文章编号:1001-1498(2015)01-0037-07

利用高通量测序分析青藏高原地区青杨的 SSR 和 SNP 特征

雷淑芸[1,2], 张发起[1,3*], Khan Gulzar[1], 王久利[1,2], 刘海瑞[1,2], 陈世龙[1]

(1. 中国科学院高原生物适应与进化重点实验室,中国科学院西北高原生物研究所,青海 西宁 810008; 2. 中国科学院大学,北京 100049; 3. 西南大学生命科学学院,重庆 400715)

摘要:利用 Illumina HiSeq™ 2000 平台对采自青海玉树的青杨进行高通量测序,SSR 分析共获得 7 067 条 SSR 序列,复合型 SSR 共 525 条,发生频率 0.149,平均跨度 4 531.87 bp。SSR 重复类型中,单核苷酸重复类型最多(33.96%);三核苷酸重复类型次之(31.00%);二核苷酸重复类型位居第三(27.69%);四核苷酸、五核苷酸、六核苷酸重复类型 SSR 含量很少(<8%)。二核苷酸重复类型中,AG 重复类型所占比例最高,GA 次之,CT、TC 则紧随其后;三核苷酸重复类型中,AAG 重复类型所占比例最高,GAA 次之,TTC、AGA、GAG、CAG、TCT、TGG 等重复类型数量相近。SNP 分析发现,L1A 中含 SNP 162 343 个,L2A 中含 SNP 229 115 个。SNP 类型中,转换类型明显高于颠换类型,L1A 中转换类型占 61.06%,颠换类型占 38.94%,L2A 中转换类型占 61.27%,颠换类型占 38.73%。转换类型中,C-T 发生频率最高,分别为 30.75%、30.66%,A-G 发生频率与 C-T 相差不大,分别为 30.31%、30.62%。L1A 和 L2A 中 SNPs 类型及其发生频率变化趋势基本一致,L2A 中与 L1A 中相对应的同一 SNPs 类型的数量之比约为 2∶1。分析表明,在青杨的遗传多样性中,SNP 标记较 SSR 标记更为可靠。

关键词:青杨;青藏高原;高通量测序;微卫星;单核苷酸多态性

中图分类号:S718.46 **文献标识码**:A

Characteristic Analysis of SSR and SNP in *Populus cathayana* on the Qinghai-Tibetan Plateau by High-Throughput Sequencing

LEI Shu-yun[1,2], *ZHANG Fa-qi*[1,3], *Khan Gulzar*[1], *WANG Jiu-li*[1,2], *LIU Hai-rui*[1,2], *CHEN Shi-long*[1]

(1. Key laboratory of Adaption and Evolution of Plateau Biota, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining 810008, Qinghai, China; 2. University of Chinese Academy of Sciences, Beijing 100039, China; 3. School of Life Sciences, Southwest University, Chongqing 400715, China)

Abstract: Two samples of *Populus cathayana* from Yushu were sequenced with high-throughput sequencing technology (Illumina HiSeq™2000). A total of 7076 sequences were hunted for microsatellites analysis, including 525 compound microsatellite sequences. The results showed that the mononucleotide repeats were the highest (33.96%), followed by trinucleotide repeats (31.00%) and dinucleotide repeats 27.69%. The tetranucleotide, pentanucleotide and hexanucleotide repeats were all less than 8%. Among the dinnucleotide repeats, the AG repeats were with highest frequency followed by GA, CT and TC respectively. Similarly in trinucleotide repeats, the AAG repeats were the highest followed by GAA, TTC, AGA, GAG, CAG, TCT and TGG. In the single nucleotide polymorphism, the transition types were higher than transversion in both samples. In sample L1A, the transition types of SNPs were 61.06% while the transversion types were 38.94%. In sample L2A, the transition types were 61.27%,

收稿日期: 2014-06-27
基金项目: 国家自然科学基金(31270270)、青海省国际科技合作项目(2014-HZ-812)
作者简介: 雷淑芸,硕士研究生。主要从事植物遗传多样性学研究。E-mail:871320418@qq.com
* 通讯作者:博士,助理研究员。E-mail:fqzhang@nwipb.cas.cn

while the transversion types were 38.73%. Among the transitions of two samples, C-T occurrences frequencies were the most, 30.75% and 30.66% respectively. A-G occurrences frequencies were similar to C-T, list to the second (30.31% and 30.62%). In both samples, the single nucleotide polymorphism trends were similar. The ratios of the same single nucleotide polymorphism in both samples were 2∶1. The results of analysis indicated that the SNP was a more reliable maker in geneticdiversity than SSR.

Key words: *Populus cathayana*; Qinghai-Tibetan Plateau; high-throughput sequencing; SSR; SNP

青藏高原(Qinghai-Tibetan Plateau)位于73°～104° E、26°～39° N之间,平均海拔超过4 000 m,总面积约250万 km^2[1],是世界上面积最大、海拔最高、地质年代最年轻的高原,被誉为"世界屋脊"、地球的"第三极"。该地区处于长江、黄河、澜沧江等主要水系的上游,水源涵养的地位十分突出,植树造林和草场恢复建设都是目前我们所面临的最急迫的任务。高原上有着植树造林和草场建设所需要的广阔土地。然而,青藏高原位于40° N以南,由于海拔高地面气温比同纬度平原地区低得多,7月份平均气温低于同纬度15～20℃,海拔4 500 m以上的高原腹地年平均气温0 ℃以下,高原面上最冷月平均气温低达－10～－15℃,很多地区最暖月平均气温低于10℃[2]。高海拔和高寒特殊环境极大地限制了优良树种在青藏高原地区的引种。然而即便在如此恶劣的自然环境下,青藏高原依然是我国植物资源重要的"基因库",保存了极高丰富度的生物多样性,是全球生物多样性热点地区之一[3]。由于高原地形复杂、环境多样、冰川活动、地理位置及地史条件等原因,青藏高原具有丰富而特殊的杨树资源。经过系统的分类学研究,青藏高原地区杨属植物约有17个种、15个变种,其中多数种为高原及其邻近地区所特有,同时由于种间杂交,还有不少的天然杂种[4]。青杨(*Populus cathayana*)隶属于杨柳科(Salicace)杨属(*Populus* L.),是青杨派杨树的主要树种,也是我国的特有乡土树种,是青藏高原众多杨树中分布最广、海拔跨度最大的树种。目前已有的研究表明,分布在青藏高原东缘的青杨,由于高山、峡谷等复杂的地质地貌的影响,各群体之间存在极大的遗传分化[5]。因此,开展青杨的遗传多样性研究,有利于了解该地区植物对复杂多变生境的适应机制,为基因资源挖掘和建立完善的种子资源评价、保护系统提供理论依据。

近年来,随着分子标记的快速发展,尤其是第二代微卫星标记和第三代单核苷酸多态性标记在探究生物群体内和群体间遗传变异及种间关系和遗传育种等研究中所凸显的优越性,使其运用的越来越广泛。微卫星(Simple Sequence Repeat,SSR)由核心序列和两侧相对保守的侧翼序列构成,具有数量多、分布广且均匀、高度多态性、分析快速方便等优点[6-8],已广泛用于动植物遗传多样性分析、分子遗传育种等。早期微卫星的分离技术主要是小片段基因组DNA随机克隆技术,通过酶切构建小片段DNA文库后进行测序筛选。由于该方法耗时费力且筛选效率较低,很难获得足够用于群体遗传结构或建立图谱等研究所需的微卫星标记[9]。随后又提出利用滤膜、磁珠等方法构建微卫星富集文库。近年来,随着高通量测序技术的发展及成本的降低,通过高通量测序获得海量EST(Expressed Sequence Tag)序列,基于EST序列大量开发SSR引物标记的方法正越来越多的引起人们的关注,与传统方法相比,其具有可能与基因功能相关、种间通用性强、开发周期短等优点[10]。单核苷酸多态性(Single Nucleotide Polymorphisms, SNP)是指一物种不同个体在基因组水平上由单个核苷酸变异引起的DNA序列多态性,通常只有频率大于1%时才被称为SNP,频率等于或小于1%的变异被称为突变[11]。SNP一般只涉及碱基的转换(Transition)和颠换(Transversion),是许多真核生物中最丰富的遗传变异形式,具有数量多、分布广、双等位基因、突变率低、筛查快、易实现自动化检测等优点[12]。目前应用于林木SNP开发的方法主要有两种:一是通过对已有表达序列标签(EST)的分析,发现SNP位点;二是对林木基因组序列的直接测序,检测SNP位点。由于直接测序较为直接、方便,常用于分析已知SNP的DNA序列,检出率高达100%,被认为是目前最可靠的一种,多数林木的SNP研究均采用该方法[13]。

青杨是我国重要的乡土树种,遗传多样性的研究对于开发利用和保护杨树种质资源尤为重要。因此,本研究在denovo转录组测序、拼接组装获得大量EST序列的基础上,充分挖掘青杨SSR和SNP的信息,通过对青杨SSR特征和SNP特征的分析,以

期为杨树 SSR 标记和 SNP 标记的开发提供生物信息学基础,为进一步研究该树种的遗传多样性奠定基础。

1　材料与方法

1.1　样品采集及 Solexa 高通量测序、拼接组装

青杨样品 L1A(chen2013180)采自海拔 3 600 m 的青海省玉树市结古镇(33°00′25. 08″ N,97°08′41. 6″ E),样品 L2A(chen2013183)采自海拔 2 338 m 的青海省贵德县拉西瓦水电站(36°5′29. 68″ N,101°12′44. 01″ E)。叶片采集后迅速投入液氮中,带回实验室后于 -80℃保存。凭证标本存于中国科学院西北高原生物研究所青藏高原生物标本馆(HNWP)。从野外采集的青杨样品 L1A、L2A 中各提取 100 μg 总 RNA, DNase 消化 DNA 后,用带有 Oligo(dT)的磁珠富集真核生物 mRNA;加入打断试剂在 Thermomixer 中适温将 mRNA 打断成短片段,以打断后的 mRNA 为模板合成一链 cDNA,然后配制二链合成反应体系合成二链 cDNA,并使用试剂盒纯化回收、粘性末端修复、cDNA 的 3′末端加上碱基“A”并连接接头,然后进行片段大小选择,最后进行 PCR 扩增;构建好的文库用 Agilent 2100 Bioanalyzer 和 ABI StepOnePlus Real-Time PCR System 质检合格后,通过 Illumina HiSeq™2000 进行测序。

SSR Unigene 的处理主要是将获得的原始读序(Raw reads)经过滤后得到干净读序(Clean reads)。通过 Trinty 软件拼接后[14],再用 Tgicl 软件将其去冗余和进一步拼接,然后再对这些序列进行同源转录本聚类,得到最终的 Unigene。之后将获得的 Unigene 序列与蛋白数据库 NR、Swiss-Prot、KEGG 和 COG 进行 blastx 比对(E <0. 000 01),取比对结果最好的蛋白确定 Unigene 的序列方向。若不同库之间的比对结果有矛盾,则按 NR、Swiss-Prot、KEGG 和 COG 的优先级确定 Unigene 的序列方向;若与以上四个数据库皆比对不上的 Unigene,则用软件 EST Scan 确定序列的方向。对于能确定序列方向的 Unigene,我们标明 5′到 3′方向;对于无法确定序列方向的 Unigene,只展示组装软件得到的序列。SNP Unigene 的分析是在通过第二次质控的情况下完成的,即对 raw reads 进行质控(QC),并经过滤得到 clean reads,再用 SOAPaligner/SOAP2 将 clean reads 比对到参考序列,统计 reads 在参考序列上的分布情况及覆盖度,(给出质控标准)。之后利用 SOAPsnp 软件对 SNP 进行检测[15]。

1.2　SSR 与 SNP 检测、筛选及统计分析

基于转录组的 SSR 检测是以组装出来的 Unigene 作为参考序列,使用 MicroSatellite(MISA;http://pgrc. ipk-gatersleben. de/misa/)搜索 SSR。首先,对所有搜索到的包含 SSR 重复单元的 Unigene 进行筛选,本次试验只保留前后序列均不小于 150 bp 的 Unigene 序列。SSR 搜索标准包括精确型(perfect)及复合型(compound)SSR 重复单元[16],各重复微卫星类型重复次数设定为:两碱基(dinnucleotide repeats,DNRs)至少重复 6 次,三碱基(trinucleotide repeats,TNRs)至少重复 5 次,四碱基(tetranucleotide repeats,TTNRs)至少重复 5 次,五碱基(pentanucleotide repeats,PTNRs)至少重复 4 次,六碱基(hexanucleotide repeats,HXNRs)至少重复 4 次。此外,本研究中还对至少重复 12 次的单核苷酸重复类型(mononucleotide repeats,MNNRs)进行统计。

基于转录组的 SNP 检测,则是将个体的测序数据与毛果杨(*Populus trichocarpa*)[17]基因组数据比对,用 SOAPsnp 检测出一致性序列。再通过一致性序列与参考序列的比较,从而找到 SNPs。通过短序列比对到参考序列,以及结合每个碱基的测序质量值,程序可以算出每个位置潜在基因型的可能性,之后再通过贝叶斯模型,计算出潜在基因型的后验概率值,值最高的便被推断为此位点的基因型。

2　结果与分析

2.1　与 SSR 相关的测序产量、质量及组装结果

近年来,随着高通量测序技术的发展及成本的降低,通过高通量测序可获得海量 EST 序列。但测序得到的 reads 并不都是有效的,里面含有带接头的、重复的和测序质量很低的 reads,它们对组装及后续分析都会产生影响。因此在数据处理过程中,华大基因将 L1A 与 L2A 转录组中测得的 raw reads 去除含接头的、N 比例大于 5% 的、低质量的 reads 后,分别获得 52 910 702 条和 52 435 372 条 clean reads,4 761 963 180 个和 4 719 183 480 个碱基,质量值≥20 的碱基达 97% 和 96. 47%,没有不确定的碱基比,而 G、C 占总碱基的比例都在 40% 以上。在对 L1A、L2A raw reads 进行相应处理后,我们用短 reads 组装软件 Trinity,对 clean reads 进行了组装分析,共获得 L1A Contig 87 203 条、Unigene 41 971 条,其中 Contig 总长 20 216 036 nt、均长 232 nt,Unigene

总长 15 428 216 nt、均长 368 nt;L2A Contig 109 374 条、Unigene 57 899 条,其中 Contig 总长 35 546 496 nt、均长 325 nt,Unigene 总长 33 466 924 nt、均长 578 nt。从组装序列的长度分布来看,我们发现 contig 和 Unigene 的数量与其片段长度大致成负相关,其中 L1A Contig 在 100～300 nt 间的数量达总 Contig 的 82.64%,而 L1A Unigene 在 100～1 000 nt 间的数量则占到总 Unigene 的 95.99%。通过对 L1A、L2A Contig 和 Unigene 平均长度与其 N50 的比较,可知组装出的 Unigene 数量、质量,与其 Contig 的组装数量、质量成明显的正相关性。

2.2 与 SNP 相关的测序产量、质量及组装结果

由华大基因 IlluminaSolexaHiSeq™ 2000 对青杨 L1A、L2A 测序,在对 raw reads 进行质控、过滤后,用 SOAPaligner/SOAP2 将 clean reads 比对到参考序列上,共比对上 L1A reads 52 910 702 条,碱基4 761 963 180 个,L2A reads 52 435 372 条,碱基 4 719 183 480 个,它们比对到参考基因、参考基因组上的比例都是 100%。样品 L1A、L2A 比对到参考基因组上的结果,与比对到参考基因上的结果相似(表 1)。通过对基因测序覆盖度的比较分析,可知 L1A 在 60%～90%间的 reads 数相对较高,但各覆盖度内的 reads 数相差并不大。L2A 在 60%～100%间的 reads 数相对较高,而 80%～100%间的 reads 数最多,与其它覆盖度内的 reads 数差异较明显。

表 1 青杨样品 L1A、L2A Reads 比对到参考基因和参考基因组上的结果分析

	比对到参考基因上的 reads 条数(%)		比对到参考基因组上 reads 条数(%)	
	样品 L1A	样品 L2A	样品 L1A	样品 L2A
总 reads 数	52 910 702(100)	52 435 372(100)	52 910 702(100)	52 435 372(100)
总碱基数	4 761 963 180(100)	4 719 183 480(100)	4 761 963 180(100)	4 719 183 480(100)
比对到参考序列上的总 reads 数	41 074 095(77.63)	35 563 814(73.76)	39 025 272(73.76)	38 160 720(72.78)
完美比对上的 reads 数	18 724 450(33.59)	13 856 011(32.67)	17 288 184(32.67)	16 044 275(30.60)
错配≤5bp 的 reads 数	22 349 645(42.24)	21 707 803(41.08)	21 737 088(41.08)	22 116 445(42.18)
比对到参考序列唯一位置的 reads 数	37 542 656(70.95)	32 003 567(67.86)	35 907 661(67.86)	31 478 930(60.03)
比对到参考序列多个位置的 reads 数	3 531 439(6.67)	3 560 247(5.89)	3 117 611(5.89)	6 681 790(12.74)
未能比对到参考序列上的 reads 数	11 836 607(22.37)	16 871 558(26.24)	13 885 430(26.24)	14 274 652(27.22)

2.3 青藏高原青杨 SSR 及 SNP 特征分析

通过 Solexa 高通量测序、运用相应软件拼接并去除重复序列后,对青杨 L1A、L2A 进行 SSR 分析,共获得 47 521 条 clean reads,总长度为 32 026 729 bp,去除 L1A 与 L2A 中的重复 SSR 后,共获得 7 067 条 SSR 序列,复合型 SSR525 条,包含 SSR 的序列有 5 847 条,包含有超过 1 条 SSR 的序列有 989 条,其发生频率为 0.149,平均跨度为 4 531.87 bp。

对高通量测序获得的 raw reads 进行质控、过滤后,用 SOAPaligner/SOAP2 将 clean reads 比对到参考序列上,共比对上 L1A reads 52 910 702 条,碱基 4 761 963 180个,SNP 162 343 个,其中染色体上有 160 183 个 SNP,染色体支架上有 2 160 个 SNP,平均跨度为 29 332.73 bp,simple SNPs 与 Hemi－SNPs 数量相当,分别为 82 786 个和 79 557 个,它们的平均跨度与它们的数量关系相一致。共比对上 L2A reads 52 435 372条,碱基 4 719 183 480 个,SNP 229 115 个,其中染色体上有 224 995 个,染色体支架上有 4 120 个,平均跨度为 20 597.44 bp,simple SNPs 与 Hemi-SNPs 数量相差较大,分别为 192 456 个和 36 659 个,它们的平均跨度与它们的数量关系也相一致,分别为 24 520.84 bp 和 128 731.93 bp。

2.4 青杨 SSR 与 SNP 丰度及分布密度分析

通过对青杨 SSR 数据库的检测分析,发现单核苷酸重复类型最多,占总 SSR 的 33.96%,以长度 12～21 bp 为主;三核苷酸重复类型次之,占总 SSR 的 31.00%,以长度 15～21 bp 为主;二核苷酸重复类型占总 SSR 的 27.69%,位居第三,以长度 12～20 bp 为主;而四核苷酸、五核苷酸、六核苷酸重复类型的 SSR 含量很少,三者总量不及总 SSR 数的 8%(图 1)。其中二核苷酸重复类型的微卫星有 12 种,三核苷酸重复类型的微卫星有 60 种,四核苷酸重复类型的微卫星有 60 种,五核苷酸重复类型的微卫星有 108 种,六核苷酸重复类型的微卫星有 181 种。

通过对主要核苷酸重复类型的比较分析发现,在单核苷酸重复类型中,A 重复类型所占比例最高,T 重复类型次之,分别为总 SSR 的 17.33%、15.06%;二核苷酸重复类型中,AG 重复类型所占比例最高,GA 次之,CT、TC 则紧随其后;三核苷酸重复类型中,

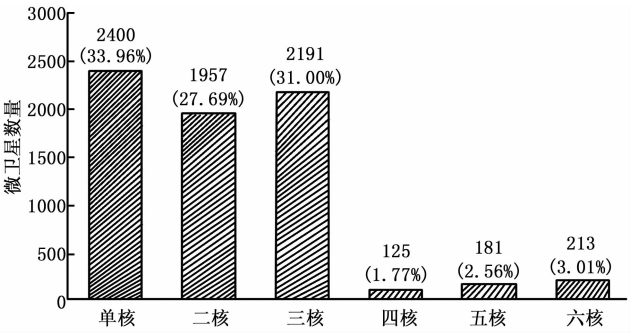


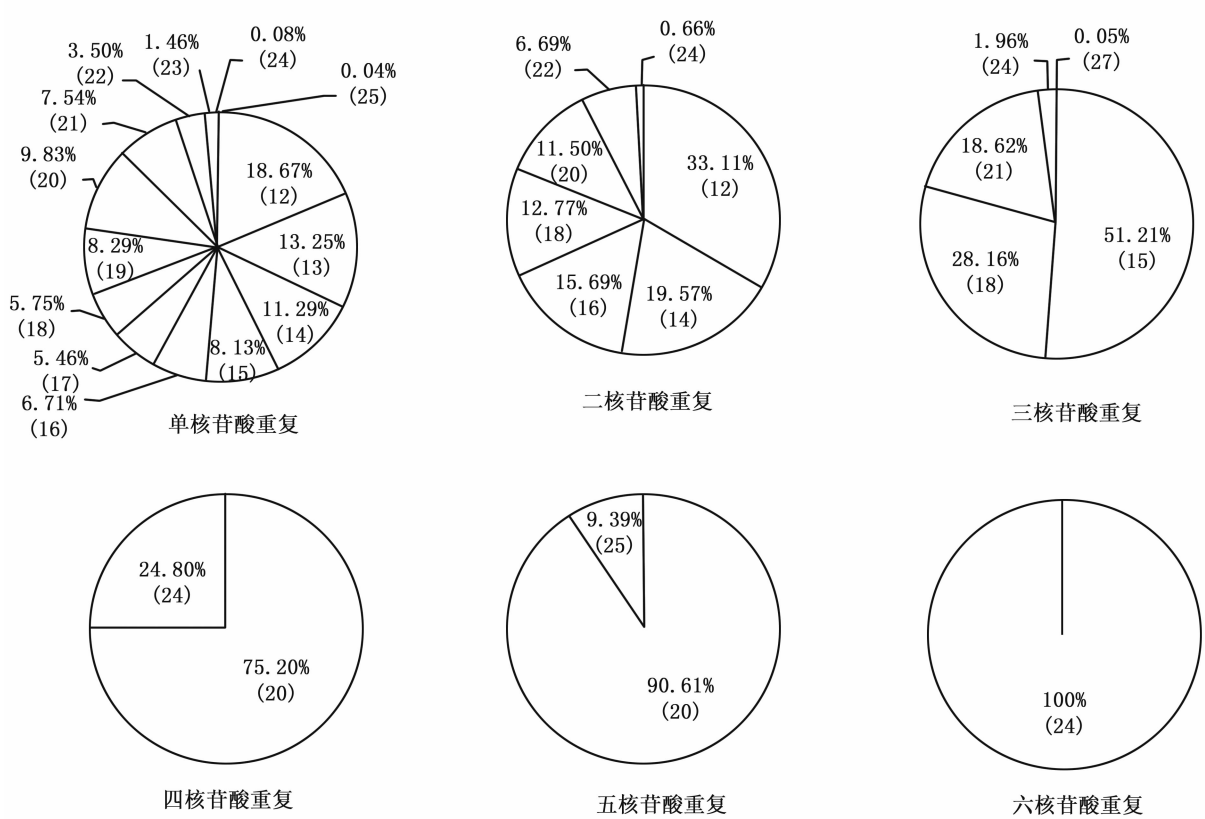
图 1　青杨不同核苷酸重复类型的 SSR 数量

AAG 重复类型所占比例最高,GAA 次之,TTC、AGA、GAG、CAG、TCT、TGG 等重复类型数量相近;四核苷酸、五核苷酸、六核苷酸重复类型的数量相对较少,不再进行进一步分析(表 2)。

通过对青杨序列不同长度重复类型微卫星长度分布及其变异的分析,有利于我们对多态性高的 SSR 标记的获得,通常来说微卫星核心序列的重复次数越多,其变异性越高[18]。通过对青杨微卫星序列的分析显示,其平均长度为 16.67 bp。除单核苷酸重复类型和六核苷酸重复类型外,其长度变化与其对应重复类型的碱基长度成反比(图 2)。

表 2　主要核苷酸重复类型在青杨序列中的百分比

主要的重复类型	重复碱基类型	SSR 总数	占总 SSR 的百分比/%
单核苷酸重复	A	1 225	17.33
	T	1 064	15.06
	G	69	0.98
	C	42	0.59
二核苷酸重复	AG	448	6.34
	GA	406	5.75
	CT	335	4.74
	TC	319	4.51
	AT	149	2.11
	TA	129	1.83
	AC	53	0.75
	其它	118	1.67
三核苷酸重复	AAG	107	1.51
	GAA	100	1.42
	TTC	91	1.29
	AGA	88	1.25
	GAG	87	1.23
	CAG	86	1.22
	TCT	77	1.09
	TGG	71	1.00
	GCT	69	0.98
	GGA	69	0.98
	TGC	65	0.92
	AGC	63	0.89
	GGT	61	0.86
	GAT	53	0.75
	GCA	53	0.75
	其它	1 051	14.87



注:每一扇形区对应不同长度微卫星所占百分比,括号内为其对应的微卫星长度。

图 2　青杨序列不同长度重复类型微卫星长度分布及其变异

对青杨 SNP 类型的分析发现,L1A 中转换类型占 61.06%(50 551 个),颠换类型占 38.94%(32 235 个),L2A 中转换类型占 61.27%(117 924 个),颠换类型占 38.73%(74 532 个)。其中转换类型明显高于颠换类型,L1A 和 L2A 中转换类型分别达到 61.06%、61.27%。在 L1A、L2A 转换类型中,C－T 发生频率最高,分别为 30.75%、30.66%,这可能与 SNPs 在 CG 序列上出现的最为频繁,而 C(胞嘧啶)常以甲基化形式存在,脱氨后即成为 T(胸腺嘧啶)有关[11]。青杨两个样品 L1A 和 L2A 中 SNPs 类型及其发生频率的变化趋势基本一致,L2A 与 L1A 中相对应的同一 SNPs 类型的数量之比约为 2:1(表 3)。

表 3 高通量测序鉴定的青杨 SNPs 类型分析

样品	SNP 类型	数量	平均跨度/bp	SNP 类型	数量	平均跨度/bp
	转换			颠换		
L1A	A<－>G	25 091	189 787.70	A<－>T	10 398	457 969.1
	C<－>T	25 460	187 037.05	G<－>T	8 005	594 873.6
	－	－	－	C<－>G	5 833	816 383.2
	－	－	－	A<－>C	7 999	595 319.8
	总数 50 551(61.06%)			总数 32 235(38.94%)		
L2A	A<－>G	58 921	80 093.40	A<－>T	23 870	197 703.5
	C<－>T	59 003	79 982.09	G<－>T	18 657	252 944.4
	－	－	－	C<－>G	13 543	348 459.2
	－	－	－	A<－>C	18 462	255 616
	总数 117 924(61.27%)			总数 74 532(38.73%)		

3 讨论

Solexa 高通量测序分析发现,青杨 SSR 的总发生频率为 0.149,平均跨度为 4 531.87 bp,明显高于毛白杨(*Populus tomentosa*)基因组中 SSR 的平均跨度(1 883 bp)[16]。这可能由 SSR 搜索标准、数据库大小及物种等因素所致。植物的 SSR 多以二、三核苷酸重复类型为主,而占优势的重复类型则在各植物中不尽相同[19－21]。除单核苷酸重复类型(33.96%)外,青杨 SSR 中同样以三核苷酸重复类型(31.00%)和二核苷酸重复类型(27.69%)为主。在青杨二核苷酸重复类型中,AG 重复类型所占比例最大(6.34%),这与油茶、花旗松等的 SSR 序列报道一致[22－23]。三核苷酸重复类型的优势分布密度因种而存在较大差异,在本研究中,三核苷酸重复类型 AAG 所占比例最多,GAA 次之,TTC、AGA、GAG 等相近,与大豆相似[24]。

青杨 L1A、L2A SNP 进行分析,发现 L1A 中含 SNP 162 343 个,平均跨度 29 332.73bp。L2A 中含 SNP 229 115 个,平均跨度 20 597.44 bp,与模式植物毛果杨基因组中 SNP 的发生频率存在较大差异(2.6/1 000 bp),比人类(*Homo sapiens*,1/1 900 bp)、大豆(*Glycine max*,1/200 bp)、水稻(*Oryza sativa*,1/268 bp)、拟南芥(*Arabidopsis thaliana*,1/3 300 bp)等的分布密度都低,这可能与物种、测序数量、SNP 分布的不均一性等有关[25－26]。青杨 SNP 类型中转换类型明显高于颠换类型,L1A 和 L2A 中转换类型分别达到 61.06%、61.27%。这是由于 DNA 序列中包含大量 CpG 位点,而 CpG 位点的胞核嘧啶极易通过脱氨和甲基化转化为胸腺嘧啶,其中该位点基因的转换频率为非 CpG 位点的 14～15 倍,颠换频率为其它位点的 3～4 倍,二者之比约为 2:1[27]。作为第三代分子标记,SNP 通常是二等位基因,主要以非连续性的方式显示,其能提供的信息量少于 SSR 标记,但其有许多不同于其它分子标记的优点,如其高密度和稳定遗传的特性弥补了信息量的不足,高效快捷[28]。SNPs 的缺失率要明显的低于 SSRs,约低于 SSR 的四倍,表明 SNP 标记比 SSR 标记更可靠,但其高昂的成本使其应用范围受限。Hamblin 等在对玉米进行近亲交配时,发现 SSRs 在聚类种质中的表现比 SNPs 好,因此他们认为当 SNP 位点的数量足够大时可以代替 SSRs 来研究遗传多样性和亲缘关系[29－30]。

综上所述,我们在 Solexa 高通量转录组测序的基础上,充分挖掘青杨 SSR、SNP 信息,对青杨微卫星重复序列特征和单核苷酸多态性特征进行的分析表明,在遗传多样性研究中 SNP 标记比 SSR 标记更为可靠,为今后杨属植物 SSR、SNP 标记的开发及在遗传多样性研究等方面的应用奠定基础,并为在其他物种中进行相关分子标记的有效开发提供了宝贵

的经验和理论指导。

参考文献:

[1] 张镱锂, 李炳元, 郑　度. 论青藏高原范围与面积[J]. 地理研究, 2002,1: 1 -8.

[2] 莫申国, 张百平, 程维明, 等. 青藏高原的主要环境效应[J]. 地理科学进展, 2004, 2: 88 -96.

[3] Myers N, Mittermeier R A, Mittermeier C G, *et al*. Biodiversity hotspots for conservation priorities[J]. Nature, 2000, 403(6772): 853 -8.

[4] Fang Z F, Zhao S D, Skvortsov A K. In: Wu Z Y, Raven P H eds. Flora of China (4)[M]. Beijing: Science Press and St. Louis: Missouri Botanical Garden Press, 1999.

[5] Lu Z, Wang Y, Peng Y, *et al*. Genetic diversity of *Populus cathayana* Rehd populations in southwestern china revealed by ISSR markers[J]. Plant Science, 2006, 170(2): 407 -412.

[6] Wright J M, Bentzen P. Microsatellites: genetic markers for the future, In Carvalho G R, Pitcher T J, eds, Molecular genetics in fisheries[M]. London: Chapman & Hall Ltd., 1995, 117 -121.

[7] Zane L, Bargelloni L, Patarnello T. Strategies for microsatellite isolation: a review[J]. Molecular ecology, 2002, 11(1): 1 -16.

[8] Schrey A W, Heist E J. Microsatellite analysis of population structure in the shortfin mako (*Isurus oxyrinchus*)[J]. Canadian Journal of Fisheries and Aquatic Sciences, 2003, 60(6): 670 -675.

[9] Rassmann K, Schlötterer C, Tautz D. Isolation of simple - sequence loci for use in polymerase chain reaction-based DNA fingerprinting [J]. Electrophoresis, 1991, 12(2/3): 113 -118.

[10] Subramanian S, Mishra R K, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions[J]. Genome Biol, 2003, 4(2): R13.

[11] Brookes A J. The essence of SNPs[J]. Gene, 1999, 234(2): 177 -186.

[12] Lewis R. SNPs as windows on evolution[J]. The Scientist, 2002, 16(1): 16 -18.

[13] 褚延广, 苏晓华. 单核苷酸多态性在林木中的研究进展[J]. 遗传, 2008(10): 1272 -1278.

[14] Grabherr M G, Haas B J, Yassour M, *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome[J]. Nat Biotech, 2011, 29(7): 644 -652.

[15] Li R, Li Y, Fang X, *et al*. SNP detection for massively parallel whole-genome resequencing[J]. Genome research, 2009, 19(6): 1124 -1132.

[16] Li S, Yin T. Map and analysis of microsatellites in the genome of *Populus*: the first sequenced perennial plant[J]. Sci China C Life Sci, 2007, 50(5): 690 -9.

[17] Tuskan G A, Difazio S, Jansson S, *et al*. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray)[J]. Science, 2006, 313(5793): 1596 -1604.

[18] Wright J, Bentzen P, Microsatellites: genetic markers for the future [M]//Carvalha G, Pitcher T. Molecular Genetics in Fisheries. Netherlands: Springer, 1995.

[19] Tuskan G A, Gunter L E, Yang Z K, *et al*. Characterization of microsatellites revealed by genomic sequencing of *Populus trichocarpa* [J]. Canadian Journal of Forest Research, 2004, 34(1): 85 -93.

[20] Kumpatla S P, Mukhopadhyay S. Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species [J]. Genome, 2005, 48(6): 985 -98.

[21] Varshney R K, Thiel T, Stein N, *et al*. In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species[J]. Cell Mol Biol Lett, 2002, 7(2a): 537 -46.

[22] Amarasinghe V, Carlson J E. The development of microsatellite DNA markers for genetic analysis in Douglas-fir[J]. Canadian Journal of Forest Research, 2002, 32(11): 1904 -1915.

[23] 温　强, 徐林初, 江香梅, 等. 基于 454 测序的油茶 DNA 序列微卫星观察与分析[J]. 林业科学, 2013(08): 43 -50.

[24] Gao L, Tang J, Li H, *et al*. Analysis of microsatellites in major crops assessed by computational and experimental approaches[J]. Molecular Breeding, 2003, 12(3): 245 -261.

[25] 吴　玲, 付凤玲, 李晚忱, 等. 利用生物信息学方法进行基于表达序列标签的玉米单核苷酸多态性标记的开发[J]. 核农学报, 2010(05): 968 -972 +1019.

[26] 赵春霞, 石先哲, 吕　申, 等. 人类基因组的单核苷酸多态性及其研究进展[J]. 色谱, 2003(02): 110 -114.

[27] Garg K, Green P, Nickerson D A. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags[J]. Genome Res, 1999, 9 (11): 1087 -92.

[28] Nianjun L, Liang C, Shuang W, *et al*. Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure[J]. BMC Genetics, 2005, 6(Suppl +1).

[29] Simic D, Ledencan T, Jambrovic A, *et al*. SNP and SSR marker analysis and mapping of a maize population[J]. Genetika, 2009, 41(3): 237 -246.

[30] Jones E S, Sullivan H, Bhattramakki D, *et al*. A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (Zea mays L.) [J]. Theoretical and Applied Genetics, 2007, 115(3): 361 -371.