《分子植物育种》网络版，2020 年，第 18 卷，第 53 篇，第 1-11 页

研究报告
Research Report

山东主要栽培苹果基因组重测序及 SNP 芯片位点挖掘

段乃彬[1*] 马玉敏[1] 王坤[2] 王效睦[1] 谢坤[1] 白静[1] 杨永义[1] 蒲艳艳[1] 宫永超[1]
1 山东省农作物种质资源中心, 济南, 250101; 2 中国农业科学院果树研究所, 兴城, 125100
* 通信作者, duannaibin@gmail.com

摘 要 为促进苹果品种快速鉴定、种质资源评价及选择利用，本研究对山东省 31 个栽培苹果开展了重测序及 SNP 位点挖掘研究。样品经 Hiseq 4000 平台建库测序，净数据量为 363 G。平均样品覆盖度达到 16.29×；充分满足重测序分析及 SNP 位点挖掘的需要。错配率比较试验发现随着错配率逐渐升高，比对率逐渐升高至饱和。其中总比对率、成对数据比对率及单端数据比对率与错配率呈现显著相关，均符合一元四阶方程(回归系数 R>0.99)。随着错配率提高，比对严谨度降低；基因组覆盖度逐渐升高，杂合位点准确度逐渐提高。采用两种算法所得到的位点，根据 ' 染色体+位点信息 ' 作为特征值取交集，得到高可靠的单碱基 SNP 位点数据集：共检测到 374 404 个变异，平均每隔 1 896 个位点能够检测到一个变异，桑格验证试验准确度高达 98.1%。SNP 的功能注释分析结果显示在全部 373 763 个位点中有 143 269 个(38.27%)位于基因间区，25 047 个(6.7%)位于基因编码区，179 426 个(47.92%)位于基因上游 - 下游的 2 kb 区域。在所有编码区 SNP 里面，有 13 422 个是非同义变异位点，11 625 个是同义变异位点。两种 SNP 比率为 1.15：1。进一步利用过滤的 4DTV 位点，采用邻接算法构建的聚类分析结果符合我省栽培苹果分类的趋势。

关键词 栽培苹果, 重测序, SNP 位点开发

SNP Mining by Genome Resequencing of 30 Apple Varieties in Shandong Province

Duan Naibin[1*] Ma Yumin[1] Wang Kun[2] Wang Xiaomu[1] Xie Kun[1] Bai Jing[1] Yang Yongyi[1] Pu Yanyan[1] Gong Yongchao[1]
1 Shandong Centre of Crop Germplasm Resources, Shandong Academy of Agricultural Sciences, Ji'nan, 250101; 2 Fruit Research Institute, Chinese Academy of Agricultural Sciences, Xingcheng, 125100
* Corresponding author, duannaibin@gmail.com
DOI: 10.5376/mpb.cn.2020.18.0053

Abstract In this article, we carried out genome resequencing and SNP mining for cultivated apples in Shandong Province, for the sake of the rapid identification of apple varieties, germplasm evaluation, and utilization. Genomic DNA was extracted immediately from leaves of each sample, and Paired-end Illumina genomic libraries were prepared and sequenced on an Illumina Hiseq 4000 platform following the manufacturer's instructions. Resequencing of the 31 apple genomes generated a total of 363 Gb high-quality cleaned sequences, with an average of 12.5 Gb per accession that represented approximately 15.9x coverage of the apple genome. The data volume fully meets the needs of downstream analysis and SNP mining. When we used the nucleotide mismatch

收稿日期：2020 年 11 月 19 日；接受日期：2020 年 11 月 20 日；发表日期：2020 年 11 月 27 日
引用格式: 段乃彬, 马玉敏, 王坤, 王效睦, 谢坤, 白静, 杨永义, 蒲艳艳, 宫永超, 2020, 山东主要栽培苹果基因组重测序及 SNP 芯片位点挖掘, 分子植物育种(网络版), 18(53): 1-11 (doi: 10.5376/mpb.cn.2020.18.0053) (Duan N.B., Ma Y.M., Wang K., Wang X.M., Xie K., Bai J., Yang Y.Y., Pu Y.Y., and Gong Y.C., 2020, SNP mining by genome resequencing of 30 apple varieties in Shandong Province, Fengzi Zhiwu Yuzhong (Molecular Plant Breeding (online)), 18(53): 1-11 (doi: 10.5376/mpb.cn.2020.18.0053))

parameter from 1-12, the mapping rate gradually increased to saturation. There was a highly significant correlation (P<0.0001) between the total mapping rate, mapping rate of pair-end data, and mismatch parameter. Univariate fourth-order equation (regression coefficient r > 0.99) were predicted. As the mismatch rate increases, the accuracy of mapping decreases; the genome coverage gradually increases, and the accuracy of heterozygous sites gradually increases. In this study, the SNP mining was obtained by the two algorithms, and the intersection was further taken based on the 'chromosome+site information' as the eigenvalues to obtain a highly reliable single nucleotide variant dataset. A total of 374 404 SNP locus were detected. On average, one mutation can be detected every 1 896 loci. The accuracy of the Sanger verification test is as high as 98.1%. Annotation analysis shows that among the 373 763 SNPs, 25 047 (6.7%) are located in the gene coding region, 143 269 (38.27%) are located in the intergenic region, and 179 426 (47.92%) are located in the 2 kb region upstream or downstream of the genes. Among the coding region SNPs, 13 422 are non-synonymous mutations, and 11 625 are synonymous mutations. The ratio of non-synonymous to synonymous SNP is 1.15: 1. Using filtered 4DTV sites, clustering analysis results constructed by neighbor-joining algorithms are in line with the trend of the classification of cultivated apples in Shandong province.

Keywords Cultivated apple, Genome resequencing, Development of SNP markers

苹果是产量位居前列的重要水果之一。2019 年全球苹果产量超过 8.314×10^{10} kg，其中中国产量 4.139×10^{10} kg (数据来自联合国粮农组织统计数据库, http://www.fao.org/faostat/zh/#search/apple)，占 50%以上。山东省苹果产量多年稳居中国前列 (占 25%以上)；同时又是苹果种质资源大省，在苹果种质资源搜集、创新及新品种选育方面全国前列。

基因组学是作物遗传育种研究的基础。苹果因其重要性，其基因组研究已经取得了长足发展。苹果基因组先后经历 4 次组装测序(Velasco et al., 2010; Li et al., 2016; Daccord et al., 2017; Zhang et al., 2019)，是全基因组组装进展最快的果树作物之一。近几年，利用重测序技术已对全球的苹果种质资源开展了群体基因组学及群体遗传学研究，由此阐明了苹果的驯化机理和进化机制；基因组学及生物信息学展现出在种质资源挖掘创新方面的强大潜力(Duan et al., 2017; 段乃彬, 2017; 贾东杰, 2018)。

从种质资源研究角度看，基于重测序的群体群体基因型鉴定为果树种质资源保护、鉴定评价及创新利用提供了新的研究思路。基因组学在高通量、大数据及全基因组关联分析方面有独特优势(陈学森等, 2015; 陈璇等, 2018)。为开展栽培苹果近缘种质资源的高通量基因型鉴定，应将基因组学及生物信息学结合，进而构建相应的 SNP 信息和相应注释信息的数据库，在苹果组学育种上面有很好的应用前景。而在后基因组时代，SNP 芯片因其独特优势，代表了低成本基因分型的方向。目前中国已有小麦、玉米、大豆、水稻和棉花等大田作物及部分十字花科蔬菜作物开展了 SNP 芯片应用或开发研究。尤其在小麦，贾继增等构建了基于 Affimatrix 平台的 660 k 高分辨率小麦 SNP 芯片，这些芯片在种质资源鉴定评价、群体基因分型、关联分析或功能基因定位及分子标记辅助育种方面均展现了不可低估的应用前景(Zhou et al., 2018)。与重测序技术相比，芯片技术分析流程简单，不需进行参考基因组比对实现高通量、检测准确性很高 99.9%以上；检测费用相对低廉：大约 100 万位点的芯片(每个样本的)检测费用在 1 000 元人民币左右。在苹果芯片研究上，有研究者先后创制了 8K、20K 及 480K 三种 SNP 芯片，并针对欧美主要的栽培苹果开展了基因分型和关联分析的应用(Chagné et al., 2012; Bianco et al., 2014; Bianco et al., 2016)。

中国已经开展 SNP 芯片位点开发的果树仅见于草莓、梨及桃；其芯片分辨率分别为 90K，200K 和 9K (Verde et al., 2012; Bassil et al., 2015; Li et al., 2019)。苹果作为重要果树，其育种急需针对性较强、低成本及高通量的基因分型手段。而针对中国特有苹果品种的 SNP 芯片研究尚未开展。本研究在已开展重测序的研究基础上，进行了 SNP 芯片位点挖掘研究。一方面可用于苹果品种快速鉴定、种质资源评价及选择利用；又可用于全基因组关联分析、功能基因定位及分子标记辅助育种。

1 结果与分析

1.1测序数据量

原始数据下机后，经过去除接头 Adapter 序列及 PCR 建库导致的重复读段，最后得到净数据量 363 G。

以苹果基因组 720 M 碱基对计算，得到基因组覆盖度最高 21.02×，基因组覆盖度最低 10.63×，平均样品覆盖度达到 16.29×；充分满足重测序分析及 SNP 位点挖掘的需要(表1)。

1.2 错配参数对数据比对率的影响

以 C18-06A 样品元帅(青岛一号)为例，针对 BWA 软件要求的错配率参数 mismatch：即数据读段与参考基因组的允许错误匹配碱基的数值，因为本研究测序读长为 150 bp，本研究该参数值从 1 (0.66%)增加到 12 (8.00%)，分别得到一系列比对文件。再利用 SAMtools 的 flagstat 功能来统计全部读段数据、成对数据及单端数据比对率的具体情况(表 2; 图 1)；首先随着错配率逐渐升高，比对率也逐渐升高，但是升高的趋势逐渐减低，直至接近饱和。

全部读段数据和 pair-end 数据对比对率呈现逼近饱和的趋势(图 1)，而单端数据比对率逐渐降低至一个最低值。其中总比对率 Total Mapping Rate 与错配率的变化趋势呈现正相关，符合四阶方程：$y=-3\times10^{-5}x^4+0.001\ 1x^3-0.014\ 5x^2+0.086\ 4x+0.741\ 8$，回归系数 $R^2=0.999\ 5$；成对读段的比对率 Paired Mapping Rate 与错配率的变化趋势呈现正相关，符合四阶方程：$y=-3\times10^{-5}x^4+0.001\ 2x^3-0.014\ 9x^2+0.086\ 3x+0.712\ 6$，

表 1 测序数据量

Table 1 Statistics of apple genome resequencing for each accession

样品名称 Sample name	有效读段数 Clean reads	有效碱基数 Clean bases	Q20 比例(%) Q20 ratio (%)	GC 比例(%) GC ratio (%)	覆盖度 Coverage
C18-01B	82 285 164	12 275 896 018	96.81	38.40	17.05
C18-02A	89 077 308	13 313 277 436	97.01	38.59	18.49
C18-03A	83 072 050	12 394 531 092	97.09	39.12	17.21
C18-04A	97 305 218	14 527 586 044	96.95	38.58	20.18
C18-05A	86 562 736	12 940 574 192	97.20	38.27	17.97
C18-06A	80 125 972	11 965 621 824	96.95	38.08	16.62
C18-07A	74 530 532	11 122 786 992	97.06	38.96	15.45
C18-08A	75 995 538	11 356 852 292	97.21	38.62	15.77
C18-09A	90 068 930	13 407 382 642	96.73	39.20	18.62
C18-10A	101 224 890	15 134 111 712	97.17	38.73	21.02
C18-11B	69 321 342	10 326 757 324	96.82	38.29	14.34
C18-12A	100 502 502	15 007 629 504	97.05	38.33	20.84
C18-13-1A	71 054 910	10 597 863 906	96.82	38.84	14.72
C18-13-2B	81 541 908	12 157 333 396	96.49	38.36	16.89
C18-14B	68 947 114	10 293 832 588	96.67	38.84	14.30
C18-15A	76 776 466	11 445 569 550	97.06	38.26	15.90
C18-16A	78 836 884	11 769 458 430	96.76	38.36	15.95
C18-17A	63 602 864	9 494 133 904	97.12	38.35	13.19
C18-18A	77 680 304	11 605 708 802	96.72	38.31	16.12
C18-19B	54 867 846	8 184 033 800	97.30	38.48	11.37
C18-20A	85 924 614	12 849 242 938	97.07	38.37	17.85
C18-21A	81 690 936	12 211 575 458	97.07	39.10	16.96
C18-22A	92 507 878	13 784 235 218	96.65	38.42	19.14
C18-23B	76 286 322	11 396 088 188	97.32	38.46	15.83
C18-24B	82 677 084	12 345 285 852	96.41	38.34	17.15
C18-25B	85 568 414	12 765 677 658	97.03	38.26	17.73
C18-26B	74 209 450	11 075 213 404	96.99	38.54	15.38
C18-27B	78 213 462	11 677 414 082	96.92	38.56	16.22
C18-28B	62 762 662	9 363 987 448	96.90	38.49	13.01
C18-29B	61 273 994	9 125 369 346	96.72	38.49	12.67
C18-30B	51 288 314	7 650 121 178	96.49	39.08	10.63

表 2 错配参数对比对率的影响

Table 2 The effect of mismatch parameters on the mapping rate

错配参数 Mismatch parameter	总比对读段数 Total mapped reads	总比对比例(%) Total mapped ratio (%)	双端读段比对数 Paired mapped reads	双端读段比对率(%) Paired mapped ratio (%)	单端读段比对数 Single mapped reads	单端读段比对率(%) Single mapped ratio (%)
1	60 771 756	0.813 9	56 597 900	0.784 1	1 207 208	0.016 7
2	64 742 356	0.867 1	60 363 941	0.836 3	968 921	0.013 4
3	67 102 418	0.898 7	62 547 086	0.866 6	824 506	0.011 4
4	68 579 887	0.918 5	63 851 818	0.884 6	737 760	0.010 2
5	69 526 126	0.931 2	64 652 170	0.895 7	686 368	0.009 5
6	70 157 055	0.939 6	65 158 530	0.902 7	651 959	0.009
7	70 594 321	0.945 5	65 487 645	0.907 3	629 492	0.008 7
8	70 909 166	0.949 7	65 708 543	0.910 4	612 792	0.008 5
9	71 144 151	0.952 8	65 862 926	0.912 5	602 451	0.008 3
10	71 325 235	0.955 2	65 976 100	0.914 1	595 493	0.008 3
11	71 468 805	0.957 2	66 060 122	0.915 2	590 686	0.008 2
12	71 584 809	0.958 7	66 124 046	0.916 1	587 561	0.008 1

回归系数 R^2=0.999 4；最后，单端比对率 Single-end Mapping Rate 与错配率的变化趋势呈现负相关，符合四阶方程：$y=2\times10^{-6}x^4-7\times10^{-5}x^3+0.000\ 9x^2-0.005\ 4x+0.021\ 2$，回归系数为 R^2=0.999 3。

1.3 错配参数对 SNP 位点真实性的影响

接下来又比较了不同错配参数对位点检测准确率的影响，以 11 号染色体 Chr11 为例(图 2)：随着允许错配率的增加，在如图的 Chr11 区域，可比对的数据逐渐增多，测序覆盖度从 11 逐渐增高至 19，在低覆盖率下呈现纯合的位点，在高覆盖度下都被检测为杂合位点，这说明错配率的增加有利于杂合位点的检测。可见随着错配率提高比对严谨度降低；基因组覆盖度逐渐升高，更有利于杂合位点的检测。很多植物具有远缘杂交、自交不亲合、较高的基因组杂合度及广泛的遗传漂变等特点；如苹果属，芸薹属，玉米等作物。对于此类作物的 SNP 位点挖掘，一方面需要提高测序在全基因组有数据覆盖，另一方面是选

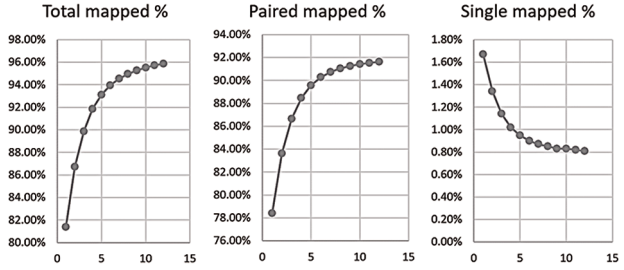

图 1 错配参数对总数据, 成对数据及单端数据比对率的影响

Figure 1 The effect of mismatch parameters on the mapping of total data, paired data and single-ended data

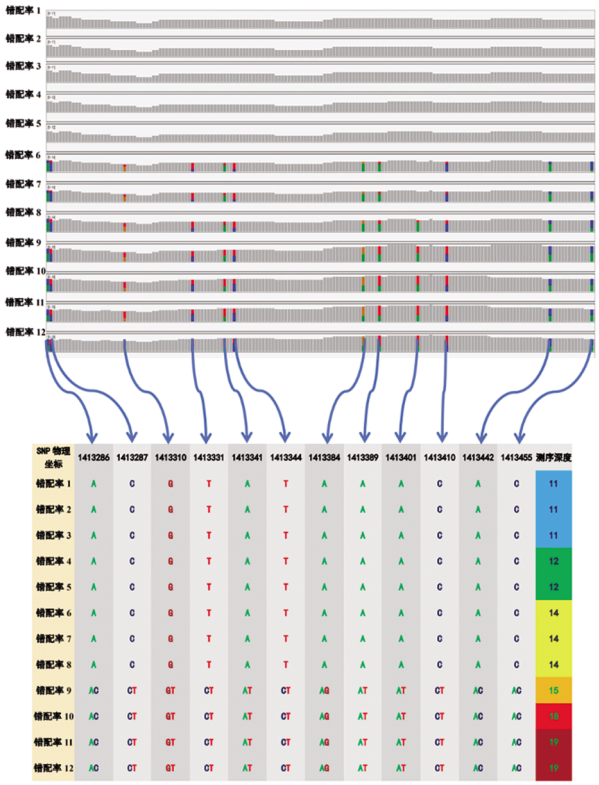

图 2 错配参数对杂合 SNP 位点检测的影响

Figure 2 The effect of mismatch parameters on accurate of heterozygous SNP genotyping

择一个最合适的错配参数；对于杂合度较高的作物的 SNP 位点挖掘有借鉴意义。

1.4 两个分析流程下的位点比较及整合

按照二代测序标准流程结合 BCFtools 工具

即：bwa-sam-bam-pileup-bcfools 算法，总共检测到 28 997 212 个变异，包括单碱基 SNP 26 758 563 个，短插入 short insert 1 060 691 个，短缺失 short deletion 1 177 95 个。该算法可检测各种类型变异，故而变异检测灵敏度较高，平均每隔 27 个位点能够检测到一个变异。

按照二代测序标准流程结合自主开发的 in-house 算法即：bwa-sam-bam-pileup-column 算法共检测到 1 147 801 个变异，该流程算法是针对单碱基 SNP 的检测，这些变异均为单碱基 SNP，故而变异检测灵敏度较低。平均每隔 618 个位点能够检测到一个变异。

结合两种算法所得到的位点，根据‘染色体 + 位点信息’作为特征值进一步取交集，则得到高可靠的单碱基 SNP 位点数据集，此位点数据集。共检测到 374 404 个变异，由于是取交集，同样这些变异均为单碱基 SNP。平均每隔 1 896 个位点能够检测到一个变异。对 1 000 个随机选择的同源 SNP 设计引物并进行 PCR 扩增，再对扩增产物 Sanger 测序，结果表明所选择的 SNP 位点在两中测序平台的符合度为 98.1%。

1.5 SNP 位点在基因组的分布

SNP 的功能注释分析表明结果显示，在全部 373 763 个 SNP 位点中有 143 269 个(38.27%)位于基因间区，25 047 个(6.7%)位于基因编码区，143 269个(38.27%)位于基因间区，179 426 个(47.92%)位于基因上游或－下游的 2 kb 区域。在所有编码区 SNP 中里面，有 13 422 个是非同义突变变异位点，11 625 个是同义变异位点突变(表 3; 图 3)。非同义与同义 SNP 的比两种 SNP 比率为 1.15∶1。非同义 SNP，又称错义 SNP，从编码一种氨基酸变为另一种氨基酸而形成表型修饰；同义 SNP 又称沉默突变，虽有碱基突变，仍编码同一种氨基酸而不能形成表型修饰。苹果与其他栽培大田作物和果树作物相比较，其基因组上可形成对应表型修饰的变异比例较低(Duan et al., 2017)。

1.6 群体聚类进化树的构建

该进化树是用最小进化法推导得到(图 4)。图上显示的是分支长度总和为 0.7665 的最佳进化树，该树是参照进化距离的比例而绘制。从整体上看，该系统进化树显示、本研究采集的山东省主要栽培苹果主要分为富士、元帅、金冠(金帅)、嘎啦四大类型及其他杂交组合。值得一提的是：(1) C18-23 样品是在新

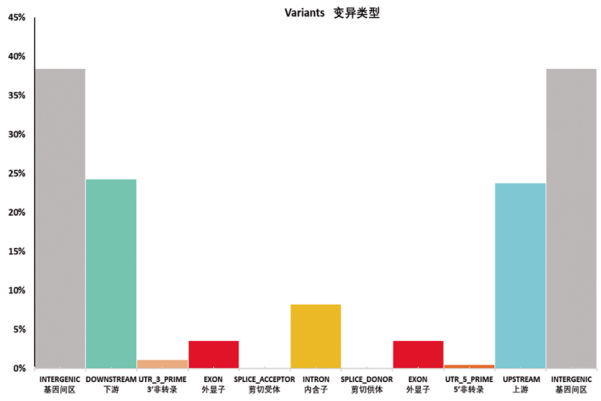

图 3 在不同基因区域 SNP 分布情况

Figure 3 Number of effects by region

疆野苹果实生选育的野生资源，在进化历史上最早发生分歧；(2) C18-2、C18-3、C18-4、C18-5 及 C18-6，依据资源圃提供的信息这些样品均为元帅系，在本试验中也成功的聚在一起。类似的，C18-8、C18-9、C18-10、C18-11、C18-12、C18-13-1 及 C18-13-2，依据资源圃提供的信息这些样品均为富士系，在本试验中也成功的聚在一起。以上样品系谱及原产地信息均来自自中国农业科学院果树研究所国家苹果资源圃(兴城)，该资源圃对这些种质资源有着多年准确的系谱资料登记在册。这个结果间接证明了本试验 SNP 数据的可靠性；(3) 其余样品的聚类结果也均符合预期。

2 讨论

2.1错配率的评估

本研究通过对一系列错配参数的比较分析，发现伴随错配率的上升，可比对到基因组的读段 Reads 数目逐渐增多，呈现逐渐饱和的趋势。错配率增至一

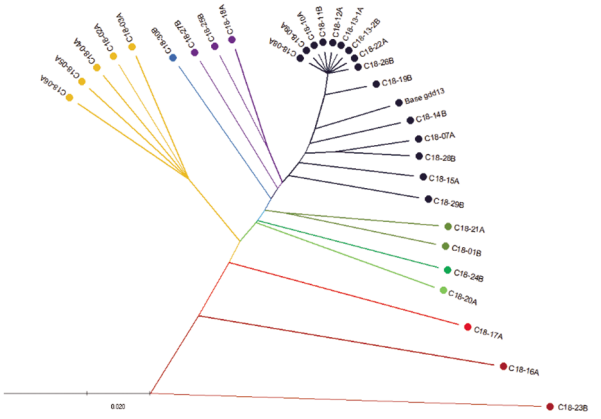

图 4 利用 4DTV 位点构建的山东栽培苹果聚类图

Figure 4 Evolutionary relationships of taxa

表 3 不同基因类型及区域的 SNP 数

Table 3 Number of effects by type and region

突变类型 Mutation type	数目 Count	比例 Percent	突变区域 Mutation region	数目 Count	比例 Percent
3' 上游非转录 3'_UTR_variant	7 906	1.09%	终止密码子缺失 Stop_lost	20	0.00%
5' 上游非转录起始密码获得 5'_UTR_premature_start_codon_gain	485	0.07%	终止密码子保留 Stop_retained_variant	19	0.00%
5' 上游非转录 5'_UTR_variant	2 773	0.38%	无义突变 Synonymous_variant	11 625	1.61%
基因下游突变 Downstream_gene_variant	174 403	24.13%	基因上游突变 Upstream_gene_variant	170 706	23.62%
起始密码子突变 Initiator_codon_variant	2	0%	下游区 Downstream	174 403	24.22%
基因间 Intergenic_region	276 549	38.27%	外显子 Exon	25 443	3.53%
内含子突变 Intron_variant	61 125	8.46%	基因间 Intergenic	276 549	38.40%
错义突变 Missense_variant	13 422	1.86%	内含子 Intron	59 074	8.20%
无编码转录本外显子突变 Non_coding_transcript_exon_variant	289	0.04%	剪接受体 Splice_site_acceptor	97	0.01%
无编码转录本突变 Non_coding_transcript_variant	356	0.05%	剪接供体 Splice_site_donor	60	0.01%
剪切受体变异 Splice_acceptor_variant	97	0.01%	剪接体区间 Splice_site_region	2 273	0.32%
剪切供体变异 Splice_donor_variant	60	0.01%	转录本 Transcript	356	0.05%
剪切体变异 Splice_region_variant	2 541	0.35%	基因上游区(5k 以内) Upstream region of gene (within 5k)	170 706	23.71%
起始密码子缺失 Start_lost	24	0.00%	3' 上游非转录区 Utr_3_prime	7 906	1.10%
终止密码子获得 Stop_gained	292	0.04%	5' 上游非转录区 Utr_5_prime	3 258	0.45%

定程度下有利于提高测序覆盖度，有利于位点挖掘。但是在接近饱和情况下无限度增高错配率，是毫无意义的。

如是，本研究为进一步确定适应于不同样本的最佳 Bwa 比对 mismatch 参数(Li and Durbin, 2009)，项目组成员自主研发了一种择定最佳比对错配率参数的方法：首先从 NCBI (Pruitt et al., 2005)下载栽培苹果参考基因组序列，建立该苹果的本地 Blast 数据库。然后在测序数据中随机抽取 1 000 条读段进行本地 Blast，在 Blast 结果中对 mapping ratio 排序，后统计第 550 条 read 的 Identity 相似度，由此确定 BWA 的 mismatch 参数。

最佳 mismach 参数的大小一定程度衡量了参试样品相对于参考基因组的亲缘关系的远近。如参试样品 23 号是由新疆野苹果(*M. sieversii* in Xinjiang)与红肉苹果(Malus domestica 'Redlove Era')杂交的新品种，遗传关系上距离参考基因组苹果 Golden Delicious (*M. domestica*)最远，相应的我们运算得到的其 mismach 参数最大，为 7；而参试样品 1 号与参考基因组同属金帅系，亲缘关系最近，相应的我们运算得到的其 mismach 参数最小，为 4。

这个最佳比对错配率择定的方法在前文中已被采用(Duan et al., 2017; 段乃彬, 2017)。择定准确的 mismatch 参数一方面能获得足够高的测序覆盖度，

保证位点准确性；另一方面能够在最小运算量下使尽量多的读段得到比对，避免了过度运算，提高分析效率。

2.2 数据整合的必要性

为增强 SNP 位点的可靠性，本研究将新取材的 31 个样品的测序数据与前文已经测序的 23 个栽培苹果测序数据进行了数据整合(Duan et al., 2017; 段乃彬, 2017)，数据整合的目的：一是增强位点的可靠性，二是可以比较省内栽培苹果群体与国外栽培苹果群体之间的多态性进行比较(另撰文发表)。

增大参试样本数量，进行数据整合具有如下优点：提高了样品的多样性；由于前文重测序所涉及全球范围内的 23 个主要栽培苹果类型；另一方面提高了挖掘位点的可靠性；通过整合实际上是用大数据集的多样性来考量子集的多样性；增强了所选位点的未来适用范围。

2.3 高杂合物种的 SNP 检测策略

首先，对于杂合比较高的物种，其基因组的组装存在一定难度，相应的组装质量普遍不高，如已发表的多个果树基因组杂合度较高。目前二代测序广泛使用，当利用读长 100~150 bp 的 reads 组装到 Contig 时，高杂合度的基因组 contig 之间的 overlap 关系不容易明确，从而导致 N50 偏低，基因组上会产生大量的不能叠连的区域(gap) (Pryszcz and Gabaldón, 2016)。相应的，当使用二代测序进行 SNP 检测的时候，则需要尽量高的测序深度、尽量长的读段 reads，如目前广泛采用的 hiseq4000 平台。本研究的平均测序深度就达到了 16 X，读长均为 150 bp。基于以上策略，本研究 SNP 位点经过一代 Sanger 测序，符合度高达 98.1%。这高于玉米、棉花群体重测序 SNP 位点的准确率，而这两种作物具有良好的遗传学研究基础，其基因组组装质量优于苹果。

再者，增加错配率下进行比对，会有更多位点被检测为杂合位点，这说明在高覆盖度有利于杂合位点的检测。很多植物因远源杂交、自交不亲和等因素，具有较高的基因组杂合度，存在明显而广泛的遗传漂变；如苹果属(Malus Mill.)，芸薹属(Brassica)，玉米等作物。对于此类作物的 SNP 位点挖掘，一方面需要提高测序在全基因组有数据覆盖，另一方面是选择一个最合适的错配参数。这对于杂合度较高的作物有借鉴意义。

最后，在位点择定方面，应采用改进的算法。尽量避免只使用一种检测流程，目前在既有的 SNP 检测流程中，其上游均采用 BWA 结合 SAMtools 的分析流程，即 BWA-Sam-bam-pileup。只是在生成的 pileup 文件之后采用不同的算法来择定 SNP，其文件格式以 VCF、hapmap 或者列表格式为主。此时采用两个或者两个以上的分析流程，将生成的数据归一化为 VCF 文件，在利用染色体的坐标信息取交集则可获得可信度较高的位点。

3 材料与方法

3.1品种搜集

本研究取材 31 个栽培苹果品种，类型广泛，涵盖了四大苹果品系富士系、元帅系、金冠系、嘎啦系及一些新的杂交品系，样品系谱及原产地信息来自中国农业科学院果树研究所国家苹果资源圃(兴城)；囊括了山东省主要栽培苹果的接穗类型。从取材地域看，取材地域范围遍及山东全省主要苹果栽培种植地区。从系谱信息看，实验取材在多样性方面具有足够的代表性(表 4)。

在 2018 年 6 月 15 日 ~20 日，其中大多数采集当年生顶梢叶片样品取回后立即液氮处理，唯有 23、25、26 三个样品的叶片为硅胶干燥。所有叶片样品按照标准 DNB 提取方法，所提取的 DNA 样品经过琼脂糖凝胶检测质量，符合测序要求后再经双末端PE150策略建库，并交付华大科技(BGI)在 Hiseq-4000 平台完成测序。

3.2 测序数据的预处理及统计分析

原始数据需先经过一个 Perl 测序脚本(由本课题组研究团队编写)去除 PCR 导致的测序重复。具体的讲，对于具有不同测序位置信息 ID 的成对 Reads，凡是 Pair1 或者 Pair2 在 15~135 bp 的区间同时出现完全一致的碱基数据即界定为 PCR 导致的测序重复，这样数据被过滤去除。命令行是："drop_dup_both_end.pl raw_fq1 raw_fq2"。

已经去除 PCR 测序重复的数据再经 Trimmomatic3.0 软件过滤去除 1、测序接头，2、低质量的读段。这样最后得到的是净数据。命令行是 "trimmomatic PE -thReads 75 fq.1 fq.2"。

包括测序总数据量统计，测序深度统计，读段比对率统计及比对 mismatch 参数的确定。命令行是 "fastqc -q trimmed_fq1 trimmed_fq2"。

3.3 错配率的确定

以 C18-06A 样品元帅(青岛一号)为例，针对 BWA

表 4 样品的取材地及品系信息

Table 4 List of variaties in this study with habitat and pedigree information

样品编号 Accession number	品种名 Variety name	品系类型 Type of breed	采样地点 Location
C18-1	金帅	金帅系	烟台牟平
	Golden Delicious	Golden Delicious	Muping, Yantai
C18-2	新红星(荷兰)	元帅系短枝	青岛胶南
	Starkrimson (Netherlands)	Delicious short spur	Jiaonan, Qingdao
C18-3	平阴短枝	元帅系	济南平阴
	Pingyin Spur	Delicious	Pingyin, Jinan
C18-4	康屯短枝	元帅系	烟台牟平
	Kangtun Spu	Delicious	Muping, Yantai
C18-5	夕阳红	元帅系	烟台农科所
	Xiyanghong	Delicious	Yantai Institute of Agricultural Sciences
C18-6	青岛一号	元帅系	烟台大泽山
	Qingdao1	Delicious	Daze Mountain, Yantai
C18-7	秋富 5 号	富士系(元帅×国光)	烟台牟平
	Akifu 5	Fuji (Yuanshuai×Guoguang)	Muping, Yantai
C18-8	秋富 1 号	富士系	青岛莱西
	Akifu 1	Fuji	Laixi, Qingdao
C18-9	长富 7 号	富士系	威海荣成
	Nagafu 7	Fuji	Rongcheng, Weihai
C18-10	秋富 2 号	富士系	威海荣成
	Akifu 2	Fuji	Rongcheng, Weihai
C18-11	长富 2 号	富士系	威海荣成
	Nagafu 2	Fuji	Rongcheng, Weihai
C18-12	烟富 5 号	富士系	烟台牟平
	Yanfu 5	Fuji	Muping, Yantai
C18-13-1	昌红 -1	富士系	德州平原
	Changhong-1	Fuji	Pingyuan, Dezhou
C18-13-2	昌红 -2	富士系	青岛即墨
	Changhong-2	Fuji	Jimo, Qingdao
C18-14	王林	金冠×印度	青岛即墨
	Orin	Golden Delicious×Indo	Jimo, Qingdao
C18-15	印度	青香蕉×?	青岛即墨
	Indo	White Winter Pearmain × ?	Jimo, Qingdao
C18-16	早捷	昆特×七月红	烟台栖霞
	Geneva Early	Quinte×Julyred	Qixia, Yantai
C18-17	藤木一号(南部魁)	待确认	山东临沂
	Fujiki 1 (Nanbusakigake)	Uknown	Linyi, Shandong
C18-18	国光	待确认	山东临沂
	RallsJanet	Uknown	Linyi, Shandong
C18-19	新世界	富士×あかぎ	山东临沂
	Shinsekai	Fuji×akagi	Linyi, Shandong
C18-20	珊夏	嘎拉×茜	山东泰安
	Sansa	Gala×Akane	Tai'an, Shandong
C18-21	乔纳金	金冠×红玉	山东泰安
	Jonagold	Golden Delicious×Jonathan	Tai'an, Shandong

续表 4
Continuing table 4

样品编号 Accession number	品种名 Variety name	品系类型 Type of breed	采样地点 Location
C18-22	王实	富士系	山东泰安
	Wangshi	Fuji	Tai'an, Shandong
C18-23	紫红一号	待确认	山东泰安
	Violetred No.1	Uknown	Tai'an, Shandong
C18-24	泰山早霞	待确认	山东泰安
	Taishan early	Uknown	Tai'an, Shandong
C18-25	国光	待确认	山东泰安
	RallsJanet	Uknown	Tai'an, Shandong
C18-26	红将军	富士系	山东聊城
	Red General Fuji	Fuji	Liaocheng, Shandong
C18-27	红玉	可口香实生	山东临沂
	Jonathan	Esopus Spitzenburg	Linyi, Shandong
C18-28	皇家嘎啦	嘎啦	山东临沂
	Royal Gala	Kidd's Orange Red×Golden Delicious	Linyi, Shandong
C18-29	美国八号	NY543	山东临沂
	Meiguo 8		Linyi, Shandong
C18-30	粉红佳人	金帅系	山东泰安
	Pink Lady	Golden Delicious	Tai'an, Shandong

软件要求的错配率参数 mismatch：即数据读段与参考基因组的允许错误匹配碱基的数值，因为苹果存在远缘杂交，杂合度较高。因此本研究将该参数值从 0.66%增加到 8.00%，对应于 150 bp 读长则为 1~12。分别得到一系列比对文件，用以比较比对率对覆盖度及 SNP 检测的影响。

为确定合适的 BWA (Li and Durbin, 2009)比对 mismatch 参数，首先从 NCBI (www.ncbi.genome.com)下载栽培苹果参考基因组序列，建立该物种的本地 Blast 数据库。在测序数据中随机抽取 1 000 条读段进行本地 Blast，对 Mapping ratio 排序后统计第 550 条的相似度，由此确定 BWA 的 mismatch 值。

3.4 测序数据比对及 SNP 位点挖掘

本研究以 2017 年发表的栽培苹果 '金帅' 的基因组(Daccord et al., 2017)作为参考序列，用本试验采集的所有 31 个及前文 (Duan et al., 2017) 采用的 23 个栽培苹果，合计 54 个栽培苹果的重测序数据与参考基因组进行 BWA (Li and Durbin, 2009)比对(mismatch 为 4~7 不等)。经由 SAMtools (Li et al., 2009)转换得到 pileup 文档。接下来采用两种不同流程检测 SNP 位点信息：(1) BWA-sam-bam-pileup-bcfools 算法，利用 SAMtools 结合 BCFtools 转换 Pileup 文件得到各个样品 VCF 文件格式的 SNP 数据集。(2)按照二代测序标准流程结合自主开发的 In-House 算法即：bwa-sam-bam-pileup-column 算法，得到类似 hapmap 的 SNP 数据集。(3)采用改进的交集算法：将以上两种方法得到的 SNP 位点信息基于染色体坐标取交集的方法，进而到更高高质量 SNP 位点。

SNP 验证试验方法：本试验选取了 6 个参试样品，在 11 号染色体随机截取 1 000 个 homogenous SNP 位点(即非杂合)，以此为中心设计两侧 50 bp 序列。进而构建引物，进行 PCR 扩增实验。再将扩增产物经 3730 毛细管电泳进行一代测序验证。

3.5 SNP 位点注释

SNPEff 是一款强大的 SNP 注释软件。与其他注释软件相比较，其不仅能得到该突变位点所在的基因区域，还能得到突变所在基因区段的类型信息，这有利于后续的功能基因挖掘和定位。由于使用 java 平台，有较强的易用性，其手册 http://SNPeff.sourceforge.net/SNPEff_manual.html 对注释方法进行了非常详细的描述；本研究注释的命令行如下：

修改 SNPEff 软件设置："vim userpath/SNPEff/SNPeff-4.3.1t-1/SNPEff.config"；添加基因组信息："# apple genome version GDDH13 GDDH13.genome :

Apple"；建立本地库："SNPEff build -gff3 -v GDDH13"；SNP 注释："SNPEff -v -stats prefix.html GDDH13 prefix.vcf >prefix.ann"；运行输出的 html 文件是以网页形式呈现的位点注释结果的图表解释，而输出的 ann 文件是则是以文本列表方式列出了每一个 SNP 注释的详细结果。

3.6 4DTV 位点的筛选及聚类分析

在基因的蛋白编码区上，有部分氨基酸所对应的第三位密码子可使用任意 4 种碱基，都不会形成氨基酸的改变，这样的位点被称作四重兼并位点(4DTV)。这种无意突变几乎没有选择压力，其突变率可以用作“时钟”来估计进化，特别适合构建进化树及群体遗传结构分析(Fazio et al., 2014)。本研究利用团队自己编写的 Perl 脚本，对整套 SNP 数据按如下规则在 CDS 区域进行位点筛选：最小等位基因频率(MAF)≥5%，且每个位点对应的数据缺失率≤10%，共筛选得到四重简并位点(4DTV) 24 326 个。最后位点输入到 Mega X 软件，在第一搜索级别上使用了接近邻居交换(close-neighbor-interchange, CNI)算法(Kumar et al., 2018)。由此构建群体的系统发育进化树。

作者贡献

段乃彬、马玉敏是本研究的实验设计和实验研究的执行人；谢坤、白静、杨永义、蒲艳艳及宫永超完成数据分析，论文初稿的写作；马玉敏、王效睦及王坤参与实验设计，试验结果分析；段乃彬是项目的构思者及负责人，指导实验设计，数据分析，论文写作与修改。全体作者都阅读并同意最终的文本。

致谢

本研究由山东省科技厅省重点研发项目(项目编号 2018GNC110031) 和山东省农业良种工程－农作物种质资源收集保护与精准鉴定(项目编号 2019LZGC017)共同资助。

参考文献

Bassil N.V., Davis T.M., Zhang H., Ficklin S., Mittmann M., Webster T., Mahoney L., Wood D., Alperin E.S., Rosyara U.R., Putten H.K.V., Monfort A., Sargent D.J., Amaya I., Denoyes B., Bianco L., van Dijk T., Pirani A., Iezzoni A., Main D., Peace C., Yang Y.L., Whitaker V., Verma S., Bellon L., Brew F., Herrera R., and van de Weg E., 2015, Development and preliminary evaluation of a 90K Axiom® SNP array for the allo-octoploid cultivated strawberry Fragaria× ananassa, BMC Genomics, 16(1): 155.

Bianco L., Cestaro A., Linsmith G., Muranty H., Denance C., Theron A., Poncet C., Micheletti D., Kerschbamer E., Di Pierro E.A., Larger S., Pindo M., van de Weg E., Davassi A., Laurens A., Velasco R., Durel C.E., and Troggio M., 2016, Development and validation of the Axiom® Apple480K SNP genotyping array, The Plant Journal, 86(1): 62-74.

Bianco L., Cestaro A., Sargent D.J., Banchi E., Derdak S., Di Guardo M., Salvi S., Jansen J., Viola R., Gut I., Laurens F., Chagné D., Velasco R., van de Weg E., and Troggio M., 2014, Development and validation of a 20K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (Malus× domestica Borkh), PLoS One, 9(10): e110377.

Chagné D., Crowhurst R.N., Troggio M., Davey M.W., Gilmore B., Lawley C., Vanderzande S., Hellens R.P., Kumar S., Cestaro A, Velasco R., Main D., Rees J.D., Iezzoni A., Mockler T., Wilhelm L., Van de Weg E., Gardiner S.E., Bassil N., and Peace C., 2012, Genome-wide SNP detection, validation, and development of an 8K SNP array for apple, PLoS One, 7(2): e31745.

Chen X., Guo R., Wang L., Liu Y H., Guo M.B., Xu Y.P., Guo H.Y., Yang M., and Zhang Q.Y., 2018, SNP analysis of wild and cultivated cannabis based on whole genome re-sequencing, Fenzi Zhiwu Yuzhong (Molecular Plant Breeding), 16 (3): 893-897. (陈璇, 郭蓉, 王璐, 柳延虎, 郭孟璧, 许艳萍, 郭鸿彦, 杨明, 张庆滢, 2018, 基于全基因组重测序的野生型大麻和栽培型大麻的多态性 SNP 分析, 分子植物育种, 16(3): 893-897.)

Chen X.S., Guo W.W., Xu J., Cong P.H., Wang L.R., Liu C.H., and Chen X.L., 2015, Genetic improvement and promotion of fruit quality of main fruit trees, Zhongguo Nongye Kexue (Scientia Agricultura Sinica), 48(17): 3524-3540. (陈学森, 郭文武, 徐娟, 丛佩华, 王力荣, 刘崇怀, 陈晓流, 2015, 主要果树果实品质遗传改良与提升实践, 中国农业科学, 48 (17): 3524-3540.)

Daccord N., Celton J.M., Linsmith G., Becker C., Choisne N., Schijlen E., Van de Geest H., Bianco L., Micheletti D., Velasco R., Di Pierro E.A., Gouzy J., Rees D.J.G., Guérif P., Muranty H., Durel C.E., Laurens F., Lespinasse Y., Gaillard S., Aubourg S., Quesneville H., Weigel D., van de Weg E., Troggio M., and Bucher E., 2017, High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development, Nat. Genet., 49(7): 1099-1106.

Duan N.B., 2017, Genomic analyses provide new insights into apple evolution domestication and genetic diversity, Dissertation for Ph.D., College of Horticulture Science and Engineering Shandong Agricultural University, Supervisor: Chen

X.S., pp.37-72. (段乃彬, 2017, 栽培苹果起源、演化及驯化机理的基因组学研究, 博士学位论文, 山东农业大学园艺科学与工程学院, 导师: 陈学森, pp.37-72.)

Duan N.B., Bai Y., Sun H.H., Wang N., Ma Y.M., Li M.J., Wang X., Jiao C., Legall N., Mao L.Y., Wan S.B., Wang K., He T. M., Feng S.Q., Zhang Z.Y., Mao Z.Q., Shen X., Chen X.L., Jiang Y.M., Wu S.J., Yin C.M., Ge S.F., Yang L., Jiang S.H., Xu H.F., Liu J.X., Wang D.Y., Qu C.Z., Wang Y.C., Zuo W. F., Xiang L., Liu C., Zhang D.Y., Gao Y., Xu Y.M., Xu K. N., Chao T., Fazio G., Shu H.R., Zhong G.Y., Cheng L.L., Fei Z.J., and Chen X.S., 2017, Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement, Nat. Commun., 8: 249.

Fazio G., Wan Y., Kviklys D., Romero L., Adams R., Strickland D., and Robinson T., 2014, Dw2, a new dwarfing locus in apple rootstocks and its relationship to induction of early bearing in apple scions, Journal of the American Society for Horticultural Science, 139(2): 87-98.

Jia D.J., 2018, Identification and validation of genes controlling apple fruit acidity and establishment of the genomic selection model, Dissertation for Ph.D., College of Horticulture China Agricultural University, Supervisor: Xu X.F., Han Z. H., and Zhang X.Z., pp.44-87. (贾东杰, 2018, 苹果果实酸度基因挖掘验证及基因组选择模型的建立, 博士学位论文, 中国农业大学, 导师: 许雪峰, 韩振海, 张新忠, pp. 44-87.)

Kumar S., Stecher G., Li M., Knyaz C., and Tamura K., 2018, MEGA X: molecular evolutionary genetics analysis across computing platforms, Mol. Biol. Evol., 35(6): 1547-1549.

Li H., and Durbin, R., 2009, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics, 25 (14): 1754-1760.

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 (16): 2078-2079.

Li X.L., Singh J., Qin M.F., Li S.W., Zhang X., Zhang M.Y., Khan A., Zhang S.L., and Wu J., 2019, Development of an integrated 200K SNP genotyping array and application for genetic mapping, genome assembly improvement and genome wide association studies in pear (Pyrus), Plant Biotechnology Journal, 17(8): 1582-1594.

Li X.W., Kui L., Zhang J., Xie Y.P., Wang L.P., Yan Y., Wang N., Xu J.D., Li C.Y., Wang W., van Nocker S., Dong Y., Ma F.W., and Guan Q.M., 2016, Improved hybrid de novo genome assembly of domesticated apple (Malus x domestica), Gigascience, 5: 35.

Pruitt K.D., Tatusova T., and Maglott D.R., 2005, NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids. Res., 33: D501-D504.

Pryszcz L.P., and Gabaldón T., 2016, Redundans: an assembly pipeline for highly heterozygous genomes, Nucleic Acids Research, 44(12): e113-e113.

Velasco R., Zharkikh A., Affourtit J., Dhingra A., Cestaro A., Kalyanaraman A., Fontana P., Bhatnagar S.K., Troggio M., Pruss D., Salvi S., Pindo M., Baldi P., Castelletti S., Cavaiuolo M., Coppola G., Costa F., Cova V., Dal Ri A., Goremykin V., Komjanc M., Longhi S., Magnago P., Malacarne G., Malnoy M., Micheletti D., Moretto M., Perazzolli M., Si-Ammour A., Vezzulli S., Zini E., Eldredge G., Fitzgerald L.M., Gutin N., Lanchbury J., Macalma T., Mitchell J. T., Reid J., Wardell B., Kodira C., Chen Z., Desany B., Niazi F., Palmer M., Koepke T., Jiwan D., Schaeffer S., Krishnan V., Wu C., Chu V.T., King S.T., Vick J., Tao Q., Mraz A., Stormo A., Stormo K., Bogden R., Ederle D., Stella A., Vecchietti A., Kater M.M., Masiero S., Lasserre P., Lespinasse Y., Allan A.C., Bus V., Chagne D., Crowhurst R.N., Gleave A.P., Lavezzo E., Fawcett J.A., Proost S., Rouze P., Sterck L., Toppo S., Lazzari B., Hellens R.P., Durel C.E., Gutin A., Bumgarner R.E., Gardiner S.E., Skolnick M., Egholm M., Van de Peer Y., Salamini F., and Viola R., 2010, The genome of the domesticated apple (Malus×domestica Borkh.), Nature Genetics, 42(10): 833-839.

Verde I., Bassil N., Scalabrin S., Gilmore B., Lawley C.T., Gasic K., Micheletti D., Rosyara U.R., Cattonaro F., Vendramin E., Main D., Aramini V., Blas A.L., Mockler T.C., Bryant D.W., Wilhelm L., Troggio M., Sosinski B., Aranzana M.J., Arús P., Iezzoni A., Morgante M., and Peace C., 2012, Development and evaluation of a 9K SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm, PLoS One, 7(4): e35668.

Zhang L.Y., Hu J., Han X.L., Li J.J., Gao Y., Richards C.M., Zhang C.X., Tian Y., Liu G.M., Gul H., Wang D.J., Tian Y., Yang C.X., Meng M.H., Yuan G.P., Kang G.D., Wu Y.L., Wang K., Zhang H.T., Wang D.P., and Cong P.H., 2019, A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour, Nat. Commun., 10(1): 1-13.

Zhou S.H., Zhang J.P., Che Y.H., Liu W.H., Lu Y.Q., Yang X.M., Li X.Q., Jia J.Z., Liu X., and Li L.H., 2018, Construction of Agropyron Gaertn. genetic linkage maps using a wheat 660K SNP array reveals a homoeologous relationship with the wheat genome, Plant Biotechnology Journal, 16 (3): 818-827.