Supplementary Material of ‘msRepDB: a comprehensive repetitive sequence database of over 80,000 species’

Xingyu Liao1,2*, Kang Hu3*, Adil Salhi3, You Zou1*, Jianxin Wang3(*), and Xin Gao3(*)

1 Computational Bioscience Research Center, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia.
2 School of Computer Science and Engineering, Central South University, ChangSha, 410083, CHINA.
3 The latest version of the Dfam database (v3.4) only contains the specific data of 552 species (https://dfam.org/home), which can be further subdivided into unique data and the data fused with RepBase. In addition, the data of other species are directly inherited from RepBase (about 61,518 species). Compared with the latest version of the Dfam database, the msRepDB database currently collects the repetitive sequences of the data of other species are directly inherited from RepBase (about 61,518 species). Compared with the latest version of the Dfam database, the msRepDB database currently collects the repetitive sequences of the data of other species are directly inherited from RepBase (about 61,518 species).

1 Background

Repetitive sequences are prevalent in the genomes of all bacteria, plants and animals, and they cover nearly half of the human genome[1],[2]. Repetitive sequences play indispensable roles in the evolution, inheritance, variation, genomic instability, and serve as substrates for chromosomal rearrangements that include disease-causing deletions, inversions, and translocations[3],[4],[5],[6],[7]. For example, the number and types of repetitive sequences vary between organisms and may reflect how rapidly an organism evolves to changes in its environment[8],[9]. Comprehensive identification, classification and annotation of repetitive elements in genome sequences for research and drug discovery are important substrates for genome evolution[10] and Dfam[11] libraries are two most often used repeat databases, but they are not sufficiently complete. For instance, in the Glycine max genome when the combination of RepBase and Dfam is used as the repetitive sequence database, only 28.47% of bases can be annotated as LTR (Long Terminal Repeat) retrotransposons cannot be accurately annotated (Table S9). Due to the lack of a comprehensive repetitive sequence database of multiple species, the current research in this field is far from being satisfactory.

2 The importance of repetitive DNA detection in human health and disease research

Repetitive sequences are abundantly distributed in the genomes of all viruses, bacteria, plants and animals[13]. For example, they constitute up to 45% of the genome in mouse and 50%-70% in human[14]. The repetitive sequences of the genome play a central role in the stability of the chromosome, the cell cycle, and the regulation of gene expression, and they are also an important substrates for genome evolution[15],[16],[17],[18]. As an example, the number and types of repetitive sequences vary between organisms and may reflect how rapidly an organism evolves to changes in its environment[19],[20]. Moreover, they are fundamental to the cooperative molecular interactions which form nucleoprotein complexes[21], and have also been implicated in molecular and cellular dysfunction associated with human diseases[22]. For instance, the tandem repeat expansion has been associated with more than 40 monogenic disorder, which has has been shown to be a major genetic contributor to frontotemporal dementia (FTD), amyotrophic lateral sclerosis (ALS) and autism spectrum disorder (ASD), the middle of which is the most common form of motor neuron disease[23],[24] and the latter of which is a group of neurodevelopmental disorders characterized by atypical social function, communication deficits, restricted interests and repetitive behaviors[25],[26],[27]. Besides, the expression of retrotransposition-competent transposable elements can lead to more insertions which can disrupt gene function or alter gene expression, contributing to complex diseases such as lung cancer, pancreatic cancer, ovarian cancer, neurological diseases, blood diseases[28],[29],[30], etc. RepBase and Dfam are two most frequently used repeat databases, but they are not sufficiently complete. For instance, in the Glycine max genome when the combination of RepBase and Dfam is used as the repetitive sequence database, only 28.47% of bases can be annotated as LTR (Long Terminal Repeat) retrotransposons cannot be accurately annotated, and in the Drosohila genome, when the combination of RepBase and Dfam is used as the repeat database, only about 20% (Table S9) of bases can be annotated as repeats, and the true proportion should be about 21-22% [31],[32]. Due to the lack of a comprehensive repeat database of multiple species, the current research in this field is far from being satisfactory.

Comprehensive identification, classification and annotation of repeats in genomes can provide accurate and targeted questions towards understanding and diagnosis of complex disorders, optimization of plant properties, development of new drugs, and individual health management. RepBase [10] and Dfam[11] libraries are two most often used repeat databases, but they are not sufficiently complete. For instance, in the Glycine max genome when the combination of RepBase and Dfam is used as the repetitive sequence database, only 28.47% of bases can be annotated as LTR (Long Terminal Repeat) retrotransposons cannot be accurately annotated (Table S9). Due to the lack of a comprehensive repetitive sequence database of multiple species, the current research in this field is far from being satisfactory.

3 The composition of database elements

The latest version of the Dfam database (v3.4) only contains the specific data of 552 species (https://dfam.org/home), which can be further subdivided into unique data and the data fused with RepBase. In addition, the data of other species are directly inherited from RepBase (about 61,518 species). Compared with the latest version of the Dfam database, the msRepDB database currently collects the repetitive sequences of 84,601 species which are obtained based on the corresponding detection results of LongRepMarker after the two processes of removing impurities and chimeras, and constructing the consensus sequences. For the partial information of the 84,601 specific species contained in msRepDB, please refer to Tables S1, S2 and S3. Furthermore, we have also constructed a tree structure-based species list on the ‘Search and download’ and ‘Online masking’ webpages of the new version of the msRepDB database. Partial display of the tree structure-based species list is shown in the Figure S1. From the point of view of data integrity, msRepDB completely covers Dfam and RepBase, while providing data on some previously unlisted species.
Fig. S1. Partial display of the tree structure-based species list.

| Species Name                  | Taxonomy id | Common species name           |
|-------------------------------|-------------|-------------------------------|
| Cyphellophora europaea CBS 101466 | 1220924     | -                             |
| Fusarium proliferatum ET1    | 1227346     | -                             |
| [Candida] duobushaemulonis    | 1231522     | -                             |
| Plasmodium inui San Antonio 1 | 1237626     | -                             |
| Acanthamoeba castellanii str. Neff | 1257118     | -                             |
| Fusarium fujikuroi IMI 58289 | 1279085     | -                             |
| Kwoniella bestiolae CBS 10118 | 1296100     | -                             |
| Kwoniella dejecticola CBS 10117 | 1296121     | -                             |
| Kwoniella mangroviensis CBS 8507 | 1296122     | -                             |
| Aspergillus campestris IBT 28561 | 1392248     | -                             |
| Aspergillus steynii IBT 23096 | 1392250     | -                             |
| Aspergillus novofumigatus IBT 16806 | 1392255   | -                             |
| Aspergillus ochraceoroseus IBT 24754 | 1392256 | -                             |
| Sporothrix schenckii 1099-18 | 1397361     | -                             |
| Sporothrix brasiliensis 5110 | 1398154     | -                             |
| Pneumocystis carinii B80     | 1408658     | -                             |
| Fonsecaea pedrosoi CBS 271.37 | 1442368     | -                             |
| Rhinocladiella mackenziei CBS 650.93 | 1442369 | -                             |
| Cebus imitator 2715852 Panamanian white-faced capuchin | | |
Table S2. Partial list B of the specific species in the msRepDB database.

| Species Name | Taxonomy id | Common species name |
|--------------|-------------|---------------------|
| Trichosporon asahii var. asahii CBS 2479 | 1186058 | Trichosporon species |
| Saprolegnia diclina VS20 | 1156394 | Saprolegnia species |
| Plasmodium cynomolgi strain B | 1120755 | Plasmodium species |
| Phytophthora sojae strain P6497 | 1094619 | Phytophthora species |
| Fusarium odoratissimum NRRL 54006 | 1089451 | Fusarium species |
| Lactococcus lactis subsp. lactis IO-1 | 1046624 | Lactococcus species |
| Kluyveromyces marxianus DMKU3-1042 | 1003335 | Kluyveromyces species |
| Leptosphaeria maculans JN3 | 985895 | Leptosphaeria species |
| Agaricus bisporus var. bisporus H97 | 936046 | Agaricus species |
| Plasmodium sp. gorilla clade G2 | 880535 | Plasmodium species |
| Exophiala dermatitidis NIH/UT8656 | 858893 | Exophiala species |
| Sordaria macrospora k-hell | 771870 | Sordaria species |
| Phycomyces blakesleeanus NRRL 1555(-) | 763407 | Phycomyces species |
| Phytophthora parasitica INRA-310 | 761204 | Phytophthora species |
| Rhizophagus irregularis DAOM 181602=DAOM 197198 | 747089 | Rhizophagus species |
| Coniophora puteana RWD-64-598 SS2 | 741705 | Coniophora species |
| Saprolegnia parasitica CBS 223.65 | 695850 | Saprolegnia species |
| Aspergillus aculeatus ATCC 16872 | 690307 | Aspergillus species |
| Batrachochytrium dendrobatidis JAM81 | 684364 | Batrachochytrium species |
| Sclerotinia sclerotiorum 1980 UF-70 | 665079 | Sclerotinia species |
| Bipolaris maydis ATCC 48331 | 665024 | Bipolaris species |
| Cryphonectria parasitica EP155 | 660469 | Cryphonectria species |
| Fusarium oxysporum NRRL 32931 | 660029 | Fusarium species |
| Tuber melanosporum Mel28 | 656061 | Tuber species |
| Spizellomyces punctatus DAOM BR117 | 645134 | Spizellomyces species |
| Colletotrichum graminicola M1.001 | 645133 | Colletotrichum species |
| Capsaspora owczarzaki ATCC 30864 | 595528 | Capsaspora species |
| Myripristis murdjan 586833 pinecone soldierfish | | |
Table S3. Partial list C of the specific species in the msRepDB database.

| Species Name | Taxonomy id | Common species name |
|--------------|-------------|---------------------|
| Bacteria     | 2           | eubacteria          |
| Firmicutes   | 1239        |                     |
| Lactococcus lactis | 1358     |                     |
| Citrus       | 2706        |                     |
| Rhodophyta   | 2763        | red algae           |
| Chondrus     | 2768        |                     |
| Chondrus crispus | 2769    | carragheen          |
| Bangiophyceae | 2797       |                     |
| Florideophyceae | 2806      |                     |
| Haptophyta   | 2830        |                     |
| Bacillariophyta | 2836     | diatoms             |
| Phaeodactylum | 2849       |                     |
| Phaeodactylum tricornutum | 2850 |                     |
| Phaeophyceae | 2870        | brown algae         |
| Ectocarpaceae | 2878       |                     |
| Emiliania    | 2902        |                     |
| Emiliania huxleyi | 2903      |                     |
| Chlorophyta  | 3041        | green algae         |
| Chlamydomonadales | 3042   |                     |
| Chlamydomonas | 3052       |                     |
| Chlamydomonas reinhardii | 3055 |                     |
| Volvocaceae  | 3065        |                     |
| Volvox       | 3066        |                     |
| Volvox carteri | 3067      |                     |
| Chlorella    | 3071        |                     |
| Chlorophyceae | 3166       |                     |
| Bryophyta    | 3208        | mosses              |
| Physcomitrium patens | 3218 |                     |
| Equisetaceae | 3256        |                     |
| Equisetum scirpoides | 3261 |                     |
| Ginkgoaceae  | 3309        |                     |
| Ginkgo       | 3310        |                     |
| Picea        | 3328        | Norway spruce       |
| Picea abies  | 3329        |                     |
| Picea glauca | 3330        | white spruce        |
| Magnoliopsida | 3398       | flowering plants    |
| Moraceae     | 3487        |                     |
| Morus        | 3497        | mulberries          |
| Spinacia oleracea | 3562     | spinach             |
| Vitaceae     | 3602        |                     |
| Vitis        | 3603        |                     |
| Malvaceae    | 3629        |                     |
| Theobroma    | 3640        |                     |
| Theobroma cacao | 3641     | cacao               |
| Malpighiales | 3646        |                     |
| Carica papaya | 3649      | papaya              |
| Cucumis      | 3655        |                     |
| Cucumis melo | 3656        | muskmelon           |
| Cucumis sativus | 3659      | cucumber            |
| Cucurbita maxima | 3661     | winter squash       |
| Cucurbita pepo | 3663       |                     |
| Populus      | 3689        | poplars             |
| Populus trichocarpa | 3694    | black cottonwood    |
| Arabidopsis thaliana | 3702 | thale cress         |
| Brassica napus | 3708       | rape                |
| Brassica rapa | 3711       | field mustard       |
| Malus domestica | 3750      | apple               |
| Prunus persica | 3760       | peach               |
| Cajanus cajan | 3821       | pigeon pea          |
| Cicer arietinum | 3827      | chickpea            |
| Glycine max | 3847        | soybean             |
| Medicago truncatula | 3880 | barrel medic        |
| Phaseolus vulgaris | 3885    |                     |
| Manihot esculenta | 3983     | cassava             |
| Ricinus communis | 3988      | castor bean         |
| Capsicum annuum | 4072      |                     |
| Solanum lycopersicum | 4081    | tomato              |
| Nicotiana sylvestris | 4096     | wood tobacco        |
| Nicotiana tabacum | 4097     | common tobacco      |
| Nicotiana tomentosiformis | 4098 |                     |
| Solanum tuberosum | 4113     | potato              |
| Erythranthe guttata | 4155     | spotted monkey flower |
| Lactuca sativa | 4236       |                     |
| Nelumbo nucifera | 4432      | sacred lotus        |
| Oryza sativa | 4530        | rice                |
| Setaria italica | 4555      | foxtail millet      |
| Sorghum bicolor | 4558       | sorghum             |
| Zea mays     | 4577        |                     |
| Ananas comosus | 4615       | pineapple           |
| Schizosaccharomyces pombe | 4896    | fission yeast       |
| Saccharomyces cerevisiae | 4932   | baker's yeast       |
| Leptosphaeria | 5021       |                     |
| Ophiostomataceae | 5152     |                     |
| Ganoderma    | 5314        |                     |
| Pleurotus ostreatus | 5322     | oyster mushroom     |
| Candida albicans | 5476      |                     |
| [Candida] glabrata | 5478     |                     |
| Alternaria alternata | 5599   |                     |
| Leishmania donovani | 5661     |                     |
| Giardia intestinalis | 5741   |                     |
| Dictyostelium purpureum | 5786    |                     |
| Hydra vulgaris | 6087       | swiftwater hydra    |
| Schistosoma mansoni | 6183    |                     |
| Caenorhabditis briggsae | 6238    |                     |
| Caenorhabditis elegans | 6239   |                     |
| Brugia malayi | 6279        |                     |
| Trichinella spiralis | 6334     |                     |
| Helobdella robusta | 6412     |                     |
| Aplysia californica | 6500     | California sea hare |
| Biomphalaria glabrata | 6526   |                     |
| Mytilus californianus | 6549    | California mussel   |
| Ostreidae    | 6563        | oysters             |
| Crassostrea virginica | 6565    | eastern oyster      |
| Mizuhopecten yessoensis | 6573   | Yesso scallop       |
| Mactridae    | 6581        | surf clams          |
| Octopoda     | 6638        |                     |
| Octopus vulgaris | 6645      | common octopus      |
| Penaeus vannamei | 6689      | Pacific white shrimp |
| Merostomata  | 6844        | horseshoe crabs     |
| Limulus polyphemus | 6850    | Atlantic horseshoe crab |
| Rhipicephalus microplus | 6941    | southern cattle tick |
| Ixodes scapularis | 6945    | black-legged tick   |
| Caledia      | 7013        |                     |
| Stauroderus  | 7018        |                     |
| Acyrthosiphon pisum | 7029     | pea aphid           |
| Bemisia tabaci | 7038       |                     |
| Tenebrio     | 7066        |                     |
| Tribolium castaneum | 7070    | red flour beetle    |
| Bombyx mori | 7091        | domestic silkworm   |
| Spodoptera frugiperda | 7108    | fall armyworm       |
| Trichoplusia ni | 7111       | cabbage looper      |
| Samia        | 7126        |                     |
| Manduca sexta | 7130       | tobacco hornworm    |
| Aedes aegypti | 7159       | yellow fever mosquito |
| Aedes albopictus | 7160      | Asian tiger mosquito |
| Anopheles albimanus | 7167    |                     |
| Anopheles arabiensis | 7173    |                     |
| Culex quinquefasciatus | 7176  | southern house mosquito |
| Loa loa      | 7209        | eye worm            |
| Ceratitis capitata | 7213      | Mediterranean fruit fly |
| Drosophila ananassae | 7217    |                     |
| Drosophila erecta | 7220      |                     |
| Drosophila grimshawi | 7222    |                     |
| Drosophila hydei | 7224      |                     |
| Drosophila mauritiana | 7226   |                     |
| Drosophila melanogaster | 7227  | fruit fly           |
| Drosophila miranda | 7229     |                     |
| Drosophila mojavensis | 7230    |                     |
| Drosophila navojoa | 7232      |                     |
| Drosophila persimilis | 7234    |                     |
| Drosophila pseudoobscura | 7237   |                     |
| Drosophila sechellia | 7238     |                     |
| Drosophila simulans | 7240     |                     |
| Drosophila subobscura | 7241    |                     |
| Drosophila virilis | 7244     |                     |
| Drosophila yakuba | 7245      |                     |
| Drosophila athabasca | 7248     |                     |
| Drosophila willistoni | 7260    |                     |
| Drosophila arizonae | 7263     |                     |
| Drosophila guanche | 7266      |                     |
4 The classes of repetitive DNA in the msRepDB database

According to the arrangement, the repetitive sequences contained in msRepDB database can be divided into two types: tandem repeats and scattered repeats (Fig. S2). Scattered repeats are commonly referred to transposons, which can also be divided into RNA transposons and DNA transposons these two categories. The RNA transposons can be roughly divided into three types according to the different transposon modes: long terminal repeat sequence (LTR), long interpersed nuclear element (LINE) and short interpersed nuclear element (SINE). The DNA transposons can be divided into four classes according to the different transposon modes: miniature inverted repeat transposable element (MITE), Crypton, Maverick and helitron. The tandem repeats commonly refer to sequences in which the repeating units are distributed in tandem, and their head and tail are connected together to form an aggregated region. The three most common types of tandem repeats in non-coding regions are satellite, minisatellite and microsatellite. The advantage of msRepDB is mainly reflected in its more complete collection of retrotransposons, especially LTRs and LINEs.

5 Approach and workflow

LongRepMarker[36] is a new framework developed recently by our group for comprehensive identification of genomic repetitive sequences. Comprehensive evaluations carried out in the study of LongRepMarker not only show that LongRepMarker can achieve more satisfactory results than the existing detection methods, but can also discover a large number of new repeat sequences and families.

1) Identification of overlap sequences.

The repetition relation is a special case of the overlap relation. Thus all possible repetition relationship can be found by searching overlap sequences. Overlap sequences occupy only a small portion of the overall sequences. By finding the overlap sequences between assemblies or chromosomes, the algorithm locates the repetitive sequences faster and more accurately.

2) Conversion of overlap sequences into unique k-mers.

The number and length of sequences will have a great impact on the efficiency of multiple sequence alignment. Generally, the more the number and the longer the length, the greater the computational resource consumption. The unique k-mers are much smaller than overlap sequences both in terms of number and length. Using unique k-mers instead of overlap sequences for mapping can greatly optimize the efficiency of multi-sequence alignment.

3) Generation of the multi-alignment unique k-mers and their coverage regions on overlap sequences.

The multi-alignment unique k-mers were first proposed in the paper of LongRepMarker, which refers to the unique k-mers that can be aligned to multiple different locations in the overlap sequences. Due to the 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. However, the multi-alignment unique k-mers are not affected by these factors. 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.

4) Classification of regions on overlap sequences that can be covered by multi-alignment unique k-mers.

Due to the short size of unique k-mers, it is easy to form a coupling alignment (coupling alignment refers to the fusion of unique k-mers that should not be fused together). To eliminate the influence of the coupling alignment, the algorithm further classifies the regions on the overlap sequences covered by the multi-alignment unique k-mers into two categories, and filters out the false repetitive sequences, thereby improving the accuracy of the detection results.

5) Merging fragments with duplication or inclusion.

The results of detection methods based on the multiple sequence alignment will inevitably contain redundant elements. In order to make the detection results as pure as possible without any impurities and redundancy, the algorithm merges the detected repetitive fragments with duplication and inclusion relationships.

6) Classification and annotation of the obtained repetitive sequences.

When the repetitive sequences are obtained, the algorithm also needs to classify and annotate them, because the repeats without classification and annotation information are meaningless. In this step, a distributed RepeatClassifier developed by our group is used to classify and annotate the obtained repetitive sequences.

Note that LongRepMarker is different from RepeatScout[37] and RepeatModeler2[38] in detection targets. RepeatScout and RepeatModeler2 both focus on the discovery of repeated families. It is well known that a repeat family is an abstraction of a type of repeat sequence (a one-to-many relationship), and its acquisition must go through the two operations of merging the obtained repeat sequences and taking the consensus sequence. However, the detection goal of LongRepMarker is not to find repeated families, but to comprehensively mine all repeated sequences of the genome, and provide a basis for accurately identifying the mutations that exist between different copies. Therefore, in the detection results of LongRepMarker, we merged duplicate copies with high consistency, and saved the duplicate copies with differences as much as possible.
possible, and at the same time analyzed the structural variation that occurred in the duplicate copies with differences. Our purpose is to provide a method to study the effect of variations that exist between different duplicate copies on the genetic, evolution and variation of organisms.

Although there will be some redundancy and chimerism in the detection results of LongRepMarker, the repetitive sequences identified by it are still the most complete compared to other existing tools. In order to remove impurities and chimeras in the detection results and output purer repetitive sequences for the database construction, three steps of impurity removal, chimerism removal and consensus sequence construction are carried out after the detection results of LongRepMarker obtained. When the purified repetitive sequences are generated, RepeatClassifier is used to classify and annotate these sequences. After that, the algorithm needs to merge the repeated sequences with its classification and annotation information, and form a file in the FASTA format[39]. In this generated file in the FASTA format, the sequence composed of A/T/G/C characters is a repeating sequence, and the sequence starting with an angle bracket above the repeating sequence is the annotation sequence, which contains the corresponding classification and annotation information[40].

6 Implementation

The data processing and analysis functions of msRepDB database have been implemented using Python v.3.6.9 (https://www.python.org/getit/) coupled with the SpringBoot integrated framework (https://spring.io/projects/spring-boot). msRepDB runs on a Linux-based Maven server 3.8.1 (Maven is a build automation tool used primarily for java projects, https://maven.apache.org/download.cgi). The database was developed using MySQL 5.7.31 (https://www.mysql.com/), and the web interface was developed using html5 markup language (https://en.wikipedia.org/wiki/HTML5) combined with Bootstrap v.5.0.2 (https://v3.bootcss.com), layUI v.2.6.8 (https://www.layui.com/) and JQuery v.1.11.1 (http://jquery.com) (Fig. S3). In the process of online masking, two aligners, bwa[41] and minimap2[42] were used, in which the short sequence fragments were aligned using bwa, and the long sequence fragments were aligned using minimap2. The screenshots of the main interfaces of the msRepDB database are shown in Figures S4-S8.

Fig. S3. Flow chart of msRepDB.

Fig. S4. A screenshot of the tree structure-based species list.
Fig. S5. A screenshot of the "Search and download" interface.

Fig. S6. A screenshot of the "Online masking" interface.
Fig. S7. A screenshot of the "Submit" interface.

Fig. S8. A screenshot of the "Team" interface.
7 Test and performance results

7.1 Comparison of the improved LongRepMarker and existing tools in detection performance

RepBase and Dian libraries are the two most frequently used repeat databases, but they are not sufficiently complete, because most of the repetitive sequences collected in these two libraries are obtained through some existing detection methods (such as RepeatMasker, RepeatScout, RepeatModeler, and RepeatModeler2). Due to the limitations of sequencing data and the defects in design of the detection principle, existing detection methods cannot accurately and comprehensively obtain the repetitive sequences of species.

LongRepMarker (DOI:10.1093/nar/gkab563, https://github.com/BioinformaticsCSU/LongRepMarker) is a new framework developed recently by our group for comprehensive identification of genomic repetitive sequences. Note that LongRepMarker is different from RepeatScout and RepeatModeler2 in detection targets. RepeatScout and RepeatModeler2 both focus on the discovery of repeated families. It is well known that a repeat family is an abstraction of a type of repeat sequence (a one-to-many relationship), and its acquisition must go through the two operations of merging the obtained repeat sequences and taking the consensus sequence. However, the detection goal of LongRepMarker is not to find repeated families, but to comprehensively mine all repeated sequences of the genome, and provide a basis for accurately identifying the mutations that exist between different copies. Therefore, in the detection results of LongRepMarker, we merged duplicate copies with high consistency, and saved the duplicate copies with differences as much as possible, and at the same time analyzed the structural variation that occurred in the duplicate copies with differences. Our purpose is to provide a method to study the effect of variations that exist between different duplicate copies on the genetic, evolution and variation of organisms.

Although there will be some redundancy and chimerism in the detection results of LongRepMarker, the specific steps of the experiment are to use RepeatMasker sequence as the annotation sequence, which contains the corresponding classification and annotation information. A file in the FASTA format is generated after the detection results of LongRepMarker obtained. When the purified repetitive sequences are generated, RepeatClassifier is used to classify and annotate these sequences. After that, the algorithm needs to merge the repeated sequences with its classification and annotation information, and form a file in the FASTA format. In this generated file in the FASTA format, the sequence composed of A/T/G/C characters is a repeating sequence, and the sequence starting with an angle bracket above the repeating sequence is the annotation information, which contains the corresponding classification and annotation information.

In order to evaluate the performance of the improved LongRepMarker, we conducted experiments on two species of drosophila and glycine max. The specific steps of the experiment are to use RepeatMasker to acquire the two species of drosophila and glycine max, respectively, and save the duplicate copies with high consistency, and saved the duplicate copies with differences as much as possible, and at the same time analyzed the structural variation that occurred in the duplicate copies with differences. In addition, we have also evaluated the performance of the newly added three steps of impurity removal, chimerism removal and consensus sequence construction in the pipeline of the improved LongRepMarker.

The evaluation results are shown in Table S6 and Figures S9-S10.

| Repeat Types | Num of bases masked: 3770909 bp (52.22%) | Num of bases masked: 3491131 bp (48.35%) | Num of bases masked: 3336440 bp (46.21%) |
|--------------|---------------------------------|---------------------------------|---------------------------------|
| Simple repeats | 18 7662 bp 0.11% | 0 0bp 0.00% | 0 0bp 0.00% |
| Satellites | 0 0bp 0.00% | 0 0bp 0.00% | 0 0bp 0.00% |
| LTR elements | 1153 75906 bp 1.05% | 1172 76644 bp 1.06% | 1224 80026 bp 1.11% |
| SINEs | 0 0bp 0.00% | 0 0bp 0.00% | 0 0bp 0.00% |
| LINEs | 227 144757 bp 2.00% | 0 0bp 0.00% | 0 0bp 0.00% |
| En-Spm | 1021 1649979 bp 22.85% | 0 0bp 0.00% | 0 0bp 0.00% |

Table S4. Comparison of the proportion of the detection results produced by the improved LongRepMarker and other two tools covering Drosophila RepBase library.
and chimeras into short repeats with higher accuracy. Added processing steps can reduce redundancy to a great extent, while cutting long fragments containing impurities after the above steps. Comparing the alignments shown in the first and second red boxes, we can see that the newly chimeras, and construction of consensus sequences. The second red box shows the alignment of repeated fragments after the above steps. Comparing the alignments shown in the first and second red boxes, we can see that the newly added processing steps can reduce redundancy to a great extent, while cutting long fragments containing impurities and chimeras into short repeats with higher accuracy.

Table S5. Comparison of the performance of the detection results produced by the improved LongRepMarker and other two tools covering Glycine max RepBase library.

| Repeat Types | Improved LongRepMarker | RepeatScout | RepeatModeler2 |
|--------------|------------------------|-------------|----------------|
| Total length: 1642293bp bases masked: 1632175 bp (99.45%) | Total length: 1642293bp bases masked: 1591794 bp (95.28%) | Total length: 1642293bp bases masked: 1592522 bp (98.56%) |
| Repeats | Number | Percentage | Number | Percentage | Number | Percentage |
|--------------|--------|-----------|--------|-----------|--------|-----------|
| ALUs:       | 0      | 0.00%     | 0      | 0.00%     | 0      | 0.00%     |
| ERLE/PELIE: | 0      | 0.00%     | 0      | 0.00%     | 0      | 0.00%     |
| LINE/SLATE: | 0      | 0.00%     | 0      | 0.00%     | 0      | 0.00%     |
| LTR:        | 512    | 78.82%    | 1030   | 75.23%    | 815    | 67.69%    |
| Satellite:  | 0      | 0.00%     | 0      | 0.00%     | 2      | 0.01%     |
| Simple repeats: | 181   | 1.84%    | 215   | 1.96%    | 255   | 2.05%     |
| DNA elements: | 0      | 0.00%     | 0      | 0.00%     | 0      | 0.00%     |
| Other class: | 0      | 0.00%     | 0      | 0.00%     | 0      | 0.00%     |
| Total interspersed repeats: | 2789 | 2.69%   | 3530 | 2.75%   | 3753 | 2.95%     |
| Small RNA:  | 27     | 0.45%     | 26     | 0.42%     | 2      | 0.32%     |
| Simple repeats: | 1     | 0.03%    | 0      | 0.00%     | 3      | 0.02%     |
| Satellites: | 0      | 0.00%     | 0      | 0.00%     | 2      | 0.01%     |
| Other (Misc): | 0       | 0.00%  | 0      | 0.00%     | 0      | 0.00%     |
| Consensus sequence: | 758    | 0.00%   | 758   | 0.00%   | 758   | 0.00%     |

Table S6. The performance evaluation of the three steps of removing impurities, removing chimerism and constructing a consensus sequence.

| Step | Improved LongRepMarker | RepeatScout | RepeatModeler2 |
|------|------------------------|-------------|----------------|
| Length occupied of sequence | Num of bases masked | Percentage | Num of bases masked | Percentage | Num of bases masked | Percentage |
| Improved LongRepMarker | 1,060 | 515,930bp | 30.49% | 0 | 0bp | 0.00% | 0 | 0bp | 0.00% |
| RepeatScout | 218 | 101,835bp | 47.50% | 0 | 0bp | 0.00% | 0 | 0bp | 0.00% |
| RepeatModeler2 | 215 | 101,335bp | 47.49% | 0 | 0bp | 0.00% | 0 | 0bp | 0.00% |

Fig. S9. The first effect demonstration of removing impurities, removing chimerism and constructing common sequences. Comparison of sequences alignment before and after removal of impurities and chimeras, and construction of consensus sequences. The first red box shows the alignment of repeated fragments without removal of impurities and chimeras, and construction of consensus sequences. The second red box shows the alignment of repeated fragments after the above steps. Comparing the alignments shown in the first and second red boxes, we can see that the newly added processing steps can reduce redundancy to a great extent, while cutting long fragments containing impurities and chimeras into short repeats with higher accuracy.

Fig. S10. The second effect demonstration of removing impurities, removing chimerism and constructing common sequences. Comparison of sequences alignment before and after removal of impurities and chimeras, and construction of consensus sequences. The first red box shows the alignment of repeated fragments without removal of impurities and chimeras, and construction of consensus sequences. The second red box shows the alignment of repeated fragments after the above steps. Comparing the alignments shown in the first and second red boxes, we can see that the newly added processing steps can reduce redundancy to a great extent, while cutting long fragments containing impurities and chimeras into short repeats with higher accuracy.
We have conducted various experimental evaluations on the comprehensiveness of the msRepDB database. For example, we used the latest version of RepeatMasker (V.4.1.2) to classify and annotate the repeats of genomes Human, Mouse, Drosophila, Rice and Glycine max based on msRepDB and the combination of the latest RepBase (V.26.06) and Dfam (V.3.3).

The frequency and length distribution, the multiple alignment ratio, the proportion of coverage over the reference genome and the duplication ratio of the repetitive sequences contained in msRepDB and the combination of Dfam and RepBase databases are shown in Table S7. We can see that the repetitive sequences collected in the msRepDB database have a higher repetition frequency and larger size as a whole. Furthermore, from the perspective of multiple alignment ratio, coverage of the reference genome, and duplication ratio, the repetitive sequences contained in msRepDB are usually more accurate and less redundant than those contained in the combination of Dfam and RepBase databases. Here, the duplication ratio represents the total number of aligned bases in the repetitive sequences divided by the total number of those in the reference. If there are too many repetitive sequences that cover the same regions, the duplication ratio will be greatly increased. This occurs due to multiple reasons, including overestimating repeat multiplicities and overlaps between repetitive sequences.

The experimental results in Tables S8, S9, S10, S11, S12 and Figure S11 show that RepeatMasker annotated 3,852,568 Retroelements-type repeats (1,291,793,390bp in length) on the human genome based on msRepDB, as compared to 2,800,814 Retroelements-type repeats (1,236,215,277bp in length) for the combination of the two other databases (Table S11), and annotated 3,443,145 DNA-transposons-type repeats (42,789,484bp in length) on the rice genome based on msRepDB, as compared to 241,722 DNA-transposons-type repeats (41,514,301bp in length) for the combination of the two other databases (Table S10), and annotated 241,722 DNA-transposons-type repeats (69,072,660bp in length) on the rice genome based on msRepDB, as compared to 241,722 DNA-transposons-type repeats (68,736,938bp in length) for the combination of the two other databases (Table S9), and annotated 61,139 DNA-transposons-type repeats (42,789,484bp in length) on the glycine max genome based on msRepDB, as compared to 58,468 DNA-transposons-type repeats (41,514,301bp in length) for the combination of the two other databases (Table S8).

For example, we used the latest version of RepeatMasker (V.4.1.2) to classify and annotate the repeats of five genomes Human, Mouse, Drosophila, Rice and Glycine max based on msRepDB and the combination of Dfam and RepBase databases. Here, the duplication ratio represents the number of repetitive sequences contained in msRepDB have a higher repetition frequency and larger size as a whole.

| Species          | Database          | Num    | Max    | Non-MAR | MAR | MAR(%) | Non-MAR(%) | Duplication ratio | Reference(%)
|------------------|-------------------|--------|--------|---------|-----|--------|------------|-------------------|----------------
| Human            | msRepDB           | 1,743  | 2,018  | 1,268   | 449 | 33.85% | 15.69%     | 31.11%            | 90.05%
| Human            | Dfam+RepBase      | 1,363  | 1,540  | 964     | 380 | 70.31% | 29.69%     | 45.92%            | 71.11%
| Mouse            | msRepDB           | 1,340  | 1,521  | 974     | 346 | 72.17% | 27.83%     | 43.14%            | 66.86%
| Mouse            | Dfam+RepBase      | 1,133  | 1,396  | 726     | 317 | 68.73% | 31.27%     | 42.14%            | 87.86%
| D.melanogaster   | msRepDB           | 477    | 2,014  | 2,571   | 115 | 96.95% | 3.05%      | 21.86%            | 2.40%
| D.melanogaster   | Dfam+RepBase      | 258    | 1,557  | 1,036   | 103 | 89.77% | 10.22%     | 43.90%            | 5.11%
| Glycine max      | msRepDB           | 1,226  | 2,050  | 2,210   | 116 | 99.65% | 0.35%      | 100.00%           | 94.41%
| Glycine max      | Dfam+RepBase      | 91     | 1,489  | 839     | 99  | 99.15% | 0.85%      | 100.00%           | 95.68%
Table S9. Partial comparison of the proportion and detailed classification of detected repeats generated based on two databases of the Drosophila genome.

| Repeat Type | Number of elements | Length occupied | Percentage of sequences |
|-------------|--------------------|-----------------|-------------------------|
| Simple repeats | 88,676 | 3,867,177bp | 1.03% |
| Satellites | 426 | 1,368,174bp | 0.37% |
| Small RNA | 4,631 | 704,192bp | 0.19% |
| Total interspersed repeats | 165,842,275bp | 44.29% |
| ++Gypsy/DTRS1 | 32,899 | 73,328,202bp | 19.58% |
| ++BEL/Pao | 0 | 0bp | 0.00% |
| +LINEs | 11,557 | 5,568,202bp | 1.49% |
| +SINEs | 6,826 | 987,304bp | 0.26% |
| Retroelements | 65,791 | 95,531,185bp | 25.51% |
| ++Gypsy/DTRS1 | 126,690 | 195,309,037bp | 19.95% |
| ++BEL/Pao | 2,326 | 3,123,105bp | 2.17% |
| +LINEs | 5,293 | 5,447,560bp | 4.49% |
| +SINEs | 0 | 0bp | 0.00% |
| Retroelements | 241,722 | 68,736,938bp | 18.36% |
| ++Gypsy/DTRS1 | 140,926 | 225,546,399bp | 23.04% |
| ++BEL/Pao | 2,937 | 3,118,973bp | 2.17% |
| +LINEs | 6,134 | 6,416,652bp | 4.46% |
| +SINEs | 0 | 0bp | 0.00% |

The test results were obtained by using RepeatMasker based on the msRepDB database and the combination of Dfam and RepBase respectively under the default parameter settings.

Table S10. Partial comparison of the proportion and detailed classification of detected repeats generated based on two databases of the Glycine max genome.

| Repeat Type | Number of elements | Length occupied | Percentage of sequences |
|-------------|--------------------|-----------------|-------------------------|
| Simple repeats | 88,676 | 3,867,177bp | 1.03% |
| Satellites | 1,372 | 1,804,199bp | 1.26% |
| Total interspersed repeats | 22,997,746bp | 16.00% |
| ++Gypsy/DTRS1 | 7,211 | 10,737,388bp | 7.47% |
| ++BEL/Pao | 2,326 | 3,123,105bp | 2.17% |
| +LINEs | 5,293 | 5,447,560bp | 4.49% |
| +SINEs | 0 | 0bp | 0.00% |
| Retroelements | 56,791 | 95,531,185bp | 25.51% |
| ++Gypsy/DTRS1 | 13,243 | 12,190,939bp | 8.48% |
| ++BEL/Pao | 2,937 | 3,118,973bp | 2.17% |
| +LINEs | 6,134 | 6,416,652bp | 4.46% |
| +SINEs | 0 | 0bp | 0.00% |

The test results were obtained by using RepeatMasker based on the msRepDB database and the combination of Dfam and RepBase respectively under the default parameter settings.

Table S11. Partial comparison of the proportion and detailed classification of detected repeats generated based on two databases of the Rice genome.

| Repeat Type | Number of elements | Length occupied | Percentage of sequences |
|-------------|--------------------|-----------------|-------------------------|
| Simple repeats | 82,139 | 4,344,053bp | 0.44% |
| Satellites | 1,372 | 1,804,199bp | 1.26% |
| Total interspersed repeats | 22,997,746bp | 16.00% |
| ++Gypsy/DTRS1 | 7,211 | 10,737,388bp | 7.47% |
| ++BEL/Pao | 2,326 | 3,123,105bp | 2.17% |
| +LINEs | 5,293 | 5,447,560bp | 4.49% |
| +SINEs | 0 | 0bp | 0.00% |
| Retroelements | 56,791 | 95,531,185bp | 25.51% |
| ++Gypsy/DTRS1 | 13,243 | 12,190,939bp | 8.48% |
| ++BEL/Pao | 2,937 | 3,118,973bp | 2.17% |
| +LINEs | 6,134 | 6,416,652bp | 4.46% |
| +SINEs | 0 | 0bp | 0.00% |

The test results were obtained by using RepeatMasker based on the msRepDB database and the combination of Dfam and RepBase respectively under the default parameter settings.

Table S12. Partial comparison of the proportion and detailed classification of detected repeats generated based on two databases of the Mouse genome.

| Repeat Type | Number of elements | Length occupied | Percentage of sequences |
|-------------|--------------------|-----------------|-------------------------|
| Simple repeats | 89 | 567,100bp | 0.26% |
| Satellites | 1,214 | 2,404,268bp | 0.98% |
| Total interspersed repeats | 22,997,746bp | 16.00% |
| ++Gypsy/DTRS1 | 7,211 | 10,737,388bp | 7.47% |
| ++BEL/Pao | 2,326 | 3,123,105bp | 2.17% |
| +LINEs | 5,293 | 5,447,560bp | 4.49% |
| +SINEs | 0 | 0bp | 0.00% |
| Retroelements | 56,791 | 95,531,185bp | 25.51% |
| ++Gypsy/DTRS1 | 13,243 | 12,190,939bp | 8.48% |
| ++BEL/Pao | 2,937 | 3,118,973bp | 2.17% |
| +LINEs | 6,134 | 6,416,652bp | 4.46% |
| +SINEs | 0 | 0bp | 0.00% |

The test results were obtained by using RepeatMasker based on the msRepDB database and the combination of Dfam and RepBase respectively under the default parameter settings.
The frequency of repetitive sequence in genome

There are several screenshots of the unique repetitive sequences of the human genome contained in msRepDB and the combination of Dfam and RepBase. There are some visual examples to illustrate the advantages of msRepDB database in terms of completeness of the repeated sequences (Figs S12-S26). These visualization examples are obtained by IGV (Integrative Genomics Viewer, https://software.broadinstitute.org/software/igv/) based on the genomes of human, mouse, rice, drosophila, and glycine max.

1) Human genome

There are several screenshots of the unique repetitive sequences of the human genome contained in msRepDB database (Figs S12-S14). The unique sequence means that the sequences only exists in msRepDB database.
Fig. S12. The first screenshot of the unique repetitive sequence of the human genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.

Fig. S13. The second screenshot of the unique repetitive sequence of the human genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.

Fig. S14. The third screenshot of the unique repetitive sequence of the human genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.
There are several screenshots of the unique repetitive sequences of the drosophila genome contained in msRepDB database (Figs S15-S17). The unique sequence means that the sequences only exists in msRepDB database.

Fig. S15. The first screenshot of the unique repetitive sequence of the drosophila genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.

Fig. S16. The second screenshot of the unique repetitive sequence of the drosophila genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.

Fig. S17. The third screenshot of the unique repetitive sequence of the drosophila genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.
3) Mouse genome

There are several screenshots of the unique repetitive sequences of the mouse genome contained in msRepDB database (Figs S18-S20). The unique sequence means that the sequences only exists in msRepDB database.

**Fig. S18.** The first screenshot of the unique repetitive sequence of the mouse genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.

**Fig. S19.** The second screenshot of the unique repetitive sequence of the mouse genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.

**Fig. S20.** The third screenshot of the unique repetitive sequence of the mouse genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.
4) Rice genome

There are several screenshots of the unique repetitive sequences of the rice genome contained in msRepDB database (Figs S21-S23). The unique sequence means that the sequences only exists in msRepDB database.

Fig. S21. The first screenshot of the unique repetitive sequence of the rice genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.

Fig. S22. The second screenshot of the unique repetitive sequence of the rice genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.

Fig. S23. The third screenshot of the unique repetitive sequence of the rice genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.
5) Glycine max genome

There are several screenshots of the unique repetitive sequences of the glycine max genome contained in msRepDB database (Figs S29-S31). The unique sequence means that the sequences only exists in msRepDB database.

Fig. S24. The first screenshot of the unique repetitive sequence of the glycine max genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.

Fig. S25. The second screenshot of the unique repetitive sequence of the glycine max genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.

Fig. S26. The third screenshot of the unique repetitive sequence of the glycine max genome contained in msRepDB database. The unique sequence means that the sequence only exists in msRepDB database.
7.2.2 The false positive evaluation of the detection results of LongRepMarker, RepeatScout and improved LongRepMarker

Following the respective reviewer’s previous suggestion, we conducted the experiments on the simulated sequencing data for Drosophila, and then we used RepeatMasker to annotate the repeats, and used the annotated set as the ground-truth set to compare with the annotation from RepeatScout and from LongRepMarker. The sequencing read simulator used in the study of PMID: 31890048 was ART (https://www.niehs.nih.gov/research/resources/software/bioinformatics/art/index.cfm), the aligners used in this study were bowtie2 and minimap2, and masking tool used in this study was RepeatMasker. In order to follow reviewer’s comments and evaluate the false positive rate of detection results generated by LongRepMarker, we have designed the following experiment.

1) Taking the Drosophila reference genome as input, and use the ART tool to generate Illumina short paired-end reads with the average coverage of 100 folds (-f 100). The average length of the reads is set to 100bp (-l 100), the average insert sizes of the paired-end reads is set to 350bp (-m 350), the standard deviation of insert sizes is set to ±30 (-o ±30), and the error rate is set to default (about 1%).

Command:
```
./art_illumina -ss HS25 -sam -i /data/Ref/Drosophila_melanogaster.BDGP6.32.dna.toplevel.fa -p -1 100 -l 100 -n 350 -s ±30 -o /data/Ref/Drosophila/read1
```

2) Taking the Illumina short paired-end reads generated in the first step as input, and use the de novo mode of LongRepMarker to obtain the repetitive sequences (Figure 9).

Command:
```
java LongRepMarker -k 49 -t 48 -q1 /data/Ref/Drosophila/read11.fq -q2 /data/Ref/Drosophila/read12.fq -o /data/Ref/Drosophila/LongRepMarker/
```

3) Taking the Drosophila reference genome as input, and use the RepeatMasker tool to get the annotation of the repetitive sequences.

Command:
```
./RepeatMasker -parallel 30 -lib /data/Ref/Drosophila/LongRepMarker/RepeatLibDrosophila.fa -html -gff -dir /data/Ref/Drosophila/RepeatMasker/ -species Drosophila
```

4) Extracting the masked repetitive sequences from the annotation results of RepeatMasker

5) Mapping the repetitive sequences obtained by the de novo mode of LongRepMarker to the masked repetitive sequences extracted in step 4).

Commands:
```
./minimap2 -d /data/Ref/Drosophila/LongRepMarker/masked.mmi /data/Ref/Drosophila/RepeatMasker/RepeatMasker-masked.fa
./minimap2 -a /data/Ref/Drosophila/LongRepMarker/RepeatLibDrosophila.fa > /data/Ref/Drosophila/LongRepMarker/masked.sam
./samtools fasta -f 4 -0 /data/Ref/Drosophila/LongRepMarker/unmapped.fa /data/Ref/Drosophila/LongRepMarker/masked.sam
./bowtie2-build /data/Ref/Drosophila/RepeatMasker/RepeatMasker-masked.fa /data/Ref/Drosophila/RepeatMasker/RepeatMasker-masked

./bowtie2 -x /data/Ref/Drosophila/LongRepMarker/RepeatMasker-masked -f -a -p 32 -U /data/Ref/Drosophila/LongRepMarker/unmapped.fa -S /data/Ref/Drosophila/LongRepMarker/RepeatScout-drosophila.sam
```

6) Generating the new detection results based on the detected fragments produced by the original LongRepMarker and the three steps of impurity removal, chimeras removal and consensus sequence construction.

7) Counting the false positive rates of the detection results produced by the original LongRepMarker, the detection results produced by RepeatScout and the new detection results generated based on the detected fragments produced by the original LongRepMarker and the three steps of impurity removal, chimeras removal and consensus sequence construction, respectively.

8) Results and analysis:

The improved LongRepMarker means that the three steps of impurity removal, chimeras removal and consensus sequence construction have been added to the pipeline of the original LongRepMarker.

In this experiment, using the masked sequences extracted from the results of RepeatMasker as the standard, the false positive rates of the results produced by the three tools were measured. There are 116,492 fragments extracted from the masking results of RepeatMasker based on the Drosophila reference genome. For the original LongRepMarker, its detection results contain 76,575 fragments, of which 59,783 fragments can be aligned to the sequences extracted from the masking results. Therefore, the false positive rate of the detection results produced by the original LongRepMarker is 21.92%. For the RepeatScout, its detection results contain 2,767 fragments, of which 1,897 fragments can be aligned to the sequences extracted from the masking results. Therefore, the false positive rate of the detection results produced by the RepeatScout is 31.44%. For the improved LongRepMarker, its detection results contain 556 fragments, of which 464 fragments can be aligned to the sequences extracted from the masking results. Therefore, the false positive rate of the detection results produced by the improved RepeatScout is 16.54% (Table S13).
Table S13. The false positive evaluation of the detection results of LongRepMarker, RepeatScout and improved LongRepMarker.

| Method          | Num   | F1 Score | Recall | Precision | ACC    |
|-----------------|-------|----------|--------|-----------|--------|
| LongRepMarker   | 556   | 0.46     | 0.59   | 0.62      | 0.57   |
| RepeatScout     | 76,575| 0.59     | 0.72   | 0.70      | 0.64   |
| Improved LongRepMarker | 463 | 0.65     | 0.78   | 0.76      | 0.72   |

In the experiment, the tool RepeatMasker was used only once, and the library it used was only RepeatBase, not the results produced by RepeatScout and LongRepMarker. All the false positives are counted by comparing the ground-truth set of annotations with that of RepeatScout or LongRepMarker. Therefore, the length-bias issue in RepeatMasker will not affect the evaluation.

References

1. Kazazian, H.H. Jr. (2004) Mobile elements: drivers of genome evolution. Science, 303, 1626-1632.
2. Treangen,T.J. and Salzberg, S. L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature reviews. Genetics, 13, 36-46.
3. Lu,Q., Walrath,L.L., Granok,H. and Elgin,S.C. (1993) (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, 13, 2802-2814.
4. Kundu,T.K. and Rao,M.R. (1999) CpG islands in chromatin organization and gene expression. J. Biochem, 125, 217-222.
5. Shapiro J.A. and von Sternberg R. (2005) Why repetitive DNA is essential to genome function. Biol. Rev., 80, 227-50.
6. Kaltenegger,E., Leng,S. and HeylA. (2018) The effects of repeated whole genome duplication events on the evolution of cytokinin signaling pathway. BMC Evol. Biol., 18, 76-95.
7. Aguilera,A. and García-Muse T. (2013) Causes of genome instability. Annu. Rev. Genet., 47,1-32.
8. George,C.M. and Alani, E. (2012) Multiple cellular mechanisms prevent chromosomal rearrangements involving repetitive DNA. Current reviews in biochemistry and molecular biology, 47,297-313.
9. Hall,A.C., Ostrowski,L.A., Pietrobon,V. and Mekhail,K. (2017) Repetitive DNA loci and their modulation by the non-canonical nucleic acid structures R-loops and G-quadruplexes. Nature, 8, 162-181.
10. Bao,W., Kojima,K.K. and Kohany,O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA, 6, 11-17.
11. Storer, J., Hubley, R., Rosen, J. et al. (2021) The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA, 12, 2-16.
12. Schmutz J., Cannon S., Schlueter J., Ma J., Mitros T., Nelson W., Hyten D.L., Song Q., Thelen J.J., Cheng J. (2010) Genome sequence of the palaeopolyploid soybean. Nature, 463, 178–183.
13. Cox R. and S. M. Mirkin. (1997) Characteristic enrichment of DNA repeats in different genomes. Proceedings of the National Academy of Sciences of the United States of America, 94, 5237-5242.
14. J. Yuyang Lu, Wen Shao, Lei Chang, Yanfei Yin, Tong Li, Hui Zhang, Yantao Hong, Michelle Perchaude, Lerui Gao. Zhangwu Yang et al. (2020) Genomic Repetes Categorize Genes with Distinct Functions for Orchestration. Regulation. Cell Reports, 30, 3296-3311.e5.
15. Ahmad S. F., Singchat W., Jehangir M., Suntoppong A., Panthum T., Malaviyontinond S. and Srikulnath K. (2020) Dark Matter of Primate Genomes: Satellite DNA Repeats and Their Evolutionary Dynamics. Cells, 9, 2714.
16. Shapiro J.A. and von Sternberg R. (2005) Why repetitive DNA is essential to genome function. Biol. Rev., 80, 227-250.
17. Kaltenegger E., Leng S. and Heyl A. (2018) The effects of repeated whole genome duplication events on the evolution of cytokinin signaling pathway. BMC Evol. Biol., 18, 76-95.
18. Hall,A.C., Ostrowski,L.A., Pietrobon,V. and Mekhail,K. (2017) Repetitive DNA loci and their modulation by the non-canonical nucleic acid structures R-loops and G-quadruplexes. Nature, 8, 162-181.
19. Shweta Mehrotra and Vinod Goyal. (2014) Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. Genomics, Proteomics & Bioinformatics, 12, 164-171.
20. Hall A.C., Ostrowski L.A., Pietrobon V. and Mekhail,K. (2017) Repetitive DNA loci and their modulation by the non-canonical nucleic acid structures R-loops and G-quadruplexes. Nature, 8, 162-181.
21. Shweta Mehrotra and Vinod Goyal. (2014) Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. Genomics, Proteomics & Bioinformatics, 12, 164-171.
22. Mariely DeJesus-Hernandez, Ian R. Mackenzie, Bradley F. Boeve, Adam L. Boxer, Matt Baker, Nicola J. Rutherford, Alexandra M. Nicholson, Nicole A. Finch, Heather Flynn, Jennifer Adamson et al. (2011) Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron, 72, 245-258.
23. Prokopy, Zhongyang Wu, Guo, Dov A. et al. (2020) Genome-wide detection of tandem DNA repeats that are expanded in autism. Nature, 586, 80-95.
24. Belancio V. P., Roy-Engel A. M. and Deininger P. L. (2010) All y’all need to know ‘bout retroelements in cancer. Front. Cell Dev. Biol., 8, 246–250.
25. Young Bun Kim, Jung Hun Oh, Lauren J. et al. (2014) Divergence of Drosophila repeatomes in different genomes. Mobile DNA, 6, 3296-3311.e5.
26. Chenais B. (2013) Transposable elements and human cancer: a causal relationship? Biochim. Biophys. Acta, 1835, 28-35.
27. Mariely DeJesus-Hernandez, Ian R. Mackenzie, Bradley F. Boeve, Adam L. Boxer, Matt Baker, Nicola J. Rutherford, Alexandra M. Nicholson, Nicole A. Finch, Heather Flynn, Jennifer Adamson et al. (2011) Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron, 72, 245–258.
28. Trost B., Englandh W., Nguyen C.M., Thiruvahindrapuram B., Dolzenko E., Backstrom L., Mirceta M., Mojarad B. A., Y. Yin, Dov A. et al. (2020) Genome-wide detection of tandem DNA repeats that are expanded in autism. Nature, 586, 80-95.
29. Citron, Journal of Molecular Evolution, 2010, 1-17.
30. Belancio V. P., Roy-Engel A. M. and Deininger P. L. (2010) All y’all need to know ‘bout retroelements in cancer. Semin. Cancer Biol., 20, 200-210.
31. Young Bun Kim, Jung Hun Oh, Lauren J. et al. (2014) Divergence of Drosophila melanogaster repeat expansion in C9ORF72 is the cause of chromosome 9p-linked FTD. Neuron, 72, 245-258.
32. Nature, 586, 80-95.
33. Zhang X., Zhang R. and Yu J. (2020) New Understanding of the Relevant Role of LINE-1 Retrotransposition in Human Disease and Immune Modulation. Front. Cell Dev. Biol., 8, 246–250.
34. Solomon S., Ewing A.D., Rahmnaan E.P., Doucet T., Nelson H.H. Burns M.B, Harris R.S., Sigmou D.F., Casella A., Erlanger B. et al. (2012) Extensive somatic Li retrotransposition in colorectal tumors. Genome Res, 22, 2328-2338.
35. Scott E.C., Gardner E.J., Masaood A., Chuaung N.T., Vertino P.M., Devine S.E. (2016) A hot Li retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res, 26, 745-755.

*Num* represents the number of fragments contained in detection results.
36. Xingyu Liao, Min Li, Kang Hu, Fang-Xiang Wu, Xin Gao and Jianxin Wang. (2021) A sensitive repeat identification framework based on short and long reads. Nucleic Acids Research, gkab563.
37. Price A.L., Jones N.C., Pevzner P.A. (2005) De novo identification of repeat families in large genomes. Bioinformatics, 21, i351-i358.
38. Jullien M. Flynn, Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte and Arian F. Smit. (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences, 117, 9451-9457.
39. Page A.J., Taylor B., Delaney A.J., Soares J., Seemann T., Keane J.A. and Harris S.R. (2016) SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom, 2, e000056.
40. Bao Z. and Eddy SR. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res, 12, 1269-1276.
41. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25, 1754-60.
42. Li Heng. (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34, 3094-3100.