大规模基因组重复序列识别与分类研究进展

葛瑞泉[1,2] 王 普[1,3] 李 烨[1,4] 蔡云鹏[1,4,5]

[1](中国科学院深圳先进技术研究院 深圳 518055)
[2](杭州电子科技大学计算机学院 杭州 310018)
[3](景德镇陶瓷大学机械电子工程学院 景德镇 333403)
[4](中国科学院健康信息学重点实验室 深圳 518055)
[5](深圳市健康大数据分析技术工程实验室 深圳 518055)

摘 要 重复序列在基因组中普遍存在，大量实验证实其在生物进化过程中起着重要作用。目前，重复序列的发现与识别技术已经成为基因组学的研究热点，文章分类总结了有关这方面的研究进展，并对相关工具的功能特点进行了简要分析，同时对重复序列发展趋势进行了总结和展望。

关键词 重复序列；转座子；长末端重复序列
中图分类号 TP 391 **文献标志码** A

Research Progress on Large Scale Repeat Identification and Classification in Genomes

GE Ruiquan[1,2] WANG Pu[1,3] LI Ye[1,4] CAI Yunpeng[1,4,5]

[1](*Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[2](*School of Computer Science and Technology*, *Hangzhou Dianzi University*, *Hangzhou* 310018, *China*)
[3](*School of Mechanical and Electronic Engineering*, *Jingdezhen Ceramic Institute*, *Jingdezhen* 333403, *China*)
[4](*Key Lab for Health Informatics*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[5](*Shenzhen Engineering Lab for Health Big Data Analysis*, *Shenzhen* 518055, *China*)

Abstract Repetitive sequences are prevalent in genomes. A large number of experiments have confirmed that they play an important role in biological evolution. At present, the discovery and detection of the repetitive sequences have been becoming a hot topic of genomics. This paper summarizes the research progress in this regard, and briefly analyses the associated tools. Finally, the development of repetitive sequences in future is prospected.

Keywords repetitive sequence; transposons; long terminal repeats

收稿日期：2017-01-19 **修回日期**：2017-06-01
基金项目：广东省应用型研发专项(2015B010129012)；中国科学院科技服务网络项目(KFJ-SW-STS-161)；杭州电子科技大学启动基金(ZX160203305002/011)
作者简介：葛瑞泉，博士，讲师，研究方向为数据挖掘与生物信息学；王普，博士，讲师，研究方向为模式识别与生物信息学；李烨（通讯作者），研究员，研究方向为健康信息学，E-mail：ye.li@siat.ac.cn；蔡云鹏（通讯作者），研究员，研究方向为生物信息学与机器学习，E-mail：yp.cai@siat.ac.cn。

1 引 言

实验已经证实，重复序列在基因组中普遍存在，尤其是在真核生物中。重复序列片段在生物进化、遗传调控和基因表达等方面起着重要作用。重复可以由多种机理引起，如染色体易位、转座子、复制与重组时出错等[1]。

如何快速准确地识别基因中的重复序列已经成为生物信息学的一个重要研究课题。目前已经有很多这方面的研究成果。本文分析了一些常见的查找重复序列的工具软件，旨在为从事这方面研究的人员提供一些参考。

2 研究内容概述

按照重复的排列方式，重复序列可以分为散在重复序列和串联重复序列两大类[2]。其中，散在重复序列是散在分布于基因组中的序列，一般为中度重复序列；串联重复序列是重复序列以各自的核心序列(重复单元)首尾相连多次重复组成的序列。外源基因岛通常也表现为重复元件，不过因其重组频繁，需要采取完全不同的算法进行检测[3]。表 1 为重复序列的细分类型。

一般来说，散在重复序列就是通常所说的转座子。它可以分为反转录转座子和 DNA 转座子。其中，反转录转座子的复制和移动是在 RNA 介导情况下，通过拷贝-粘贴机制在反转录酶的作用下完成的。而 DNA 转座子只通过剪切-粘贴或复制-粘贴机制完成，不需要 RNA 参与[4]。串联重复在编码区和非编码区都会存在。其中，在编码区有 rRNA、tRNA 基因与组蛋白基因，而非编码区在真核生物中广泛存在。卫星序列重复单位一般在 150～500 bp，位于染色体的异染色质区，其功能可能与细胞分裂时染色体的运动及基因组的稳定性有关[5]。根据串联重复序列的位置特征，可以寻找其前具有一定催化功能的序列，进而分析非编码区的功能结构特征。小卫星重复单位一般在 10～100 bp，分布于常染色质区。微卫星序列也叫短串联重复序列(Short Tandem Repeats)或简单重复序列(Simple Sequence Repeats)，重复单位一般在 2～10 bp，多分布于非编码区和内含子中，呈共显性遗传[6]。其作为遗传标记，用于群体遗传学、人类遗传图谱制作、筛选目的基因与基因诊断、法医鉴定等方面[7]。

简单重复序列进化机制有两种解释模型，分别为 UCO(Unequal Crossing-Over)模型和 SSM(Slip-Strand Mispairing)模型[8]。其中，UCO 模型认为简单重复序列内不等交换律升高。不等交换是由不配对同源染色体重组造成的，而重复片段增加了其可能性。SSM 模型认为 DNA 复制时滑动链错配频率提高。DNA 复制时，滑动后

表 1 重复序列分类

Table 1 Repetitive sequence classification

重复序列类型			长度（bp）	分布区域
散在重复序列	RNA 转座子	长末端重复序列（LTR）	100～5 000	反转录病毒两端
		长散布元件 (LINE)	500～4 000	散在分布
		短散布元件 (SINE)	＜500	散在分布
	DNA 转座子	微型反向重复转座元件（MITE）	＜500	细菌、植物和动物基因
		旋束管等	＜500	散在分布
串联重复序列	卫星（Satellite）		150～500	异染色质
	小卫星（Minisatellite）		10～100	常染色质
	微卫星（Microsatellite）		2～10	非编码区，内含子

容易生成发夹和茎环等结构，从而导致模板链或新生链重复序列减少或增加[9]。目前普遍认为后者是主要成因。

重复序列一般含有遗传调控信息，可促使核酸包装成各种高级结构，并使染色体异染色质化而关闭基因表达。其分布出现 power-law 关系，大部分重复序列只出现少数几次[10]。这是进化的必然结果，可以起到保护编码序列的作用，并产生进化的动力，形成新基因。

目前，在基因序列中自动准确识别并发现重复序列非常困难。这是因为重复序列可以比较灵活地复制、插入、删除、切割，序列会产生遗传变异和重排，容易造成重复序列的不一致，因此很难确定重复序列的边界并对其进行家族分类。

3　研究现状

由于基因在进化过程中存在缺失和插入等情况，及计算机科学中的字符串匹配问题，因此很难精确查找出基因组中的重复序列。目前研究重复序列类型的方法有很多种，表 2 为目前比较流行的一些查找重复序列的工具。

相似性查找一般可分为精确查找和启发式方法两类。例如，按照海明距离或编辑距离阈值确定的 REPuter 为精确查找重复方法，而串联重复发现(Tandem Repeats Finder，TRF)等工具大多基于统计分析的为启发式方法。目前转座元件查找有基于库的方法、基于特征的方法、从头计算的方法、管道程序和家族分类方法。此外，还有一些非转座元件方法和转座元素变异查找的分析工具。本论文通过这几个方面对目前的相关工具进行分析总结。

3.1　转座元件查找

3.1.1 基于库的方法

基于库识别重复序列的方法基本以某个数据库为标准进行同源搜索，找出重复序列并掩盖相同序列，其中大多工具以 RepeatMasker[11]为主要参考库。其搜索引擎可以用 CROSSMATCH、AB-BLAST 或 WU-BLAST。RepeatMasker 具有有较高的效率和搜索速度，可以发现低拷贝数量的家族，但只能搜索同源序列，不能产生新的元素。这类方法被认为是黄金准则，通常作为查找重复序列的第一步。

除了 RepeatMasker 外，还有几个基于同源搜索的工具，如 Censor[12]、TESeeker[13]、Greedier[14,15]和 T-lex[16]等。其中，Censor[12] 主要用 NCBI Blast 做同源搜索库。TESeeker[13]用 Tefam、NCBI 和 Repbase 做同源搜索库，使用迭代的多序列比对查找重复序列，其容易使用和配置，但不能识别非编码区 SINEs 和 MITEs，转座子分类也需人工审查。Greedier[14,15]则使用了贪婪算法和局部比对的方法，有效地解决了嵌入重复问题。T-lex[16]使用高通量测序数据，可以查找串联重复的侧翼区、non-LTR 和片段重复区。RTclass1[17]基于系统进化使用改进的邻接法进行 non-LTR 的分类识别。RetroSeq[18]对高通量测序数据分析发现转座子，它用配对末端序列检测转座元件，参考使用了 BreakDancer[19]工具。Winmaster[20]用 DUST[21]搜索低复杂度序列、TRF 搜索串联重复、全局重复用统计 Nmers($N \leqslant 15$)的方法，结果比 RepeatMasker、MaskerAid[14,15]具有一定的优势。另外，还存在不少重复序列数据库。如 SINEBase[22]，它包括了迄今为止的 SINE 家族系列，但不能识别库中不存在的 SINEs。此外，还有几个可以图形化显示的工具，如 PLOTREP、TinT 和 TE Displayer[23]。其中，PLOTREP[24]使用了 Censor 工具，可显示重复序列的变化区域，解决了片段重复问题；TinT[25]用矩阵显示重复序列进化史；TE Displayer 需要的输入较多，用了类似 PCR 实验查找转座子。

表 2 常见重复序列检测与分类工具

Table 2 The usual repetitive sequence detection and classification tools

分类		工具名称	网址
基于库的方法		RepeatMasker	http://www.repeatmasker.org
		Censor	http://www.girinst.org/censor/download.php
		TESeeker	http://repository.library.nd.edu/view/16/index.html
		MaskerAid; Greedier	无
		T-lex	http://petrov.stanford.edu/cgi-bin/Tlex_manual.html
		RTclass1	http://www.girinst.org/RTphylogeny/RTclass1/
		RetroSeq	https://github.com/tk2/RetroSeq
		winmaster	http://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/lxr/source/src/app/winmasker/
		SINEBase	http://sines.eimb.ru/
		TinT	http://www.bioinformatics.uni-muenster.de/tools/tint/index.hbi?lang=en&mscl=0&cscl=0
		PLOTREP	http://repeats.abc.hu/cgi-bin/plotrep.pl
		TE Displayer	http://labs.csb.utoronto.ca/yang/TE_Displayer/TE_Displayer_1.0.2/
基于特征的方法		LTR_STRUC	http://www.mcdonaldlab.biology.gatech.edu/ltr_struc.htm
		RetroTector	http://retrotector.neuro.uu.se/pub/queue.php?show=submit
		LTR_FINDER	http://tlife.fudan.edu.cn/ltr_finder
		LTRharvest	http://www.zbh.uni-hamburg.de/LTRharvest
		LTR_par	http://www.eecs.wsu.edu/Bananth/software.htm
		MGEScan-LTR	http://darwin.informatics.indiana.edu/cgi-bin/evolution/ltr.pl
		MASiVE	http://tools.bat.infspire.org/masive/#news
		RTAnalyzer	http://www.riboclub.org/cgi-bin/RTAnalyzer/index.pl
		TSDfinder	http://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/TSDfinder
		SINEDR, FINDMITE	无
		P-MITE	http://pmite.hzau.edu.cn/django/mite/
		MITE-Hunter	http://target.iplantcollaborative.org/mite_hunter.html
		MAK	http://perl.idmb.tamu.edu/mak.htm
		MUST	http://csbl1.bmb.uga.edu/ffzhou/MUST/
		TRANSPO	http://alggen.lsi.upc.es/recerca/search/transpo/transpo.html
		HelitronFinder	http://limei.montclair.edu/HF.html
		IRF	http://tandem.bu.edu/irf/irf.download.html
		ISA-System	http://csbl.bmb.uga.edu/~ffzhou/isa_web/
从头计算的方法	自身比较法	Repeat Pattern Toolkit	无
		RECON	http://selab.janelia.org/recon.html
		PILER	http://www.drive5.com/piler/
		LTRdigest	http://www.zbh.uni-hamburg.de/?id=207
		popoolationte	https://code.google.com/p/popoolationte/
		Adplot	无
		BLASTER suite	https://urgi.versailles.inra.fr/Tools/Blaster
	K-mer 和空位种子法	RepeatScout	http://repeatscout.bioprojects.org/
		ReAS	ftp://ftp.genomics.org.cn/pub/ReAS/software/
		REPuter	http://bibiserv2.cebitec.uni-bielefeld.de/reputer
		RepSeek	http://wwwabi.snv.jussieu.fr/public/RepSeek/

续表 2

分类		工具名称	网址
从头计算的方法	K-mer 和空位种子法	Repeat-match	http://mummer.sourceforge.net/
		SMaRTFinder	http://services.appliedgenomics.org/software/smartfinder/
		Tallymer	http://www.zbh.uni-hamburg.de/Tallymer/
		Vmatch	http://www.vmatch.de/
		mer-engine	无
		FORRepeats	http://al.jalix.org/FORRepeats/
		P-Clouds	http://www.evolutionarygenomics.com/PClouds.html
	周期方法	SRF	http://www.imtech.res.in/raghava/srf/
	定义重复家族	RepeatFinder	http://cbcb.umd.edu/software/RepeatFinder/
		RepeatGluer	http://nbcr.sdsc.edu/euler/intro_tmp.htm
管道程序		RepeatModeler	http://www.repeatmasker.org/RepeatModeler.html
		RepeatRunner	http://www.yandell-lab.org/software/repeatrunner.html
		RepeatExplorer	http://galaxy.umbr.cas.cz:8080/
		REannotate	http://www.bioinformatics.org/reannotate/index.html
		ReRep,RetroPred	无
		RISCI	http://www.ccmb.res.in/rakeshmishra/tools/RISCI_Readme.htm
		Tea-TE analyzer	http://compbio.med.harvard.edu/Tea/
		ELAN	http://nldsps.jnu.ac.in/elan.html#pub
		DAWG-PAWS	http://dawgpaws.sourceforge.net/
分类方法		ModuleOrganizer	http://mobyle.genouest.org/cgi-bin/Mobyle/portal.py?#forms::moduleorganizer
		TCF	http://research.mssm.edu/warbup01/paper/files.html
		REPET	https://urgi.versailles.inra.fr/Tools/REPET
		Dfam	http://dfam.janelia.org/
		TEnest	http://www.plantgdb.org/tool/TE_nest/
		TARGeT	http://target.iplantcollaborative.org/
		RepClass	http://sourceforge.net/projects/repclass/
		TEclass	http://www.compgen.uni-muenster.de/teclass
		DomainOrganizer	www.irisa.fr/symbiose/DomainOrganizer/
		RetroMap	http://www.burchsite.com/bioi/RetroMapHome.html
		LTR_MINER	http://genomebiology.com/content/5/10/R79
非转座元件检测		Mreps	http://bioinfo.lifl.fr/mreps
		OMWSA	http://www.hy8.com/~tec/sw01/omwsa01.zip
		TRAP	http://www.coccidia.icb.usp.br/trap/
		TRF	http://tandem.bu.edu/trf/trf.html
		TROLL	http://finder.sourceforge.net

3.1.2 基于特征的方法

与基于库的方法不同，基于特征的方法一般根据序列结构特征来识别重复序列，用基于先验信息的启发式算法来查找分析。部分方法在处理时也会参照使用基于库的方法。这种方法可以发现新的重复元素，但不能发现新类型。其类型是根据掌握的特征多少来确定的。

早期研究比较多的是长末端重复序列（Long Terminal Repeats，LTR）识别方法，大多根据 LTR 序列的特征来识别重复。例如，LTR_STRUC[26]、RetroTector[27]、LTR_par[28]、LTR_FINDER[29]、LTRharvest[30]、MASiVE[31]

和 MGEScan-LTR[32]等。这些工具基本都是根据 LTR 结构特征，设计几个相关参数来查找重复。其中，LTR_STRUC 不允许修改参数，使用启发式种子和扩展策略查找重复，根据 LTR 各个子结构特征进行评分确定最终的 LTR，其结果存在丢失。RetroTector 除了序列比较外，增加了相邻 LTR 的空间关系分析。LTR_par 改进了 LTR_STRUC，使用后缀数组存储，序列比较用改进的动态规划代替贪婪的启发式算法。LTR_FINDER 基于 LTR_STRUC 和 LTR_par 原则建立了网页工具。LTRharvest 根据结构特征，实现了几步过滤，确定 LTR 的边界位置，使用 LTRdigest 对 LTR 进行注释。MASiVE 专门分析植物基因组中的特殊 LTR 转座子。MGEScan-LTR 使用了近似字符串匹配和蛋白质结构域分析方法，确定完好的 LTR 反转录转座子。实验证实，某些参数需要根据被测对象来确定才能产生比较好的结果，但这样就缺乏了灵活性，给用户使用带来一定难度。

此外，还有几个研究 non-LTR 反转录转座子和 MITE 的工具。如 RTAnalyzer[33]和 TSDfinder[34]用于检测长散布元件，SINEDR[35,36]用于检测已知的短散布元件。其中，RTAnalyzer 通过查找目标节点重复、3′端的聚腺苷酸和 5′端的酶切位点来识别长散布元件。TSDfinder 则通过查找目标节点重复、3′端的聚腺苷酸和相关的反转断点来识别长散布元件。另外，还有几个 MITE 的检测工具，如 P-MITE[37]、MITE-Hunter[38]、FINDMITE[35,36]、MAK[39]、MUST[40]和 TRANSPO[41]。其中，MITE 是 DNA 转座子的一种，一般小于 500 bp，通常根据终端反向重复和目标节点重复及其间距等结构特征来确定。P-MITE 是关于植物的 MITE 数据库，用 MITE Digger[42]和 MITE-Hunter 识别 MITE。MITE-Hunter 使用了管道机制，成对比较查找的转座序列，然后用多序列比对过滤所选模板的假阳性，更新模板后分组到不同家族。FINDMITE 用了满足一定间距的串匹配技术来查找 MITE。MAK 认为同一家族的 MITE 序列同源，可以用 BLAST 检索和预测新的 MITE 元素。MUST[40,43]搜索给定窗口中终端反向重复的所有出现，用序列比对方法分类，存在较高的假阳性。TRANSPO 在给定窗口大小下用近似串匹配算法搜索与已知 MITE 中相似的终端反向重复，该方法不能发现新的 MITE 元素。此外，还有 HelitronFinder[44]用于检测玉米中的螺线管元素，IRF[45]专门用于检测反向重复序列。反向重复无茎环大小限制，使用多次采样策略，对每个 K 元组匹配确定一个中心位置，然后检测有相同或相近中心的元组簇，反向重复用类似史密斯-沃特曼比对算法判定。

3.1.3 从头计算的方法

从头计算发现重复序列方法不必依赖已知重复元素，可以发现新的重复元素。随着高通量基因组测序的发展，这种方法已经得到广泛的应用。目前主要有两种方法：序列自身比对法和短序列重复出现搜索法。

使用自身比对方法的工具有：Repeat Pattern Toolkit[46,47]、RECON[48]、PILER[49]、LTRdigest[50]、popoolationte[51]、Adplot[46, 47]和 BLASTER suite[52]。其中，Repeat Pattern Toolkit 使用双序列比对相似性计分和单链聚类，图形化体现家族。RECON 是一个使用较多的从头检测重复序列的工具，它基于 BLAST 的非组装多序列段比对，用无向图表示单链聚类，可有效解决嵌套重复问题。PILER 使用长序列比对排序(Pairwise Alignment of Long Sequences，PALS)方法，可以发现和区分不同类型的重复，包括串联数组、散在家族、终端重复和假卫星。Popoolationte 用于分析转座插入位点，通过序列比较搜索转座子库，结合搜索 GO 数据库图形化显示基因注释。LTRdigest 使用局部比对与隐马

尔可夫链模型，根据序列特征识别家族，注释长末端重复序列反转录转座子的内部结构特征。Adplot 基于伯努利实验的概率统计方法可视化重复序列。BLASTERsuite 中除了 BLAST，还有 MATCHER 和 GROUPER 工具，分别映射匹配和聚类重复到相应的家族。

此外，还有部分工具基于 K 元组统计或其衍生的空位种子方法来检测重复序列。其中，基于 K 元组的方法在序列中查找长度较小的相同序列出现频率；而空位种子方法在 K 元组基础上允许有一定的变异，通过设定一定的相似比和长度变化来检测重复序列。这样的工具有 RepeatScout[53]、ReAS[54]、REPuter[55]、RepSeek[56]、Repeat-match[57]、SMaRTFinder[58]、Tallymer[59]、Vmatch[60]、mer-engine[61]、FORRepeats[62]和 P-Clouds[63]等。其中，RepeatScout 使用 K 元组检测方法，程序自动设置了 K 的默认大小，并过滤了低复杂度和串联重复序列，可以过滤小于一定数量的重复序列，使用贪婪搜索统计方法确定重复序列边界，比 RECON 具有更高的敏感性和速率。ReAS 处理未组装序列片段，使用不同的相似比限制的贪婪算法查找重复片段，使用 ClustalW[64]比对；为防止遗漏，通过减少序列片段长度和增加限制条件来优化查找重复序列。REPuter 包含 RepeatGluer 和 RepeatFinder，基于后缀树识别最长精确匹配，但不能识别存在分歧的重复。其中 RepeatFinder 也使用了 MUMmer[65]软件包的一部分。RepSeek 可查找任何类型重复，但无法识别家族。Repeat-match 可进行大序列片段的重复检测，是 MUMmer3.0 的一部分。SMaRTFinder 使用了后缀树搜索和近似匹配算法，但内存消耗严重，搜索速度较慢。Tallymer 使用了后缀数组，占用的空间仅为后缀树的 1/3～1/5，输出结果有助于基因组注释，但不能直接产生重复家族。Vmatch 工具除使用后缀数组存储，还对序列建立了索引，它有两种种子扩展策略：贪婪策略和允许最大错误匹配策略。其中 REPuter 使用了本工具。Mer-engine 用于查找精确匹配，使用了字符压缩技术，有助于基因组注释，但也不能直接产生重复家族。FORRepeats 使用基于 factor oracle 数据结构的启发式算法：首先查找完全匹配重复，然后通过两两比较找近似重复序列。P-Clouds 对出现频率较少的重复，创建类似重复的聚类，避免了序列比对和相似性搜索，对重复度高的区域使用滑动窗口机制。

另外，还有一些文献根据重复序列的特点研究其周期性和重复家族定义。其中 SRF[66]根据 DNA 序列频谱图分析发现重复，使用傅里叶变换可识别任意类型重复，但是有长度限制，每次变换的序列长度大于 10 kbp 时耗时较长。RepeatFinder[67]聚类定义重复家族，同时需要 REPuter 或 Repeatmatch 确定完全的重复。RepeatGluer[68]使用启发式方法，擅长重复边界的确定，基于 deBruijn 图发现子重复，用部分规则比对算法进行多序列比对。

3.1.4 管道程序

随着重复序列分析工具的增多，其优缺点逐渐被挖掘，随之出现许多综合使用各种工具的管道程序。例如，RepeatModeler[69]结合了 RECON 和 RepeatScout 软件，用于识别重复序列边界和家族关系。RepeatRunner[70]用 BLASTX 搜索编码蛋白数据库查找存在分叉的重复，弥补了 RepeatMasker 的缺陷，同时结合 PILER-DF 进行重复识别。RepeatExplorer[71]从头开始查找重复序列，基于图形化的 Louvain 方法聚类，用 RepeatMasker、BLAST 等软件进行物种间重复序列比对。REannotate[72]使用 RepeatMasker 的输出，注释存在的重复序列，自动分析注释分散的重复元件。ReRep[73,74]进行序列组装前的重复识别，使用了 BLAST 或 MUMmer 软件。RetroPred[73,74]识别非 LTR 反转录转座子，结合

使用 PALS 和 PILER 软件，并通过人工神经网络模型进行分类。RISCI[75]使用 RepeatMasker、BLAST 软件进行序列比对得分统计，它可以进行整个基因组或部分基因组分析，进行长度或基因过滤。Tea[76]是序列组装前重复识别工具，使用了 CAP3、BWA 序列组装软件。ELAN[77]可以跨基因组分析，通过滑动窗口扫描 DNA，用支持向量机学习与分类发现插入节点。DAWG-PAWS[78]是一个对植物注释基因和转座子的管道程序，用到以下几个识别不同重复类型的工具：LTR_struc、LTR_Finder、LTR_par、Repseek、FINDMITE、TRF、RepeatMasker 和 TENest 等。

3.1.5 分类方法

这里所说的分类有两种含义：一种是有些工具能够自动识别出各种类型的重复，另一种是有一些专门识别某种重复类型的工具将重复类型聚类。例如，ModuleOrganizer[79]引入转座元件模块，不用多序列比对方法，而用后缀树选择模块，通过递归查找最大重复进入同一模块，并进行分层聚类。TCF[80]可以使用 RepeatMasker 中已知重复数据，创建片段重复中断矩阵查找转座子。REPET[81]用 BLASTER 进行序列比较，用工具 GROUPER、RECON 和 PILER 进行序列聚类，综合使用 RepeatMasker，TRF 和 Mreps 查找简单重复序列，对聚类的每簇使用多序列比对方法，根据相似性去除冗余。Dfam[82]基于 RepeatMasker 真核基因组数据库使用隐马尔可夫链模型，获得的结果较详细，具有全部重复序列信息，可进行灵活的搜索。TEnest[83]重构了插入嵌套引起的分离转座子，嵌入重复可图形化显示，可以从中看出基因进化过程，并且程序可实现并行化处理。TARGeT[84]使用 muscle[85]多序列比对、用 TreeBest[86]产生系统发育树、使用邻接法预测同族、使用同源序列查询，因此不能实时更新已有数据库，但使用进化树可以看出基因结构特征和系统进化。RepClass[87]可以基于同源进行分布式计算，根据 TSD 和其他结构特征分类后整合分析。TEclass[88]可分类出 DNA 转座、长末端重复、长散在重复和短散在重复，采用了基于低聚物重复频率的支持向量机、随机森林等机器学习方法分类。DomainOrganizer[89]基于区域识别分类，ClustalW 多序列比对产生序列同类区域，使用 HMMER[90]匹配某类区域的出现频率，图形映射出现碎片区域，用频率矩阵的分类可以用系统进化分析代替。RetroMap[91]是一个反转录因子映射工具，通过定位两个重复的侧翼序列来推断长末端重复序列。LTR_MINER[92]使用 RepeatMasker 的输出识别 LTR 反转录转座子，包括插入变异和碎片相似性序列片段识别 LTR 反转录转座子。

3.2 非转座元件查找

目前主要是串联重复序列的查找方法，其中以简单重复序列即小卫星的查找为研究热点。Mreps[93]从头计算使用穷举法识别串联重复序列，用调整两个参数来识别模糊重复，它不能直接处理存在插入删除的重复序列，因此某些存在这样插删重复的序列可能会丢失。OMWSA[94]使用自回归模型分析序列谱，用优化的移动窗口频谱分析识别串联重复序列；启发式方法不能保证搜到所有重复，计算复杂度随着序列中串联重复序列拷贝的长度呈指数增长，并且傅里叶变换容易出现假顶点。TRAP[95]使用了 TRF 的输出，分析卫星内容，标识发现的微卫星，注释重复区。TRF[96]识别串联重复序列，把序列比对看作 N 阶独立同分布伯努利试验，通过确定 4 个参数：模式长度、匹配成功率、变异率和 K 元组匹配，找出候选序列位置，然后用动态规划法确定串联重复序列。但由于它是在具有大量重复片段区选取重复，当每个重复大小不相同时，重复序列可能会报告多次。TROLL[97]识别完全的串联重复序列，它基于修改的 Aho-Corasick 自动机算法建立

关键词树，不能识别存在变异的重复序列。

4　总结与展望

自 20 世纪 90 年代开始，重复序列的研究一直是基因组学的研究热点。随着高通量技术的发展，越来越多的物种信息被人类认知，重复序列会得到更广泛和深入的研究及发展。相信在不久的将来，人类会有更多的发现，从而能更进一步认识世界和自身。未来重复序列的研究可能会有以下几个方向：

(1)新的重复序列类型发现。近年来，基于重复序列又发现很多具有特殊结构或功能的片段区。如发现了具有很多成簇的、规律间隔的短回文重复序列(Clustered Regularly Interspaced Short Palindromic Repeats，CRISPR)及相关基因，它促进了基因组改造新技术的发展[98]。目前已经出现一些 CRISPR 检测程序，如：PILER-cr[99]、CRT[100]、 CRISPRFinder[101]和 CRISPRdigger[102]等。它们可以从一个给定的基因组或基因片段文件识别出其中的 CRISPRs。

(2)插入删除变异的进一步研究。例如，imapper[103]插入位点序列分析，其基于聚合酶链式反应，将序列映射到 Ensembl 数据库，标签序列、酶切位点和污染序列通过图形显示；InFiRe[104]工具用于细菌转座子诱变识别；TESD[105]用于转座子结构的物种间动态演化，迁移可以解释自然界中很多转座分布模式，它用种群间迁移、大陆岛屿迁移、有限种群间岛屿迁移 3 个模型表明转座子在物种迁移中的重要作用。插入删除也作为新一代遗传标记，兼有短串联重复序列和单核苷酸多态性(Single Nucleotide Polymorphisms)等优点，在法医 DNA 分析中将进一步得到应用。

(3)重复序列功能发现研究及应用。重复序列的保留机制、基因的多向性效应和进化选择通过重复产生新功能基因形成稳定的保留等都有新的研究和应用。重复基因可以在特定时间和空间表达，促使基因表达水平多样性的提高，增强抗突变能力。近些年，虽然重复序列研究有很大进展，但仍存在很多问题需要解决。如重复序列对蛋白相互作用及表达网络的影响没有得到合理的解释。

(4)微环境重复序列簇研究。肠道细菌基因间重复序列与某些插入序列有一定相似性，可以调节两端基因的表达，影响基因的稳定性，为细胞拟核染色体合成提供相关位点。其内序列高度保守，可以进行系统研究，测定疾病与正常图谱，找出恢复正常图谱的最佳疗法。

(5)重复序列检测算法研究。转座子研究热点之一仍然在其发现和检测算法上，更多地体现在性能和可用性的提高上。目前比较流行的做法是整合多种已有检测算法的结果。一种是通过选票的方法确定，另外也有些使用管道机制确定。但整合中出现的检测冲突仍需寻找新方法和证据来解释。另外，转座子的结构特征和存储形式也是需要考虑的内容。同时，可以利用领域知识和数据挖掘新技术加速重复序列模式发现。

随着第三代测序技术的发展，新类型的重复元件不断出现。从头发现序列中的重复片段技术显得更为重要。如何开发敏捷准确的技术，对重复片段进行更好地分类和注释将是未来一个重要研究方向。其中需要研究的内容很多，如设计参数自动优化方法、重复新家族的发现与分类、重复元素的位置定位等。同时，考虑提供并行加速处理技术，以应对越来越多的序列数据。这些对进一步分析相关基因和功能有重要参考意义。

参 考 文 献

[1] Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions [J]. Nature Reviews

Genetics, 2012, 13(1): 36-46.

[2] Shu HW, Liu TT, Chan HI, et al. Genome sequence of the repetitive-sequence-rich mycoplasma fermentans strain M64 [J]. Journal of Bacteriology, 2011, 193(16): 4302-4303.

[3] Wang G, Zhou F, Olman V, et al. Prediction of pathogenicity islands in enterohemorrhagic Escherichia coli O157 : H7 using genomic barcodes [J]. Febs Letters, 2010, 584(1): 194-198.

[4] Mindlin S, Kholodii G, Gorlenko Z, et al. Mercury resistance transposons of gram-negative environmental bacteria and their classification [J]. Research in Microbiology, 2001, 152(9): 811-822.

[5] Cyril F, Olivia P, Hélène R, et al. Diversity of mycoplasma hominis clinical isolates from Bordeaux, France, as assessed by multiple-locus variable-number tandem repeat analysis [J]. BMC Microbiology, 2013, 13(1): 120.

[6] Kumar RP, Krishnan J, Pratap SN, et al. Gata simple sequence repeats function as enhancer blocker boundaries [J]. Nature Communications, 2013, 4(5): 1844.

[7] Sukumaran S, Grant A. Effects of genotoxicity and its consequences at the population level in sexual and asexual artemia assessed by analysis of inter-simple sequence repeats (ISSR) [J]. Mutation Research, 2013, 757(1): 8-14.

[8] Madesis P, Ganopoulos I, Tsaftaris A. Microsatellites: evolution and contribution [M] // Methods in Molecular Biology, 2013: 1-13.

[9] Kupriyanova NS, Shibalev DV, Voronov AS, et al. Enhanced heterogeneity of the LR2 segment in the human ribosomal intergenic spacer [J]. Gene, 2008, 425(1-2): 44-47.

[10] Wilson MK, Lane AB, Law BF, et al. Analysis of the pan genome of Campylobacter jejuni isolates recovered from poultry by pulsed-field gel electrophoresis, multilocus sequence typing (MLST), and repetitive sequence polymerase chain reaction (rep-PCR) reveals different discriminatory capabilities [J]. Microbial Ecology, 2009, 58(4): 843-855.

[11] Tempel S. Using and understanding RepeatMasker [M] // Mobile Genetic Elements, 2012: 29-51.

[12] Jurka J, Klonowski P, Dagman V, et al. Censor—a program for identification and elimination of repetitive elements from DNA sequences [J]. Computers & Chemistry, 1996, 20(1): 119-121.

[13] Kennedy RC, Unger MF, Christley S, et al. An automated homology-based approach for identifying transposable elements [J]. BMC Bioinformatics, 2011, 12(1): 130.

[14] Bedell JA, Korf I, Gish W. MaskerAid: a performance enhancement to repeatmasker [J]. Bioinformatics, 2000, 16(11): 1040-1041.

[15] Li X, Kahveci T, Settles AM. A novel genome-scale repeat finder geared towards transposons [J]. Bioinformatics, 2008, 24(4): 468-476.

[16] Fistonlavier AS, Carrigan M, Petrov DA, et al. T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data [J]. Nucleic Acids Research, 2011, 39(6): e36.

[17] Kapitonov VV, Tempel S, Jurka J. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences [J]. Gene, 2009, 448(2): 207-213.

[18] Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing data [J]. Bioinformatics, 2013, 29(3): 389-390.

[19] Chen K, Wallis JW, Mclellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation [J]. Nature Methods, 2009, 6(9): 677-681.

[20] Morgulis A, Gertz EM, Schaffer AA, et al. WindowMasker: window-based masker for sequenced genomes [J]. Bioinformatics, 2006, 22(2): 134-141.

[21] Morgulis A, Gertz EM, Schaffer AA, et al. A fast and symmetric DUST implementation to mask low-complexity DNA sequences [J]. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology, 2006, 13(5): 1028-1040.

[22] Vassetzky NS, Kramerov DA. SINEBase: a database and tool for SINE analysis [J]. Nucleic

Acids Research, 2013, 41: D83-D89.

[23] Rooke R, Yang G. TE displayer for post-genomic analysis of transposable elements [J]. Bioinformatics, 2011, 27(2): 286-287.

[24] Toth G, Deak G, Barta E, et al. Plotrep: a web tool for defragmentation and visual analysis of dispersed genomic repeats [J]. Nucleic Acids Research, 2006, 34: 708-713.

[25] Wojciech M, Jürgen B, Andrej K, et al. A novel web-based TinT application and the chronology of the Primate Aluretroposon activity [J]. BMC Evolutionary Biology, 2010, 10(1): 376.

[26] Mccarthy EM, Mcdonald JF. LTR_STRUC: a novel search and identification program for LTR retrotransposons [J]. Bioinformatics, 2003, 19(3): 362-367.

[27] Sperber G, Lovgren A, Eriksson NE, et al. RetroTector online, a rational tool for analysis of retroviral elements in small and medium size vertebrate genomic sequences [J]. BMC Bioinformatics, 2009, 10(6): S4.

[28] Kalyanaraman A, Aluru S. Efficient algorithms and software for detection of full-length LTR retrotransposons [C] // Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference, 2005: 56-64.

[29] Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons [J]. Nucleic Acids Research, 2007, 35: W265-W268.

[30] Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons [J]. BMC Bioinformatics, 2008, 9(1): 1-14.

[31] Darzentas N, Bousios A, Apostolidou V, et al. MASiVE: mapping and analysis of sirevirus elements in plant genome sequences [J]. Bioinformatics, 2010, 26(19): 2452-2454.

[32] Rho M, Choi JH, Kim S, et al. De novo identification of LTR retrotransposons in eukaryotic genomes [J]. BMC Genomics, 2007, 8(1): 90.

[33] Jean-François L, Jonathan P, Jean-François N, et al. RTAnalyzer: a web application for finding new retrotransposons and detecting L1 retrotransposition signatures [J]. Nucleic Acids Research, 2007, 35: W269-W274.

[34] Szak ST, Pickeral OK, Makalowski W, et al. Molecular archeology of L1 insertions in the human genome [J]. Genome Biology, 2002, 3(10): 1-18.

[35] Tu Z, Li S, Mao C. The changing tails of a novel short interspersed element in aedesaegypti: genomic evidence for slippage retrotransposition and the relationship between 3′ tandem repeats and the poly (dA) tail [J]. Genetics, 2004, 168(4): 2037-2047.

[36] Tu Z. Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, anopheles gambiae [J]. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(4): 1699-1704.

[37] Chen J, Hu Q, Zhang Y, et al. P-MITE: a database for plant miniature inverted-repeat transposable elements [J]. Nucleic Acids Research, 2014, 42: 1176-1181.

[38] Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences [J]. Nucleic Acids Research, 2010, 38(22): e199.

[39] Yang G, Hall TC. MAK, a computational tool kit for automated MITE analysis [J]. Nucleic Acids Research, 2003, 31(13): 3659-3665.

[40] Chen Y, Zhou F, Li G, et al. MUST: a system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi [J]. Gene, 2009, 436(1-2): 1-7.

[41] Santiago N, Herraiz C, Goni JR, et al. Genome-wide analysis of the emigrant family of MITEs of arabidopsis thaliana [J]. Molecular Biology and Evolution, 2002, 19(12): 2285-2293.

[42] Yang G. MITE digger: an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements [J]. BMC Bioinformatics, 2013, 14(1): 186.

[43] Chen Y, Zhou F, Li G, et al. A recently active miniature inverted-repeat transposable element, Chunjie, inserted into an operon without disturbing the operon structure in Geobacter uraniireducens

Rf4 [J]. Genetics, 2008, 179(4): 2291-2297.

[44] Du C, Fefelova N, Caronna J, et al. The polychromatic helitron landscape of the maize genome [J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(47): 19916-19921.

[45] Warburton PE, Giordano J, Cheung F, et al. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes [J]. Genome Research, 2004, 14(10A): 1861-1869.

[46] Agarwal P, States DJ. The repeat pattern toolkit (RPT): analyzing the structure and evolution of the *C. elegans* genome [C] // Proceedings of the International Conference on Intelligent Systems for Molecular Biology, 1994: 1-9.

[47] Taneda A. Adplot: detection and visualization of repetitive patterns in complete genomes [J]. Bioinformatics, 2004, 20(5): 701-708.

[48] Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes [J]. Genome Research, 2002, 12(8): 1269-1276.

[49] Edgar RC, Myers EW. PILER: identification and classification of genomic repeats [J]. Bioinformatics, 2005, 21(Suppl 1): i152-i158.

[50] Steinbiss S, Willhoeft U, Gremme G, et al. Fine-grained annotation and classification of de novo predicted LTR retrotransposons [J]. Nucleic Acids Research, 2009, 37(21): 7002-7013.

[51] Kofler R, Betancourt AJ, Schlotterer C. Sequencing of pooled DNA samples (pool-seq) uncovers complex dynamics of transposable element insertions in drosophila melanogaster [J]. PloS Genetics, 2012, 8(1): e1002487.

[52] Ganguly S, Mitchell AP. Mini-blaster-mediated targeted gene disruption and marker complementation in candida albicans [M] // Host-Fungus Interactions, 2012: 19-39.

[53] Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes [J]. Bioinformatics, 2005, 21(Suppl 1): i351-i358.

[54] Li R, Ye J, Li S, et al. ReAS: recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun [J]. PloS Computational Biology, 2005, 1(4): e43.

[55] Kurtz S, Choudhuri JV, Ohlebusch E, et al. REPuter: the manifold applications of repeat analysis on a genomic scale [J]. Nucleic Acids Research, 2001, 29(22): 4633-4642.

[56] Achaz G, Rocha EP, Viari A, et al. Repseek, a tool to retrieve approximate repeats from large DNA sequences [J]. Bioinformatics, 2007, 23(1): 119-121.

[57] Kurtz S, Phillippy A, Delcher AL, et al. Versatile and open software for comparing large genomes [J]. Genome Biology, 2004, 5(2): R12.

[58] Morgante M, Policriti A, Vitacolonna N, et al. Structured motifs search [J]. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology, 2005, 12(8): 1065-1082.

[59] Kurtz S, Narechania A, Stein JC, et al. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes [J]. BMC Genomics, 2008, 9(1): 517.

[60] Abouelhoda MI, Kurtz S, Ohlebusch E. Replacing suffix trees with enhanced suffix arrays [J]. Journal of Discrete Algorithms, 2004, 2(1): 53-86.

[61] Healy J, Thomas EE, Schwartz JT, et al. Annotating large genomes with exact word matches [J]. Genome Research, 2003, 13(10): 2306-2315.

[62] Lefebvre A, Lecroq T, Dauchel H, et al. FORRepeats: detects repeats on entire chromosomes and between genomes [J]. Bioinformatics, 2003, 19(3): 319-326.

[63] Gu W, Castoe TA, Hedges DJ, et al. Identification of repeat structure in large genomes using repeat probability clouds [J]. Analytical Biochemistry, 2008, 380(1): 77-83.

[64] Thompson JD, Gibson TJ, Higgins DG. Multiple sequence alignment using ClustalW and ClustalX [M] // Current Protocols in Bioinformatics, 2002, doi: 10.1002/0471250953.

[65] Delcher AL, Salzberg SL, Phillippy AM. Using MUMmer to identify similar regions in large sequence sets [M] // Current Protocols in

Bioinformatics, 2003, doi: 10.1002/0471250953.bi1003s00.

[66] Sharma D, Issac B, Raghava GPS, et al. Spectral repeat finder (SRF): identification of repetitive sequences using Fourier transformation [J]. Bioinformatics, 2004, 20(9): 1405-1412.

[67] Volfovsky N, Haas BJ, Salzberg SL. A clustering method for repeat analysis in DNA sequences [J]. Genome Biology, 2001, 2(8): 1-11.

[68] Pevzner PA, Tang H, Tesler G. De novo repeat classification and fragment assembly [J]. Genome Research, 2004, 14(9): 1786-1796.

[69] Smit A, Hubley R. RepeatModeler Open-1.0 [OL]. 2008-2015. http://www.repeatmasker.org.

[70] Smith CD, Edgar RC, Yandell MD, et al. Improved repeat identification and masking in dipterans [J]. Gene, 2007, 389(1): 1-9.

[71] Novak P, Neumann P, Pech J, et al. RepeatExplorer: a galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads [J]. Bioinformatics, 2013, 29(6): 792-793.

[72] Pereira V. Automated paleontology of repetitive DNA with REANNOTATE [J]. BMC Genomics, 2008, 9(1): 614.

[73] Marcelo AF, Gomes Leonardo HF, Otto TD, et al. ReRep: computational detection of repetitive sequences in genome survey sequences (GSS) [J]. BMC Bioinformatics, 2008, 9(1): 366.

[74] Naik PK, Mittal VK, Gupta S. RetroPred: a tool for prediction, classification and extraction of non-LTR retrotransposons (LINEs & SINEs) from the genome by integrating PALS, PILER, MEME and ANN [J]. Bioinformation, 2008, 2(6): 263-270.

[75] Singh V, Mishra RK. RISCI—repeat induced sequence changes identifier: a comprehensive, comparative genomics-based, in silico subtractive hybridization pipeline to identify repeat induced sequence changes in closely related genomes [J]. BMC Bioinformatics, 2010, 11(1): 609.

[76] Lee E, Iskow R, Yang L, et al. Analysis of somatic retrotransposition in human cancers [J]. BMC Proceedings, 2012, 337(6097): 967-971.

[77] Mandal PK, Rawal K, Ramaswamy R, et al. Identification of insertion hot spots for non-LTR retrotransposons: computational and biochemical application to entamoebahistolytica [J]. Nucleic Acids Research, 2006, 34(20): 5752-5763.

[78] Estill JC, Bennetzen JL. The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes [J]. Plant Methods, 2009, 5(1): 8.

[79] Tempel S, Rousseau C, Tahi F, et al. ModuleOrganizer: detecting modules in families of transposable elements [J]. BMC Bioinformatics, 2010, 11(1): 474.

[80] Giordano J, Ge Y, Gelfand Y, et al. Evolutionary history of mammalian transposons determined by genome-wide defragmentation [J]. PloS Computational Biology, 2007, 3(7): e137.

[81] Flutre T, Duprat E, Feuillet C, et al. Considering transposable element diversification in de novo annotation approaches [J]. PloS One, 2011, 6(1): e16526.

[82] Wheeler TJ, Clements J, Eddy SR, et al. Dfam: a database of repetitive DNA based on profile hidden Markov models [J]. Nucleic Acids Research, 2013, 41(Database issue): D70-D82.

[83] Kronmiller BA, Wise RP. TEnest 2.0: computational annotation and visualization of nested transposable elements [M] // Plant Transposable Elements, 2013: 305-319.

[84] Han Y, Burnette JM, Wessler SR. TARGeT: a web-based pipeline for retrieving and characterizing gene and transposable element families from genomic sequences [J]. Nucleic Acids Research, 2009, 37(11): e78.

[85] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput [J]. Nucleic Acids Research, 2004, 32(5): 1792-1797.

[86] Vilella AJ, Severin J, Ureta-Vidal A, et al. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates [J]. Genome Research, 2009, 19(2): 327-335.

[87] Feschotte C, Keswani U, Ranganathan N, et al. Exploring repetitive DNA landscapes using

REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes [J]. Genome Biology & Evolution, 2009, 1(1): 205-220.

[88] Grundmann N, Demester L, Makalowski W. TEclass—a tool for automated classification of unknown eukaryotic transposable elements [J]. Bioinformatics, 2009, 25(10): 1329-1330.

[89] Tempel S, Giraud M, Lerman IC, et al. Domain organization within repeated DNA sequences: application to the study of a family of transposable elements [J]. Bioinformatics, 2006, 22(16): 1948-1954.

[90] Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching [J]. Nucleic Acids Research, 2011, 39: 29-37.

[91] Gao X, Havecker ER, Baranov PV, et al. Translational recoding signals between gag and pol in diverse LTR retrotransposons [J]. RNA, 2003, 9(12): 1422-1430.

[92] Pereira V. Insertion bias and purifying selection of retrotransposons in the arabidopsis thaliana genome [J]. Genome Biology, 2004, 5(10): R79.

[93] Kolpakov R, Bana G, Kucherov G. Mreps: efficient and flexible detection of tandem repeats in DNA [J]. Nucleic Acids Research, 2003, 31(13): 3672-3678.

[94] Du L, Zhou H, Yan H. OMWSA: detection of DNA repeats using moving window spectral analysis [J]. Bioinformatics, 2007, 23(5): 631-633.

[95] Sobreira TJ, Durham AM, Gruber A. TRAP: automated classification, quantification and annotation of tandemly repeated sequences [J]. Bioinformatics, 2006, 22(3): 361-362.

[96] Benson G. Tandem repeats finder: a program to analyze DNA sequences [J]. Nucleic Acids Research, 1999, 27(2): 573-580.

[97] Castelo AT, Martins W, Gao GR. TROLL—tandem repeat occurrence locator [J]. Bioinformatics, 2002, 18(4): 634-636.

[98] Mai G, Ge R, Sun G, et al. A comprehensive curation shows the dynamic evolutionary patterns of prokaryotic CRISPRs [J]. Biomed Research International, 2016: 7237053.

[99] Edgar RC. PILER-CR: fast and accurate identification of CRISPR repeats [J]. BMC Bioinformatics, 2007, 8(1): 18.

[100] Bland C, Ramsey TL, Sabree F, et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats [J]. BMC Bioinformatics, 2007, 8(1): 209.

[101] Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats [J]. Nucleic Acids Research, 2007, 35: 52-57.

[102] Ge R, Mai G, Wang P, et al. CRISPRdigger: detecting CRISPRs with better direct repeat annotations [J]. Scientific Reports, 2016, 6: 32942.

[103] Kong J, Zhu F, Jim S, et al. iMapper: a web application for the automated analysis and mapping of insertional mutagenesis sequence data against Ensembl genomes [J]. Bioinformatics, 2008, 24(24): 2923-2925.

[104] Shevchuk O, Roselius L, Gunther G, et al. InFiRe—a novel computational method for the identification of insertion sites in transposon mutagenized bacterial genomes [J]. Bioinformatics, 2012, 28(3): 306-310.

[105] Deceliere G, Letrillard Y, Charles S, et al. TESD: a transposable element dynamics simulation environment [J]. Bioinformatics, 2006, 22(21): 2702-2703.