

Distribution Characteristics of Repeat Sequences in the Genome of the Yeast Recombinant Han0458 Obtained by Low-Energy Ion Beam Implantation

Hanyu Zhang¹, Chao Tang¹, Ting Wang², Guangwen Feng¹, Weidong Qian², Peihong Mao^{1,3*}, Changlong Cai^{3*}

¹Research Center of Ion Beam Biotechnology, College of Physics Science and Technology, Xinjiang University, Urumqi Xinjiang

²School of Food and Biological Engineering, Shaanxi University of Science & Technology, Xi'an Shaanxi

³Research Center of Ion Beam Biotechnology and Biodiversity, Xi'an Technological University, Xi'an Shaanxi

Email: *phmao@china.com, *changlongcai@126.com

Received: Sep. 5th, 2018; accepted: Sep. 18th, 2018; published: Sep. 25th, 2018

Abstract

Repeat sequence is an important part of the genome and play an important role in the biological evolution, genetics, and regulation of gene expression. In order to further understand the biological effects of low-energy ion implantation on the yeast genome structure, in this study, we used bioinformatics methods to study the distribution characteristics of the genomic repeat sequences on the basis of *de novo* sequencing of the Han0458 genome. The results shows that the length of tandem repeats of the recombinant Han0458 genome is 227,825 bp and interspersed repeats is 107,439 bp, which account for 2.46% of the total genome length, and the frequency of repeats in the genome is 0.25 per Kb. The copy number of repeat units in microsatellite DNA are mostly less than 15 times, the dominant base type is trinucleotide repeats, and the number of microsatellite sequences with repeat unit motif AAC is the largest; the copy number of repeat units in minisatellite DNA is smaller than that of microsatellite DNA. The number of minisatellite DNA with a repeat unit greater than 15bp is decreasing with the increase of the repeat unit length. The number of long-terminal repeats (LTRs) is the largest in interspersed repeats, and the average length of rolling coils (RCs) is the longest. The results of this study provide molecular evidence for low-energy ion implantation-mediated yeast genome mutation and evolution.

Keywords

Ion Beam Recombinant Yeast, Genome, Minisatellite DNA, Microsatellite DNA, Interspersed Repeats

*通讯作者。

离子束重组酵母菌Han0458基因组重复序列的分布特征

张寒玉¹, 唐朝¹, 王婷², 冯光文¹, 钱卫东², 毛培宏^{1,3*}, 蔡长龙^{3*}

¹新疆大学, 物理科学与技术学院, 离子束生物技术中心, 新疆 乌鲁木齐

²陕西科技大学, 食品与生物工程学院, 陕西 西安

³西安工业大学, 离子束生物工程与生物多样性研究中心, 陕西 西安

Email: phmao@china.com, changlongcai@126.com

收稿日期: 2018年9月5日; 录用日期: 2018年9月18日; 发布日期: 2018年9月25日

摘要

重复序列是基因组的重要组成部分, 对生物的进化、遗传和基因的表达调控有重要作用。为了进一步认识低能离子注入对酵母菌基因组结构的生物学效应, 本研究利用生物信息学方法, 在离子束重组酵母菌株Han0458全基因组*de novo*测序的基础上, 对其基因组重复序列的分布特征进行了研究。结果表明, 重组菌株Han0458基因组串联重复序列227,825 bp, 散在重复序列107,439 bp, 共占基因组长度的2.46%, 重复序列在基因组中出现的频率为0.25个/Kb。微卫星DNA中重复单元拷贝数大多低于15次, 其优势碱基类型为三碱基重复, 重复单元基序为AAC的微卫星序列数目最多; 小卫星DNA重复单元拷贝数小于微卫星DNA, 主要分布1~3次, 重复单元大于15 bp的小卫星序列数目随着重复单位长度的增加呈下降趋势。散在重复序列中末端重复序列(LTR)数目最多, 滚环(RC)平均长度最长。本研究结果为低能离子注入介导的酵母菌基因组突变与进化提供了分子证据。

关键词

离子束重组酵母菌, 基因组, 微卫星DNA, 小卫星DNA, 散在重复序列

Copyright © 2018 by authors and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

重复序列是真核生物基因组中重要的组成部分, 按其在基因组中的分布方式, 分为串联重复序列(Tandem Repeat Sequences)和散在重复序列(Interspersed Repeat Sequences) [1]。串联重复序列又可根据其重复单元长度划分为卫星DNA (Satellite DNA) [2]、小卫星DNA (Minisatellite DNA) [3]和微卫星DNA (Microsatellite DNA) [4]。其中微卫星DNA又称为短串联重复序列(Short Tandem Repeat, STR)或简单重复序列(Simple Sequence Repeats, SSRs), 随机分布于生物体整个基因组中, 微卫星标记作为理想的分子遗传标记, 被广泛地用于目的基因筛选、基因诊断多样性分析及遗传连锁图谱构建等工作中。而散在重复序列(又称转座子元件, Transposable Element, TE)分为RNA介导的转座元件(又称RNA转座子)和DNA介导的转座元件(又称DNA转座子), 不仅可以影响基因组的大小, 还能直接或间接促成基因组重排, 并可

影响基因表达水平、改写基因调控网络。因此 TEs 对基因组的进化中具有重要意义[5]。本研究在离子束重组酵母菌株 Han0458 全基因组 *de novo* 测序的基础上, 利用生物信息学方法分析其基因组中各种重复序列的类型及分布特点, 以期深入了解低能离子注入对酵母菌基因组结构特征的影响, 并为重组菌的分子育种及 SSR 分子标记的开发提供理论依据。

2. 材料与方法

2.1. 菌株重复序列的获取

低能 N^+ 注入介导外源基因转化异常汉逊酵母, 经选择性筛选获得重组菌株 Han0458 [6] [7]。应用 PacBio 单分子测序技术对其进行基因组 *De novo* 测序, 获得的基因组 DNA 序列, 作为本研究的基本数据。

2.2. 生物信息学方法

应用 TRF (Tandem Repeat Finder)方法 (<http://tandem.bu.edu/trf/trf404.linux64.download.html>)获取重组菌株 Han0458 基因组 DNA 序列中的串联重复序列, 最大的重复单元 bp 数设置为 2000 bp。

对 TRF 获取的结果进行细分, 设置微卫星 DNA 序列重复单位为 2~6 bp, 小卫星 DNA 序列重复单位为 10~60 bp。

使用 RepeatMasker 方法(<http://www.repeatmasker.org/RMDownload.html>)获取重组菌株 Han0458 基因组 DNA 序列中的散在重复序列。

3. 结果与分析

3.1. 微卫星 DNA 序列的分布特征

利用 TRF 方法在重组菌株 Han0458 基因组重复序列中发现了 175 个 SSR, 总长为 8310 bp, 占基因组序列总长度的 0.06%, 平均每 77.98 Kb 就能检测到一个 SSR。

SSR 在三碱基(Trinucleotide)模体中的数目最多, 为 97 条, 占重复序列总数的 55.43%; 其次是六碱基(Hexanucleotide)模体, 为 47 条, 占 26.86%; 五碱基(Pentanucleotide)和四碱基(Tetranucleotide)模体数目相对较少, 分别在 13~16 条之间, 占比约为 7.43%~9.14%; 二碱基(Dinucleotide)模体的重复序列最少, 仅有 3 条, 占 1.71%(图 1)。

碱基类型的重复基序分析结果(表 1)表明, 在 4 种两碱基类型重复中, 仅有 AT 重复基序。

三碱基类型重复基序有 10 种, 其中数量较多的碱基类型依次是 AAC (51 条, 52.58%)、ACT (13 条, 13.40%)、ATG (12 条, 12.37%)。累积长度最长的依然是上述三个类型的重复序列: AAC (2176 bp)、ACT (756 bp)、ATG (595 bp)。

四碱基类型重复序列中含有 AAAT、AACT、ATTA、GTTA 和 TTCA 重复类型, 且前两种类型数量较多, 占四碱基重复序列数目的 76.92%, 长度较长, 约占四碱基重复序列累积长度的 66.42%;

五碱基类型重复序列共有 15 条, 其中 AATTC 和 TATAC 重复单元的序列各有 2 条, 其余各类型基序重复序列均只有 1 条。

六碱基类型重复序列共 47 条, 其每种重复单元基序数目, 均为 1~2 个。

各种重复类型的拷贝数分析结果(表 2)表明, 微卫星序列均在低拷贝区出现频率较高, 拷贝数低于 15 次的微卫星序列占比 73.14%; 拷贝数在 15~27 之间的微卫星序列, 占比 18.29%; 拷贝数在 27~39 之间的占比 7.43%; 拷贝数大于 39 次的微卫星序列最少, 仅占 1.14%。

从图 2 可以看出, 拷贝数越大, 微卫星序列数目越少, 微卫星平均拷贝数随着重复单位长度的增加而减少。

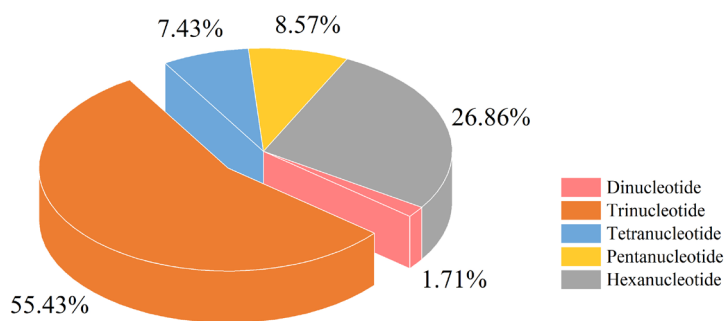


Figure 1. Distribution of the SSR of the recombinant yeast strain Han0458 genome in different motif types

图 1. 重组酵母菌株 Han0458 基因组 SSR 在不同模体类型中的分布

Table 1. Repeat motif distribution of Microsatellite DNA in the genome of yeast recombinant Han0458

表 1. 重组菌株 Han0458 基因组微卫星 DNA 的重复基序分布

重复类型 Repeat type	重复序列数目 repeat sequence No.	占 SSR 总数的百 分比 Percentage of total SSRs	累积长度(bp) Total length (bp)	占 SSR 总长度百 分比 Percentage of total SSR length	拷贝数 范围 Copy number range	平均 拷贝数 Average copy No.
AT	3	1.71	180	2.17	33.0	33.0
Subtotal	3	1.71	180	2.17	33.0	33.0
AAC	51	29.14	2 176	26.19	8.3~44.0	14.3
AAG	9	5.14	423	5.09	8.3~22.7	15.7
AAT	6	3.43	288	3.47	11.7~25.0	15.8
ACC	1	0.57	35	0.42	11.70	11.7
ACG	1	0.57	25	0.30	8.3	8.3
ACT	13	7.43	756	9.10	9.0~37.3	19.4
AGC	2	1.14	96	1.16	9.7~22.0	15.9
AGG	2	1.14	94	1.13	15.70	15.7
ATG	12	6.86	595	7.16	8.3~37.3	16.5
Subtotal	97	55.43	4 488	54.01	8.3~44.0	15.4
AAAT	7	4.00	374	4.50	6.2~17.8	13.1
AACT	3	1.71	77	0.93	6.2~6.8	6.4
ATTA	1	0.57	73	0.88	18.8	18.8
GTTA	1	0.57	128	1.54	34.0	34.0
TTCA	1	0.57	27	0.32	6.8	6.8
Subtotal	13	7.43	679	8.17	6.2~34.0	13.1
AATAC	1	0.57	44	0.53	9.6	9.6
AATAG	1	0.57	38	0.46	7.6	7.6
AATTC	2	1.14	59	0.71	5.2~6.8	6.0
ATTAA	1	0.57	43	0.52	8.2	8.2
ATTCa	1	0.57	30	0.36	6.0	6.0
ATTGT	1	0.57	29	0.35	5.8	5.8
TATAC	2	1.14	95	1.14	7~12.4	9.7
TCAAT	1	0.57	33	0.40	7.0	7.0
TGAAT	1	0.57	38	0.46	7.6	7.6

Continued

TTCTC	1	0.57	56	0.67	11.6	11.6
TTGTA	1	0.57	58	0.70	11.8	11.8
TTGTT	1	0.57	32	0.39	6.4	6.4
TTTTG	1	0.57	35	0.42	7.0	7.0
Subtotal	15	8.57	590	7.10	5.2~12.4	8.0
AACAGG	1	0.57	29	0.35	4.8	4.8
AAGCAA	2	1.14	60	0.72	4.5~5.5	5.0
AATCAG	1	0.57	32	0.39	5.3	5.3
ACAAAA	1	0.57	34	0.41	5.7	5.7
ACCTGA	1	0.57	33	0.40	5.5	5.5
ATCAGA	1	0.57	43	0.52	7.2	7.2
ATGAAG	1	0.57	67	0.81	10.7	10.7
CAAGGT	1	0.57	35	0.42	5.8	5.8
CAATAG	1	0.57	30	0.36	5.0	5.0
CAGGTT	1	0.57	27	0.32	4.5	4.5
CATCAC	1	0.57	35	0.42	5.8	5.8
CATCTT	1	0.57	337	4.06	56.2	56.2
CCTGAT	1	0.57	38	0.46	6.3	6.3
CTGTTG	1	0.57	36	0.43	6.0	6.0
CTTTTG	1	0.57	34	0.41	5.8	5.8
GAAGAC	1	0.57	33	0.40	5.5	5.5
GAGCTT	1	0.57	25	0.30	4.2	4.2
GATGAA	2	1.14	76	0.91	4.3~8.3	6.3
GATGAC	1	0.57	41	0.49	6.8	6.8
GATGAG	1	0.57	47	0.57	7.8	7.8
GTGAAA	1	0.57	112	1.35	19.0	19.0
TAAACA	1	0.57	33	0.40	5.5	5.5
TAGTAA	1	0.57	52	0.63	9.3	9.3
TAGTAT	2	1.14	106	1.28	8.8	8.8
TATGAA	1	0.57	25	0.30	4.2	4.2
TCAAGA	1	0.57	26	0.31	4.3	4.3
TCATTT	1	0.57	29	0.35	4.8	4.8
TCTTCA	1	0.57	60	0.72	10.0	10.0
TCTTCC	1	0.57	69	0.83	11.7	11.7
TGAAGT	1	0.57	32	0.39	5.3	5.3
TGAGGT	2	1.14	84	1.01	5.5~8.5	7.0
TGATGT	1	0.57	30	0.36	5.0	5.0
TGATTA	1	0.57	39	0.47	6.5	6.5
TGATTT	1	0.57	45	0.54	8.0	8.0
TGCTGT	1	0.57	32	0.39	5.3	5.3
TGTTGC	1	0.57	127	1.53	21.2	21.2
TTCATC	2	1.14	80	0.96	6.5~6.8	6.7
TTCTGA	1	0.57	28	0.34	4.7	4.7

Continued

TTCTTC	1	0.57	183	2.20	30.5	30.5
TTGCTG	1	0.57	27	0.32	4.5	4.5
TTTGTG	1	0.57	29	0.35	4.8	4.8
TTTTGT	1	0.57	33	0.40	5.5	5.5
Subtotal	47	26.86	2 373	28.56	4.2~56.2	8.4

Table 2. Copy number distribution of different motifs of Microsatellite DNA in the genome of yeast recombinant Han0458
表 2. 重组菌株 Han0458 基因组微卫星 DNA 不同模体的拷贝数分布

拷贝数范围 Copy No. range	3~15	15~27	27~39	>39	Total
Dinucleotide			3		3
Trinucleotide	61	27	8	1	97
Tetranucleotide	9	3	1		13
Pentanucleotide	15				15
Hexanucleotide	43	2	1	1	47
Total	128	32	13	2	175
Percent (%)	73.14	18.29	7.43	1.14	100

注: Dinucleotide: 两碱基; Trinucleotide: 三碱基; Tetranucleotide: 四碱基; Pentanucleotide: 五碱基; Hexanucleotide: 六碱基。

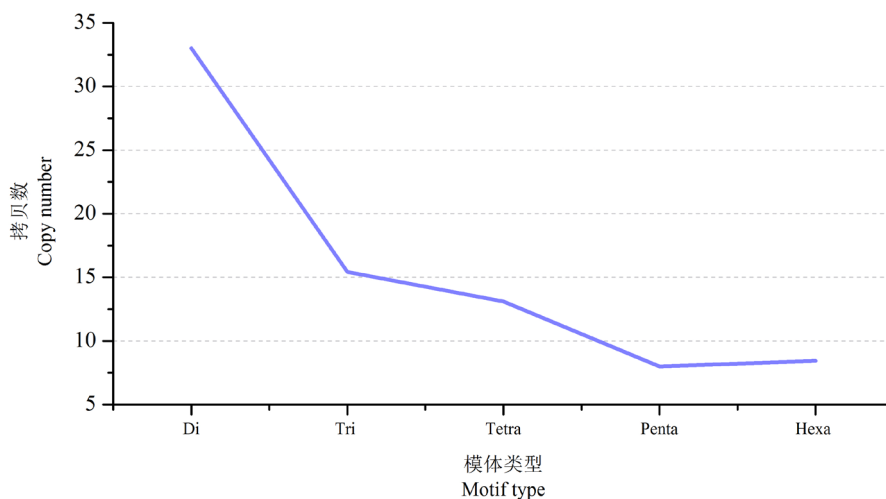


Figure 2. Relationship between repeat unit length and copy number of Microsatellite DNA in the genome of yeast recombinant Han0458

图 2. 重组菌株 Han0458 基因组微卫星 DNA 的重复单元长度与拷贝数关系

3.2. 小卫星 DNA 序列的分布特征

利用 TRF 方法在重组菌株 Han0458 基因组的串联重复序列中发现了 1361 条小卫星 DNA 序列, 总长分别为 74,362bp, 占串联重复序列 72.71%, 占基因组序列总长的 0.54%, 平均每 10 Kb 出现一个小卫星序列。

小卫星 DNA 序列集中分布于 25 bp~74 bp, 占小卫星 DNA 序列总数的 78.55%, 其中长度为 30 bp 的小卫星序列数目最多, 为 76 条; 重复单位为 15 bp 的序列数目最多, 高达 192 条, 占小卫星序列总数的 14.11%; 以 18 bp 为重复单元的序列长度最长, 8258 bp, 占小卫星序列总长的 11.11%; 各重复单元的

拷贝数范围为 1.9~20.7, 平均拷贝数为 2.6 的重复序列数目最多。小卫星 DNA 序列的重复类型数目、序列长度及拷贝数如图 3 所示。

小卫星 DNA 序列数目与重复单位长度有一定关系, 随着重复单位长度的增加呈下降趋势, 这一特征在重复单元大于 15 bp 的小卫星序列中尤为显著; 与微卫星 DNA 类似, 小卫星 DNA 序列重复单位拷贝数较低, 主要分布在 1~3 次; 重复单元拷贝数与小卫星 DNA 序列之间无显著相关关系(图 4)。

3.3. 散在重复序列分布特征

运用 RepeatMasker 方法, 获得了重组菌株 Han0458 的多种散在重复序列(表 3), 其在基因组中占比很小, 仅为 0.79% 左右。其中长末端重复序列(LTR)数目最多, 为 690 条, 占总数的 46.34%; 其次是 DNA 转座子, 为 440 个; 长散在重复序列(LINE)共 315 条; 而短散在重复序列(SINE)只有 25 条; 滚环(RC) 13 个。

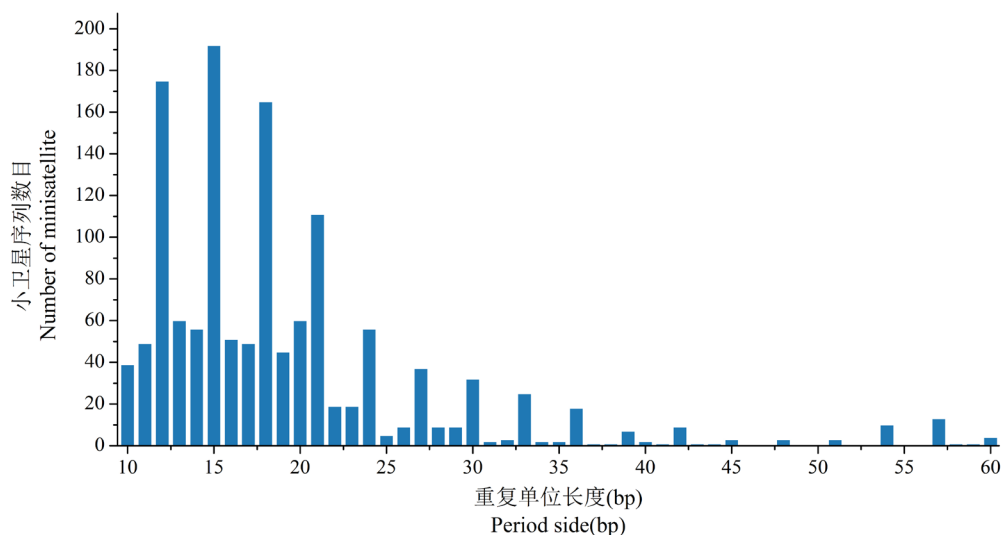


Figure 3. Relationship between period side and quantity of minisatellite DNA sequence in the genome of yeast recombinant Han0458

图 3. 重组菌株 Han0458 基因组小卫星 DNA 序列的重复单元长度与其数量关系

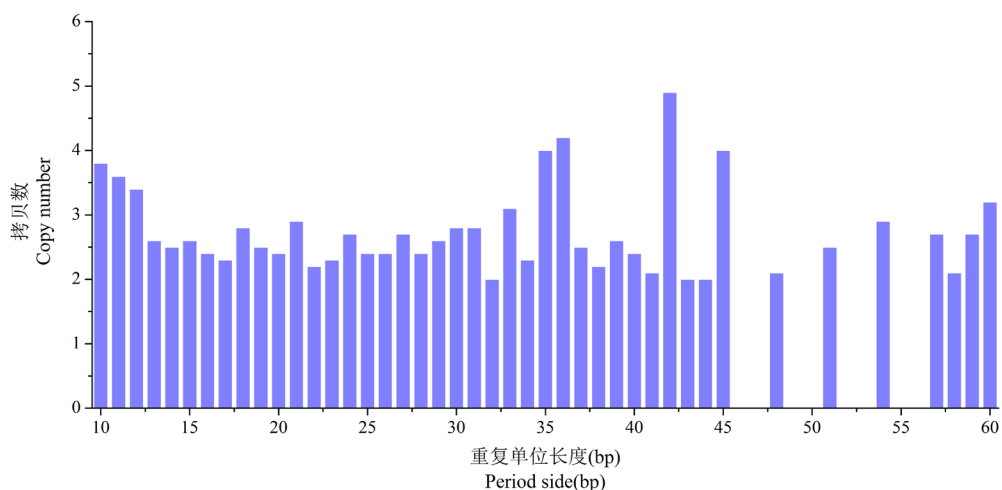


Figure 4. Relationship between period side and copy number of minisatellite DNA sequence in the genome of yeast recombinant Han0458

图 4. 重组菌株 Han0458 基因组小卫星 DNA 序列的重复单元长度与其拷贝数关系

Table 3. Distribution of interspersed repeats in the genome of the yeast recombinant Han0458
表 3. 重组菌株 Han0458 基因组中散在重复序列的分布

重复类型 Repeat type	数目 No.	总长度(bp) Total length (bp)	占基因组大小百分比(%) Repeat length/Genome (%)	平均长度(bp) Average length (bp)
LTR	690	51,858	0.3800	77
DNA	440	31,782	0.2329	77
LINE	315	26,919	0.1973	89
SINE	25	1444	0.0106	61
RC	13	1428	0.0105	110
Unknown	6	643	0.0047	107
Total	1489	107,439	0.7873	79

分析结果表明,各类型散在重复的总长度分布与数目分布保持一致,其长度大小关系为 LTR > DNA > LINE > SINE > RC, 各占散在重复序列总长的 48.27%、29.58%、25.01%、1.34%和 1.33%。值得注意的是,虽然 RC 的重复序列数目较少,但其平均长度约为 SINE 的两倍。

4. 讨论与结论

重复序列是基因组的重要组成部分,对生物的进化、遗传和基因的表达与调控有重要作用。重复序列的出现,说明基因组中的遗传物质在不断地进行自我复制,并进行水平交换和垂直交换,对丰富生物的遗传信息具有重要作用[8]。生物体中许多关键基因是单拷贝的,重复序列的存在能保护这些重要的基因结构不受破坏,同时也是新基因产生的物质基础,是驱动生物进化的重要因素之一[9]。

本研究利用生物信息学方法分析了来源于异常汉逊酵母菌的离子束重组菌株 Han0458 基因组重复序列在其基因组中的分布及特征,结果表明,重组酵母菌株 Han0458 基因组微卫星 DNA 序列在其基因组中的占比不到千分之一,重复单元的拷贝数大多低于 15 个,重复单位长度与其拷贝数间存在着负相关;优势重复类型为三碱基重复, AAC 为所有微卫星 DNA 类型中数目最多的基序;两碱基重复序列数目最少,且仅有 AT 重复。在微卫星 DNA 和小卫星 DNA 中,AT 含量均大于 50%。这与 Edwards 等人的研究结果一致,AT 类型重复的在植物、酵母和真菌类串联重复序列中的频率最高[10]。

重组酵母菌株 Han0458 基因组重复序列在基因组中含量较少,约为 2.5%左右,对其转座元件的分析发现, RNA 转座子的数目与长度均高于 DNA 转座子,与其他真菌的研究结果一致[11]。

三碱基类型重复和六碱基类型重复是重组酵母菌株 Han0458 基因组微卫星 DNA 序列的优势碱基类型,因此,在设计 SSR 引物时,可优先考虑从这两种重复类型入手。

低能离子注入对微生物的系统发育与进化具有重要的驱动作用[12],本研究结果为进一步认识低能离子注入对酵母菌基因组结构的生物学效应提供了素材,同时为低能离子注入介导的酵母菌基因组突变与进化提供了分子证据。

致 谢

本研究由国家自然科学基金(11575149 和 31760016)和陕西科技大学博士科研项目(126021759)资助。北京诺禾致源生物信息科技股份有限公司微生物事业部协助进行离子束重组酵母菌株 Han0458 全基因组 *de novo* 测序。

参考文献

- [1] Li, X., Kahveci, T. and Settles, A.M. (2007) A Novel Genome-Scale Repeat Finder Geared towards Transposons. *Bio-*

- informatics*, **24**, 468-476. <https://doi.org/10.1093/bioinformatics/btm613>
- [2] Férandon, C., Peuchant, O., Renaudin, H., *et al.* (2013) Diversity of *Mycoplasma hominis* Clinical Isolates from Bordeaux, France, as Assessed by Multiple-Locus Variable-Number Tandem Repeat Analysis. *BMC Microbiology*, **13**, 120. <https://doi.org/10.1186/1471-2180-13-120>
- [3] 杨惠. 条斑紫菜功能基因组及重复序列特征研究[D]: [博士学位论文]. 青岛: 中国海洋大学, 2011: 10.
- [4] Sukumaran, S. and Grant, A. (2013) Effects of Genotoxicity and Its Consequences at the Population Level in Sexual and Asexual *Artemia* Assessed by Analysis of Inter-Simple Sequence Repeats (ISSR). *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, **757**, 8-14. <https://doi.org/10.1016/j.mrgentox.2013.03.015>
- [5] 韩丽娟. 比较研究啮齿类动物基因组中的散在重复序列[J]. 科技通报, 2012, 28(12): 30-31.
- [6] Lü, J., Jin, X., Mao, P.H., *et al.* (2009) Transfer of Ephedra Genomic DNA to Yeasts by Ion Implantation. *Applied Biochemistry and Biotechnology*, **158**, 571-581. <https://doi.org/10.1007/s12010-009-8597-9>
- [7] 毛培宏, 吕杰. 离子束重组酵母菌 Ar_Han0458 的 RAPD 与 SSH 的初步研究[J]. 基因组学与应用生物学, 2015, 34(3): 449-453.
- [8] Eichler, E.E. and Sankoff, D. (2003) Structural Dynamics of Eukaryotic Chromosome Evolution. *Science*, **301**, 793-797. <https://doi.org/10.1126/science.1086132>
- [9] 艾对元. 基因组中重复序列的意义[J]. 生命的化学, 2008, 28(3): 343-345.
- [10] Edwards, K.J., Barker, J.H., Daly, A., *et al.* (1996) Microsatellite Libraries Enriched for Several Microsatellite Sequences in Plants. *Biotechniques*, **20**, 758-760. <https://doi.org/10.2144/96205bm04>
- [11] Kim, J.M., Vanguri, S., Boeke, J.D., *et al.* (1998) Transposable Elements and Genome Organization: A Comprehensive survey of Retrotransposons Revealed by the Complete *Saccharomyces cerevisiae* Genome Sequence. *Genome Research*, **8**, 464-478. <https://doi.org/10.1101/gr.8.5.464>
- [12] 蔡长龙, 吕杰, 毛培宏, 等. 低能氮离子注入驱动 DOB 的 16S rRNA 基因突变与进化[J]. 西安工业大学学报, 2017, 27(11): 288-293.

知网检索的两种方式:

1. 打开知网页面 <http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD>
下拉列表框选择: [ISSN], 输入期刊 ISSN: 2327-0810, 即可查询
2. 打开知网首页 <http://cnki.net/>
左侧“国际文献总库”进入, 输入文章标题, 即可查询

投稿请点击: <http://www.hanspub.org/Submission.aspx>

期刊邮箱: amb@hanspub.org