植物研究 2019,39(1):156~160
Bulletin of Botanical Research

小黑杨基因组的初步组装及 SSR 信息分析

周玉敏[1] 王 遂[2] 刘 轶[2] 李开隆[2] 由香玲[2*]

(1. 湖北生态工程职业技术学院,武汉 430200;2. 东北林业大学林木遗传育种国家重点实验室,哈尔滨 150040)

摘 要 小黑杨是人们以小叶杨和欧洲黑杨为亲本培育出的杂交种,兼具双亲生长速度快、抗性强的优点。本研究通过二代测序技术,对小黑杨进行全基因组测序,初步组装出小黑杨基因组,并以此为基础,识别分析其 SSR 序列,为小黑杨品种划分、表型性状关联等提供参考。结果表明,共组装得到了 366 876 条总计 368.96 Mbp 的 contig 序列;对其中不小于 2 000 bp 的 21 788 条的非冗余 contig 进行 SSR 分析,共识别出 18 111 条 SSR 序列,其中一、二、三核苷酸重复基序较多。对得到的 SSR 序列设计引物,共得到 12 838 对引物,供今后实验使用。

关键词 小黑杨;基因组;SSR

中图分类号:S792.119 **文献标志码**:A **doi**:10.7525/j.issn.1673-5102.2019.01.019

Sequencing and Assembly of *Populus simonii* × *P. nigra* Genome and SSR Analysis

ZHOU Yu-Min[1] WANG Sui[2] LIU Yi[2] LI Kai-Long[2] YOU Xiang-Ling[2*]

(1. Hubei Ecology Polytechnic College, Wuhan 430200; 2. Northeast Forestry University, Harbin 150040)

Abstract *Populus simonii* × *P. nigra*, which is the hybrid crossed by *P. simonii* and *P. nigra*, inherits the advantages of fast growth and strong resistance from their parents. In this study, the genome sequence of *Populus simonii* × *P. nigra* was preliminarily sequenced and assembled by the next generation sequencing (NGS) technology. The SSR sequences were identified and analyzed. We hope this study can provide a reference for classification and phenotypic correlation of *Populus simonii* × *P. nigra*. The results show that we total assembled 368.96 Mbp genome sequence, which contain 366 876 contigs. SSR analysis was performed on 21 788 non-redundant contigs which were not less than 2 000 bp. In total, 18 111 SSR sequences were identified, and most of them are one, two or three nucleosides acid repeat motif. The primers were designed for the obtained SSR sequences, and a total of 12 838 primers were obtained for future experiments.

Key words *Populus simonii* × *P. nigra*; genome; SSR

小黑杨(*Populus simonii* × *P. nigra*)是中国林业科学院林业研究所黄东森等人于 1960 年以取自北京地区的小叶杨(*Populus simonii* Carr)为母本,取自前苏联巴什基尔共和国首都乌法的花枝的欧洲黑杨(*P. nigra* L.)为父本,人工杂交培育的新品种。小黑杨喜光,喜冷湿气候,常生长于土壤肥沃、排水良好的沙质土壤上,在我国黄河以北各省区均有分布;其生长速度快,树干通直圆满,树高可达 20 m;同时,其适应能力较强,对低温、干旱、盐碱,营养亏缺等逆境均有一定程度的抗性。小黑杨木材材质优良,均匀细致、色白、心材不明显,适做造纸、纤维等工业原料,又可供建筑、家具及农业使用,是我国北方地区重要的经济绿化树种[1~2]。

基金项目:国家自然科学基金项目(31670675)资助
第一作者简介:周玉敏(1972—),女,教授,主要从事园林植物的教学和研究。
* 通信作者:E-mail:yxiangling@yahoo.com
收稿日期:2018-03-04
Foundation item: Supported by the National Natural Science Foundation of China (31670675)
First author introduction: ZHOU Yu-Min (1972—), female, professor, Mainly engaged in teaching and research garden plants.
* Corresponding author: E-mail: yxiangling@yahoo.com
Received date: 2018-03-04

由于小黑杨的最初来源为集团选择的两个集团杂交种子,在杂交的过程中,基因的自由组合与染色体的连锁交换,使其子代间产生了丰富的基因型与表型。经过半个多世纪的引种与推广,我国不同省份区域的科研工作者筛选出了多个适合当地立地条件的小黑杨品种[3~4]。然而,随着种植区域的扩大,许多地区对引进的小黑杨品种信息缺失;有的地区并未进行引种试验而将其他区域的所谓优树直接引入;同时,已有的小黑杨良种经过半个多世纪种植,品质与抗性均有所下降。近年来,小黑杨的新增种植面积逐年减少,市场占有率持续下降。因此,重新对已有小黑杨品种进行划分,将品种与表型等性状相关联,建立小黑杨品种数据库,具有重要的现实意义。

简单重复序列(Simple Sequence Repeats,SSR)标记是近年来被广泛使用的一种以微卫星序列多态性为基础的分子标记技术,人们利用 SSR 两端序列的高度保守性,设计特异引物,通过 PCR 将其扩增出来进而利用电泳区分不同个体序列长度的差异,具有高度重复性、丰富的多态性、共显性、高度可靠性等优点。在林木的生产实践中,良种的选育是育种工作者主要追求的目标,常规的良种选育方法主要针对表型进行选育,其结果周期长,稳定性差,易受环境影响;而利用分子标记将表型选择转换为基因型选择,做到有的放矢,可以极大地缩短育种周期[5]。但由于林木基因组信息相对匮乏,准确可靠标记的获得并不容易。目前,SSR 分子标记技术已广泛用于杨树品种鉴定及遗传多样性分析[6~8]。黄烈健等人在 132 对 SSR 引物中筛选出了多对与杨树木材密度,纤维长、宽,纤维丝角等相关联的 SSR 标记,为标记辅助育种奠定了基础[9];梁海永等人利用 10 对 SSR 引物,将 10 个杨树品种分为 3 大类[8];张新叶等人基于 EST 序列,设计了 48 对全新的 SSR 引物以区分杨树品种[10];宋跃朋等人利用 16 个杨树无性系比较了 10 对 Genomic-SSR 引物和 10 对 EST-SSR 引物的遗传差异[11]。从技术的角度讲,早期杨树的 SSR 鉴定分析多使用通用引物,特异性较差,而随着二代测序技术的普及和三代测序价格的下降,生物体基因组测序拼接的成本显著降低,通过全基因组测序,人们可以较为精准地了解物种的基因组序列信息,这在很大程度上推动了 SSR 的快速发展[12]。

为了对小黑杨进行 SSR 序列识别和信息分析,本研究将首先利用二代测序技术对小黑杨基因组进行 *de novo* 测序,获得小黑杨基因组组装的初步结果,进而分析 SSR 序列信息,为今后利用 SSR 标记进行小黑杨品种划分、表型性状关联等奠定基础。由于小黑杨是由青杨派的小叶杨和黑杨派的小黑杨杂交而来,其基因组含有两种杨树派系的遗传信息,因而得到的 SSR 序列信息也可以用来进行青杨派和黑杨派杨树的遗传分析。同时,拼接得到的小黑杨基因组信息,也为今后小黑杨的研究提供了参考。

1　材料与方法

1.1　样品取材与基因组 DNA 提取

于 2017 年 6 月 25 日在黑龙江省哈尔滨市东北林业大学校园内选择一株长势良好、无病虫害的小黑杨,取其成熟叶片若干,存于液氮中备用。参考 BioTeke 新型快速植物基因组 DNA 提取试剂盒(BioTeke,DP3111)说明书进行小黑杨基因组 DNA 提取操作。将得到的 gDNA 送华大基因科技服务有限公司(武汉,中国),构建 insert size 约为 250 bp 的小片段文库,基于 Illumina HiSeq X Ten 平台,进行 PE151 测序。

1.2　数据质控

利用 FastQC(v0.11.5)软件,对公司返回的去除了接头和引物序列的 raw data 进行测序质量统计。根据得到的结果,通过 NGSQCtoolkit(v2.3.3)套件对原始数据进行过滤,同时使用 FastUniq(v1.1)去除 PCR 重复[13],最终得到符合拼接要求的 clean data。

1.3　基因组序列拼接

使用 Edena(v3.131028)对小黑杨基因组进行初步组装,设定组装得到的 contig 长度不小于 500 bp,同时对得到的 contig 序列进行统计[14]。选取长度不小于 2 000 bp 的 contig 与 NCBI 的 Nt 库(更新于 2017 年 9 月 17 日)进行 Blastn 比对,其中 max_target_seqs 设定为 20,evalue 为 1e-5,相似性阈值设定为不小于 60%,对比对到的物种进行统计分析。

1.4　SSR 序列识别与分析

将 Edena 组装得到的小黑杨基因组进行过滤,保留长度大于等于 2 000 bp 的 contig,用 cd-hit 去除冗余,再利用 MIcroSAtellite identification tool(MISA)软件进行 SSR 序列的识别和统计。对 SSR 的限制条件设定为 1 个碱基重复不小于 10

次;2 个碱基重复不小于 6 次;3 个碱基重复不小于 5 次;4 个碱基重复不小于 5 次;5 个碱基重复不小于 5 次;6 个碱基重复不小于 5 次。同时,两个微卫星之间距离小于 100 bp 时,2 个微卫星组成 1 个复合微卫星。

1.5 计算资源

本研究计算平台为东北林业大学高性能计算机集群。

2 结果与分析

2.1 数据质控

FastQC 对 raw data 的统计结果显示,华大基因实际交付的去除接头和引物的 raw data 信息采集大小为 42.49 Gbp,reads 长度 150 bp,GC 含量为 40%,碱基整体质量较好,达到了合同要求。reads 单碱基质量分布盒形图结果显示,reads 前几个碱基质量较差,这可能是测序引物刚刚与 reads 结合,测序不稳定的结果;而 reads 后几个碱基质量也下降较快,这主要是随着 reads 的延伸,酶效率的下降,造成复制错误累积而造成的。而每个位点的碱基含量统计结果也显示,前几个碱基 A 与 T,G 与 C 含量并不相等,说明 reads 前几个碱基准确性较低。因此,在数据过滤时,我们截去了 reads 5′端 10 个碱基和 3′端 5 个碱基,进而以 4 个碱基为窗口,从 5′端向 3′端滑动,当平均质量小于 15 时,将其切除。由于在小黑杨 DNA 文库构建的过程中经过了 PCR 来提升 DNA 浓度,测序结果中会含有 PCR 重复,这对基因组的拼接并没有帮助,因此使用 FastUniq 将重复去掉。最终,我们得到了 29.64 Gbp 的 clean data,reads 长度 135 bp,GC 含量依然为 40%。

2.2 基因组组装

前期的流式细胞仪检测和 k-mer 分析均显示,小黑杨基因组与毛果杨(*Populus trichocarpa*)相近,约为 418 Mbp。即使是过滤后的 clean data 其测序深度也达到了 70x,远高于一般的简化基因组和重测序,使其拼接结果可信度更高。根据小黑杨基因组小于 500 Mbp,用于拼接的 reads 质量较好,且拼接结果主要用于 SSR 等分析的特点,因此选用 Edena 进行组装。Edena 是一款基于 overlaps-graph-based 的 *de novo* 组装软件,其使用简便,运行速度快,无需输入插入片段长度和 k-mer 等参数,避免了对不同 k-mer 值的循环尝试,特别适合小基因组的初步组装。由于本研究主要是为 SSR 分析,且仅构建了一个小片段文库,因此在组装基因组时,直接将小于 500 bp 的 contig 忽略。经过拼接,最终得到了 366 876 条 contig,总计大小为 368.96 Mbp,其中最长的 contig 为 49.87 Kbp,平均 contig 为 1.01 Kbp,N50 为 1.05 Kbp,GC 含量为 37.09%。

为了检测 gDNA 提取时是否混有细菌等污染,同时对拼接结果进行初步分析,我们将 contig 长度不小于 2 000 bp 的 22 634 条序列与最新的 Nt 数据库进行比对。对所有的 query 物种注释信息进行统计。结果显示,比对注释到的物种共有 240 个,总计 17 804 次。其中注释得到最多的是毛果杨,共有 12 010 次成功比对,其次是胡杨(*Populus euphratica*),有 3 038 次成功比对,而比对数最多的前 10 个物种的总注释数占全部注释物种次数的 92.08%(表 1)。从注释的结果看,小黑杨基因组与毛果杨高度相似,其次是胡杨,与其为黑杨派与青杨派的杂交种起源相符。

表 1 小黑杨基因组比对注释统计

Table 1 Genome alignments and annotations statistics of *Populus simonii* × *P. nigra*

物种名称 Species name	注释次数 Number of annotations
毛果杨 *Populus trichocarpa*	12 010
胡杨 *Populus euphratica*	3 038
葡萄 *Vitis vinifera*	387
毛白杨 *Populus tomentosa*	202
美洲黑杨 *Populus deltoides*	168
核桃 *Juglans regia*	156
巴西橡胶树 *Hevea brasiliensis*	136
大叶钻天杨 *Populus balsamifera*	111
蓖麻 *Ricinus communis*	96
桃 *Prunus persica*	90

2.3 SSR 序列的识别分析

由于小黑杨为青杨派与黑杨派的杂交种,因此具有较高的杂合度。这有可能导致在基因组组装的过程中,有的姐妹染色单体不能合并,造成组装结果偏大,序列可能存在冗余。同时,长度较短的 contig 可能是重复区域,且在引物设计上存在困难。因此,在进行 SSR 识别之前,最好先进行序列过滤,对非冗余的序列进行分析。利用 cd-hit 对长度不小于 2 000 bp 的 contigs 合并,得到 21 788 条非冗余 contig,再利用 MISA 进行识别分析。结果

显示，在 10 969 条含有 SSR 的 contig 中，共识别得到 18 111 条 SSR。其中 SSR 数量较多的基序类型是一、二、三核苷酸重复，数量分别是 13 207，2 960，1 644，依次占总 SSR 数目的 72.92%，16.34%，9.08%。而五、六核苷酸重复类型所占比例较少，仅有 53 条和 44 条，分别占 SSR 总数的 0.29% 和 0.24%（表 2）。

表 2　小黑杨 SSR 序列信息

Table 2　Information of SSR sequences in *Populus simonii* × *P. nigra*

SSR 类型 SSR type	数量 Number	百分比 Percentage (%)	重复基序种类数 Motif type number
单核苷酸 Mononucleotide	13 207	72.92	4
二核苷酸 Dinucleotide	2 960	16.34	12
三核苷酸 Trinucleotide	1 644	9.08	57
四核苷酸 Tetranucleotide	203	1.12	67
五核苷酸 Pentanucleotide	53	0.29	40
六核苷酸 Hexanucleotide	44	0.24	44

从得到的结果看，不同核苷酸基序种类及重复次数差异较大。在考虑到碱基互补配对原则的情况下，单核苷酸基序主要为 A/T 重复，且重复次数多为 10～13 次，而 C/G 基序出现频率较低；二核苷酸基序重复次数最多的是 AG/CT，重复次数多在 6 与 10 之间；三核苷酸基序重复次数最多的为 AAT/ATT，多数重复 5～7 次；四核苷酸重复次数最多的为 AAAT/ATTT，其重复 5 次的共有 68 条；五、六核苷酸基序重复次数最多的分别是 AAAAG/CTTTT 与 AAAAAT/ATTTTT。同时，根据 SSR 位点信息，利用 Primer3 批量设计了 12 838 对引物，供实验使用。由于篇幅原因，拼接组装得到的基因组文件，过滤后的去冗余 contig 序列文件及 SSR 位点详细信息与其统计文件及引物相关文件，均保存在 http://www.wangsui.net.cn/resource/database/public/plant/Populus/xiaohei/survey/SSR/目录下，供下载。

3　讨论

以 SSR 为分子标记进行品种鉴定和多样性分析已经有二十余年的历史了，早期人们多使用通用引物在不同品种甚至不同物种中进行鉴定，引物特异性较差，常常合成的大多数引物不能很好地扩增目的序列；随着技术的进步，人们可以对一些片段的两端序列进行测定，EST-SSR 开始兴起，但其序列信息也仅局限于 cDNA 两端的短片段；近年来，测序技术的突飞猛进，使得通过全基因组测序而获得大片段的物种基因组序列，进而根据序列信息筛选 SSR 成为可能。从本研究的结果来看，构建价格极低的小片段文库进行全基因组测序，拼接得到准确度较高的 contig，再分析 SSR 信息，不仅可以得到更多的信息，准确性也大大提高。本研究提供的序列过滤，拼接，SSR 识别及引物设计构成了一个较为完整的 pipeline，使用方便，对计算资源的要求也不高，适合有条件的实验室依据自身平台在更多的物种上展开分析。

杨树是在全球广泛分布的重要经济树种，更作为木本模式植物，于 2006 年率先完成了全基因组测序[15]。杨树相关的研究因此得到了快速发展。然而，由于杨树派系众多，不同派系，不同品种的杨树之间基因组差异巨大，仅依据毛果杨（*Populus trichocarpa*）基因组序列进行分析，可能会存在一定偏差。本研究通过二代测序技术，对小黑杨基因组进行了初步组装，利用去冗余的 contig 序列进行 SSR 分析，设计引物并将全部信息公布在实验室网站。为今后小黑杨的品种划分、表型性状关联及基因组相关研究奠定基础，同时也为青杨或黑杨派杨树的遗传分析提供了一定的参考。

参　考　文　献

1. 李国梁. 小黑杨引种初报[J]. 科技简报，1981(1)：8.
   LI G L. Preliminary report on introduction of *Populus simonii* × *P. nigra*[J]. Technology Bulletin，1981(1)：8.
2. 沈清越，康忠信，刘亚芹. 杨树良种—小黑杨[J]. 林业科技通讯，1979(7)：7－8.
   Shen Q Y，Kang Z X，Liu Y Q. Poplar fine varieties-*Populus simonii* × *P. nigra*[J]. Forest Science and Technology，1979(7)：7－8.
3. 鹿学程. 昭盟小黑杨的选种[J]. 内蒙古林业科技，1983(1)：1－12.
   Lu X C. Seletion seed of *Populus simonii* × *P. nigra*[J]. Journal of Inner Mongolia Forestry Science and Technology，1983(1)：1－12.
4. 崔景山，赵庆武，张贵，等. 双辽县沙地杨树品种比较试验研究[J]. 吉林林业科技，1991(5)：1－5.
   Cui J S，Zhao Q W，Zhang G，et al. Comparative experimental study on Poplar Varieties in sandy land of Shuangliao County[J]. Jilin Forestry Science and Technology，1991(5)：1－5.

5. 黄秦军,苏晓华,张香华. SSR 分子标记与林木遗传育种[J]. 世界林业研究,2002,15(3):14 – 21.
Huang Q J,Su X H,Zhang X H. Microsatellite marker and its application in tree genetics and breeding[J]. World Forestry Research,2002,15(3):14 – 21.

6. 藕丹,樊军锋,高建社,等. SSR 和 SCoT 标记在美洲黑杨×青杨派杂种无性系遗传差异性分析上的比较[J]. 西北农林科技大学学报:自然科学版,2017,45(4):79 – 85,93.
Ou D,Fan J F,Gao J S,et al. Comparison of genetic diversity of *Populus deltoides* × *Section tacamahaca* hybrids based on SSR and SCoT markers[J]. Journal of Northwest A&F University:Natural Science Edition,2017,45(4):79 – 85,93.

7. 韩志校,张军,左力辉,等. 3 个不同派别杨树资源遗传多样性的 SSR 分析[J]. 河南农业科学,2017,46(4):99 – 103.
Han Z X,Zhang J,Zuo L H,et al. Analysis of genetic diversity of populus from three different factions by SSR markers[J]. Journal of Henan Agricultural Sciences,2017,46(4):99 – 103.

8. 梁海永,刘彩霞,刘兴菊,等. 杨树品种的 SSR 分析及鉴定[J]. 河北农业大学学报,2005,28(4):27 – 31.
Liang H Y,Liu C X,Liu X J,et al. Simple sequence repeat (SSR) analysis and identify of different cultivars in *Populus* L. [J]. Journal of Agricultural University of Hebei,2005,28(4):27 – 31.

9. 黄烈健,苏晓华,张香华,等. 与杨树木材密度、纤维性状相关的 SSR 分子标记[J]. 遗传学报,2004,31(3):299 – 304.
Huang L J,Su X H,Zhang X H,et al. SSR molecular markers related to wood density and fibre traits in poplar[J]. Acta Genetica Sinica,2004,31(3):299 – 304.

10. 张新叶,宋丛文,张亚东,等. 杨树 EST-SSR 标记的开发[J]. 林业科学,2009,45(9):53 – 59.
Zhang X Y,Song C W,Zhang Y D,et al. Development of EST-SSR in *Populus deltoides* and *P. euramericana*[J]. Scientia Silvae Sinicae,2009,45(9):53 – 59.

11. 宋跃朋,江锡兵,张曼,等. 杨树 Genomic-SSR 与 EST-SSR 分子标记遗传差异性分析[J]. 北京林业大学学报,2010,32(5):1 – 7.
Song Y P,Jiang X B,Zhang M,et al. Genetic differences revealed by Genomic-SSR and EST-SSR in poplar[J]. Journal of Beijing Forestry University,2010,32(5):1 – 7.

12. Zalapa J E,Cuevas H,Zhu H Y,et al. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences[J]. American Journal of Botany,2012,99(2):193 – 208.

13. Xu H B,Luo X,Qian J,et al. FastUniq:a fast de novo duplicates removal tool for paired short reads[J]. PLoS One,2012,7(12):e52249.

14. Hernandez D,François P,Farinelli L,et al. De novo bacterial genome sequencing:millions of very short reads assembled on a desktop computer[J]. Genome Research,2008,18(5):802 – 809.

15. 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.