甜瓜基因组GAPs特征及形成原因探究*

王孟文，王贤磊†，彭 媛，李泽玉

(新疆大学 生命科学与技术学院，新疆 乌鲁木齐 830017)

摘 要： 未知区域（GAPs）是基因组中未被测序或组装的区域. 经单分子三代测序，甜瓜基因组中GAPs由基因组V3.6.1中的79.68 Mb减少至基因组V4.0中的0.12 Mb. 以甜瓜基因组V3.6.1和V4.0数据为研究对象，获取并分析基因组中GAPs内部及侧翼序列特征及规律，并探究GAPs形成原因，为组装高质量甜瓜基因组提供参考. 结果表明：与基因组整体相比，GAPs内部两侧150 bp区域，简单重复序列（Simple Sequence Repeats, SSR）密度较高、非SSR区域GC含量较高，GAPs外侧150 bp含有较高的多拷贝序列. 较高GC含量和微卫星密度会影响PCR扩增，多拷贝序列的存在会影响下游序列的拼接组装，且在GAPs两侧150 bp与GAPs全序列比较发现，越临近GAPs边界，其GC含量与微卫星密度越高. 认为甜瓜基因组V3.6.1中GAPs形成的主要原因是含有较高GC含量、微卫星密度及较多的多拷贝序列；对V4.0基因组GAPs两侧序列比对分析发现，多拷贝序列占比为98.24%，多拷贝序列可能是V4.0中GAPs形成的重要原因.

关键词： 甜瓜；基因组；重复序列；SSR；GAPs

DOI： 10.13568/j.cnki.651094.651316. 2023.07.27.0002

中图分类号： S652 **文献标识码：** A **文章编号：** 2096-7675(2025)01-0066-07

引文格式： 王孟文，王贤磊，彭媛，李泽玉. 甜瓜基因组GAPs特征及形成原因探究[J]. 新疆大学学报(自然科学版中英文)，2025，42(1)：66-72.

英文引文格式： WANG Mengwen，WANG Xianlei，PENG Yuan，LI Zeyu. Exploring the causes of the formation of melon genome GAPs[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English)，2025，42(1)：66-72.

Exploring the Causes of the Formation of Melon Genome GAPs

WANG Mengwen, WANG Xianlei, PENG Yuan, LI Zeyu

(*School of Life Science and Technology, Xinjiang University, Urumqi Xinjiang 830017, China*)

Abstract： Unknown regions (GAPs) are regions of the genome that have not been sequenced or assembled. After single molecule real-time (SMRT) sequencing, the size of GAPs in melon genome was reduced from 79.68 Mb (genome V3.6.1) to 0.12 Mb (genome V4.0). Based on the genome V3.6.1 and V4.0 data, the internal and flanking sequence characteristics and rules of GAPs in the genome were obtained and analyzed, and the reasons for the formation of GAPs were explored, so as to provide a reference for assembling high-quality melon genomes. The results showed that the inner 150 bp region of GAPs had a higher density of simple sequence repeats (SSR), a higher GC content in the non-SSR region, and the outer 150 bp of GAPs contained more multi-copy sequences compared to the whole genome. Higher GC content and microsatellite density will affect PCR amplification, and the presence of multi-copy sequences will affect the splicing and assembly of downstream sequences, and a comparison of the 150 bp on both sides of the GAPs with the full sequence of the GAPs revealed that the closer to the GAPs boundary, the higher the GC content and the higher the microsatellite density. Therefore, the main reason for the formation of V3.6.1 GAPs in melon genome is that it contains higher GC content, microsatellite density and more multi-copy sequences. Sequence comparison analysis of both sides of GAPs in V4.0 genome revealed that the high ratio of multi-copy sequences (98.24%) may be the important reason for the formation of GAPs in V4.0.

Key words： melon; genome; repetitive sequence; SSR; GAPs

0 引 言

甜瓜（*Cucumis melo* L.）是葫芦科一年生蔓生草本，是开展果实发育[1]、植物性别分化[2]等研究的模式植物，尤其是甜瓜基因组数据的公布和完善，有力促进了甜瓜基因层面的研究进展. 2008年，454生命科学公司推

* **收稿日期：** 2023-07-27

基金项目： 国家自然科学基金“甜瓜抗果斑病基因的定位克隆与功能验证”（32160068）.

作者简介： 王孟文（1997—），女，硕士生，从事生物化学与分子生物学的研究，E-mail：2393189241@qq.com.

† **通讯作者：** 王贤磊（1981—），男，博士，主要从事甜瓜分子育种的研究，E-mail：wangxianlei2000@163.com.

出第一个基于焦磷酸测序原理的超高通量基因组测序系统（Genome Sequencer FLX System, GS FLX）[3]. 2012年，西班牙农业科学研究院基于该测序系统对厚皮甜瓜DHL92双单倍体进行测序，组装出第一版甜瓜基因组V3.5.1[4]；2018年，该研究院在V3.5.1基础上利用optical mapping方法，构建染色体物理图谱，发表了序列注释信息更完整、GAPs（未知区，未测通区）大小确定的甜瓜基因组V3.6.1[5]；2019年，其又利用PacBio的单分子实时（Single Molecule Real-Time, SMRT）测序技术，将甜瓜基因组序列覆盖率提高至94.31%，组装出甜瓜基因组V4.0[6].

二代测序技术使得基因组测序通量提高，且成本大幅降低，生命科学研究从单个基因扩展到全基因组[7]，从人类应用扩展到植物科学[8]. 完整的植物基因组，可进行基因分型，研究不同细胞和组织在发育过程中对环境条件反应的基因组结构变化和序列修改，可高效开展重要农艺性状基因定位研究，以及设计新的育种策略等[9]. 高通量技术在甜瓜中也有较多应用，周慧文[10]基于公布的甜瓜基因组信息开发了CAPS标记，定位了甜瓜果皮性状相关基因. 张乔玲[11]基于甜瓜基因组信息开发了甜瓜单性雌花分子标记用于辅助育种，胡倩梅[12]通过全基因组关联分析（Genome-Wide Association Study, GWAS）发现了与甜瓜重要农艺性状显著相关的SNP位点及相关基因，Liu等[13]获得了甜瓜在遗传驯化过程中影响其果实大小及果肉厚度的关键位点. 高质量的基因组数据结合丰富的遗传资源有力推动了从结构基因组学到功能基因组学的研究进程[8].

大多数真核生物的核基因组中重复序列占DNA总含量的一半以上[14]，有些重复序列还很长，这必然导致早期公布的基于读长较短的二代测序技术难以组装出高质量的基因组序列[7]，如基于二代测序技术的甜瓜基因组V3.6.1中含有44 650个GAPs. 基于SMRT技术的单分子三代测序技术的出现，有效克服了二代测序技术短读长的缺点，其通过染色体构象捕获[15]、DNA稀释技术[16]等，有力推动了高质量基因组的组装，促进植物多样性、表观遗传修饰和重要性状基因定位等方面的研究[17]. 二代测序技术和SMRT技术的结合，可以提升基因组组装的效率和质量[18]. 如在二代测序技术基础上，利用SMRT技术组装出的甜瓜基因组V4.0中仅含有1 169个GAPs. 即便如此，甜瓜基因组V4.0中的GAPs可能仍然包含与性状相关的基因，导致基因组数据无法用于GWAS或某些基因的精细定位研究[6]. 因此，本文通过分析甜瓜基因组V3.6.1和V4.0中的GAPs序列特征，进而探索甜瓜基因组GAPs形成原因，为组装高质量甜瓜基因组提供参考.

1　材料与方法

1.1　研究材料

从葫芦科基因组数据库（http://cucurbitgenomics.org/）下载甜瓜基因组V3.6.1、V4.0. 基因组V3.6.1是在V3.5.1基础上通过Argus系统的optical mapping完善后组装而成，大小为417 Mb，含腺嘌呤（Adenine, A）、胸腺嘧啶（Thymine, T）、鸟嘌呤（Guanine, G）、胞嘧啶（Cytosine, C）碱基共337.32 Mb，未知碱基（N）共79.68 Mb. 基因组V4.0通过SMRT测序技术和HGAP4（Hierarchical Genome Assembly Process4）组装而成，大小为358 Mb，A、T、G、C碱基共357.74 Mb，N共0.12 Mb.

1.2　研究对象

通过生物信息学方法分析，获取甜瓜基因组V3.6.1、V4.0的N所处位置. 利用TBtools软件[19]中的Fasta Extract工具提取V3.6.1中的GAPs上下游各2 000 bp序列，以甜瓜基因组V4.0为参考，将提取序列通过TBtools软件中BLAST工具进行同源序列比对，选取最优匹配区间、Size小于80 kb的18 012条GAPs序列，每条序列所在区间记为N区域，简记为GAPs，并以之代表已知序列的GAPs区域进行分析. 以GAPs区间位置为两侧端点，向上游平移10 kb，记为基因组区，并以之代表基因组随机区域进行分析. 取GAPs上游端点向下150 bp和下游端点向上150 bp序列，以之为GAPs内侧150 bp进行分析. 基因组内侧150 bp同上.

1.3　研究方法

通过TBtools软件中的Fasta Stat工具，对比GAPs 150 bp和基因组区150 bp、GAPs和基因组，分析GAPs序列的碱基组成；通过TBtools软件中的SSR miner工具，获取并统计以上区域序列SSR的碱基组成和重复类型特征；通过TBtools软件中的BLAST工具，对比GAPs和基因组区150 bp、GAPs和基因组，获取并统计该区域序列的重复序列；通过TBtools软件中的Fasta Stat工具，统计基因组V3.6.1中不同区域的N含量.

参照张国庆[20]统计SSR的方法，单碱基重复、二碱基重复、三碱基重复、四碱基重复和六碱基重复的最小

长度为12 bp，五碱基重复最小长度为15 bp，七碱基重复最小长度为14 bp，基因组1 Mb区域内的微卫星位点数为SSR密度. 使用TBtools软件获取1~7 bp基本单元重复序列位点和区域序列大小，计算SSR密度.

2 结果与分析

2.1 甜瓜基因组V3.6.1和V4.0的碱基组成特征

V3.6.1、V4.0的碱基组成如表1所示，已知区域由A、T、G、C碱基组成，GAPs由N组成. 与V3.6.1相比，V4.0染色体区（Chr1~12）A、T、G、C碱基数量减少17.62 Mb，N数量减少57.09 Mb；与V3.6.1未组装染色体区（Chr0）相比，V4.0中Chr0区A、T、G、C碱基数量减少41.52 Mb，N数量减少22.46 Mb. V4.0中的N含量为0.03%，远低于V3.6.1中的19.11%.

表 1 甜瓜基因组V3.6.1和V4.0染色体的碱基组成

	V3.6.1				V4.0			
染色体	大小/bp	N/bp	N含量/%		大小/bp	N/bp	N含量/%	
Chr0	41 641 883	22 464 661	53.95	19.11	121 315	1 800	1.48	0.03
Chr1~12	375 360 399	57 212 307	15.24		357 736 055	117 400	0.03	

V3.6.1与V4.0的12条染色体中N数量如图1所示. 与V3.6.1中12条染色体上N的数量相比，V4.0中12条染色体上N的数量均显著减少. 这也验证了随着测序技术的进步，基因组序列信息更加完整.

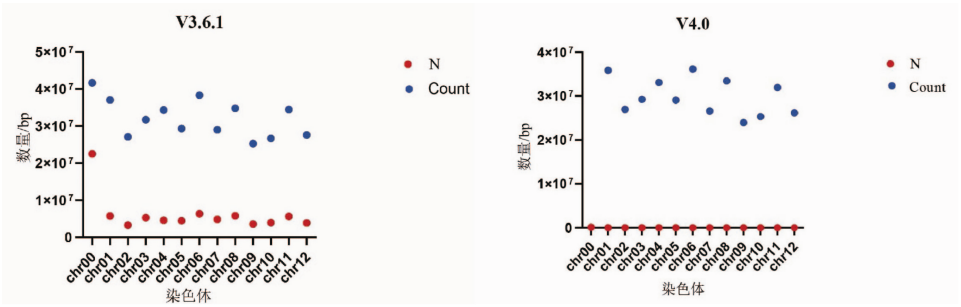

图 1 甜瓜基因组V3.6.1和V4.0 12条染色体中碱基组成

2.2 甜瓜基因组V3.6.1和V4.0的GAPs特征

V3.6.1和V4.0的GAPs特征如表2所示. V3.6.1中AT、GC碱基含量分别为66.84%、33.16%，V4.0中AT、GC碱基含量分别为66.46%、33.54%. GAPs中AT碱基含量为64.28%，比基因组区内侧150 bp（66.22%）、基因组V3.6.1（66.84%）、基因组V4.0（66.46%）的AT碱基含量低；GAPs内侧150 bp中AT碱基含量为69.23%，比基因组区内侧150 bp（66.22%）、基因组V3.6.1（66.84%）、基因组V4.0（66.46%）的AT碱基含量高.

表 2 甜瓜基因组不同区域的碱基组成

碱基	GAPs内侧150 bp	基因组区内侧150 bp	GAPs	基因组V3.6.1	基因组V4.0
AT/bp	3 766 099	32 325 513	31 387 636	225 461 987	237 760 228
大小/bp	5 439 624	48 814 078	48 827 859	337 325 314	357 738 170
AT含量/%	69.23	66.22	64.28	66.84	66.46

统计GAPs内侧150 bp、基因组区内侧150 bp、GAPs、基因组V3.6.1、基因组V4.0区域中SSR的AT碱基含量如表3所示. 根据非SSR区域AT含量=（总AT碱基数－SSR中AT碱基数）/总AT碱基数，对5个区域非SSR区域AT含量进行统计. 5个区域的非SSR区域AT含量分别为60.76%、66.01%、61.34%、64.93%、64.49%，可见GAPs内侧150 bp的非SSR区域AT含量偏低，GC含量偏高.

表 3 甜瓜基因组不同区域中微卫星重复序列的碱基组成

碱基	GAPs内侧150 bp	基因组区内侧150 bp	GAPs	基因组V3.6.1	基因组V4.0
AT/bp	460 840	104 828	1 438 320	6 428 389	7 052 046
大小/bp	544 097	131 048	1 858 078	7 999 567	8 860 941
AT含量/%	84.70	79.99	77.41	80.36	79.59
非SSR区域AT含量/%	60.76	66.01	61.34	64.93	64.49

使用TBtools软件中BLAST工具对GAPs内侧150 bp、基因组区内侧150 bp、GAPs和基因组区4个区域的序列进行BLAST比对分析，统计其重复序列的含量，如表4所示. 4个区域的多拷贝序列含量依次为57.11%、50.53%、64.47%、63.10%. GAPs内侧150 bp区域比基因组区内侧150 bp区域的多拷贝序列含量高6.58%；GAPs的多拷贝序列含量比基因组区的多拷贝序列含量高1.37%，表明GAPs区域内序列在基因组其它位置含有相同或类似片段的序列多，且内侧150 bp区域具有更多的重复序列.

表 4　甜瓜基因组不同区域重复序列的含量

重复类型	GAPs内侧150 bp	基因组区内侧150 bp	GAPs	基因组区
单拷贝/数量	15 449	17 818	6 400	6 968
多拷贝/数量	20 575	18 202	11 612	11 042
多拷贝含量/%	57.11	50.53	64.47	63.10

使用TBtools软件中SSR miner工具对GAPs内侧150 bp、基因组区内侧150 bp、GAPs、基因组V3.6.1、基因组V4.0的SSR各重复类型的数量、密度和比率（SSR数量/总SSR数量）进行统计，如表5所示. 结果表明：GAPs中含有114 122个SSR位点，GAPs内侧150 bp中含有24 864个SSR位点. GAPs区域SSR密度为2 354.29 SSR/Mb，比基因组V3.6.1（1 368.28 SSR/Mb）、基因组V4.0（1 735.12 SSR/Mb）区域的SSR密度高. GAPs内侧150 bp区域的SSR密度为6 622.52 SSR/Mb，比基因组区内侧150 bp区域的SSR密度（1 694.97 SSR/Mb）高. 表明GAPs区域的SSR密度偏高，尤其是GAPs内侧150 bp区域的SSR密度更高.

表 5　甜瓜基因组不同区域SSR各重复类型的总数、密度和比率

重复类型	项目	GAPs内侧150 bp	基因组区内侧150 bp	GAPs	基因组V3.6.1	基因组V4.0
单碱基	数量	4 726	928	11 067	53 321	62 755
	比率/%	19.01	10.07	9.70	9.35	10.11
二碱基	数量	5 742	772	11 374	45 225	45 843
	比率/%	23.09	8.37	9.97	7.93	7.39
三碱基	数量	4 130	1 053	19 339	56 883	67 111
	比率/%	16.61	11.42	16.95	9.97	10.81
四碱基	数量	1 778	829	7 890	52 252	55 000
	比率/%	7.15	8.99	6.91	9.16	8.86
五碱基	数量	301	177	4 673	9 021	11 067
	比率/%	1.21	1.92	4.09	1.58	1.78
六碱基	数量	6 766	4 441	46 874	285 892	305 914
	比率/%	27.21	48.17	41.07	50.11	49.28
七碱基	数量	1 421	1 019	12 905	67 980	73 031
	比率/%	5.72	11.05	11.31	11.91	11.77
总数	数量	24 864	9 219	114 122	570 574	620 721
	大小/Mb	3.75	5.44	48.47	417	357.86
	密度/(SSR/Mb)	6 622.52	1 694.97	2 354.29	1 368.28	1 735.12

分析基因组不同区域SSR中各重复类型数量及占比（表5）. 以区间内侧两端上下游各150 bp为对象，GAPs比基因组区的各碱基重复序列多，尤其是GAPs中二碱基SSR位点显著增多，说明GAPs内侧150 bp区域的SSR密度高可能是由二碱基类型重复多造成的. GAPs三碱基重复比率（16.95%）比基因组V3.6.1（9.97%）、V4.0（10.81%）区域的三碱基重复比率高，表明GAPs的SSR密度较高可能是由三碱基重复位点多造成的.

由表5可知，甜瓜各重复中GAPs和基因组区内侧两端上下游各150 bp区域相比，二碱基重复序列较多. 对GAPs、基因组区及其内侧两端上下游各150 bp区域二碱基微卫星重复序列中的重复类型分布进行统计（表6）. 甜瓜二碱基微卫星重复类型中，4个区域中各重复类型占比不同，最高的均为AT重复，最低的均为GC重复. 4个区域的AT重复含量为GAPs内侧150 bp>GAPs>基因组区>基因组区内侧150 bp. GAPs内侧150 bp处偏高的SSR密度主要是由该区域二碱基重复序列偏多造成，而二碱基重复中占比最高的是AT重复类型（≥70%）.

表 6 甜瓜基因组4个区域二碱基SSR的重复类型数量和比率

重复类型	项目	GAPs内侧150 bp	基因组区内侧150 bp	GAPs	基因组区
AT	数量	5 070	559	9 831	4 728
	比率/%	88.30	72.41	86.43	77.17
AC	数量	75	33	189	204
	比率/%	1.31	4.27	1.66	3.33
AG	数量	260	83	587	483
	比率/%	4.53	10.75	5.16	7.88
TC	数量	244	69	553	482
	比率/%	4.25	8.94	4.86	7.87
TG	数量	92	28	210	224
	比率/%	1.60	3.63	1.85	3.66
GC	数量	1	0	4	6
	比率/%	0.02	0.00	0.04	0.10
二碱基重复总数	大小/bp	5 742	772	11 374	6 127

对基因组V3.6.1的信使RNA（messenger RNA, mRNA）上游0～2 000 bp（记为mRNA上游）、mRNA、编码区（Coding Sequence, CDS）、外显子（Exon）、5′非翻译区（5′ Untranslated Region, 5′ UTR）、3′非翻译区（3′ Untranslated Region, 3′ UTR）的未知碱基进行统计，如表7所示. N含量分别为7.654 9%、4.696 7%、0.000 8%、0.001 3%、0.004 0%、0.002 0%，排序为mRNA上游>mRNA>5′ UTR>3′ UTR>Exon>CDS. 6个区域的N含量均低于基因组染色体中的N含量（15.24%），表明mRNA中的N含量更多，编码区N含量较低，推测是由于非功能基因具有更多重复序列所致.

表 7 基因组V3.6.1不同区域未知碱基统计

碱基	mRNA上游	mRNA	编码区	外显子	5′非翻译区	3′非翻译区
N/bp	4 590 243	5 063 542	234	539	173	132
大小/bp	59 965 134	107 811 060	29 200 683	40 225 117	4 364 691	6 659 743
N含量/%	7.654 9	4.696 7	0.000 8	0.001 3	0.004 0	0.002 0

基因组V4.0中含有1 192个GAPs，基因组V3.6.1中含有44 652个GAPs. 取基因组V4.0 N区域、基因组V3.6.1中GAPs和基因组区两端上下游外侧150 bp，对其碱基组成、多拷贝序列含量、SSR密度进行统计（表8）.

表 8 甜瓜基因组GAPs外侧150 bp序列组成

类型	V3.6.1 GAPs外侧150 bp	V4.0 GAPs外侧150 bp	基因组区外侧150 bp
AT碱基含量/%	65.69	60.92	66.22
多拷贝含量/%	72.13	98.24	50.53
SSR密度/(SSR/Mb)	4 096.14	2 605.47	1 694.97

统计不同区域外侧150 bp的AT碱基含量，由高至低依次为基因组区>V3.6.1 GAPs>V4.0 GAPs. 基因组V4.0 GAPs外侧150 bp中AT碱基含量为60.92%，比基因组区外侧150 bp（66.22%）和基因组V3.6.1 GAPs外侧150 bp（65.69%）的AT碱基含量低，表明N区域含有较少的AT碱基，即GAPs区域的GC碱基偏多.

统计不同区域外侧150 bp的SSR密度，由高至低依次为V3.6.1 GAPs>V4.0 GAPs>基因组区，基因组区的SSR密度为1 694.97 SSR/Mb，远低于基因组V4.0 GAPs（2 605.47 SSR/Mb）和V3.6.1 GAPs（4 096.14 SSR/Mb）区域，表明GAPs外侧150 bp含有较多的SSR位点.

统计不同区域外侧150 bp的多拷贝含量，显示V4.0 GAPs>V3.6.1 GAPs>基因组区，V4.0 GAPs多拷贝含量为98.24%，比基因组区（50.53%）、V3.6.1 GAPs（72.13%）区域的多拷贝含量高，表明GAPs外侧较高的多拷贝含量使该区域序列与其它区域序列相同或类似的序列多，基因组组装过程中，由于该区域在基因组内其它区域存在较多重复序列导致下游序列无法正确拼接而形成GAPs，基因组V4.0 GAPs主要由多拷贝序列引起.

3　讨 论

单分子实时测序技术可以有效克服二代测序技术短读长的不足，解决基因组组装的高错误率问题. 通过PacBio的单分子测序技术组装的甜瓜基因组V4.0的质量明显高于甜瓜基因组V3.6.1，GAPs碱基总数和数量均大幅减少[6]，本文也获得了相似结果：GAPs碱基总数由V3.6.1中的79.68 Mb减少至V4.0中的0.12 Mb，GAPs数量由V3.6.1中的44 652个减少至V4.0中的1 192个. 对甜瓜基因组的SSR密度进行统计，发现基因组V3.5.1的SSR密度为1 455.8个/兆碱基对[20]，V3.6.1的SSR密度为1 368.28个/兆碱基对、V4.0的SSR密度为1 735.12个/兆碱基对，表明随着基因组测序技术的不断进步，大量未知区域被扩增组装，获得的基因组序列信息也更加完善.

PCR扩增过程中，富含GC的模板通常会受到发卡二级结构形成和更高熔解温度的阻碍[21]，影响PCR合成. 本文发现GAPs非SSR区域含有较高的GC含量，前期研究证实较高的GC含量可能会影响基因组测序，阻碍基因组组装，进而形成GAPs.

微卫星重复序列又称短串联重复序列（Short Tandem Repeats, STR）或简单重复序列（SSR），属于串联重复序列[22]. 使用ABI 3730X1全自动测序仪对含有重复序列的样品进行测序时，会产生测序信号终止、减弱或出现重叠峰的情况，阻止测序或影响测序结果[23]. 而且重复序列在变性和退火循环过程中，于聚合方向前形成发卡结构，使DNA聚合酶与模板解离，停止聚合反应，终止测序[24]. 本文发现GAPs含有较高的SSR密度，前期研究表明基因组某些含有较高微卫星密度区域会阻碍该区域测序而形成GAPs. 本文还发现GAPs内侧150 bp的GC含量、SSR密度均高于GAPs，表明越临近GAPs边界，其GC含量与SSR密度越高.

除较高的SSR密度外，GAPs区域含有较高的多拷贝含量，表明该区域含有较多与基因组其它区域相近和类似的序列. 使用二代测序技术对基因组进行测序组装时，由于二代测序序列较短，会导致重复序列形成相同片段，而相同片段会映射到染色体其它位置或被错误地叠加在一起，导致复杂的、错误组装的基因组区域[25]，形成GAPs.

Tóth等[26]对酵母、灵长类动物、拟南芥微卫星标记的频率进行统计，发现与所有分类群中的外显子序列相比，基因间区域和内含子的总SSR密度更高. 本文对mRNA上游、mRNA、CDS、Exon、5′ UTR、3′ UTR进行统计，发现6个区域的N含量均低于基因组平均N含量，表明GAPs在基因间出现频率较高.

4　结 论

通过对GAPs内外侧150 bp、基因组区内外侧150 bp、GAPs、基因组区、基因组V3.6.1和基因组V4.0及其GAPs区域进行分析，表明与基因组整体相比，GAPs内部两侧150 bp区域具有较高的SSR密度（6 622.52 SSR/Mb）、非SSR区域具有较高的GC含量（39.24%），GAPs外侧150 bp含有较高的多拷贝序列含量. 较高GC含量和微卫星密度的区域会影响PCR扩增，较高多拷贝含量区域序列的存在会影响下游序列的拼接组装，且在GAPs内侧150 bp与GAPs全序列比较发现，越临近GAPs边界，其GC含量与微卫星密度越高. 因此，甜瓜基因组V3.6.1 GAPs形成的主要原因是：基因组测序组装过程中，含有较高GC含量、微卫星密度及较多多拷贝含量的区域无法正常测序和组装. 比对分析基因组V4.0 GAPs两侧序列，发现V4.0中的多拷贝序列占比为98.24%，可能是GAPs形成的重要原因，且GAPs更容易出现在基因间.

参考文献：

[1] 刘畅. 甜瓜*CmNAC34*和*CmHsp83*基因在果实发育中功能的初步分析[D]. 呼和浩特：内蒙古大学，2018.
LIU C. The preliminarily functional analysis of melon *CmNAC34* and *CmHsp83* genes in fruit development[D]. Hohhot：Inner Mongolia University，2018. (in Chinese)

[2] 张慧君，王学征，高鹏，等. 甜瓜性别分化的研究进展[J]. 园艺学报，2012，39(9)：1773-1780.
ZHANG H J，WANG X Z，GAO P，et al. Progress of study on sex differentiation in melon[J]. Acta Horticulturae Sinica，2012，39(9)：1773-1780. (in Chinese)

[3] DROEGE M，HILL B. The Genome Sequencer FLX System：Longer reads，more applications，straight forward bioinformatics and more complete data sets[J]. Journal of Biotechnology，2008，136(1/2)：3-10.

[4] GARCIA-MAS J，BENJAK A，SANSEVERINO W，et al. The genome of melon (*Cucumis melo* L.)[J]. Proceedings of the National Academy of Sciences of the United States of America，2012，109(29)：11872-11877.

[5] RUGGIERI V，ALEXIOU K G，MORATA J，et al. An improved assembly and annotation of the melon (*Cucumis melo* L.)

reference genome[J]. Scientific Reports, 2018, 8: 8088.

[6] CASTANERA R, RUGGIERI V, PUJOL M, et al. An improved melon reference genome with single-molecule sequencing uncovers a recent burst of transposable elements with potential impact on genes[J]. Frontiers in Plant Science, 2019, 10: 1815.

[7] SCHUSTER S C. Next-generation sequencing transforms today's biology[J]. Nature Methods, 2008, 5: 16-18.

[8] MORRELL P L, BUCKLER E S, ROSS-IBARRA J. Crop genomics: Advances and applications[J]. Nature Reviews Genetics, 2012, 13: 85-96.

[9] BICKHART D M, ROSEN B D, KOREN S, et al. Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome[J]. Nature Genetics, 2017, 49: 643-650.

[10] 周慧文. 甜瓜果实性状表型分析及其CAPS标记的研究[D]. 哈尔滨: 东北农业大学, 2016.
ZHOU H W. The research of phenotype analysis and CAPS marker for fruit traits in melon[D]. Harbin: Northeast Agricultural University, 2016. (in Chinese)

[11] 张乔玲. 甜瓜重要性状分子标记辅助育种体系的建立[D]. 天津: 天津大学, 2019.
ZHANG Q L. Establishment of marker-assisted breeding system for important traits in melon[D]. Tianjin: Tianjin University, 2019. (in Chinese)

[12] 胡倩梅. 甜瓜重要农艺性状全基因组关联分析[D]. 郑州: 河南农业大学, 2019.
HU Q M. Genome-wide association study of important agronomic traits in melon[D]. Zhengzhou: Henan Agricultural University, 2019. (in Chinese)

[13] LIU S, GAO P, ZHU Q L, et al. Resequencing of 297 melon accessions reveals the genomic history of improvement and loci related to fruit traits in melon[J]. Plant Biotechnology Journal, 2020, 18(12): 2545-2558.

[14] BISCOTTI M A, OLMO E, HESLOP-HARRISON J S. Repetitive DNA in eukaryotic genomes[J]. Chromosome Research, 2015, 23(3): 415-420.

[15] BURTON J N, ADEY A, PATWARDHAN R P, et al. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions[J]. Nature Biotechnology, 2013, 31: 1119-1125.

[16] AMINI S, PUSHKAREV D, CHRISTIANSEN L, et al. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing[J]. Nature Genetics, 2014, 46: 1343-1349.

[17] ZHENG G X Y, LAU B T, SCHNALL-LEVIN M, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing[J]. Nature Biotechnology, 2016, 34: 303-311.

[18] 柳延虎, 王璐, 于黎. 单分子实时测序技术的原理与应用[J]. 遗传, 2015, 37(3): 259-268.
LIU Y H, WANG L, YU L. The principle and application of the single-molecule real-time sequencing technology[J]. Hereditas, 2015, 37(3): 259-268. (in Chinese)

[19] CHEN C J, CHEN H, ZHANG Y, et al. TBtools: An integrative toolkit developed for interactive analyses of big biological data[J]. Molecular Plant, 2020, 13(8): 1194-1202.

[20] 张国庆. 五种葫芦科物种简单重复序列的分析及应用[D]. 泰安: 山东农业大学, 2018.
ZHANG G Q. Analysis and application of Simple Repeat Sequences in Cucurbitaceae[D]. Tai'an: Shandong Agricultural University, 2018. (in Chinese)

[21] FREY U H, BACHMANN H S, PETERS J, et al. PCR-amplification of GC-rich regions: "Slowdown PCR" [J]. Nature Protocols, 2008, 3: 1312-1317.

[22] LANDER E S, LINTON L M, BIRREN B, et al. Initial sequencing and analysis of the human genome[J]. Nature, 2001, 409: 860-921.

[23] 刘松梅. 重复序列的特征对测序结果的影响分析[D]. 大连: 大连理工大学, 2015.
LIU S M. Analysis of the effect of repetitive DNA sequence characteristics on sequencing results[D]. Dalian: Dalian University of Technology, 2015. (in Chinese)

[24] CANCEILL D, VIGUERA E, EHRLICH S D. Replication slippage of different DNA polymerases is inversely related to their strand displacement efficiency[J]. Journal of Biological Chemistry, 1999, 274(39): 27481-27490.

[25] POP M, SALZBERG S L. Bioinformatics challenges of new sequencing technology[J]. Trends in Genetics, 2008, 24(3): 142-149.

[26] TÓTH G, GÁSPÁRI Z, JURKA J. Microsatellites in different eukaryotic genomes: Survey and analysis[J]. Genome Research, 2000, 10(7): 967-981.

责任编辑: 张自强