Hans Journal of Agricultural Sciences 农业科学, 2018, 8(7), 699-704
Published Online July 2018 in Hans. http://www.hanspub.org/journal/hjas
https://doi.org/10.12677/hjas.2018.87105

Hans汉斯

SNP Distribution Characteristic of Chinese Wolfberry Based on RAD Sequencing

Guanghui Fan[1,2], Hang Yu[1], Zhanlin Wang[1,2]

[1]Qinghai Academy of Agriculture and Forestry, Qinghai University, Xining Qinghai
[2]Qinghai Plateau Key Laboratory of Tree Genetics and Breeding, Xining Qinghai
Email: qhfgh@163.com

Received: Jun. 20th, 2018; accepted: Jul. 3rd, 2018; published: Jul. 10th, 2018

Abstract

The single-nucleotide polymorphisms (SNPs) in the genome of *Lycium barbarum* were identified using the high throughput sequencing technology based on the Illumina HiSeq2500 platform. A total of 5,780,671,000 bp high quality data were produced. All of the reads were assembled into 880,315 contigs with 295 bp average length. Using the contig assemblies as a reference, 721,813 SNPs were identified. Among the SNPs, transitions were 454,827, transversions were 266,986, and the value of Ts/Tv was 1.70. Among the SNPs, A/G (31.69%) was the most abundant, followed with C/T (31.32%), A/C (10.78%), G/T (10.75%), A/T (10.27%) and C/G (5.18%).

Keywords

***Lycium*, SNPs, RAD, High Throughput Sequencing**

基于RAD测序的枸杞SNP分布特征分析

樊光辉[1,2]，虞　杭[1]，王占林[1,2]

[1]青海大学农林科学院林业研究所，青海　西宁
[2]青海高原林木遗传育种实验室，青海　西宁
Email: qhfgh@163.com

收稿日期：2018年6月20日；录用日期：2018年7月3日；发布日期：2018年7月10日

摘　要

利用illumina hiseq2500平台，对枸杞进行了RAD测序并对其SNPs的数目和分布特征进行了分析和比较。

文章引用: 樊光辉, 虞杭, 王占林. 基于 RAD 测序的枸杞 SNP 分布特征分析[J]. 农业科学, 2018, 8(7): 699-704.
DOI: 10.12677/hjas.2018.87105

测序后共得到5,780,671,000 bp的高质量数据，经过组装后得到平均长度为295 bp的contig 880,315个。采用软件进行SNP的检测后得到721,813个SNPs。其中转换替换共有454,827个，占总数的63.01%，颠换共有266,986个，占总数的36.98%，转换和颠换的比例Ts/Tv为1.70。所有的替换类型中A/G占的比例最高，为31.69%，其余类型所占的比例依次为C/T (31.32%), A/C (10.78%), G/T (10.75%), A/T (10.27%)和C/G (5.18%)。

关键词

枸杞，SNPs，RAD标记，高通量测序

1. 引言

枸杞为茄科枸杞属的灌木树种，生长在干旱和半干旱地区，盐碱生境和海岸带也有分布(Fukuda *et al*., 2001) [1]。我国枸杞的分布地区基本集中在西北和其它干旱、半干旱地区以及一些盐碱地区(Jia *et al*., 2009) [2]。枸杞对于恶劣的土壤和气候环境条件具有很强的适应性，能在极度干旱、含盐量高的土壤中正常生长，因此在北方，尤其是我国青海、内蒙、甘肃、宁夏和新疆地区作为于防沙、治沙和固沙的先锋树种(Zhao *et al*., 2004) [3]。

枸杞果实中含有很多活性物质，例如枸杞多糖，生物碱，黄酮和类胡萝卜素等物质，因而其具有促进免疫、抗衰老、抗肿瘤、清除自由基、抗疲劳、抗辐射、保肝、生殖功能保护和改善等多种作用(Inbaraj *et al*., 2010; Duan *et al*., 2010) [4] [5]。枸杞果实和根很早就被用于治疗眼疾和炎症，也是治疗肝胆疾病和肾脏方面的疾病的传统药物(Hitchcock, 1932) [6]。由于枸杞的生态和经济的双重作用，其在我国北方被广泛种植，甚至在宁夏，青海和新疆等地区已经成为地区经济收入的支撑产业。

SNP (Single-nucleotide polymorphisms)，即单核苷酸多态性，是基因组序列中最丰富的 DNA 多态性，它们能够影响到蛋白的功能，因而也是很多疾病和表型特征变异的基础。

一般情况下，物种基因组内 SNPs 的分布和其分布特征是不均匀的，编码区的 SNP 分布频率要低于非编码区。自然选择，遗传重组，突变率以及其他的因素都能影响到 SNPs 的分布密度(Nachman 2001) [7]。

SNPs 中有一部分是能够通过影响蛋白的功能而影响到物种表型的变异，另一部分是对于表型没有任何影响的变异，这类变异称为沉默突变，或者同义突变。这类突变数量巨大，并且具有稳定遗传的特点，因而在全基因组关联分析(GWAS, genome-wide association studies)，遗传图谱的构建(Thomas, 2011) [8]，QTL 分析(Garrett *et al*., 2012) [9]，分子标记辅助育种(Thavamanikumar *et al*., 2011) [10]等反面都被广泛利用。

基于其在遗传学和基因组学方面的重要性，检测和研究 SNP 在基因组上的分布和特征也具有重要意义(Steele *et al*., 2008) [11]。随着第二代通量测序技术的发展，其高通量、省时和高效的特点使得对 SNP 的检测也步入新的阶段。迄今为止，基于高通量测序的方法进行 SNP 的检测已经在很多物种中加以利用。

枸杞的植物化学，药理学和育种等方面研究已经比较深入，但其分子水平上的研究目前还处于起步阶段，其基因组 SNPs 方面的研究尚未见报道。基础研究尤其是分子生物学方面的滞后在一定程度上影响了枸杞育种工作，所以其遗传学和基因组学方面的研究亟待进行。基于枸杞在经济和生态方面的重要性，本文利用第二代高通量测序技术用 RAD 标记对枸杞的基因组进行了简化分析后查找了其基因组水平

上的 SNPs 标记，这些标记可以用于下一步枸杞高密度遗传图谱，关联分析的标记开发和使用，为枸杞遗传学和基因组学研究奠定基础。

2. 材料和方法

2.1. 植物材料和 DNA 提取

宁夏枸杞(*L. barbarum* L.)新鲜叶子于 2014 年八月份采自青海诺木洪农场枸杞种质资源圃。基因组 DNA 利用 kitDP305 (天根，北京)提取。提取后利用 NanoDrop 2000 (Wilmington, DE, USA)和琼脂糖凝胶电泳进行质量检测。

2.2. 建库和测序

用 RAD (Restriction-site associated DNA-sequencing)测序方法将枸杞基因组进行简化(Baird, 2008) [12]。取基因组 DNA1ug，利用 EcoRI 内切酶进行消化(G|AATTC)，然后加 P1 接头(可与 EcoRI 酶切 DNA 缺口互补)；将连接有接头的所有片段混合后随即打断，电泳回收 300 bp~700 bp 的片段，然后末端平化后加 A；加 Solexa P2 Adapter，P2 为局部双链分叉 Y 型 DNA，可实现选择性的扩增同时含有 P1 和 P2 接头的 RAD 标记；PCR 扩增两端分别含有 P1 和 P2 接头的 tag 序列。制备好的测序库利用 Qubit 2.0kit (Life Technologies, Carlsbad, CA, USA)检测质量，Agilent 2100 (AgilentTechnologies, Palo Alto, Calif)检测片段的大小。检测后的测序库利用 Illumina HiSeq2500 (Illumina Inc., San Diego, Calif)根据程序进行测序。

2.3. 数据质量控制和组装

利用 In-House scripts 将低质量和重复的测序数据的去除，使用 EcoRI (G|AATTC)酶切位点，对 Clean reads 进行去除重复处理后，统计去重后 EcoRI 捕获的 Reads 数。利用 Velvet Optimiser software (Zerbino D R and Birney E, 2008) [13]，根据默认参数进行数据组装。将带有酶切识别序列的 reads 进行聚类，并按照深度由大到小进行排序。将深度高的 reads 作为种子进行聚类。根据深度信息对聚类后的 reads 进行纠错、过滤重复区域等。根据聚类的结果，将另一端的 reads 进行 contig 拼接，结合插入片段的大小和 overlap 的关系，将拼接好的 contig 与另一端聚类的 reads 进行连接，组装成最终的 contig 序列。

2.4. SNP 查找

利用 BWA 软件(Li and Durbin 2009) [14]将测序的所有 reads 比对到组装好的序列上，比对结果经 SAMTOOLS (Li *et al*., 2009) [15] [16]去除重复(参数：rmdup)。Candidate sequence variation were filtered according to the following criteria：利用贝叶斯模型检测群体中的多态性位点，通过以下过滤和筛选得到高质量的 SNPs：1) Q20 质量控制(将质量值 Q20 即测序错误率大于 1%的 SNPs 过滤掉)；2) SNP 的支持数(覆盖深度)在 2~1000 范围内；3) 缺失控制(将群体内 SNP 位点缺失率大于 0.1 的位点过滤掉)。

3. 实验结果

3.1. 测序结果和质量

利用 Illumine HiSeq2500 平台，测序共得到 5,868,777,750 bp 的碱基，去除低质量的测序数据后得到 5,780,671,000 bp 的高质量数据，数据有效率达到 98.5%，错误率为 0.04%，Q20 和 Q30 分别为 92.87%和 86.9%，GC 含量为 37.97%。

3.2. RAD tag 统计和聚类及局部组装结果

捕获的 Reads 数为 21,317,936，占 Clean reads 去重后的 96.76%。经过组装后得到平均长度为 295 bp

的 contig 880,315 个，共涉及 260,163,757 bp。

3.3. 比对到参考序列结果

将测序得到的reads对比到组装好的序列，共有41,912,578条reads成功比对，占总reads数目的90.63%，这些比对的 reads 的平均测序深度为 14.65。

3.4. SNP 检测结果

检测过滤和筛选后得到 721,813 个 SNPs。其中转换替换共有 454,827 个，占总数的 63.01%，颠换共有 266,986 个，占总数的 36.98%，转换和颠换的比例 Ts/Tv 为 1.70。这个比值远远大于 Ts/Tv 的理论比值 0.5，这种 Ts/Tv 实际值与理论值不一致的情况称为“转换偏差”(Collins *et al*., 1994) [17]。转换偏差的产生可能是物种在长期的进化过程基于进化选择中形成减少有害突变形成的方式(Li *et al*., 1984; Wakeley, 1996) [18] [19]，也有可能是 DNA 分子内嘌呤和嘧啶的结构以及代谢等内在特征决定(Tang *et al*., 2008) [20]。转换偏差现象在很多动植物中都有发现，例如核桃基因组中 Ts/Tv 比例为 2.79，远远大于 0.5 的理论值(廖卓毅，2015) [21]，玉米中也有这种现象的发现(Morton, 1995; Batley *et al*., 2003) [22] [23]，野鸭和火鸡的基因组研究中也有类似的报道(Kraus, 2011; Aslam, 2012) [24] [25] (表 1)。

所有的替换类型中 A/G 占的比例最高，为 31.69%，其余类型所占的比例依次为 C/T (31.32%)，A/C (10.78%)，G/T (10.75%)，A/T (10.27%)和 C/G (5.18%) (表 1)。

4. 结论与讨论

根据前人的研究，DNA 中的 5-甲基胞嘧啶(5-methylcytosine, 5mC)突变成 T 的频率较高，因此 C/T 突变在 SNPs 突变中占最高的比率(Bird, 1980) [26]。但本研究中的结果却不符合该规律，A/G 类型的替换略高于 C/T 类型的替换。青杨的基因组 SNP 分布中也有 A/G 和 C/T 数量相差极小的报道，而在其他物种的研究中 C/T 类型占有最高的比例(Chao *et al*., 2009; Kraus *et al*., 2011) [27]。

随着高通量测序技术的发展和基因组测序，转录组测序在多个物种中的开展，对于 SNPs 的研究

Table 1. SNPs type and quantity
表 1. SNPs 类型和数量

替换类型		数量	百分比(%)
转换		454,827	63.01
	A/G	228,731	31.69
	C/T	226,096	31.32
颠换		266,986	36.98
	A/C	77,843	10.78
	A/T	74,163	10.27
	C/G	37,378	5.18
	G/T	77,602	10.75
Ts/Tv		1.70	
Total		721,813	100

也不断深入。本文中对于枸杞 SNPs 的发掘和分布特性研究可以为后续的高密度遗传图谱以及一些性状和基因的关联分析(association analysis)提供有效数据，同时为枸杞中的重要基因的功能分析和挖掘奠定基础。

基金项目

本研究由青海省自然科学基金青年项目(2015-ZJ-926Q)和青海省重大科技专项(2015-NK-A2)基金共同资助。

参考文献

[1] Fukuda, T., Yokoyama, J. and Ohashi, H. (2001) Phylogeny and Biogeography of the Genus *Lycium* (Solanaceae): Inferences from Chloroplast DNA Sequences. *Molecular Phylogenetics and Evolution*, **19**, 246-258. https://doi.org/10.1006/mpev.2001.0921

[2] Jia, H.T., Zhao, C.Y., Sheng, Y., *et al*. (2009) Screening of Tree Species with High-Efficiency Carbon Sequestration for Returning Cultivated Land to Forest in Arid Area. *Journal of Xinjiang Agricultural University*.

[3] Zhao, C., Wang, Y., Song, Y., *et al*. (2004) Biological Drainage Characteristics of Alakalized Desert Soils in North-Western China. *Journal of Arid Environments*, **56**, 1-9. https://doi.org/10.1016/S0140-1963(03)00005-3

[4] Inbaraj, B.S., Lu, H., Kao, T.H., *et al*. (2010) Simultaneous Determination of Phenolic Acids and Flavonoids in *Lycium barbarum* Linnaeus by HPLC-DAD-ESI MS. *Pharmaceutical and Biomedical Analysis*, 549E-556E.

[5] Duan, H., Chen, Y. and Chen, G. (2010) Far Infrared-Assisted Extraction Followed by Capillary Electrophoresis for the Determination of Bioactive Constituents in the Leaves of *Lycium barbarum* Linn. *Chromatography*, 4511-4516.

[6] Hitchcock, C.L. (1932) A Monographic Study of the Genus *Lycium* of the Western Hemisphere. *Annals of the Missouri Botanical Garden*, **19**, 179-348, 350-366. https://doi.org/10.2307/2394155

[7] Nachman, M.W. (2001) Single Nucleotide Polymorphisms and Recombination Rate in Humans. *Trends in Genetics*, **17**, 481-485. https://doi.org/10.1016/S0168-9525(01)02409-X

[8] Thomas, P.E., Klinger, R., Furlong, L.I., Hofmann-Apitius, M. and Friedrich, C.M. (2011) Challenges in the Association of Human Single Nucleotide Polymorphism Mentions with Unique Database Identifiers. *BMC Bioinformatics*, **12**, S4. https://doi.org/10.1186/1471-2105-12-S4-S4

[9] Garrett, M., Moylan, C., Gibson, J., *et al.* (2012) Expression QTL Analysis of a Gene Expression Signature Which Predicts Advanced Non-Alcoholic Fatty Liver Disease. *Hepatology*, **56**, 267A-268A.

[10] Thavamanikumar, S., Mcmanus, L.J., Tibbits, J.F.G., *et al*. (2011) The Significance of Single Nucleotide Polymorphisms (SNPs) in Breeding Programs. *Australian Forestry*, **74**, 23-29. https://doi.org/10.1080/00049158.2011.10676342

[11] Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., *et al.* (2001) A Map of Human Genome Sequence Variation Containing 1.42 Million Single Nucleotide Polymorphisms. *Nature*, **409**, 928-933. https://doi.org/10.1038/35057149

[12] Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. and Johnson, E.A. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS One*, **3**, e3376. https://doi.org/10.1371/journal.pone.0003376

[13] Zerbino, D.R. and Birney, E. (2008) Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs. *Genome Research*, **18**, 821-829. https://doi.org/10.1101/gr.074492.107

[14] Li, H. and Durbin, R. (2009) Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

[15] Li, H. and Durbin, R. (2009) Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

[16] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map Format and SAMtools. *Bioinformatics*, **25**, 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

[17] Collins, D.W. and Jukes, T.H. (1994) Rates of Transition and Transversion in Coding Sequences since the Human-Rodent Divergence. *Genomics*, **20**, 386-396. https://doi.org/10.1006/geno.1994.1192

[18] Li, W.H., Wu, C.I. and Luo, C.C. (1984) Nonrandomness of Point Mutation as Reflected in Nucleotide Substitutions in Pseudogenes and Its Evolutionary Implications. *Journal of Molecular Evolution*, **21**, 58-71.

https://doi.org/10.1007/BF02100628

[19] Wakeley, J. (1996) The Excess of Transitions among Nucleotide Substitutions: New Methods of Estimating Transition Bias Underscore Its Significance. *Trends in Ecology & Evolution*, **11**, 158-162. https://doi.org/10.1016/0169-5347(96)10009-4

[20] Tang, P., Wang, Q. and Chen, J.Q. (2008) The Patterns and Influences of Insertions, Deletions and Nucleotide Substitutions in Solanaceae Chloroplast Genome. *Hereditas*, **30**, 1506-1511.

[21] 廖卓毅. 基于 454 测序核桃基因组微卫星和核苷酸变异序列的特征分析[D]: [硕士学位论文]. 南京: 南京林业大学, 2015.

[22] Morton, B.R. (1995) Neighboring Base Composition and Transversion/Transition Bias in a Comparison of Rice and Maize Chloroplast Noncoding Regions. *Proceedings of the National Academy of Sciences USA*, **92**, 9717-9721. https://doi.org/10.1073/pnas.92.21.9717

[23] Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J. and Edwards, D. (2003) Mining for Single Nucleotide Polymorphisms and Insertions/Deletions in Maize Expressed Sequence Tag Data. *Plant Physiology*, **132**, 84-91. https://doi.org/10.1104/pp.102.019422

[24] Kraus, R.H., Kerstens, H.H., Hooft, V.P., Crooijmans, R.P., DerPoel, J.J., Elmberg, J., Vignal, A., Huang, Y., Li, N., Prins, H.H. and Groenen, M.A. (2011) Genome Wide SNP Discovery, Analysis and Evaluation in Mallard (*Anas platyrhynchos*). *BMC Genomics*, **12**, 2191-2198. https://doi.org/10.1186/1471-2164-12-150

[25] Aslam, M.L., Bastiaansen, J.W., Elferink, M.G., Megens, H.J., Crooijmans, R.P., Blomberg, L.A., Fleischer, R.C., Tassell, C.P.V., Sonstegard, T.S., Schroeder, S.G., Groenen, M.A.M. and Long, J.A. (2012) Whole Genome SNP Discovery and Analysis of Genetic Diversity in Turkey (*Meleagris gallopavo*). *BMC Genomics*, **13**, 391. https://doi.org/10.1186/1471-2164-13-391

[26] Bird, A.P. (1980) DNA Methylation and the Frequency of CpG in Animal DNA. *Nucleic Acids Research*, **8**, 1499-1504. https://doi.org/10.1093/nar/8.7.1499

[27] Chao, S.M., Zhang, W.J., Akhunov, E., Sherman, J., Ma, Y.Q., Luo, M.C.M. and Dubcovsky, J. (2009) Analysis of Gene-Derived SNP Marker Polymorphism in US Wheat (*Triticum aestivum* L.) Cultivars. *Molecular Breeding*, **23**, 23-33. https://doi.org/10.1007/s11032-008-9210-6