

# 棉属四倍体 AD<sub>1</sub> 与二倍体 A<sub>2</sub>、D<sub>5</sub> 基因组的同源 SSR 分析

孙高飞<sup>1,2</sup>, 何守朴<sup>1</sup>, 潘兆娥<sup>1</sup>, 杜雄明<sup>1</sup>

1. 中国农业科学院棉花研究所, 棉花生物学国家重点实验室, 安阳 455000;
2. 安阳工学院计算机科学与信息工程学院, 安阳 455000

**摘要:** SSRs(Simple sequence repeats)是一类广泛存在于动植物基因组的 DNA 短串联重复序列, 是重要的基因组分子标记。

比较不同基因组同源 SSR 的差异, 有利于了解相近物种间的进化过程。文章使用雷蒙德氏棉基因组(D<sub>5</sub>)、亚洲棉基因组(A<sub>2</sub>)全基因组序列和陆地棉(AD<sub>1</sub>)的限制性酶切基因组测序数据, 进行全基因组 SSR 扫描, 比较了 A 组和 D 组的 SSR 分布情况, 通过识别 3 个基因组之间的同源 SSR, 比较它们之间同源 SSR 重复序列的差异。结果发现, A 组和 D 组同源 SSR 的分布规律非常相似, 但 A 组与 AD 组的同源 SSR 保守性比 D 组与 AD 组同源 SSR 的保守性强。与 AD 组同源 SSR 相比, A 组中重复序列长度增长的 SSR 数量约为长度缩短的 SSR 数量的 5 倍, 在 D 组中这一比值约为 3 倍。可以推测, 四倍体 AD 组在与 A 组、D 组的平行进化过程中, 由于基因组融合, 导致 SSR 的重复序列长度变化速率与二倍体 A、D 组有差异, 同时这种差异可能导致了 AD 组 SSR 重复序列长度在进化过程中与二倍体相比有变短的趋势。文章首次对 3 个棉花基因组的同源 SSR 进行了系统地比较, 发现了同源 SSR 在棉属四倍体基因组和二倍体基因组中的显著差异, 为进一步揭示棉属基因组的进化规律提供了基础。

**关键词:** SSR; 棉花基因组; 同源 SSR; 进化

## Homologous simple sequence repeats (SSRs) analysis in tetraploid (AD<sub>1</sub>) and diploid (A<sub>2</sub>, D<sub>5</sub>) genomes of *Gossypium*

Gaofer Sun<sup>1,2</sup>, Shoupu He<sup>1</sup>, Zhaoe Pan, Xiongming Du<sup>1</sup>

1. State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang 455000, China;
2. School of Computer Science and Information Engineering, Anyang Institute of Technology, Anyang 455000, China

**Abstract:** Simple sequence repeats (SSRs) are a class of repetitive DNA sequences, which are commonly used for genome analysis. Comparison of the homologous SSRs among different genomes is helpful to understand the evolutionary process in relative species. In this study, SSR scanning was performed to investigate their distribution and length variation among the genomes of *G. raimondii* (D<sub>5</sub>), *G. arboreum* (A<sub>2</sub>) and *G. hirsutum* (AD<sub>1</sub>). The results demonstrated that the distribution of SSRs in A genome was very similar with that in D genome, while the length variation of homologous SSRs between A and AD genome was more conserved than that between D and AD genome.

收稿日期: 2014-08-15; 修回日期: 2014-09-24

基金项目: “十二五”国家支撑计划项目(编号: 2013BAD01B03)和科技部、财政部国家科技基础条件平台项目(编号: 2012-014)资助

作者简介: 孙高飞, 博士, 副教授, 研究方向: 棉花生物信息学。E-mail: sungaofei@sina.com

何守朴, 硕士, 助理研究员, 研究方向: 棉花种质资源学。E-mail: zephyr0911@126.com

孙高飞和何守朴并列第一作者。

通讯作者: 杜雄明, 博士, 研究员, 研究方向: 棉花种质资源学。E-mail: dujeffrey8848@hotmail.com

DOI: 10.16288/j.ycz.14-274

网络出版时间: 2014-12-15 8:49:34

URL: <http://www.cnki.net/kcms/detail/11.1913.R.20141215.0849.001.html>

Compared with SSRs in AD genome, the number of SSRs with longer motif length in A genome was about five times of those with shorter motif length, while it was about three times in D genome. This implied that the length variation rates of homologous SSRs between diploid cotton and tetraploid cotton were different during the parallel evolution due to the subgenome fusion, and the motif length of most SSRs in tetraploid genome tended to become shorter than homologous SSRs in diploid genome during the process of evolution. This study comprehensively compared the SSRs in three cotton genomes and revealed the significant difference among them, providing a foundation for further evolutionary study of *Gossypium* genome.

**Keywords:** SSR; cotton genome; homologous SSR; evolution

简单序列重复(Simple sequence repeats, SSRs) 又称微卫星(Microsatellites), 是一类由几个碱基组成的基序串联重复而成的 DNA 序列, 基序长度一般为 1~6 bp, 总长一般大于或等于 10 bp, 广泛分布于动植物基因组, 是重要的基因组分子标记。SSR 具有多态性强、长度小、易于快速检测等特点, 主要应用于动植物的分子标记开发、遗传图谱构建、基因定位等理论和应用研究<sup>[1,2]</sup>。

SSR 在生物进化研究中也扮演着重要角色, 在大规模基因组测序开始之前, 对于 SSR 的研究是通过 PCR 等实验方法获得同源位点 SSR, 通过检测序列长度的差异来分析物种间的遗传关系, 研究范围覆盖动物<sup>[3~6]</sup>和植物<sup>[7~10]</sup>。随着测序技术的发展, 完整测序的基因组越来越多, 对于 SSR 的研究也就更加的全面和深入, 基于物种之间、群体之间的 SSR 的比较研究不断涌现<sup>[11~16]</sup>。

棉花 SSR 分子标记的开发在近年获得了快速进步<sup>[17~19]</sup>, 成为棉花分子生物学研究中应用最为成功的分子标记之一, 广泛地应用于棉花种质资源的遗传多样性<sup>[20~22]</sup>、重要农艺性状的 QTL 定位<sup>[23~25]</sup>和全基因组关联分析等领域<sup>[26~28]</sup>。然而由于缺乏参考基因组, 同时已开发的 SSR 标记来源相对单一(大多数来源于纤维 EST 库), 因而对于棉花 SSR 在整个基因组上的分布和变化规律依然缺乏宏观的认识和研究。

通常认为, 现在栽培上所用的异源四倍体陆地棉(AD<sub>1</sub>)的两个亚基因组供体种为二倍体亚洲棉(A<sub>2</sub>)和雷蒙德氏棉(D<sub>5</sub>)<sup>[29]</sup>。近年来随着这两个二倍体基因组草图<sup>[30,31]</sup>和部分陆地棉基因组测序原始序列的公布<sup>[32]</sup>, 人们对这 3 个基因组的结构了解地更加深入, 但作为序列变异重要来源之一的 SSR, 特别是同源 SSR 在 3 个基因组之间的比较尚未见报道。本研究利用生物信息学方法, 对亚洲棉、雷蒙德氏棉

和陆地棉 3 个基因组同源 SSR 进行系统的比较和分析, 重点探讨了 3 个基因组之间同源 SSR 的差异, 以及产生这些差异的可能原因。

## 1 材料和方法

### 1.1 数据来源

雷蒙德氏棉基因组序列(以下简称 D 组)以及相关注释下载自 NCBI (<http://www.ncbi.nlm.nih.gov/assembly/519268/>), 亚洲棉基因组(以下简称 A 组)序列以及相关注释信息下载自 <http://cgp.genomics.org.cn/>。由于之前报道的陆地棉遗传图谱编号和两个二倍体全基因组测序编号存在差异, 为了更直观分析 SSR 的同源性, 本文首先根据相关文献对这些编号进行了整合<sup>[33]</sup>(表 1)。陆地棉基因组(以下简称 AD 组)测序数据来源于 NCBI 中 <http://www.ncbi.nlm.nih.gov/bioproject/168346>。该基因组测序使用 6 种不同的陆地棉品种和两种不同的酶切方法, 形成 12 个不同的测序序列样本, 在本文中以下划线连接样本名称和内切酶名称做为陆地棉序列库的名称。

另外根据全基因组测序的染色体和基因注释信息, 本文将基因组上的基因区域划分为外显子区(CDS)、内含子区(intron)、5'UTR 区、3'UTR 区、基因上游 1000 bp(1 K)、基因下游 1000 bp(1K)和非编码区(由于注释的原因, A 组没有 5'UTR 区、3'UTR 区)。使用 perl 语言编程, 定位每个 SSR 在 A 组和 D 组上所在的基因区域, 以便对各基因区域内包含的 SSR 数量进行比较。

### 1.2 全基因组 SSR 扫描

基于 Perl 的 MISA 程序(<http://pgrc.ipk-gatersleben.de/misa/>)对 A 组、D 组和 AD 组的基因组序列进行扫描, 按照默认参数, 识别最少为 10 次的单碱

表 1 本研究使用的棉花染色体编号与遗传图谱和基因组测序的染色体编号对比

遗传图谱编号	全基因组测序编号	本研究编号	遗传图谱编号	全基因组测序编号	本研究编号
Chr.1	Ca1	A01	Chr.15	Chr02	D01
Chr.2	Ca2	A02	Chr.14	Chr05	D02
Chr.3	Ca3	A03	Chr.17	Chr03	D03
Chr.4	Ca4	A04	Chr.22	Chr12	D04
Chr.5	Ca5	A05	LGD02/Chr.19	Chr09	D05
Chr.6	Ca6	A06	Chr.25	Chr10	D06
Chr.7	Ca7	A07	Chr.16	Chr01	D07
LGA02/Chr.8	Ca8	A08	LGD03/Chr.24	Chr04	D08
Chr.9	Ca9	A09	Chr.23	Chr06	D09
Chr.10	Ca10	A10	Chr.20	Chr11	D10
LGA03/Chr.11	Ca11	A11	LGD02/Chr.21	Chr07	D11
Chr.12	Ca12	A12	Chr.26	Chr08	D12
LGA01/Chr.13	Ca13	A13	Chr.18	Chr13	D13

基重复和最少 5 次的 2、3、4、5、6 碱基重复为 SSR。作为一个 SSR 认定的两个重复序列之间的碱基数不超过 100 bp。

### 1.3 同源 SSR 识别

将 MISA 扫描获得的数据分别建立 SSR 序列库。使用 perl 编写脚本分别提取 A 组、D 组 SSR 基序两侧各 100 bp 碱基序列,形成长度超过 210 bp 的 SSR 识别序列。利用 BLAST 工具<sup>[34]</sup>,将同源 SSR 识别序列映射到目标基因组进行匹配。由于 BLAST 的比对算法会产生不同长度的比对结果,在这些结果中,只有匹配长度达到一定的长度才能认为两个 SSR 识别序列是同源的。对于 A 组到 D 组、A 组到 AD 组、D 组到 AD 组的匹配记录,以匹配长度为观察参数,以 10 bp 为区间,分别计算匹配长度的分布(图 1),可见,匹配长度在 190 bp 位置的分布数量迅速上升。考虑到侧翼序列有 200 bp,取 190 bp 作为匹配长度,这时序列匹配度可达 95%,因此选择 190 bp 作为认定 SSR 识别序列同源的阈值。如果 A 组的一个 SSR (A\_SSR)识别序列和 D 组的一段序列同源,则称这段序列是 A\_SSR 识别序列在 D 组的同源匹配序列,如果该同源匹配序列中同时也存在一个 SSR(D\_SSR),则认为 A\_SSR 和 D\_SSR 为同源 SSRs。从 A 组到 D 组识别同源 SSRs,以及从 A 组到 AD 组、D 组到 AD 组识别同源 SSRs,均采用这一标准。

### 1.4 同源 SSR 的重复序列比对

通过 MISA 扫描获得的 SSR 重复类型有 8 种  $c$  和  $c^*$ ,

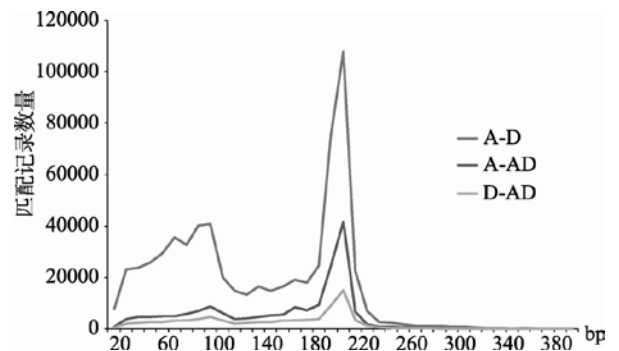


图 1 3 个基因组之间 SSR 识别序列 Blast 结果中不同匹配长度的数量分布

$p1, p2, p3, p4, p5, p6$ 。其中  $c$  和  $c^*$  是组合型的 SSR,即由两个或两个以上的重复序列组成。其中  $c^*$  重复序列之间没有或只有一个其它碱基,而  $c$  类型的 SSR 重复序列之间包含若干个碱基。相对于组合型 SSR,简单类型 SSR 是指由一个重复基序经过多次重复形成的重复序列, $p_n$  中的  $n$  是指基序的碱基数,例如  $p2$  值基序为 2 bp 的重复序列。

同源 SSR 中的重复序列总体上可以分为:重复类型不同和重复类型相同两种情况。本文将针对这两种不同的情况分别进行分析。

### 1.5 同源 SSR 类型差异统计

SSR 变化的过程和机制目前还不清楚,因此,本文对于基因组之间同源 SSR 类型的差异情况只通过比较不同位置的 SSR 类型差异数量进行统计分析。

## 1.6 同源 SSR 类型相同的情况

对于类型相同的同源 SSR, 本文对匹配序列的 A 组和 D 组的 SSR 进行了分析, 将两个 SSR 序列进行如下的分类: (1) 两个 SSR 的基序没有发生变化, 只是重复基序的次数有变化; (2) 两个 SSR 的基序不同, 但是这个不同是由于重复序列的起始位置的若干个碱基发生突变, 导致重复基序产生了碱基的滑动, 这种情况本文定义为基序移位, 例如 AGC 和 GCA, 这种基序移位在和统计中认定为是基序相同; (3) 两个 SSR 的基序完全不同, 例如 AAT 和 GTC。

由于 BLAST 比对时, 主要是以 SSR 两端的侧翼序列为识别序列, 序列比对时已经确定了两个同源序列的匹配方向, 因此 SSR 的方向应该和同源序列的方向是一致的。如果匹配序列的方向是反向的, 则需将其中一个 SSR 基序的互补序列与另外一个 SSR 基序进行比较。例如基序为 AAG 和 CTT 的两个 SSR, 如果两翼序列匹配方向是反向的, 则是完全相同的基序序列, 如果两翼序列是正向匹配, 则这两个 SSR 基序是完全不同的。在对同源 SSR 进行长度对比时, 将基序不同的记录剔除, 因为这样的记录中的两个 SSR 不是真正同源的 SSR, 很可能是在同源位点分别进化的 SSR。

## 2 结果与分析

### 2.1 A 基因组和 D 基因组 SSR 分布

以 1Mb 长度作为区间来研究 SSR 在 A 组和 D 组中的分布情况, 同时根据两个基因组的注释, 研究同区间的基因分布。结果发现, SSR 和注释基因无论在位置还是数量上均高度一致, 并且单位区间内 SSR 数量和基因数量高度相关(图 2), 相关系数分别为 A ( $r=0.913^{***}$ ) 和 D ( $r=0.939^{***}$ )。

本文在 A 组中定位到 326664 个 SSR (平均~4.69 kb 一个), 在 D 组中定位到 191377 个 SSR (平均约~3.92 kb 一个)。每条染色体中 SSR 的数量和其所在区域的长度高度相关(表 2)。结果显示, 外显子区的 SSR 数量明显低于其他区域, A 组中的数量为 2141 个, 而 D 组中 SSR 数量为 2326 个, 但在与外显子区长度相当的内含子区域, 却有将近 10 倍数量的 SSR(A 组为 26 231 个, D 组为 23 892 个)。由于 A 组的 UTR

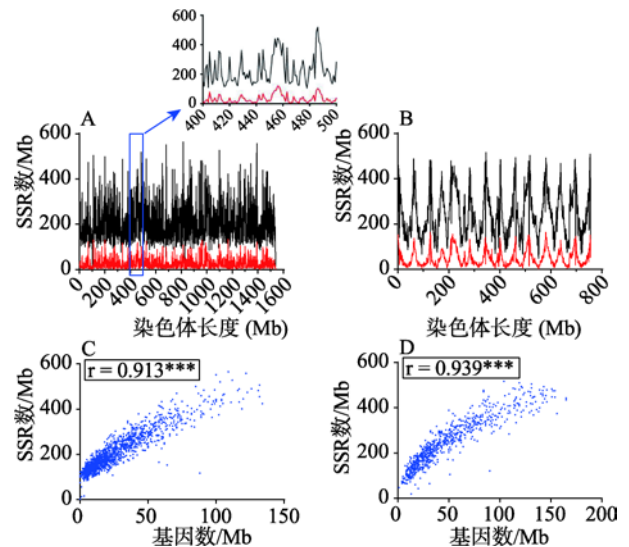


图 2 A 组、D 组 SSR 分布图

A: A 组 SSR 和基因在染色体上的分布; B: D 组 SSR 和基因在染色体上的分布; C: A 组单位区间内 SSR 数量和基因数量的相关性; D: D 组单位区间内 SSR 数量和基因数量的相关性。

区未注释, 在进行比较时, 将 D 组的 5' UTR 数据合并至基因上游 1 K, 将 3' UTR 数据合并至基因下游 1 K。

由于 SSR 的数量和所在区域的长度有密切关系, 为了更准确地了解 SSR 的分布规律, 本文计算了 A 组、D 组各基因区域的长度, SSR 的数量和分布密度(表 2)。由表 2 可知, CDS 区 SSR 的密度最小, 而在其他区域中, 上游 1 K 的 SSR 密度最高, 基因内含子和下游 1 K 的 SSR 密度也较高, 这很可能是与 SSR 参与基因区域的转录调控有关<sup>[35]</sup>。

为了进一步了解不同基序类型 SSR 在基因组不同区域的分布特点, 本文对 A 组、D 组在各区域内不同基序类型 SSR 的数量所占比例进行了比较(图 3)。A 组、D 组在基因上游 1 K、基因下游 1 K、基因内含子和非编码区的 SSR 类型分布非常一致, 不同类型 SSR 所占比例从高到低依次是 p1、p2、c、p3、p4、c\*、p5 和 p6。CDS 区的类型分布则和其他区域完全不同, p3 类型占据了绝对的优势, 这是因为在 SSR 的多态性类型中最常见的是重复序列的扩增和缩减, 在 CDS 中除了 p3 和 p6 外, 其他类型 SSR 的基序重复次数的变化, 都可能导致基因阅读框的变化, 从而改变基因翻译后的蛋白结构, 因此会受到进化选择的限制。而且由于基序长度的限制, p6 出现的概率明显小于 p3, 所以 p3 在 CDS 区的比例最大。另外一个略为突出的现象就是在非编码区,

表 2 A 组、D 组各基因区域 SSR 的数量与密度

基因相关区域	区域长度(bp)		SSR 数量		SSR 密度(个/Mb)	
	A 组	D 组	A 组	D 组	A 组	D 组
基因外显子	43703865	41536944	2141	2326	49.0	56.0
基因内含子	53177480	49745507	26231	23892	493.3	480.3
基因上游 1 K	40134000	43532929	22364	21299	557.2	489.3
基因下游 1 K	40134000	47744978	17977	17747	447.9	371.7
非编码区	1354888248	549835825	282562	128160	208.6	233.1

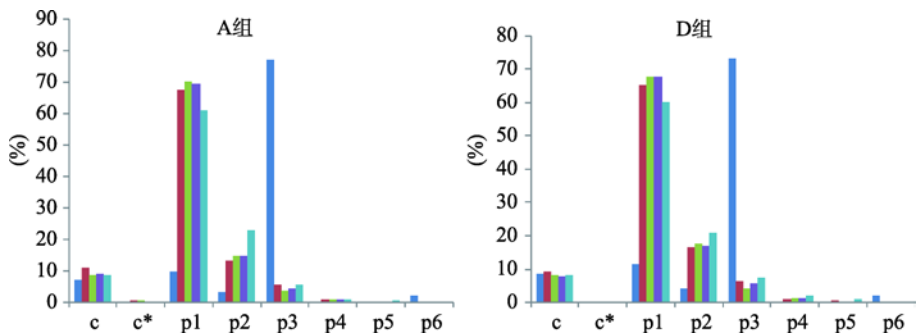


图 3 不同基序类型 SSR 在 A 组、D 组各基因区域的分布  
■ 基因外显子；■ 基因上游 1k；■ 基因内含子；■ 基因下游 1k；■ 非编码区。

p2 类型的比例比其他区域略高。

2.2 A、D 两个基因组中同源 SSR 的分析

2.2.1 A、D 基因组同源 SSR 的染色体分布

通过同源 SSR 识别方法,在 A 组和 D 组中获得 60037 个同源 SSR 记录,每个记录都包含来自 A 组和 D 组的两个同源 SSR,其中 57275 个同源记录都定位在组装的染色体上(部分定位在未组装的 scaffold 上)。根据同源记录在两个基因组染色体上的数量对应关系,本文构建了 A 组和 D 组 SSR 分布的关联图(图 4)。

由于 SSR 在染色体上的分布相对均匀,SSR 及侧翼序列可以看成基因组序列的部分抽样,同源 SSR 的对应关系和其在两个基因组上的分布能够部分反映两个基因组中染色体的同源性。从图 4 可以看出,在 A 组和 D 组的 13 个染色体中,大部分同源 SSR 分布于相应的同源染色体上,说明先前认定的同源染色体是可靠的。但是 A01、A03、D01、D034 个染色体之间有较大的同源 SSR 交换,尤其是 A03 和 D01 的同源 SSR 数量(1709)超过了 A03 和 D03 同源 SSR 的数量(1015),这说明染色体 A03 与 D01 之间 SSR 发生了大量的移位,从而使其同源性更高。这 4 条染色体的同源关系还有待进一步确认。

2.2.2 A、D 基因组同源 SSR 的比较

在 A 组和 D 组 60037 对同源 SSR 中,有 10903 对的重复类型不同。为了比较清晰地比较 A 组和 D 组之间的 SSR 的差异,本文将同源 SSR 记录分为两类:一是同源 SSR 重复类型不同的情况,一种是 SSR 重复类型相同的情况。在前者中重点比较 SSR 不同类型变化的数量,在后者中重点比较基序的长度比较。

在 10903 个同源 SSR 记录中,根据两个基因组不同的重复类型数量,比较重复类型差异比例(表 3),可以看到组合型 SSR 的重复类型差异比例最大,简单型 SSR 基序碱基数量越大,重复类型差异的比例越大。A 基因组中组合型 SSR 的重复类型差异比例高于 D 组,而 D 组的简单型 SSR 的重复类型差异比例均高于 A 组。

本文将基因组不同位置的 SSR 重复类型差异进行了统计(表 4),CDS 区 SSR 重复类型差异的比例明显偏低,不足其他区域类型差异比例的一半。

根据上述获得同源 SSR 的方法和基序比较方法,A 组和 D 组 SSR 同源且类型相同的同源记录有 49134 条,去掉其中基序不同的 2441 条记录,对剩余的 46693 条基序相同的同源 SSR 记录进行重复序列长度的比较(表 5)。比对 A 组和 D 组各基因区域同源 SSR 的重复序列长度,可以发现,在 CDS 区的



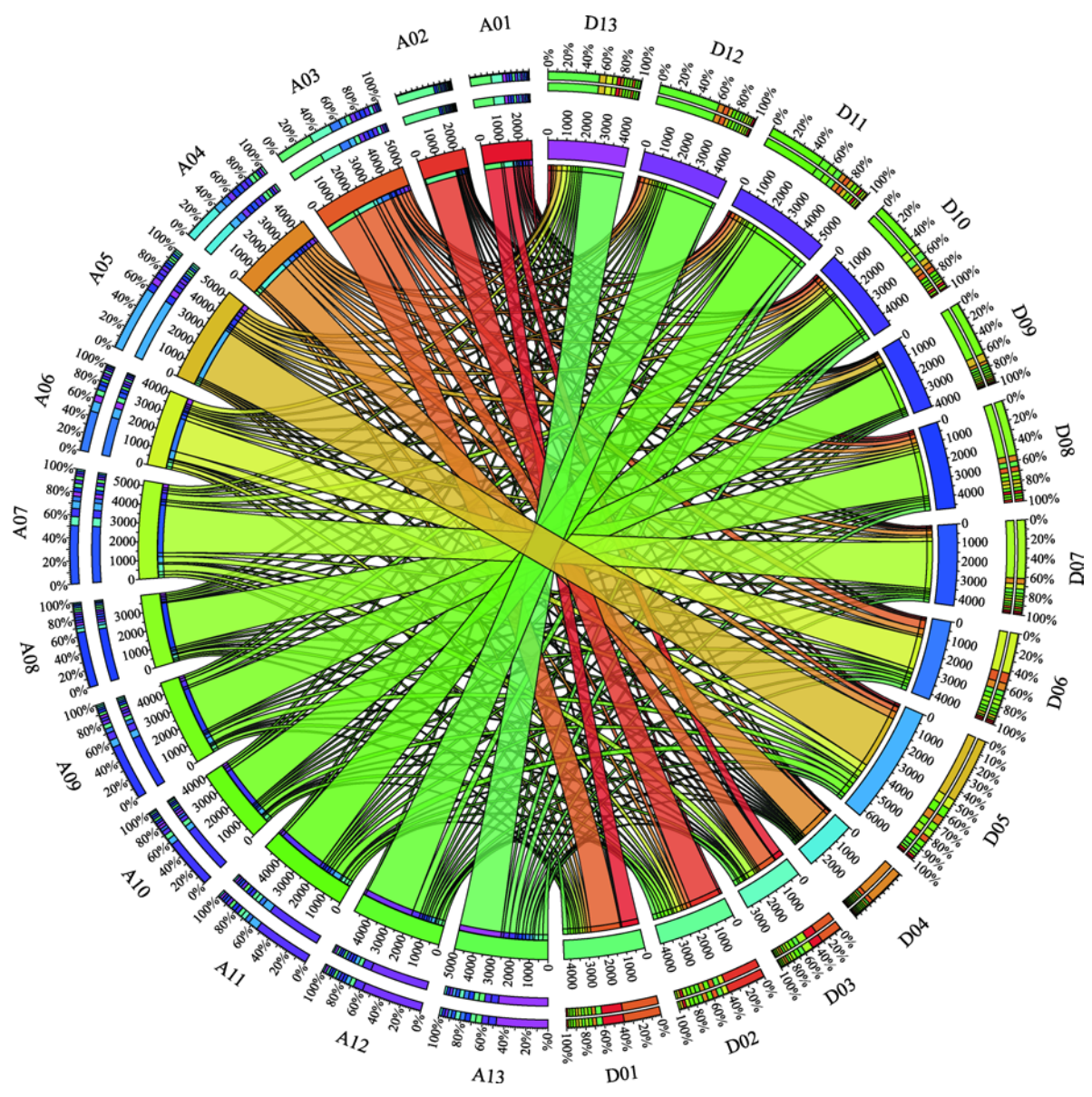


图 4 A 组和 D 组 SSR 分布关联图

表 3 A 组和 D 组重复类型差异的同源 SSR 统计

SSR 类型	A 组			D 组		
	同源 SSR 数量	重复类型差异数量	重复类型差异比例(%)	同源 SSR 数量	重复类型差异数量	重复类型差异比例(%)
c	6951	4598	66.10	5943	3590	60.40
c*	284	263	92.60	187	166	88.80
p1	35374	3583	10.10	35854	4063	11.30
p2	11172	1283	11.50	11558	1669	14.40
p3	5125	863	16.80	5218	956	18.30
p4	844	203	24.10	889	248	27.90
p5	203	72	35.50	291	160	55.00
p6	84	38	45.20	97	51	52.60

表 4 A 组和 D 组各基因区域重复类型差异的同源 SSR 分布

位置	A 组			D 组		
	重复类型差异数量	同源 SSR 数量	重复类型差异比例(%)	重复类型差异数量	同源 SSR 数量	重复类型差异比例(%)
CDS	123	1362	9.0	163	1479	11.0
基因下游 1 K	930	5164	18.0	688	3778	18.2
基因内含子	1798	11526	15.6	2014	12312	16.4
非编码区	6492	34713	18.7	5913	31968	18.5
基因上游 1 K	1560	7272	21.5	1143	5335	21.4
5'UTR				603	2700	22.3
3'UTR				379	2465	15.4

表 5 A 组与 D 组各基因区域的基序相同的同源 SSR 重复序列长度比较

A 组位置	A 组小于 D 组			相等		A 组大于 D 组			总数量
	数量	长度差 <sup>a</sup>	比例(%)	数量	比例(%)	数量	长度差	比例(%)	
基因外显子	310	-5.67	26.5	571	48.8	290	5.40	24.8	1171
基因内含子	3165	-4.67	33.7	2079	22.1	4147	4.62	44.2	9391
基因上游 1 K	1772	-4.68	33.1	1333	24.9	2242	4.31	41.9	5347
基因下游 1 K	1373	-4.68	33.8	895	22.0	1795	4.33	44.2	4063
非编码区	8140	-4.32	30.5	8606	32.2	9975	4.04	37.3	26721

注：<sup>a</sup>表中长度差为均值。

SSR 重复序列长度相等的比例是最大的,达到 48.8%,在其他的位 置,A 组 SSR 重复序列长度大于 D 组 SSR 重复序列的数量,要比小于 D 组的比例多 10%左右。

为了解基序相同的同源 SSR 的重复序列长度和重复类型是否存在关系,本研究比较了 A 组和 D 组各重复类型基序相同的同源 SSR 重复序列长度和重复类型(表 6)。组合型的 SSR 长度相等的比例远低于其他类型,这说明组合型类型 SSR 在两个基因组间变化很大。在简单型 SSR 中,A 组中 p1 类型 SSR 重复序列长度大于 D 组的数量远超过长度小于的数

量(41.02%:28.52%),其他类型在 A 组 SSR 重复序列长度大于 D 组 SSR 和 A 组 SSR 重复序列长度小于 D 组,在比例上总体相当。不同重复类型 SSR 所表现的重复长度差异没有明显的规律。

2.3 A 组、D 组和 AD 组 SSR 同源关系比较

2.3.1 总体比较

本研究使用 A 组 SSR 识别序列和 D 组 SSR 识别序列,分别对陆地棉(AD 组)12 个样本的测序序列进行了匹配定位,确定 A 组、D 组和 AD 组的同源

表 6 A 组与 D 组各重复类型的基序相同的同源 SSR 长度比较

SSR 类型	A 组小于 D 组			相等		A 组大于 D 组			总数
	数量	长度差	比例(%)	数量	比例(%)	数量	长度差	比例(%)	
c	572	-8.67	43.60	84	6.40	656	11.21	50.00	1312
c*	4	-5.75	26.67	1	6.67	10	6.10	66.67	15
p1	8759	-2.38	28.52	9354	30.46	12600	2.85	41.02	30713
p2	3846	-7.13	39.48	2073	21.28	3823	6.63	39.24	9742
p3	1384	-8.19	33.45	1628	39.34	1126	7.33	27.21	4138
p4	154	-8.00	25.16	259	42.32	199	6.09	32.52	612
p5	29	-12.07	23.39	74	59.68	21	6.19	16.94	124
p6	12	-16.50	32.43	11	29.73	14	12.86	37.84	37

SSR。由于 12 个样本中的 MCU\_5\_HpaII 读段数量不足其他样本数量的 1/3,导致与其他数据比较有较大差异,因此在进行统计时,剔除了样本 MCU\_5\_HpaII 的数据。

为了便于说明,本文将 3 个基因组之间的同源 SSR 关系绘制成图(图 5),A 组中和 AD 同源的 SSR 称为 A-AD 同源 SSR ;D 组中和 AD 组同源的 SSR,称为 D-AD 同源 SSR。同理,AD 组中和 A 组同源的 SSR,称为 AD-A 同源 SSR,AD 组中和 D 组同源的 SSR,称为 AD-D 同源 SSR。

将 AD 组 SSR 分成 3 个部分:只和 A 组 SSR 同源的,称为 AD-A 特有同源 SSR;只和 D 组 SSR 同源的,称为 AD-D 特有 SSR;和 A、D 组均有同源 SSR 的称为共有同源 AD 组 SSR。

AD-A 同源 SSR 数量=AD-A 特有同源 SSR 数量+共有同源 AD 组 SSR 数量;

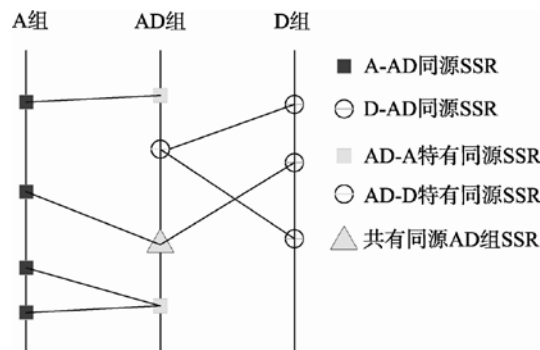


图 5 3 个基因组 SSR 同源关系图

AD-D 同源 SSR 数量=AD-D 特有同源 SSR 数量+共有同源 AD 组 SSR 数量;

依据以上的定义,A-AD 同源 SSR 数量平均是 D-AD 同源 SSR 数量的 2.76 倍,而 AD-A 同源 SSR 的数量平均是 AD-D 组同源 SSR 数量的 1.3 倍,AD-A 特有 SSR 的数量平均是 AD-D 特有 SSR 数量的 1.61 倍。

2.3.2 3 个基因组基序相同的同源 SSR 长度比较

分别对 A 组和 AD 组、D 组和 AD 组的基序相同的同源 SSR 的重复序列长度进行了比较,计算了比较后 A 组、D 组 SSR 重复序列长度小于、等于和大于 AD 组同源 SSR 记录的数量比例(表 8),并分别计算了同源 SSR 之间的重复序列长度差的平均值。

为验证 3 个基因组之间基序相同的同源 SSR 长度差异是否显著,分别对 3 个基因组之间的 SSR 长度做了 *t* 检验,其中 A 组和 D 组的同源 SSR 数量为 60036 对,*t* 检验 *P* 值为 3.97E-20,说明两个基因组 SSR 长度差异极为显著,A 组和 AD 组、D 组和 AD 组的 SSR 长度差异的显著性检验结果同样是极为显著(表 8)。

A 组和 AD 组同源 SSR 重复序列长度相等的数量高于 D 组和 AD 组同源 SSR 长度相等的数量比例,同时,A 组和 AD 组同源 SSR 重复序列长度不同时的长度差小于 D 组和 AD 组同源 SSR 重复序列长度不同时的长度差。

本文对 A、D 组各基因区域的 SSR 和 AD 组同源 SSR 的重复序列长度比较结果进行分别统计(图 6)。

表 7 A 组、D 组和 AD 组同源 SSR 数量统计

AD 组样本名称	A-AD	D-AD	A-AD/D-AD	AD-A	AD-D	AD-A/AD-D	AD 共有	AD-A 特有	AD-D 特有	AD-A/AD-D
JKC_703_claI	18825	7109	2.65	6242	4660	1.34	2413	3829	2247	1.70
JKC_703_HpaII	14139	5350	2.64	4826	4083	1.18	2229	2597	1854	1.40
JKC_725_claI	22505	8850	2.54	7939	5822	1.36	2860	5079	2962	1.71
JKC_725_HpaII	17229	5299	3.25	4817	3784	1.27	2066	2751	1718	1.60
JKC_737_claI	23099	8799	2.63	8201	6063	1.35	2952	5249	3111	1.69
JKC_737_HpaII	16362	5525	2.96	4513	3547	1.27	1723	2790	1824	1.53
JKC_770_claI	18749	6764	2.77	6268	4744	1.32	2276	3992	2468	1.62
JKC_770_HpaII	15600	6145	2.54	4882	3820	1.28	1971	2911	1849	1.57
LRA_5166_claI	18857	6738	2.8	5966	4629	1.29	2348	3618	2281	1.59
LRA_5166_HpaII	18399	7494	2.46	6249	5004	1.25	2643	3606	2361	1.53
MCU_5_claI	14230	4504	3.16	3536	2649	1.33	1431	2105	1218	1.73
平均			2.76			1.3				1.61



表 8 基序相同的同源 SSR 长度比较

AD 组样本名称	A-AD 同源 SSR 长度比较							D-AD 同源 SSR 长度比较						
	同源 SSR 数	<i>t</i> 检验 <i>P</i> 值	平均长度差(bp)		百分比(%)			同源 SSR 数	<i>t</i> 检验 <i>P</i> 值	平均长度差(bp)		百分比(%)		
			A<AD	A>AD	A<AD	A=AD	A>AD			D<AD	D>AD	D<AD	D=AD	D>AD
JKC_703_claI	15845	1.1E-113	-4.6	6.36	9.8	41.0	49.2	5844	1.1E-29	-5.01	7.72	15.9	36.9	47.3
JKC_703_HpaII	12076	2.9E-76	-4.31	5.96	10.4	43.6	45.9	4384	3.2E-20	-5.39	7.59	17.6	36.6	45.8
JKC_725_claI	19149	7.5E-72	-4.36	7.99	11.2	43.0	45.7	7104	8.2E-42	-4.63	7.88	15.5	37.6	46.9
JKC_725_HpaII	14874	6.7E-95	-3.18	5.89	14.6	40.4	45.1	4325	3.0E-21	-4.97	7.79	15.8	39.0	45.2
JKC_737_claI	19434	9.5E-92	-4.62	6.04	10.8	44.6	44.6	7196	1.2E-35	-4.36	7.96	15.3	37.1	47.6
JKC_737_HpaII	13906	4.9E-88	-4.39	7.06	8.1	42.3	49.4	4348	1.7E-28	-5.11	9.34	17.1	35.7	47.2
JKC_770_claI	15904	3.1E-82	-4.02	6.65	10.8	40.0	49.2	5573	3.2E-29	-4.72	7.22	14.9	37.5	47.6
JKC_770_HpaII	13175	5.0E-89	-4.19	6.35	9.4	39.0	51.6	4872	2.9E-26	-4.81	7.51	16.8	37.4	45.8
LRA_5166_claI	16222	8.9E-103	-3.85	5.83	11.0	43.4	45.6	5563	1.3E-22	-4.74	7.24	17.3	37.6	45.1
LRA_5166_HpaII	15071	1.7E-86	-4.06	5.86	10.8	38.0	51.0	6021	1.7E-31	-4.65	7.25	16.1	37.1	46.7
MCU_5_claI	11959	1.9E-22	-4.06	6.44	11.0	44.2	44.8	3592	3.4E-19	-4.57	7.43	16.2	40.0	43.8
平均	15238		-4.15	6.4	10.7	41.8	47.5	5347		-4.81	7.72	16.2	37.5	46.3

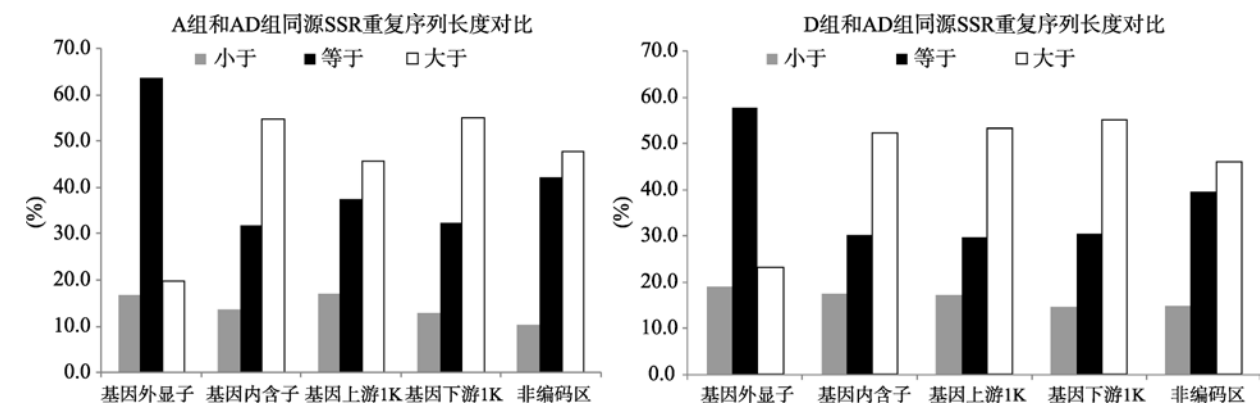


图 6 A 组、D 组和 AD 组各基因区域同源 SSR 重复序列长度对比

首先，无论 A 组和 D 组，重复序列长度大于 AD 组同源 SSR 的数量比例，远超过重复序列长度小于 AD 组同源 SSR 的数量比例；其次，在外显子区域，A、D 组和 AD 组重复序列长度相等的同源 SSR 比例最高，说明 SSR 在外显子区域在 3 个棉种之间都非常保守；最后，A、D 组重复序列长度小于 AD 组同源 SSR 的数量在各基因区域都比较接近。

3 讨 论

本研究中 3 个基因组虽然同属于棉属，但是在基因组大小上有较大的差异，且 AD 基因组的测序方法与 A 组、D 组的测序方法有所不同。要保证 3 个基因组间的 SSR 能够进行可信的比较，需要选取 3 个基因组中高度同源的 SSR 记录。

本研究使用 SSR 及两侧侧翼序列各 100 bp 作为识别序列到目标基因组进行匹配，并要求匹配长度大于等于 190 bp，这一匹配要求是非常严格的(实验使用的 PCR 引物长度一般为 20~30 bp)，满足匹配条件的两个识别序列具有高度的同源性。同时对于匹配到的同源序列，要求其必须同样具有 SSR 重复序列，这进一步保证了匹配序列为高度同源的 SSR 序列。由于比较只在同源 SSR 间进行，因此能够避免基因组大小、染色体倍数和测序手段差异对研究结果的影响，最大限度保证了研究结果的可靠性。

在对同源 SSR 记录进行重复序列长度比较之前，先对 SSR 基序进行了比对，剔除了基序不同的同源 SSR 记录，因此，最后进行 SSR 重复序列长度比较的同源 SSR 记录，其最主要差异就是重复序列长度

的差异,这部分同源记录中的 SSR 重复序列长度的变化保留了3个基因组在进化过程中留下的 SSR 重复序列长度变化的痕迹。

### 3.1 3个基因组 CDS 区的 SSR 非常保守

比较 A 组和 D 组 SSR 发现,无论在不同基因区域的 SSR 密度,还是不同重复类型 SSR 在基因区域所占的比例都非常接近。A 组和 D 组 CDS 区重复类型有差异的 SSR 比例不到其他基因区域比例的一半(表 4),在基序相同的同源 SSR 中,CDS 区域长度相等的比例是其他区域的两倍多(表 5),这些数据均说明 A 组和 D 组在 CDS 区域的同源 SSR 比其他区域保守得多。

A 组和 AD 组比较,CDS 区域长度相等的 SSR 所占比例为 63.6%,D 组和 AD 组比较,同源长度相等 SSR 所占比例为 57.8%,约是其他区域的两倍(图 6)。这也说明,3 个基因组 CDS 区域 SSR 的保守型都要远高于其他区域。

同时,各基因组 CDS 区域的 SSR 数量明显小于其他区域。以上现象可能是因为 SSR 的高度可变性容易导致外显子区阅读框的改变,从而影响基因的功能,由于进化的选择,大量 CDS 区域的 SSR 被淘汰并趋于稳定,从而导致外显子区 SSR 的高度保守<sup>[36,37]</sup>,这与已有研究的结果是相符的<sup>[8,38]</sup>。

### 3.2 就同源 SSR 数量和长度变化而言, A 组在同源性上比 D 组更接近 AD 组

A 组的 SSR 数量约是 D 组 SSR 数量的 1.8 倍(表 2),A-AD 同源 SSR 数量约是 D-AD 同源 SSR 数量的 2.76 倍(表 7),这说明和 D 组相比,A 组中 SSR 中有更高的比例与 AD 组同源的 SSR。

同样,AD-A 组同源 SSR 是 AD-D 同源 SSR 的 1.3 倍,而特有的 AD-A 同源 SSR 是特有 AD-D 同源 SSR 的 1.6 倍(表 7),这说明 AD 组和 A 组同源的 SSR 数量明显高于和 D 组同源 SSR 数量。A 组和 D 组的 SSR 的分布规律是非常相似的,因此,如果把 SSR 所在序列看成基因组的抽样,可以推测,A 组和 AD 组的 SSR 同源性高于 D 组和 AD 组的 SSR 同源性。另外,在同源且类型相同的 SSR 记录中,A 组和 AD 组同源 SSR 长度相同的比例(41.8%)要高于 D 组和 AD 组同源 SSR 长度相同的比例(37.5%)(表 8),这说明 A 组和 AD 组的同源 SSR 保守性比 D 组和 AD 组

的同源 SSR 保守性更高,这从另一个方面也印证了上述结论。

A 组小于 AD 组同源 SSR 的平均长度差为-4.15,而 D 组小于 AD 组同源 SSR 的平均长度差为-4.81;A 组大于 AD 组同源 SSR 的平均长度差为 6.4,而 D 组大于 AD 组同源 SSR 的平均长度差为 7.72(表 8)。D 组与 AD 组的同源 SSR 长度差均大于 A 组和 AD 组同源 SSR 的长度差,这也是 A 组与 AD 组的 SSR 同源性高于 D 组与 AD 组的 SSR 同源性的又一个佐证。

### 3.3 AD 组和 A、D 组相同基序的同源 SSR 长度变化的数量差异

A 组和 D 组来自共同的祖先,而 AD 组是 A 组和 D 组融合加倍而成,3 个基因组是平行进化的<sup>[30]</sup>。根据同源 SSR 的长度比较,A 组小于 D 组和 A 组大于 D 组的 SSR 数量比例基本是相等的(表 6),但是,A 组 SSR 长度大于 AD 组同源 SSR 长度的 SSR 数量,约是 A 组 SSR 长度小于 AD 组同源 SSR 长度的 SSR 数量的 5 倍,D 组和 AD 组比较的结果约为 3 倍(表 8)。产生这种情况,有两种可能:一种情况是 A 组和 D 组相对于 AD 组,大量 SSR 的长度增长了;另一种是 AD 组和 A 组、D 组相比,大量 SSR 的长度缩短了。无论是哪一种情况,都可以认为,AD 组的 SSR 长度变化速率与 A、D 组是不同的,而且,这种变化具有倾向性(多数 SSR 倾向于增长,或者缩短),因为只有这种倾向性,才有可能导致目前本文获得的 3 个基因组之间 SSR 长度差异情况。

关于人和黑猩猩<sup>[39]</sup>、羊和牛<sup>[3]</sup>的相近物种 SSR 长度差异现象很早就被提出,之后这种差异被解释为是由于测量偏差(Ascertainment bias)而导致的,同时也有观点认为测量偏差不能完全解释人和黑猩猩之间 SSR 的长度差异<sup>[40]</sup>。后来又有研究发现,人类除了两碱基重复的 SSR 的长度明显长于黑猩猩之外,其他重复类型的 SSR 没有发现明显的差异,而单碱基重复还发现了相反的趋势<sup>[41]</sup>。上述研究从不同的角度研究了相近物种之间的同源 SSR 长度差异,认为相近物种之间导致 SSR 长度差异的原因很可能是因为突变速率存在差异<sup>[11]</sup>。

由于本研究直接使用 SSR 侧翼序列匹配来获得同源 SSR,侧翼两端序列长度取值达到 200 bp,匹配长度不小于 190 bp,因此获得的同源 SSR 序列的

同源碱基比例超过 95%。同时,本研究是使用统一的条件来获得 3 个基因组内所有的同源 SSR,不属于抽样调查,因此本身可以排除测量偏差。在高度同源的 SSR 对比中,AD 组的 SSR 长度与 A 组、D 组相比,变长的 SSR 数量远低于变短的 SSR 数量。考虑到 A、D 组的 SSR 长度差异的数量比例比较接近,而且它们都是二倍体,而 AD 组是四倍体,这是和 A、D 组最为明显的差异。因此这种 SSR 长度变化在数量的差异,很有可能与 AD 组是四倍体而 A、D 基因组是二倍体有关。本文推测由于 AD 组是 A、D 基因组的融合,而这种融合过程导致了 SSR 长度变化速率的差异,而且这种差异导致了 SSR 长度变化的倾向性(大部分的 SSR 倾向于增长或缩短),进而形成了现有的 AD 组中大量的 SSR 长度小于 A、D 组同源 SSR 的现象。

在进化过程中,SSR 的长度受复制滑动事件和点突变等多因素影响<sup>[41,42]</sup>,因此 A、D 组中大量 SSR 相对于 AD 组同源 SSR 同步增长的概率要比 AD 组部分 SSR 的长度缩短的概率小,因此本文推测,四倍体的棉花 AD 组 SSR 相对于二倍体的 A、D 组同源 SSR 具有长度变短的倾向性。

## 参考文献

- [1] Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet*, 2004, 5(6): 435–445. [\[DOI\]](#)
- [2] Morgante M, Olivieri AM. PCR-amplified microsatellites as markers in plant genetics. *Plant J*, 1993, 3(1): 175–182. Ellegren H, Moore S, Robinson N, Byrne K, Ward W, Sheldon BC. Microsatellite evolution—a reciprocal study of repeat lengths at homologous loci in cattle and sheep. *Mol Biol Evol*, 1997, 14(8): 854–860. [\[DOI\]](#)
- [3] Webster MT, Smith NGC, Ellegren H. Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. *Proc Natl Acad Sci USA*, 2002, 99(13): 8748–8753. [\[DOI\]](#)
- [4] Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 1994, 368(6470): 455–457. [\[DOI\]](#)
- [5] 杨弘, 李大宇, 曹祥, 邹芝英, 肖炜, 祝璟琳. 微卫星标记分析罗非鱼群体的遗传潜力. *遗传*, 2011, 33(7): 768–775. [\[DOI\]](#)
- [6] Peakall R, Gilmore S, Keys W, Morgante M, Rafalski A. Cross-species amplification of soybean (*Glycine max*) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. *Mol Biol Evol*, 1998, 15(10): 1275–1287. [\[DOI\]](#)
- [7] Morgante M, Hanafey M, Powell W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet*, 2002, 30(2): 194–200. [\[DOI\]](#)
- [8] Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res*, 2001, 11(8): 1441–1452. [\[DOI\]](#)
- [9] 谢文刚, 张新全, 马啸, 彭燕, 黄琳凯. 鸭茅种质遗传变异及亲缘关系的 SSR 分析. *遗传*, 2009, 31(6): 654–662. [\[DOI\]](#)
- [10] Vowles EJ, Amos W. Quantifying ascertainment bias and species-specific length differences in human and chimpanzee microsatellites using genome sequences. *Mol Biol Evol*, 2006, 23(3): 598–607. [\[DOI\]](#)
- [11] Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res*, 2008, 18(1): 30–38. Chistiakov DA, Hellemans B, Volckaert FAM. Microsatellites and their genomic distribution, evolution, function and applications: a review with special reference to fish genetics. *Aquaculture*, 2006, 255(1–4): 1–29. [\[DOI\]](#)
- [12] La Rota M, Kantety RV, Yu JK, Sorrells ME. Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics*, 2005, 6(1): 23. [\[DOI\]](#)
- [13] Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 2008, 319(5866): 1100–1104. [\[DOI\]](#)
- [14] Sonah H, Deshmukh RK, Sharma A, Singh VP, Gupta DK, Gacche RN, Rana JC, Singh NK, Sharma TR. Genome-wide distribution and organization of microsatellites in plants: an insight into marker development in *Brachypodium*. *PLoS One*, 2011, 6(6): e21298. [\[DOI\]](#)
- [15] Wang HT, Li XM, Gao WH, Jin X, Zhang XL, Lin ZX. Comparison and development of EST-SSRs from two 454 sequencing libraries of *Gossypium barbadense*. *Euphytica*, 2014, 198(2): 277–288. Han ZG, Wang CB, Song XL, Guo WZ, Gou ZY, Li CH, Chen XY, Zhang TZ. Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton. *Theor Appl Genet*, 2006, 112(3): 430–439. [\[DOI\]](#)
- [16] Wang CB, Guo WZ, Cai CP, Zhang TZ. Characterization, development and exploitation of EST-derived microsatellites in *Gossypium raimondii* Ulbrich. *Chin Sci Bull*, 2006, 51(5): 557–561. [\[DOI\]](#)
- [17] Lacape JM, Dessauw D, Rajab M, Noyer JL, Hau B. Microsatellite diversity in tetraploid *Gossypium germplasm*: assembling a highly informative genotyping set of cotton

- SSRs. *Mol Breeding*, 2007, 19(1): 45–58. [\[DOI\]](#)
- [18] Alves MF, Barroso PA, Ciampi AY, Hoffmann LV, Azevedo VC, Cavalcante U. Diversity and genetic structure among subpopulations of *Gossypium mustelinum* (Malvaceae). *Genet Mol Res*, 2013, 12(1): 597–609. [\[DOI\]](#)
- [19] Liu DQ, Guo XP, Lin ZX, Nie YC, Zhang XL. Genetic diversity of Asian cotton (*Gossypium arboreum* L.) in China evaluated by microsatellite analysis. *Genet Resour Crop Ev*, 2006, 53(6): 1145–1152. [\[DOI\]](#)
- [20] Shen XL, Zhang TZ, Guo WZ, Zhu XF, Zhang XY. Mapping fiber and yield QTLs with main, epistatic, and QTL× environment interaction effects in recombinant inbred lines of upland cotton. *Crop Sci*, 2006, 46(1): 61–66. [\[DOI\]](#)
- [21] Mei M, Syed NH, Gao W, Thaxton PM, Smith CW, Stelly DM, Chen ZJ. Genetic mapping and QTL analysis of fiber-related traits in cotton (*Gossypium*). *Theor Appl Genet*, 2004, 108(2): 280–291. [\[DOI\]](#)
- [22] Jiang CX, Wright RJ, Woo SS, DelMonte TA, Paterson AH. QTL analysis of leaf morphology in tetraploid *Gossypium* (cotton). *Theor Appl Genet*, 2000, 100(3–4): 409–418. [\[DOI\]](#)
- [23] Jia YX, Sun XW, Sun JL, Pan Z, Wang XW, He SP, Xiao SH, Shi WJ, Zhou ZL, Pang BY, Wang LR, Liu JG, Ma J, Du XM, Zhu J. Association mapping for epistasis and environmental interaction of yield traits in 323 cotton cultivars under 9 different environments. *PLoS One*, 2014, 9(5): e95882. [\[DOI\]](#)
- [24] Abdurakhmonov IY, Kohel RJ, Yu JZ, Pepper AE, Abdullaev AA, Kushanov FN, Salakhutdinov IB, Buriev ZT, Saha S, Scheffler BE, Jenkins JN, Abdulkarimov A. Molecular diversity and association mapping of fiber quality traits in exotic *G. hirsutum* L. germplasm. *Genomics*, 2008, 92(6): 478–487. [\[DOI\]](#)
- [25] Kantartzi SK, Stewart JM. Association analysis of fibre traits in *Gossypium arboreum* accessions. *Plant Breeding*, 2008, 127(2): 173–179. [\[DOI\]](#)
- [26] Wendel JF, Brubaker C, Alvarez I, Cronn R, Stewart JM. Evolution and natural history of the cotton genus. In: Paterson AH, ed. *Genetics and Genomics of Cotton*. US: Springer, 2009: 3–22. [\[DOI\]](#)
- [27] Li FG, Fan GY, Wang KB, Sun FM, Yuan YL, Song GL, Li Q, Ma ZY, Lu CR, Zou CS, Chen WB, Liang XM, Shang HH, Liu WQ, Shi CC, Xiao GH, Gou CY, Ye WW, Xu X, Zhang XY, Wei HL, Li ZF, Zhang GY, Wang JY, Liu K, Kohel RJ, Percy RG, Yu JZ, Zhu YX, Wang J, Yu SS. Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat Genet*, 2014, 46(6): 567–572. [\[DOI\]](#)
- [28] Wang KB, Wang ZW, Li FG, Ye WW, Wang JY, Song GL, Yue Z, Cong L, Shang HH, Zhu SL, Zou CS, Li Q, Yuan YL, Lu CR, Wei HL, Gou CY, Zheng ZQ, Yin Y, Zhang XY, Liu K, Wang B, Song C, Shi N, Kohel RJ, Percy RG, Yu JZ, Zhu YX, Wang J, Yu SX. The draft genome of a diploid cotton *Gossypium raimondii*. *Nat Genet*, 2012, 44(10): 1098–1103. [\[DOI\]](#)
- [29] Rai KM, Singh SK, Bhardwaj A, Kumar V, Lakhwani D, Srivastava A, Jena SN, Yadav HK, Bag SK, Sawant SV. Large-scale resource development in *Gossypium hirsutum* L. by 454 sequencing of genic-enriched libraries from six diverse genotypes. *Plant Biotechnol J*, 2013, 11(8): 953–963. [\[DOI\]](#)
- [30] Wang K, Song XL, Han ZG, Guo WZ, Yu JZ, Sun J, Pan JJ, Kohel RJ, Zhang TZ. Complete assignment of the chromosomes of *Gossypium hirsutum* L. by translocation and fluorescence in situ hybridization mapping. *Theor Appl Genet*, 2006, 113(1): 73–80. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, 2004, 32(Web Server issue): W20–W25. [\[DOI\]](#)
- [31] Sawaya S, Bagshaw A, Buschiazzi E, Kumar P, Chowdhury S, Black MA, Gemmell N. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PLoS One*, 2013, 8(2): e54710. [\[DOI\]](#)
- [32] Tóth G, Gáspári Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res*, 2000, 10(7): 967–981. [\[DOI\]](#)
- [33] Loire E, Higuier D, Netter P, Achaz G. Evolution of coding microsatellites in primate genomes. *Genome Biol Evol*, 2013, 5(2): 283–295. [\[DOI\]](#)
- [34] Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol*, 2004, 21(6): 991–1007. [\[DOI\]](#)
- [35] Garza JC, Slatkin M, Freimer NB. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol Biol Evol*, 1995, 12(4): 594–603. [\[DOI\]](#)
- [36] Cooper G, Rubinsztein DC, Amos W. Ascertainment bias cannot entirely account for human microsatellites being longer than their chimpanzee homologues. *Hum Mol Genet*, 1998, 7(9): 1425–1429. [\[DOI\]](#)
- [37] Kruglyak S, Durrett RT, Schug MD, Aquadro CF. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci USA*, 1998, 95(18): 10774–10778. [\[DOI\]](#)
- [38] Santibáñez-Koref MF, Gangeswaran R, Hancock JM. A relationship between lengths of microsatellites and nearby substitution rates in mammalian genomes. *Mol Biol Evol*, 2001, 18(11): 2119–2123. [\[DOI\]](#)
- [39] Garza JC, Slatkin M, Freimer NB. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol Biol Evol*, 1995, 12(4): 594–603. [\[DOI\]](#)
- [40] Cooper G, Rubinsztein DC, Amos W. Ascertainment bias cannot entirely account for human microsatellites being longer than their chimpanzee homologues. *Hum Mol Genet*, 1998, 7(9): 1425–1429. [\[DOI\]](#)
- [41] Kruglyak S, Durrett RT, Schug MD, Aquadro CF. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci USA*, 1998, 95(18): 10774–10778. [\[DOI\]](#)
- [42] Santibáñez-Koref MF, Gangeswaran R, Hancock JM. A relationship between lengths of microsatellites and nearby substitution rates in mammalian genomes. *Mol Biol Evol*, 2001, 18(11): 2119–2123. [\[DOI\]](#)