MolBreed Publisher

Molecular Plant Breeding 2020, Vol.11, No.27, 1-12
http://genbreedpublisher.com/index.php/mpb

Research Article **Open Access**

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, China

2 Fruit Research Institute, Chinese Academy of Agricultural Sciences, Xingcheng, 125100, China

✉ Corresponding author Email: duannaibin@gmail.com

Molecular Plant Breeding, 2020, Vol.11, No.27 doi: 10.5376/mpb.2020.11.0027

Received: 19 Nov., 2020

Accepted: 30 Nov., 2020

Published: 31 Dec., 2020

Preferred citation for this article:

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, Molecular Plant Breeding, 11(27): 1-12 (doi:10.5376/mpb.2020.11.0027)

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 4 000 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 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 heterozygous sites' accuracy gradually increases. In this study, two algorithms were used in SNP mining. 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 variation can be identified from 1 896 bp. 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 corresponding genes. Among the coding region SNPs, 13 422 are non-synonymous, while 11 625 are synonymous variations. The ratio of non-synonymous to synonymous SNP is 1.15: 1. Using the filtered 4DTV sites, population clustering analysis results constructed using 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

Apple is one of the most important fruits in the forefront of production. Global apple yealid in 2019 exceeded 8.314×10^{10} kg, of which China's output was 4.139×10^{10} kg (data from the FAO database, http://www.fao.org/faostat/zh/#search/apple), accounting for 50% the above. For many years, apple production of Shandong province has been at the forefront of China (accounting for more than 25%). Simultaneously, it is also a province with extensive apple germplasm resources, leading the country in apple germplasm resource collection, innovation, and new varieties selection.

Genomics research is the basis of crop genetic breeding. Due to its importance, apple genome research has made significant progress. The apple genome has been assembled and sequenced four times (Velasco et al., 2010; Li et al., 2016; Daccord et al., 2017; Zhang et al., 2019). It is one of the fruit trees crops with the fastest progress in whole-genome assembly. Using genome-resequencing technology, researchers have carried out population genomics and population genetics study on the global apple germplasm resources. These studies clarified the domestication and evolutionary mechanism of modern apples. Genomics and bioinformatics showed that the germplasm resources tap into innovation's strong potential (Duan et al., 2017; Duan, 2017; Jia, 2018).

Regarding the germplasm resources research, population genotyping based on resequencing provides novel methods for protecting, identifying, evaluating, and innovating fruit tree germplasm resources. Genomics studies provide unique advantages such as high-throughput, big data, and GWAS analysis among phenotype and genotype (Chen et al., 2015; Chen et al., 2018). Genomics and bioinformatics studies should be combined to carry out the high-throughput genotyping of the related germplasm resources of cultivated apples. The underlying SNP and corresponding annotation database should be constructed together. There will be an excellent job in apple omics (Genomic and Proteomic) breeding. In the post-genome era, SNP array represents the future direction of low-cost genotyping due to its unique advantages. China has carried out the research, development, and application of SNP array in field crops such as wheat, corn, soybean, rice, cotton, and some cruciferous vegetable crops. Especially in wheat, Jia Jizeng et al. built a high-resolution (Affymetrix 660k) SNP array. These arrays have demonstrated well in identifying and evaluating germplasm resources, population genotyping, association analysis or functional gene mapping, and molecular marker-assisted breeding. The application prospects should not be underestimated (Zhou et al., 2018). Compared with resequencing, the tech of chip or array is easier to perform in that no reference genome comparison is required to achieve high throughput. Also, the chip or array has high detection accuracy of more than 99.9%, while the detection cost is relatively as low as about 1 000 CNY per sample. Researchers have designed three resolution of SNP array for apple breeding such as 8 K, 20 K, and 480 K, using which the application of genotyping and association analysis for the popular cultivated apples in Europe and America have been developed (Chagné et al., 2012; Bianco et al., 2014; Bianco et al., 2016).

In China, the fruit trees for which SNP array have been developed were strawberries, pears, and peaches; the density is 90 K, 200 K and 9 K, respectively (Verde et al., 2012; Bassil et al., 2015; Li et al., 2019). As apple is an important fruit tree, it's breeding urgently needs targeted, low-cost, and high-throughput genotyping methods. However, the SNP array research for the unique apple varieties in China has not been carried out yet. In this study, based on the resequencing research that has been carried out, the SNP array site mining research was carried out. On the one hand, it can be used for rapid identification of apple varieties, evaluation and selection of germplasm resources; it can also be used for genome-wide association analysis, functional gene positioning and molecular marker-assisted breeding.

1 Results

1.1 Statistics of apple genome resequencing for each accession

Upon the raw data was got, the adapter sequence and duplicated reads caused by PCR library building was removed. The volume of cleaned data is 363 G. Calculated with 720 M base pairs of the apple genome, the highest genome coverage was 21.02 × , the lowest genome coverage was 10.63×, and the average sample coverage reached 16.29×; it fully met the needs of re-sequencing analysis and SNP site mining (Table 1).

1.2 Effect of mismatch parameters on the mapping rate

Taking C18-06A sample Marshal (Qingdao No.1) as an example, the mismatch rate parameter mismatch required by the BWA software is the number of allowable mismatched bases between the data read and the reference genome, because the sequencing read length in this study is 150 bp in this study, the parameter value was increased from 1 (0.66%) to 12 (8.00%), and a series of comparison files were obtained. Then use the flag stat function of SAMtools to count the specific conditions of the comparison rate of all read data, paired data and single-ended data (Table 2; Figure 1); first, as the mismatch rate gradually increases, the comparison rate also gradually rises. High, but the upward trend gradually decreases until it is close to saturation.

The mapping rate of all reads and pair-end reads showed a trend of approaching saturation (Figure 1), while the mapping rate of single-ended data gradually decreased to a minimum value. As is shown in Figure 1, the total mapping rate is positively correlated with the mismatch rate, which is in line with the fourth-order equation: $y=-3E-05x^4+0.0\ 011x^3-0.0\ 145x^2+0.0\ 864x+0.7\ 418$ (regression coefficient $R^2 = 0.9\ 995$). The paired mapping rate of the reading segment is positively correlated with the mismatch rate, which is in line with the fourth-order equation: $y = -3E-05x^4 + 0.0\ 012x^3-0.0\ 149x^2 + 0.0\ 863x + 0.7\ 126$ (the regression coefficient $R^2 = 0.9\ 994$). The single-end mapping rate is negative correlated with the mismatch rate, which is in line with the fourth-order equation: $y = 2E-06x^4-7E-05x^3 + 0.0\ 009x^2-0.0\ 054x + 0.0\ 212$ (Regression coefficient is $R^2 = 0.9\ 993$).

Table 1 Statistics of apple genome resequencing for each accession

Sample name	Clean reads	Clean bases	Q20 ratio (%)	GC ratio (%)	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

Table 2 Effect of mismatch parameters on the mapping rate

Mismatch parameter	Total mapped reads	Total mapped ratio (%)	Paired mapped reads	Paired mapped ratio (%)	Single mapped reads	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

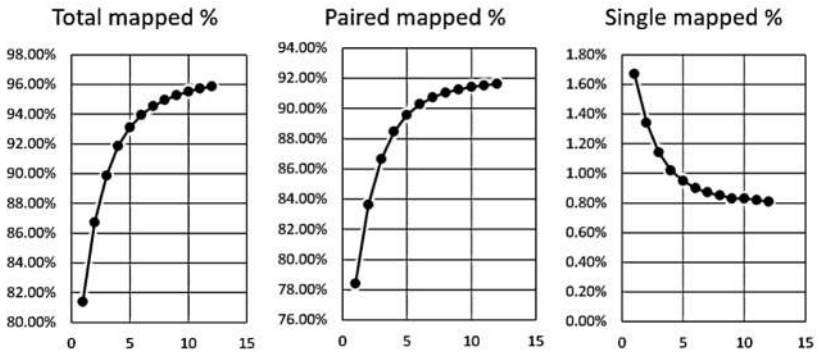

Figure 1 Effect of mismatch parameters on the mapping of total data, paired data and single-ended data

1.3 Effect of mismatch parameters on accurate of heterozygous SNP calling

Next, we compared the influence of different mismatch parameters on the accuracy of locus detection. Take Chr11 on chromosome 11 as an example (Figure 2): As the allowable mismatch rate increases, in the Chr11 area in the figure, comparable data Gradually increase, the sequencing coverage gradually increased from 11 to 19, and homozygous loci under low coverage were identified as heterozygous loci under high coverage, indicating that the increase in mismatch rate is beneficial to heterozygous Site detection. It can be seen that as the mismatch rate increases, the comparison stringency decreases; the genome coverage gradually increases, which is more conducive to the detection of heterozygous sites. Many plants have the characteristics of distant hybridization, self-incompatibility, high genomic heterozygosity and extensive genetic drift, such as apple, Brassica, corn and other crops. For the SNP site mining of such crops, on the one hand, it is necessary to improve the data coverage of the whole genome by sequencing, and on the other hand, it is to select the most suitable mismatch parameter; it is of reference significance for the SNP site mining of crops with higher heterozygosity.

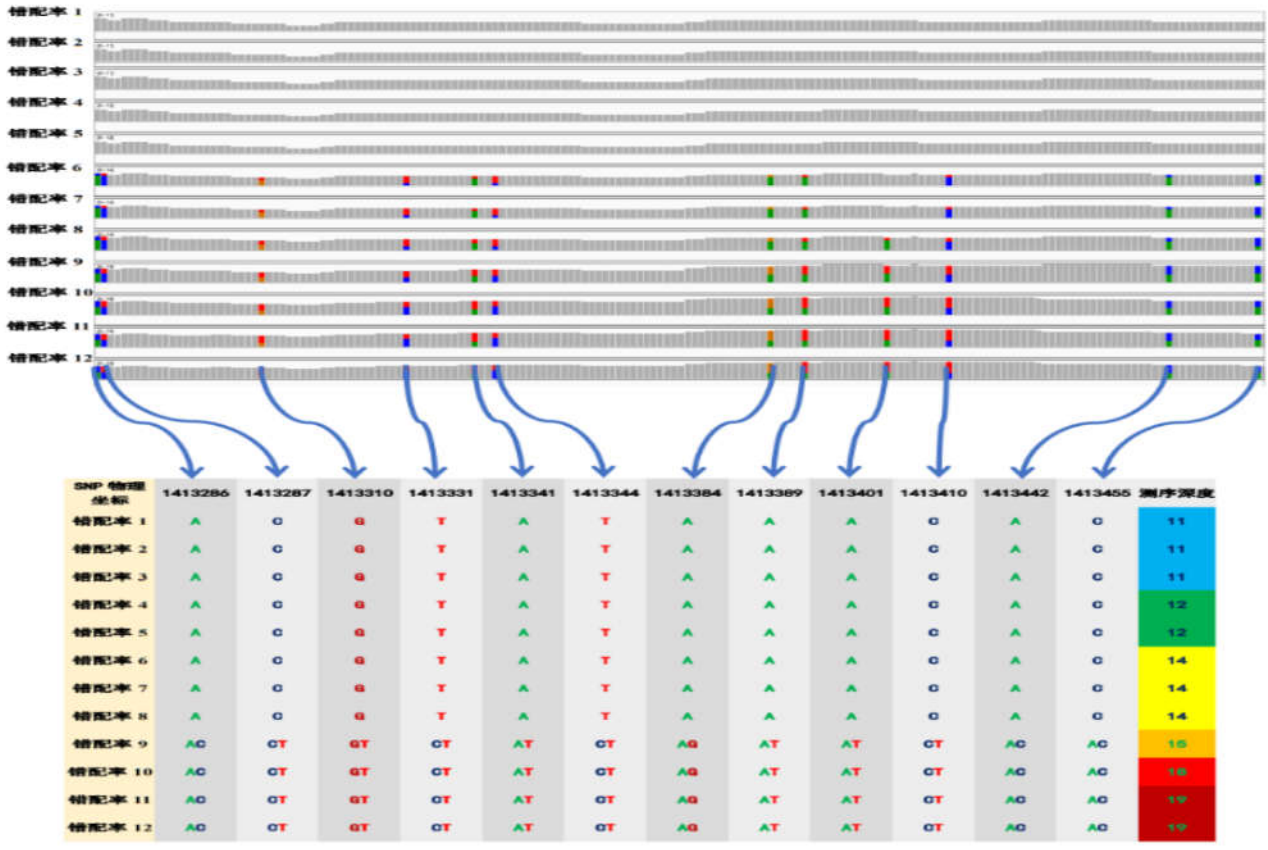

Figure 2 Effect of mismatch parameters on accurate of heterozygous SNP genotyping

1.4 Comparison and integration of sites under the two analysis procedures

According to the standard procedure of next-generation sequencing combined with the BCF tools: bwa-sam-bam-pileup-bc fools' algorithm, a total of 28 997 212 SNVs were identified, including 26 758 563 single-base SNPs, and 1 060 691 short inserts. 117 795 short deletions. This algorithm can detect 3 types of variation so the sensitivity of variation detection is as high as one variation can be identified from every 27 loci on average.

According to the standard procedure of next-generation sequencing combined with the in-house algorithm: the bwa-sam-bam-pileup-column algorithm identified a total of 1 147 801 variation. This algorithm is designed for the detection of single-base SNPs, whereas the sensitivity of variation detection is pretty low. On average, one variation can be identified from every 618 loci.

Combining the obtained by the two algorithms, and further taking the intersection based on the "chromosome + site coordinate" as the feature value, a highly reliable single-base SNP variation data set is obtained. A total of 374 404 variation were identified. Because of the intersection, these variations are all single-base SNPs. On average, one variation can be identified from every 1 896 loci. Primers were designed for 1 000 randomly selected homologous SNPs and amplified by PCR, and then the amplified products were sequenced by Sanger. The results showed that the coincidence of the selected SNP sites on the two platforms was as high as 98.1%.

1.5 Distribution of SNPs among the apple genomes

Annotation analysis of SNPs showed that out of the total 373 763 SNP loci, 143 269 (38.27%) were located in the intergenic region, 25 047 (6.7%) were located in the gene coding region, and 143 269 (38.27%) were located in the gene coding region. Located in the intergenic region, 179 426 (47.92%) were located in the 2 kb region upstream or downstream of the gene. Among all the SNPs in the coding region, 13 422 are non-synonymous variations and 11 625 are synonymous variations (Table 3; Figure 3). The ratio of non-synonymous and synonymous SNPs is 1.15: 1. Non-synonymous SNPs, also called missense SNPs, change from encoding one amino acid to another to form a phenotypic modification; synonymous SNPs are also called silent variations, although there are base variations, they still encode the same amino acid and cannot be formed Phenotypic modification. Compared with other cultivated field crops and fruit tree crops, apples have a lower percentage of variation in the genome that can form corresponding phenotypic modifications (Duan et al., 2017).

Table 3 Number of effects by type and region

Variation type	Count	Percent	Variation region	Count	Percent
3'_UTR_variant	7 906	1.09 %	Downstream	174 403	24.22 %
5'_UTR_premature_start_codon_gain	485	0.07 %	Exon	25 443	3.53 %
5'_UTR_variant	2 773	0.38 %	Intergenic	276 549	38.40%
Downstream_gene_variant	174 403	24.13 %	Intron	59 074	8.20%
Initiator_codon_variant	2	0 %	Splice_site_acceptor	97	0.01%
Intergenic_region	276 549	38.27 %	Splice_site_donor	60	0.01%
Intron_variant	61 125	8.46 %	Splice_site_region	2 273	0.32%
Missense_variant	13 422	1.86 %	Transcript	356	0.05%
Non_coding_transcript_exon_variant	289	0.04 %	Upstream region of gene (within 5k)	170 706	23.71%
Non_coding_transcript_variant	356	0.05 %	Utr_3_prime	7 906	1.10%
Splice_acceptor_variant	97	0.01 %	Utr_5_prime	3 258	0.45%
Splice_donor_variant	60	0.01 %	-	-	-
Splice_region_variant	2 541	0.35 %	-	-	-
Start_lost	24	0.00%	-	-	-
Stop_gained	292	0.04 %	-	-	-
Stop_lost	20	0.00%	-	-	-
Stop_retained_variant	19	0.00 %	-	-	-
Synonymous_variant	11 625	1.61 %	-	-	-
Upstream_gene_variant	170 706	23.62 %	-	-	-

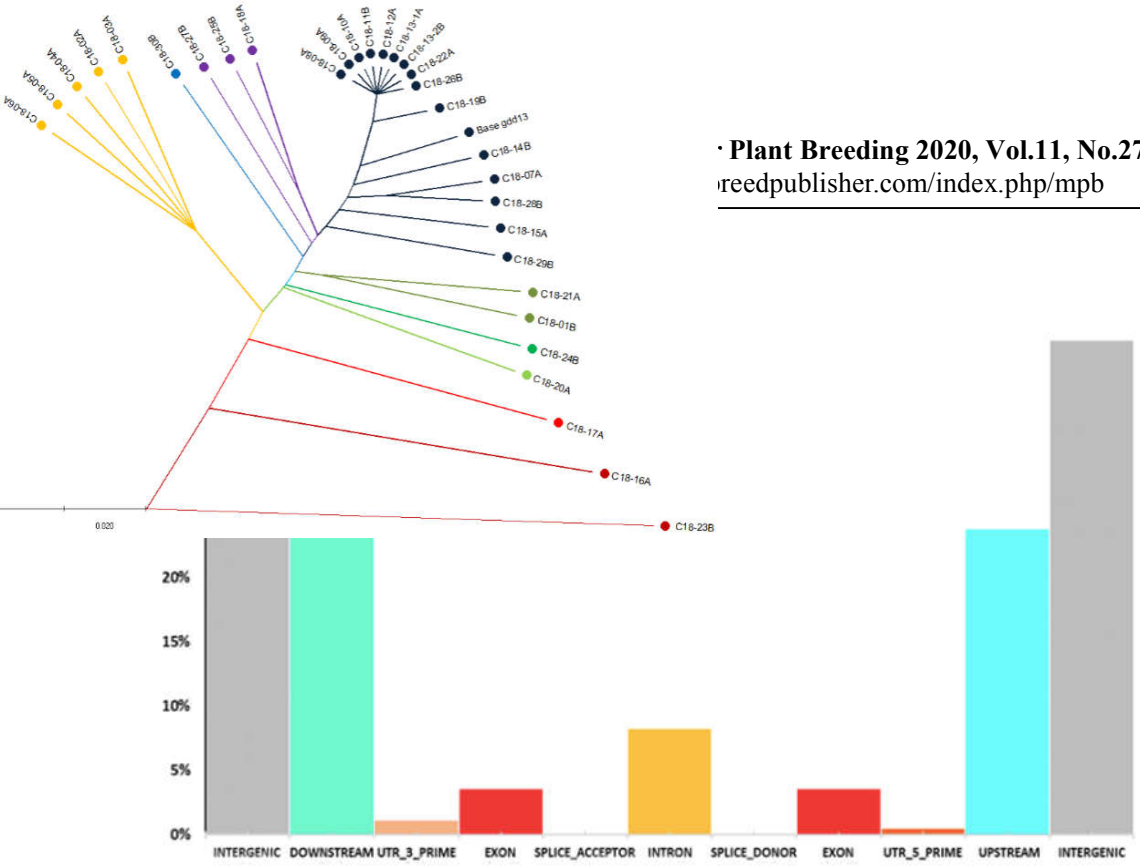

Figure 3 Number of effects by region

1.6 Construction of population clustering evolutionary tree

The evolutionary tree was derived using the minimal evolution method (Figure 4) (Rzhetsky and Nei, 1992). The figure shows the best evolutionary tree with a total branch length of 0.7 665, which is drawn with reference to the ratio of evolutionary distance. On the whole, the phylogenetic tree shows that the main cultivated apples collected in this study in Shandong Province are mainly divided into four types: Fuji, Marshal, Golden Crown (Jin Guan), Gala, and other hybrid combinations. It is worth mentioning that: (1) C18-23 samples are wild resources that were bred in Xinjiang wild apples, and the earliest divergence occurred in the evolutionary history; (2) C18-2, C18-3, C18-4, C18- 5 and C18-6, according to the information provided by the resource nursery, these samples are all marshals, and they were successfully gathered together in this experiment. Similarly, C18-8, C18-9, C18-10, C18-11, C18-12, C18-13-1 and C18-13-2. According to the information provided by the resource nursery, these samples are all Fuji-series. We successfully got together during the experiment. The above samples are all taken from the National Apple Resource Nursery (Xingcheng) of the Institute of Fruit Trees, Chinese Academy of Agricultural Sciences, which has many years of accurate genealogical data registered for these germplasm resources. This result indirectly proves the reliability of the SNP data in this experiment; (3) The clustering results of the remaining samples are also in line with expectations.

Figure 4 Evolutionary relationships of taxa

2 Discussion

2.1 How to determine the ideal mismatch parameters

Through the comparative analysis of a series of mismatch parameters, this study found that with the increase in the mismatch rate, the number of Reads that can be compared to the genome gradually increased, showing a trend of gradual saturation. Increasing the mismatch rate to a certain level is conducive to improving sequencing coverage and facilitating site mining. But it is meaningless to increase the mismatch rate indefinitely when it is close to saturation.

If so, in order to further determine the best Bwa alignment mismatch parameters for different samples (Li and Durbin, 2009), the project team members independently developed a method to select the best alignment mismatch rate parameters: NCBI (Pruitt et al., 2005) downloaded the reference genome sequence of cultivated apples and established a local Blast database of the apples. Then randomly select 1 000 reads from the sequencing data for local Blast, sort the mapping ratio in the Blast results, and then calculate the Identity similarity of the 550 th read, thereby determining the BWA mismatch parameters.

The size of the best mismatch parameter measures to a certain extent the distance of the test sample relative to the reference genome. For example, the test sample No. 23 is a new hybrid of Xinjiang wild apple (M. sieversii in Xinjiang) and red meat apple (Malus domestica 'Redlove Era'), and the genetic relationship is far from the reference genome apple Golden Delicious (M. domestica) The farthest, correspondingly, the mismatch parameter obtained by our calculation is the largest, which is 7; while the reference sample No. 1 and the reference genome belong to the Jinshuai line, the relationship is closest, and the corresponding mismatch parameter obtained by our calculation is the smallest, which is 4.

This method of selecting the best alignment mismatch rate has been adopted in the previous article (Duan et al., 2017; Duan, 2017). Choosing accurate mismatch parameters on the one hand can obtain a sufficiently high sequencing coverage and ensure the accuracy of the site; on the other hand, it can compare as many reads as possible with a minimum amount of calculation, avoiding excessive calculations and improving analysis efficiency.

2.2 The necessity of data integration

In order to enhance the reliability of SNP loci, this study integrated the sequencing data of 31 newly collected samples with the sequencing data of 23 cultivated apples sequenced previously (Duan et al., 2017; Duan, 2017)), the purpose of data integration: one is to enhance the reliability of the site, and the other is to compare the polymorphisms between the cultivated apple populations in the province and the cultivated apple populations abroad (a separate article is published).

Increasing the number of reference samples and data integration has the following advantages: it increases the diversity of samples; because the previous re-sequencing involves 23 main cultivated apple types worldwide; on the other hand, it improves the reliability of mining sites; Through the integration, the diversity of the large data set is actually used to test the diversity of the quantum set; the future application scope of the selected site is enhanced.

2.3 SNP calling strategies for highly heterozygous species

First, for species with high heterozygosity, the assembly of their genomes is difficult, and the corresponding assembly quality is generally not high. For example, multiple published fruit tree genomes have high heterozygosity. At present, next-generation sequencing is widely used. When reads with a read length of 100~150 bp are used to assemble Contigs, the overlap relationship between contigs of highly heterozygous genomes is not easy to clarify, resulting in a low N50 and a large number of failures in the genome. The overlapping area (gap) (Pryszcz and Gabaldón, 2016). Correspondingly, when second-generation sequencing is used for SNP detection, the highest possible sequencing depth and as long as possible read reads are required, such as the hiseq 4 000 platform that is currently widely used. The average sequencing depth of this study reached 16 X, and the read length was 150 bp. Based on the above strategy, the SNP loci in this paper have been sequenced by Sanger, and

the coincidence rate is as high as 98.1%. This is higher than the accuracy of re-sequencing SNP loci in maize and cotton populations, and these two crops have a good genetic research foundation, and their genome assembly quality is better than that of apples.

Furthermore, when the mismatch rate is increased, more sites will be identified as heterozygous sites, which indicates that high coverage is beneficial to the detection of heterozygous sites. Many plants have high genomic heterozygosity due to factors such as distant hybridization, self-incompatibility, etc., and there are obvious and extensive genetic drift, such as Malus Mill., Brassica, Maize, etc. crop. For the SNP site mining of such crops, on the one hand, it is necessary to improve the data coverage of the sequencing in the whole genome, and on the other hand, it is to select the most suitable mismatch parameter. This has reference significance for crops with higher heterozygosity.

Finally, an improved algorithm should be used in site selection. Try to avoid using only one detection process. At present, in the existing SNP detection process, the upstream analysis process is BWA combined with SAM tools, namely BWA-Sam-bam-pileup. Different algorithms are used to select SNP after the generated pileup file, and the file format is mainly VCF, hapmap or list format. At this time, two or more analysis processes are used to normalize the generated data into a VCF file. By using the coordinate information of the chromosome to take the intersection, a more reliable site can be obtained.

3 Materials and Methods

3.1 Variety collection

This study selected 31 cultivated apple varieties with a wide range of types, covering the four major apple lines Fuji, Marshal, Golden Crown, Gala and some new hybrid lines. The pedigree and origin of the samples are from the Institute of Fruit Research, Chinese Academy of Agricultural Sciences National Apple Resource Nursery (Xingcheng); covers the scion types of the main cultivated apples in Shandong Province. From the point of view of the area where the materials are obtained, the area of the materials is all over the main apple cultivation areas in Shandong Province. From the pedigree information, the experimental materials are sufficiently representative in terms of diversity (Table 4).

From June 15th to 20th, 2018, most of the leaves were treated with liquid nitrogen immediately after they were collected from the raw top tip leaves of the year. Only the leaves of the 23, 25, and 26 samples were dried with silica gel. All leaf samples were extracted according to the standard DNB extraction method. The extracted DNA samples were tested on agarose gel for quality, and after meeting the sequencing requirements, the library was constructed by the paired-end PE150 strategy and delivered to BGI to complete the sequencing on the Hiseq-4 000 platform.

3.2 Preprocessing and statistical analysis of sequencing data

The original data needs to go through a Perl sequencing script (written by the research team of this research group) to remove sequencing duplication caused by PCR. Specifically, for paired Reads with different sequencing position information IDs, any pair 1 or pair 2 with identical base data at the same time in the 15~135 bp interval is defined as a sequencing duplication caused by PCR, and the data is filtered out. The command line is: "drop_dup_both_end.pl raw_fq1 raw_fq2".

The data from which PCR sequencing duplicates have been removed is filtered by Trimmomatic3.0 software to remove 1, sequencing adapters, and 2, low-quality reads. In this way, the net data is finally obtained. The command line is "trimmomatic PE -thReads 75 fq.1 fq.2".

Including total sequencing data statistics, sequencing depth statistics, read comparison rate statistics and the determination of comparison mismatch parameters. The command line is "fastqc -q trimmed_fq1 trimmed_fq2".

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

Accession number	Variety name	Type of breed	Location
C18-1	Golden Delicious	Golden Delicious	Muping, Yantai
C18-2	Starkrimson (Netherlands)	Delicious short spur	Jiaonan, Qingdao
C18-3	Pingyin Spur	Delicious	Pingyin, Jinan
C18-4	Kangtun Spu	Delicious	Muping, Yantai
C18-5	Xiyanghong	Delicious	Yantai Institute of Agricultural Sciences
C18-6	Qingdao1	Delicious	Daze Mountain, Yantai
C18-7	Akifu 5	Fuji (Yuanshuai×Guoguang)	Muping, Yantai
C18-8	Akifu 1	Fuji	Laixi, Qingdao
C18-9	Nagafu 7	Fuji	Rongcheng, Weihai
C18-10	Akifu 2	Fuji	Rongcheng, Weihai
C18-11	Nagafu 2	Fuji	Rongcheng, Weihai
C18-12	Yanfu 5	Fuji	Muping, Yantai
C18-13-1	Changhong-1	Fuji	Pingyuan, Dezhou
C18-13-2	Changhong-2	Fuji	Jimo, Qingdao
C18-14	Orin	Golden Delicious×Indo	Jimo, Qingdao
C18-15	Indo	White Winter Pearmain × ?	Jimo, Qingdao
C18-16	Geneva Early	Quinte×Julyred	Qixia, Yantai
C18-17	Fujiki 1 (Nanbusakigake)	Uknown	Linyi, Shandong
C18-18	RallsJanet	Uknown	Linyi, Shandong
C18-19	Shinsekai	Fuji×akagi	Linyi, Shandong
C18-20	Sansa	Gala×Akane	Tai'an, Shandong
C18-21	Jonagold	Golden Delicious×Jonathan	Tai'an, Shandong
C18-22	Wangshi	Fuji	Tai'an, Shandong
C18-23	Violetred No.1	Uknown	Tai'an, Shandong
C18-24	Taishan early	Uknown	Tai'an, Shandong
C18-25	RallsJanet	Uknown	Tai'an, Shandong
C18-26	Red General Fuji	Fuji	Liaocheng, Shandong
C18-27	Jonathan	Esopus Spitzenburg	Linyi, Shandong
C18-28	Royal Gala	Kidd's Orange Red×Golden Delicious	Linyi, Shandong
C18-29	Meiguo 8	NY543	Linyi, Shandong
C18-30	Pink Lady	Golden Delicious	Tai'an, Shandong

3.3 Determine of mismatch values

Take the sample C18-06A (Qingdao 1, Golden delicious) as an example, for the mismatch rate parameter mismatch required by the BWA software: the value of allowable mismatched bases between the data read and the reference genome, because apples have distant hybridization and heterozygosity Higher, therefore, increase the parameter value from 0.66% to 8.00%, which corresponds to a read length of 150 bp and is 1~12. A series of comparison files were obtained to compare the comparison rate to coverage and SNP detection.

In order to determine the appropriate BWA (Li and Durbin, 2009) alignment mismatch parameters, firstly download the reference genome sequence of cultivated apple from NCBI (www.ncbi.genome.com) and establish a local Blast database of this species. Randomly extract 1 000 reads from the sequencing data for local Blast, and then count the similarity of the 550th read after sorting the Mapping ratio to determine the mismatch value of BWA.

3.4 Data mapping and SNP mining

In this study, the genome of the cultivated apple 'Jinshuai' published in 2017 (Daccord et al., 2017) was used as the reference sequence, and all the 31 collected in this experiment and the 23 cultivated in the previous article (Duan et al., 2017) were used. For apples, the resequencing data of a total of 54 cultivated apples were compared with the reference genome by BWA (Li and Durbin, 2009) (mismatch ranging from 4 to 7). The pileup file is converted by SAM tools (Li et al., 2009). Next, two different processes are used to detect SNP site information: (1)

BWA-sam-bam-pileup-bc fools' algorithm, using SAM tools combined with BCF tools to convert Pileup files to obtain SNP data sets in VCF file format for each sample. (2) According to the second-generation sequencing standard process combined with the self-developed In-House algorithm, namely: the bwa-sam-bam-pileup-column algorithm, a SNP data set similar to hapmap is obtained. (3) Using an improved intersection algorithm: the SNP site information obtained by the above two methods is based on the method of taking the intersection of chromosome coordinates, and then to a higher quality SNP site.

SNP validation method: 6 samples were selected, and 1,000 homogenous SNP sites (that is, non-heterozygous) were randomly intercepted from chromosome 11, and 50 bp sequences on both 3' and 5' flank were designed based on this. Then primers were constructed, and PCR amplification experiments were performed. Finally, the amplified products were testified by 3730 capillary electrophoresis.

3.5 Annotation of SNP loci

Compared with other annotation software, SNPEff is powerful. It can get the gene region where the variation site anchor and the detailed gene region information conducive to subsequent functional gene mining and mapping. Due to the use of the java platform, it is to learn and use. The manual http://SNPeff.sourceforge.net/SNPEff_manual.html describes the annotation method in detail. The command line for the annotation in this study is as follows:

To modify SNPeff software settings: "vim user path/SNPEff/SNPeff-4.3.1t-1/SNPEff.config"; To add genome information: "# apple genome version GDDH13 GDDH13.genome: Apple"; To build local library: "SNPEff build -gff3 -v GDDH13"; SNP annotation: "SNPEff -v -stats prefix.html GDDH13 prefix.vcf> prefix.ann"; the output html file is a graphical interpretation of the site annotation results presented in the form of a web page, and the output an file is a text listing the detailed results of each SNP annotation.

3.6 4DTV loci filtering and cluster analysis

In the protein coding region of the gene, there are some amino acids corresponding to the third codon that can use any 4 kinds of bases, and no amino acid changes will be formed. Such a site is called a quadruple degeneration site (4DTV). This kind of unintentional variation has almost no selective pressure, and its variation rate can be used as a "clock" to estimate evolution, which is particularly suitable for building evolutionary trees and analyzing population genetic structure (Fazio et al., 2014). This study used the Perl script written by the team to screen the entire set of SNP data in the CDS region according to the following rules: minimum allele frequency (MAF) $\geqslant$ 5%, and the data loss rate corresponding to each locus$\leqslant$10%. Finally, 24 326 quadruple degenerate sites (4DTV) were screened. Finally, the location is input into the Mega X software, and the close-neighbor-interchange (CNI) algorithm (Kumar et al., 2018) is used on the first search level. Thus, the phylogenetic tree of the population is constructed.

Authors' contributions

Duan N.B., and Ma Y.M. are the executives of the experimental design and experimental research of this study; Xie K., Bai J., Yang Y.Y., Pu Y.Y., and Gong Y.C. completed DNA extraction, PCR amplification and Sanger sequencing; Duan Naibin completed sequencing data preprocessing and genome comparison , SNP site mining, gene annotation and paper writing; Ma Y.M., Wang K., and Wang X.M. are responsible for sample collection and revision of some paper chapters; Duan Naibin is the project designer and person in charge, directing experimental design, data analysis, paper writing and Modification. All authors read and approved the final manuscript.

Acknowledgements

This research was jointly funded by the Provincial Key Research and Development Project of the Department of Science and Technology of Shandong Province (Project No. 2018GNC110031) and the Shandong Agricultural Good Seed Project-Collection, Protection and Accurate Identification of Crop Germplasm Resources (Project No. 2019LZGC017).

References

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
https://doi.org/10.1186/s12864-015-1310-1
PMid:25886969 PMCid:PMC4374422

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
https://doi.org/10.1111/tpj.13145
PMid:26919684

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
https://doi.org/10.1371/journal.pone.0110377
PMid:25303088 PMCid:PMC4193858

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
https://doi.org/10.1371/journal.pone.0031745
PMid:22363718 PMCid:PMC3283661

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

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.

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
https://doi.org/10.1038/ng.3886
PMid:28581499

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

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 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
https://doi.org/10.1038/s41467-017-00336-7
PMid:28811498 PMCid:PMC5557836

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
https://doi.org/10.21273/JASHS.139.2.87

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

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
https://doi.org/10.1093/molbev/msy096
PMid:29722887 PMCid:PMC5967553

Li H., and Durbin R., 2009, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics, 25(14): 1754-1760
https://doi.org/10.1093/bioinformatics/btp324
PMid:19451168 PMCid:PMC2705234

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
https://doi.org/10.1093/bioinformatics/btp352
PMid:19505943 PMCid:PMC2723002

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
https://doi.org/10.1111/pbi.13085
PMid:30690857 PMCid:PMC6662108

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
https://doi.org/10.1186/s13742-016-0139-0
PMid:27503335 PMCid:PMC4976516

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: 501-504
https://doi.org/10.1093/nar/gki025
PMid:15608248 PMCid:PMC539979

Pryszcz L.P., and Gabaldón T., 2016, Redundans: an assembly pipeline for highly heterozygous genomes, Nucleic Acids Research, 44(12): e113-e113
https://doi.org/10.1093/nar/gkw294
PMid:27131372 PMCid:PMC4937319

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
https://doi.org/10.1038/ng.654
PMid:20802477

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
https://doi.org/10.1371/journal.pone.0035668
PMid:22536421 PMCid:PMC3334984

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
https://doi.org/10.1038/s41467-019-09518-x
PMid:30940818 PMCid:PMC6445120

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
https://doi.org/10.1111/pbi.12831
PMid:28921769 PMCid:PMC5814592