockey	107	-	-	-	LINE	L2	3
LINE	I-Jockey	156	-	-	-	-	-	-
LINE	CR1	22	-	-	-	-	-	-
LINE	I	29	-	-	-	-	-	-
LINE	LOA	19	-	-	-	-	-	-
LINE	R2	1	-	-	-	-	-	-
LINE	L2	1	-	-	-	-	-	-
DNA	TcMar-Pogo	5	DNA	Maverick	151	DNA	MULE-MuDR	1
DNA	P	97	DNA	Kolobok-T2	90	DNA	hAT-Charlie	3
DNA	hAT-Ac	4	DNA	TcMar-Mariner	136	DNA	PiggyBac	1
DNA	MULE-NOF-NOF	4	DNA	TcMar-Tc1	182	DNA	CMC-EnSpm	2
DNA	hAT-hATm	2	-	-	-	DNA	MuLE-MuDR	1
DNA	TcMar-Tc1	18	-	-	-	DNA	Ginger	1
DNA	CMC-Transib	7	-	-	-	DNA	Other	1
DNA	hAT-hobo	5	-	-	-	-	-	-
DNA	IS	1	-	-	-	-	-	-
RC	Helitron	10	RC	Helitro	29	-	-	-
-	-	-	-	-	-	SINE	MIR	2
rRNA	-	7	-	-	-	-	-	-
-	-	-	-	-	-	scRNA	-	1
Simple_repeat	-	21	Simple_repeat	-	4	Simple_repeat	-	2
Satellite	-	382	-	-	-	Satellite	telomeric	1
Unknown	-	3568	Unknown	-	10046	Unknown	-	2551

The results in Tables S61 and S62 show that LongRepMarker can find some new repetitive sequence types that cannot be found by RepeatMasker. For example, the results in Table S61 show that the former tool found DNA transposon elements such as hAT-hATm, IS, MULE-NOF and hAT-hobo on the Drosophila dataset, but these are not found by latter tool. In addition, according to the number of repeats in some categories, LongRepMarker can find more repeats than the latter tool under the same conditions. For example, LongRepMarker found 277 DNA transposon elements with subcalss name tcmar-tc1 in ant dataset, while RepeatMasker only found 136 such elements. Further more, it can be seen from the results shown in Table S62 that LongRepMarker can find many unique repetitive sequences which do not appear in RepeatMasker's detection results at all. For example, LINE elements such as R1, R1-LOA, Jockey, I-Jockey, CR1, I, LOA, R2 and L2 on the Drosophila dataset only appear in the detection results of LongRepMarker.

4 Conclusion

Numerous studies have shown that the repetitive elements in genomes play an indispensable role in the evolution, inheritance, variation, gene expression, transcriptional regulation, chromosome construction, and physiological metabolism of organisms, and they are one of the principal causes of genomic instability. Most existing detection methods cannot achieve satisfactory performance on identifying repeats in terms of both accuracy and size, since NGS reads are too short to identify long repeats whereas SMS long reads are with high error rates.

In this study, we present a novel identification framework, LongRepMarker, based on the global *de novo* assembly of Illumina short paired-end reads and barcode linked reads or SMS long reads, and the *k-mer*-based multiple sequence alignment for precisely marking long repetitive sequences in genomes. LongRepMarker provides five different working modes: 1) the reference-assisted mode, which can quickly and accurately derive a repeat library for some large species when the reference genomes are provided. 2) the *de novo* mode, which consists of 4 sub-modes (*de novo* mode based on only NGS short reads, *de novo* mode based on NGS short reads + barcode linked reads, *de novo* mode based on NGS short reads + SMS long reads and *de novo* mode based on only SMS long reads) and can identify the repeats in the genomes to a greater extent by assembling mixed sequencing reads of different spans. Among them, the *de novo* mode based on only SMS long reads is one of the few methods that only rely on the third generation sequencing reads for repetitive sequences detection, and has the advantages of low memory consumption, high speed and high detection accuracy. The experimental results show that LongRepMarker can not only identify the repetitive sequences comprehensively, accuracy and rapidly in the reference-assisted mode, but also achieve more satisfactory results than most existing *de novo* detection methods.

5 ACKNOWLEDGEMENTS

This work has been supported by the National Natural Science Foundation of China under Grant: No.62002388, No.61772557, No.61732009, Hunan Provincial Science and technology Program (No. 2018wk4001), 111 Project (No. B18059), and King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under award numbers BAS/1/1624-01, FCC/1/1976-18-01, FCC/1/1976-23-01, FCC/1/1976-25-01, FCC/1/1976-26-01, REI/1/0018-01-01, REI/1/4216-01-01, REI/1/4437-01-01, REI/1/4473-01-01, and URF/1/4098-01-01.

References

1. Kazazian,H.H., "Mobile elements: drivers of genome evolution," science, vol. 303, no. 5664, pp. 1626-1632, 2004.
2. Liao X., Li M., Zou Y., et al., "Improving de novo assembly based on read classification," IEEE/ACM Trans Comput Biol Bioinform, vol. 17, no. 1, pp. 177-188 , 2018.
3. Treangen T. J. and Salzberg S. L., "Repetitive DNA and next-generation sequencing: computational challenges and solutions," Nature Reviews Genetics , vol. 13, no. 1, pp. 36 ,2012.
4. Lu Q., Wallrath L. L., Granok H., et al., "$(CT)_n$ $(GA)_n$ repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the Drosophila hsp26 gene," Molecular and Cellular Biology, vol. 13, no. 5, pp. 2802-2814, 1993.
5. Kundu T.K. and MRS R., "CpG islands in chromatin organization and gene expression," The Journal of Biochemistry, vol. 125, no. 2, pp. 217-222, 1999.
6. Shapiro J. A. and Von Sternberg R., "Why repetitive DNA is essential to genome function," Biological Reviews, vol. 80, no. 2, pp. 227-250, 2005.
7. Kaltenegger E., Leng S. and Heyl A., "The effects of repeated whole genome duplication events on the evolution of cytokinin signaling pathway," BMC evolutionary biology, vol. 18, no. 1, pp. 76, 2018.
8. Lu S., Wang G., Bacolla A., et al., "Short inverted repeats are hotspots for genetic instability: relevance to cancer genomes," Cell reports, vol. 10, no. 10, pp. 1674-1680, 2015.
9. Pavlicek A., Kapitonov V.V. and Jurka J., "Human Repetitive DNA," Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine, Springer Inc, Berlin, Heidelberg, 2005.
10. Gary B., "Tandem repeats finder: a program to analyze DNA sequences," Nucleic Acids Research, Vol. 27, no. 2, pp. 573-580, 1999.
11. Borstnik B. and Pumpernik D., "Tandem repeats in protein coding regions of primate genes," Genome Research, vol. 12, no. 6, pp. 909-915, 2002.
12. Achara S., "Chapter 19 - Introduction to Human Genetics," Clinical and Translational Science, Elsevier Inc, pp. 265-287, 2009.
13. Wicker, T., Sabot, F., Hua-Van, A., et al., "A unified classification system for eukaryotic transposable elements". Nat Rev Genet, Vol. 8, pp. 973-982, 2007.
14. Du, D., Du, X., Mattia, M. R., et al., "LTR retrotransposons from the Citrus x clementina genome: characterization and application". Tree Genetics & Genomes, Vol. 14, no. 43, 2018.
15. Schmidt T., "LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes," Plant Mol Biol, vol. 40, no. 6, pp. 903-910, 1999.
16. Wang W., Lin C., Lu D., et al., "Chromosomal transposition of PiggyBac in mouse embryonic stem cells," Proc Natl Acad Sci USA, vol. 105, no. 27, pp. 9290-9295, 2008.
17. Smit, A. F. A., Hubley, R. and Green, P., "RepeatMasker Open-4.0," Google Scholar, 2015.
18. Tarailo-Graovac M. and Chen N., "Using RepeatMasker to identify repetitive elements in genomic sequences," Current protocols in bioinformatics, vol. 25, no. 1, pp. 4.10.1-4.10.14, 2009.
19. Tempel S., "Using and understanding RepeatMasker," Mobile Genetic Elements, Humana Press, pp. 29-51, 2012.

20. Jurka J., Klonowski P., Dagman V., et al., "CENSOR-a program for identification and elimination of repetitive elements from DNA sequences," Computers & chemistry, vol. 20, no. 1, pp. 119-121, 1996.
21. Kennedy R. C., "Identification and annotation of transposable elements and agent-and gis-based modeling of pathogen transmission," University of Notre Dame, 2011.
22. Bedell J. A., Korf I. and Gish W., "MaskerAid: a performance enhancement to RepeatMasker," Bioinformatics, vol. 16, no. 11, pp. 1040-1041, 2000.
23. Li X., Kahveci T. and Settles A. M., "A novel genome-scale repeat finder geared towards transposons," Bioinformatics, vol. 24, no.4 , pp. 468-476, 2007.
24. Fiston-Lavier A. S., Carrigan M., Petrov D. A., et al., "T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data," Nucleic acids research, vol. 39, no. 6, pp. e36-e36, 2010.
25. Jiang N., "Overview of repeat annotation and de novo repeat identification," Plant Transposable Elements, Humana Press, Totowa, NJ, pp. 275-287, 2013.
26. Ellinghaus D., Kurtz S. and Willhoeft U., "LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons," BMC bioinformatics, vol. 9, no. 1, pp. 18, 2008.
27. Darzentas N., Bousios A., Apostolidou V., et al., "MASiVE: Mapping and Analysis of SireVirus Elements in plant genome sequences," Bioinformatics, vol. 26, no. 19, pp. 2452-2454, 2010.
28. Rho M, Choi J H, Kim S, et al., "De novo identification of LTR retrotransposons in eukaryotic genomes," BMC Genomics, vol. 8, no. 1, pp. 90, 2007.
29. Tu Z., Li S. and Mao C., "The changing tails of a novel short interspersed element in Aedes aegypti: genomic evidence for slippage retrotransposition and the relationship between 3' tandem repeats and the poly (dA) tail," Genetics, vol. 168, no. 4, pp. 2037-2047, 2004.
30. Han Y. and Wessler S. R., "MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences," Nucleic acids research, vol. 38, no. 22, pp. e199-e199, 2010.
31. Ye C., Ji G. and Liang C., "detectMITE: A novel approach to detect miniature inverted repeat transposable elements in genomes," Sci Rep, vol. 6, pp. 19688, 2016.
32. Zhijian T., "Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito Anopheles gambiae," Proceedings of the National Academy of Sciences, vol. 98, no. 4, pp. 1699-1704, 2001.
33. Chen Y., Zhou F., Li G. and Xu Y., "MUST: a system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi," Gene, vol. 436, pp. 1-7, 2009.
34. Yang, G., "MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements," BMC Bioinformatics, vol. 14, no. 186, 2013. https://doi.org/10.1186/1471-2105-14-186.
35. Crescente, J., Zavallo, D., Helguera, M. et al., "MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes," BMC Bioinformatics, vol. 19, no. 348, 2018. https://doi.org/10.1186/s12859-018-2376-y.
36. Shi J. and Liang C., "Generic Repeat Finder: A high-sensitivity tool for genome-wide de novo repeat detection," Plant physiology, vol. 180, no. 4, pp. 1803-1815, 2019.
37. Agarwal P., "The Repeat Pattern Toolkit (RPT): analyzing the structure and evolution of the C. elegans genome," International Conference on Intelligent Systems for Molecular Biology, vol. 2, pp. 1-9, 1994.
38. Chen G. L., Chang Y. J. and Hsueh C. H., "PRAP: an ab initio software package for automated genome-wide analysis of DNA repeats for prokaryotes," Bioinformatics, vol. 29, no.21, pp. 2683-2689, 2013.
39. Edgar R. C. and Myers E. W., "PILER: identification and classification of genomic repeats," Bioinformatics, vol. 21, no. suppl_1, pp. i152-i158, 2005.
40. Nicolas J., Peterlongo P. and Tempel S., "Finding and characterizing repeats in plant genomes," Plant Bioinformatics, Humana Press, pp. 293-337, 2016.
41. Ye C., Ji G. and Liang C., "detectMITE: A novel approach to detect miniature inverted repeat transposable elements in genomes," Scientific reports, vol. 6, pp. 19688, 2016.
42. Rodrigo L., Ville S., Stephen R., Asif K., Warren G., "WU-Blast2 server at the European Bioinformatics Institute," Nucleic Acids Research, vol. 31, no. 13, pp. 3795C3798, 2003.
43. Chen N., "Using Repeat Masker to identify repetitive elements in genomic sequences," Current protocols in bioinformatics, vol. 25, no.1, pp. 4.10.1-4.10.14, 2004.
44. Ou S., Su W., Liao Y., Chougule K., Agda JRA., Hellinga AJ., Lugo CSB., Elliott TA., Ware D., Peterson T., Jiang N., Hirsch CN. and Hufford MB., "Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline," Genome Biol, vol. 20, no. 1, pp. 275, 2019.
45. Surya S., Susan B., Zenaida V. M. and Daniel G. P., "Empirical comparison of ab initio repeat finding programs," Nucleic Acids Research, vol. 36, no. 7, pp. 2284C2294, 2008.
46. Price A. L., Jones N. C. and Pevzner P. A., "De novo identification of repeat families in large genomes," Bioinformatics, vol. 21, no. suppl_1, pp. i351-i358, 2005.
47. Li R., Ye J., Li S., Wang J., Han Y., Ye C., Wang J., Yang H., Yu J., Wong GK. and Wang J., "ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun," PLoS Comput Biol, vol. 1, no. 4, pp. e43, 2005.
48. Shi J. and Liang C., "Generic Repeat Finder: A High-Sensitivity Tool for Genome-Wide De Novo Repeat Detection," Plant Physiol, vol. 180, no. 4, pp. 1803-1815, 2019.
49. Flynn JM., Hubley R., Goubert C., Rosen J., Clark AG., Feschotte C. and Smit AF., "RepeatModeler2 for automated genomic discovery of transposable element families," Proc Natl Acad Sci USA, vol. 117, no.17, pp. 9451-9457, 2020.
50. Ellinghaus D., Kurtz S., Willhoeft U., "LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons," BMC Bioinformatics, vol. 9, pp. 18, 2008.
51. Ou S. and Jiang N., "LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons," Plant Physiol, vol. 176, no. 2, pp. 1410-1422, 2018.
52. Zhao X. and Hao W., "LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons," Nucleic Acids Research, vol. 35, no. suppl_2, pp. W265CW268, 2007.
53. Su W., Gu X. and Peterson T., "TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome," Mol Plant, vol. 12, no. 3, pp. 447-460, 2019.
54. Xiong W., He L., Lai J., Dooner HK. and Du C., "HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes," Proc Natl Acad Sci USA, vol. 111, no. 28, pp. 10263-8, 2014.
55. Koch P., Platzer M. and Downie B. R., "RepARK-de novo creation of repeat libraries from whole-genome NGS reads," Nucleic acids research, vol. 42, no. 9, pp. e80-e80, 2014.
56. Chu C., Nielsen R. and Wu Y., "REPdenovo: inferring de novo repeat motifs from short sequence reads," PloS one, vol. 11, no. 3, pp. e0150719, 2016.
57. Guo R, Li Y R, He S, et al., "RepLong: de novo repeat identification using long read sequencing data," Bioinformatics, vol. 34, no. 7, pp. 1099-1107, 2017.

58. Gyorgy A., Norbert G., Luc D. and Wojciech M., "TEclassa tool for automated classification of unknown eukaryotic transposable elements," Bioinformatics, vol. 25, no. 10, pp. 1329C1330, 2009.
59. Liao X., Li M., Zou Y., et al., "Current challenges and solutions of de novo assembly," Quantitative Biology, vol. 7, no. 2, pp. 90-109, 2019.
60. Boetzer M, Pirovano W., "SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information," BMC bioinformatics, vol. 15, no. 1, pp. 211, 2014.
61. Kamath G M, Shomorony I, Xia F, et al., "HINGE: long-read assembly achieves optimal repeat resolution," Genome Research, vol. 27, no. 5, pp. 747-756, 2017.
62. Bankevich A., Nurk S., Antipov D., et al. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing," Journal of computational biology, vol. 19, no. 5, pp. 455-477, 2012.
63. Luo R, Liu B, Xie Y, et al., "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler," Gigascience, vol. 1, no. 1, pp. 18, 2012.
64. Simpson J T, Wong K, Jackman S D, et al., "ABySS: a parallel assembler for short read sequence data," Genome research, vol. 19, no. 6. pp. 1117-1123, 2009.
65. Zerbino D R. and Birney E., "Velvet: algorithms for de novo short read assembly using de Bruijn graphs," Genome research, vol. 18, no. 5, pp. 821-829, 2008.
66. Peng Y, Leung H C M, Yiu S M, et al., "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth," Bioinformatics, vol. 28, no.11, pp. 1420-1428, 2012.
67. Luo J., Wang J., Zhang Z., et al., "BOSS: a novel scaffolding algorithm based on an optimized scaffold graph," Bioinformatics, vol. 33, no. 2, pp. 169-176, 2017.
68. Coombe L., Warren R. L., Jackman S. D., et al., "Assembly of the complete Sitka spruce chloroplast genome using 10X Genomics GemCode sequencing data," PLoS One, vol. 11, no. 9, pp. e0163059, 2016.
69. Yeo S., Coombe L., Warren R. L., et al., "ARCS: scaffolding genome drafts with linked reads," Bioinformatics, vol. 34, no. 5, pp. 725-731, 2017.
70. Zhang H., Jain C., Aluru S., et al., "A comprehensive evaluation of long read error correction methods," bioRxiv, 2019.
71. Morisse P., Lecroq T., Lefebvre A., et al., "Long-read error correction: a survey and qualitative comparison," bioRxiv, 2020.
72. Choudhury O, Chakrabarty A, Emrich S J, et al., "HECIL: A Hybrid Error Correction Algorithm for Long Reads with Iterative Learning," Scientific Reports, vol. 8, no. 1 , pp. 1-9, 2018.
73. Salmela L. and Rivals E., "LoRDEC: accurate and efficient long read error correction," Bioinformatics, vol. 30, no. 24, pp. 3506-3514, 2014.
74. Bao E. and Lan L., "HALC: High throughput algorithm for long read error correction," BMC Bioinformatics vol. 18, no. 204, 2017.
75. Andrey D. P., Irina V., Anton B., et al., "ExSPAnder: a universal repeat resolver for DNA fragment assembly," Bioinformatics, vol. 30, no. 12, pp. i293-i301, 2014.
76. Huptas C, Scherer S, Wenning M, et al. "Optimized Illumina PCR-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly," BMC Research Notes, vol. 9, no.1, pp. 269-269, 2016.
77. Huptas C., Scherer S., & Wenning M., "Optimized Illumina PCR-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly," BMC Res Notes vol. 9, no. 269, 2016.
78. Hamid M., Timothy J. Close S. L., "De novo meta-assembly of ultra-deep sequencing data," Bioinformatics, Vol. 31, no. 12, pp. i9-i16, 2005.
79. Souvorov A., Agarwala R. & Lipman D., "SKESA: strategic k-mer extension for scrupulous assemblies," Genome Biol vol. 19, no. 153, 2018.
80. Yahav T., and Privman E., "A comparative analysis of methods for de novo assembly of hymenopteran genomes using either haploid or diploid samples," Sci Rep, vol. 9, no. 6480, 2019.
81. M. E. J. Newman, "Modularity and community structure in networks," Proceedings of the National Academy of Sciences, vol. 103, no. 23, pp. 8577-8582, 2006.
82. Blondel, V. D. et al., "Fast unfolding of communities in large networks," J. Stat. Mech. Theory Exp., pp. P10008, 2008.
83. Yang, Z., Algesheimer, R. & Tessone, C. A., "Comparative Analysis of Community Detection Algorithms on Artificial Networks," Scientific Reports, vol. 6, no. 1, pp. 30750, 2016.
84. Coombe L., Warren R. L., Jackman S. D., et al., "Assembly of the complete Sitka spruce chloroplast genome using 10X Genomics GemCode sequencing data," PLoS One, vol. 11, no. 9, pp. e0163059, 2016.
85. Yeo S., Coombe L., Warren R. L., and et al., "ARCS: scaffolding genome drafts with linked reads," Bioinformatics, vol. 34, no. 5, pp. 725-731, 2017.
86. Luo R, Sedlazeck F J, Darby C A, et al., "LRSim: a linked-reads simulator generating insights for better genome partitioning," Computational and structural biotechnology journal, vol. 15, pp. 478-484, 2017.
87. Sedlazeck FJ., Rescheneder P., Smolka M., et al., "Accurate detection of complex structural variations using single-molecule sequencing," Nat Methods, vol. 15, no. 6, pp. 461-468, 2018.
88. Li H., "Minimap2: pairwise alignment for nucleotide sequences," Bioinformatics, vol. 34, no. 18, pp. 3094-3100, 2018.
89. Guillaume M. and Carl K., "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers," Bioinformatics, vol. 27, no. 6, pp. 764-770, 2011.
90. Rizk G., Lavenier D. and Chikhi R., "DSK: k-mer counting with very low memory usage," Bioinformatics, vol. 29, no. 5, pp. 652-653, 2013.
91. Deorowicz S, Kokot M, Grabowski S, et al. "KMC 2: fast and resource-frugal k-mer counting," Bioinformatics, vol. 31, no. 10, pp. 1569-1576, 2015.
92. Li H., Handsaker B., Wysoker A., et al., "The sequence alignment/map format and SAMtools," Bioinformatics, vol. 25, no. 16, pp.2078-2079, 2009.
93. James T. Robinson, Helga Thorvaldsdttir, Wendy Winckler, et al., "Integrative Genomics Viewer," Nature Biotechnology vol. 29, pp. 24C26, 2011.