Article

Genome Survey Sequencing and Analysis of SSR Markers of Rambutan (*Nephelium lappaceum* L.)

Xing’e Lin, Jianhong Ming, Hongmao Gao, Zheli Ding, Yu Ge, Zhaoxi Zhou\*

*Haikou Experimental Station, Chinese Academy of Tropical Agricultural Science, Haikou 570000, China*

\*ForCorrespondence: zhouzhaoxi@catas.cn

Received: date; Accepted: date; Published: date

**Abstract**

*Nephelium lappaceum* L. is one of the most popular fruits in tropical regions of the world, but little genomic information is available for this species. In this study, to provide a foundation for subsequent sequencing schemes, the Illumina Hiseq sequencing platform was used to survey the genome of *N. lappaceum.* Bioinformatics software was used to evaluate the genome size and heterozygosity ratio, to identify repeated sequences, and to estimate the GC content. 54.07 Gb clean reads were generated, and 987,898 contigs (N50 length, 923-bp) and 723,955 scaffolds (360,534 Mb total length) were obtained. Our results indicate that the genome of rambutan is about 386.95 Mb in size with a heterozygosity rate of 0.91%, based on a K-mer analysis. By screening this assembled rambutan genome, 71,291 simple sequence repeat microsatellites (SSRs) were detected. The genome of *N. lappaceum* is complex with high heterozygosity ratio and a large number of repeated sequences. We suggest that a combination of Illumina and PacBio sequencing technologies with Hi-C technology and resequencing technology should be used for subsequent large-scale sequencing.

**Keywords:** Rambutan (*Nephelium lappaceum* L.), Genome survey, Genome size, SSR

Introduction

Rambutan (*Nephelium lappaceum* L.) is a non-deciduous tropical fruit tree in the family Sapindaceae, which includes other fruit crops such as litchi (*Litchi chinensis* Sonn.), longan (*Dimocarpus longan* Lour), and pulasan (*Nephelium mutabile* Blume). Rambutan is indigenous to Indonesia and Malaysia and is now widely cultivated over all Southeast Asia, as well as in China, India, Australia, Hawaii, and Africa. The pulp of rambutan is sweet, juicy, and rich in vitamins, carbohydrates, and proteins. Rambutan fruits are usually eaten fresh, but are also canned, dried, and processed into preserves, juices, jams, jellies (Sukmandari et al 2017). The pulp, seeds, and peel of rambutan fruits contain compounds with strong antioxidant properties such as flavonoids, ellagic acid, corilagin, geraniin, β-carotene, and vitamin C (Gusman and Tsai 2015; Palanisamy *et al.* 2008; Nont *et al.* 2008; Palanisamy *et al.* 2011). Thus, there is potential to extend the applications of rambutan in the medicine and food industries.

In recent years, whole genome sequencing of fruit trees has become a research focus, and many genome maps of fruit trees have been published. Such studies have provided data that are essential for research on the molecular biology and comparative genomics of fruit trees. However, progress in molecular biology research and the whole genome sequencing of fruit trees is limited by their long growth periods, the high heterozygosity of their genomes, the large number of repeated sequences, and the unclear genetic background caused by self-incompatibility. With the development of sequencing technology, the improvement of sequencing efficiency, and the reduction of sequencing costs, there have been rapid advances in whole genome sequencing of fruit trees. Taking the plaBiPD online website (https://www.plabipd.de/index.ep) as an example, since the first genome sequence of grape was completed in 2007, whole genomes of 28 other fruit species had been sequenced by 2019. The results of whole genome sequencing of fruit trees have provided a huge resource platform for molecular biology research. Such studies have shed light on the genome structure and function of fruit trees, provided details of their origin and evolution, and allowed for the localization and cloning of important functional genes to accelerate molecular breeding.

There are 22 *Nephelium* species in the genus *Nephelium*, with five mainly distributed in Myanmar, Thailand, and Indonesia, 13 in Peninsular Malaysia, 16 in Borneo, four in the Philippines, and three in western Java (Ampthill 2010). Detailed and high-quality genetic and genomic data and a high diversity of accessions are essential to develop new and improved crop varieties. However, although rambutan is cultivated throughout the tropics, the genetic background of the existing germplasm resources is unclear, and this has greatly hampered the breeding process. Genome survey sequencing can provide information about the genomic structure of a species, including genome size, the extent of heterozygosity, and the number of repeat sequences. It also be used to detect sequences for new SSR markers in plants (Wei *et al.* 2014; Hou *et al.* 2016; An *et al.* 2017). In this study, we conducted a genome survey and determined the genome size, GC content, and heterozygosity rate of *N. lappaceum*. On the basis of our data, novel SSR markers were developed and were used to reveal the population structure of the rambutan germplasm.

Materials and Methods

**Plant Materials**

*N. lappaceum* cv. ‘BR7’, the most popular and widely cultivated rambutan cultivar in China, was selected for genome survey sequencing (Figure 1). Young leaves of BR7 plants grown at the Baoting Tropical Crop Research Institute, Baoting city, China (109.7°, 18.64N) were collected and frozen in liquid nitrogen. Genomic DNA was extracted by the CTAB method (Porebski *et al.* 1997). The DNA concentration was determined by ultraviolet-visible spectrophotometry, and its quality was confirmed by 1% agarose gel electrophoresis. The purified DNA samples were standardized at 20 ng/μl and stored at −20 °C.

**Library Construction and Genome Survey Sequencing**

The DNA sample was randomly sheared into 350-bp fragments by an ultrasonicator (Covaris, Woburn, MA, USA). The library was constructed by end-repair, polyA-tail addition, sequencing adapter addition, purification, and PCR amplification steps. The Illumina HiSeq platform (Novogene Co. Ltd., Beijing, China) was used for high-throughput paired-end sequencing of the constructed library.

**Quality Assessment of Sequencing Data**

To ensure the accuracy of bioinformatic analyses, raw reads generated by Illumina sequencing were filtered to remove low-quality reads, reads with joint contamination, and reads with >10% N (where N is an unascertainable base). Then, the reads were further blasted against the NCBI database to filter chloroplast sequences to obtain clean reads. Sequencing quality was evaluated according to three factors: GC content, sequencing data quality, and sequencing error rate.

**Genome Size Estimation by 17-mer Analysis**

A K-mer analysis was used to estimate the genome size, heterozygosity ratio, and abundance of repeated sequences. The K-value was taken as 17 for analysis. Assuming that all the K-mers extracted from the reads traverse the whole genome, and the K-mer depth frequency distribution obeys Poisson distribution, the K-mer frequency distribution can be counted from all sequenced reads, and the K-mer depth estimation value can be calculated to estimate the genome size.

**Estimation of Heterozygosity Rate and Proportion of Repeat Sequences**

The K-mer map can be used to determine the number of repeats in a genome. If the genome contains a high proportion of repeats, the map will show a thick tail. The genomic heterozygosity was evaluated by simulation data fitting. For heterozygous genomes, K-mers fall into two categories: heterozygous and homozygous. Assuming that each heterozygous site is covered by 2 × K K-mers, the expected depth of a heterozygous K-mer is 1/2. Therefore, the number of heterozygous sites can be estimated using 1/2 (percentage of heterozygous K-mer types) and nKspecies (total number of all K-mer types). The heterozygosity ratio was calculated using Eq. 1. The proportion of repeated sequences was obtained by calculating the proportion of the number of K-mers greater than 1.8-fold the homozygous peak depth.

$∅=\frac{a\_{1/2}×n\_{Kspecies}/(2×K)}{n\_{Kspecies}-a\_{1/2}×n\_{Kspecies}/2}=\frac{a\_{1/2}}{K(2-a\_{1/2})}$  (1)

**Preliminary Genome Assembly**

We used SOAP denovo software to carry out preliminary genomic assembly. Original sequences were pretreated by removing joint and low-quality sequences (Li *et al.* 2010). All available reads were first assembled into contigs, and then the contigs were connected to form large contig fragments. Longer scaffold sequences were assembled by matching the ends of library sequences through different insert sizes. The filtered reads were aligned to the assembly sequence with SOAP software to obtain base depth. With 10 KB as the window for non-repetitive advancement in the sequence, the average depth and GC content of each window were calculated, and the GC depth spot map was constructed. The GC depth distribution can provide information about an obvious GC bias or bacterial contamination of the sequence data. The heterozygosity rate and repeat distribution of the genome can also be estimated from the stratification of GC clumps.

**Development of Genomic SSR Markers**

Microsatellite repeats were identified in the rambutan genome sequence using SR search software. Only the SSR loci with SSR motif repeats of 2–6 nucleotides were considered. The minimum SSR length was defined as 12.

Results

**Sequencing Data Evaluation**

After strict filtering to remove low-quality data, 54.07 Gb clean bases were obtained by sequencing the genome survey library with 350-bp inserts. The quality of sequencing data should be >Q20 for reliable results from subsequent analyses. The higher the base mass value is, the more reliable the base recognition is and the less likely the base measurement error is. The sequencing quality evaluation showed that Q20 and Q30 were 95.90% and 90.70%, respectively, and the sequencing error rate was 0.03% (Table 1). These results indicated that the sequence quality of N. lappaceum was good.

**K-mer Analysis and Genome Size Estimation**

Based on the genome sequence data, 54.07 Gb clean bases were used for K-mer analysis with a K-value of 17. The frequency distribution is shown in Figure 2 (the X-axis is the K-mer depth and the Y-axis is the number of K-mers for the corresponding depth). The peak of depth distribution was 110 and the number of K-mers was 43,449,243,256. According to the following formula: genome size = K-mer number/K-mer depth; the genome size of *N. lappaceum* was estimated to be approximately 394.99 Mbp. After excluding the error effects due to erroneous K-mers, the revised genome size was 386.95 Mbp. As shown in Figure 2, there was a minor peak at the position of the integer multiples of the main peak, indicative of a high level of heterozygosity. The heterozygosity calculated using Eq. 1 was 0.91%. The proportion of repeated sequences (45.72%) was calculated as the proportion of 1.8-times the number of K-mers after the main peak out of the total number of K-mers.

**Preliminary Genome Assembly Results for Rambutan**

All the clean bases (54.07 Gb) were used for preliminary genome assembly. A K-value of 41 was used to construct contigs and scaffolds. We obtained the original genome sequence of *N. lappaceum* (Table 2). Statistical analyses were performed on the contigs in scaffolds with assembled lengths greater than or equal to 100 bp. A total of 987,898 contigs with a total sequence length of 356,250,469 bp were obtained. The maximum contig length was 68,350 bp. The N50 and N90 contig lengths were 923 bp and 126 bp, respectively. Only scaffolds longer than 100 bp were counted in the assembly results. In total, 723,955 scaffolds with the total length of 360,534,817 bp were assembled, and the longest scaffold was 142,353 bp. The lengths of scaffolds N50 and N90 were 1,929 bp and 147 bp, respectively. The distribution of the assembled contigs (Figure 3) showed two distinct peaks. According to the genome size obtained from the K-mer-17 survey analysis, we determined that the peak value around 85 was homozygous. About half of the peaks in the abscissa in front of the homozygous peaks were heterozygous, so the genome of *N. lappaceum* shows a certain degree of heterozygosity.

**GC Content Distribution**

The estimated GC content of the *N. lappaceum* genome was 33.52%. The GC content of assembled contigs was calculated (Figure 4). The GC depth distribution was divided into two parts. Most of the contigs were in the 20% – 60% GC content and 20× – 120× depth area. Combined with the K-mer analysis, these results implied that the two parts corresponded to homozygosity and heterozygosity peaks, respectively. Some contigs were distributed in the area with coverage above 120×, which was probably caused by repetitive sequences. A few contigs were distributed in the area with 0 – 20% GC content and 20× – 80× depth,. The GC content of this plant genome clearly differed from that of other common plant genomes, but was consistent with that of a bacterial genome. We speculated that this may have resulted from contamination, possibly by an endophyte.

To verify that the sample was contaminated, all contigs were blasted against the nucleic acid database at the NCBI. The blast results showed that majority of contigs with 10%–60% GC content and 20× – 120× depth aligned with homologous species. However, the contigs with 0–20× depth, 20% – 50% and 60% – 80% GC content aligned with bacteria in the *Methylobacterium* genus (Table 3). Therefore, we concluded that the sample was contaminated with a bacterium in this genus.

**Frequency Distribution of Different SSR Marker Types**

The assembled scaffolds were searched for SSRs and 71,291 SSRs were identified. The number of primer SSRs was 59,259 and the proportion of primer SSRs out of total SSRs was 83.12%. Among the SSR repeats, dinucleotide repeats (44.00%) and trinucleotide repeats (44.45%) were the main types of motifs, accounting for more than three-quarters of the total SSRs. Less commonly found repeats were tetranucleotide repeats (8.10%), pentanucleotide repeats (2.47%), and hexanucleotide repeats (0.99%) (Figure 5a). Among the dinucleotide repeats, AT/AT accounted for 67.38%, AG/CT for 24.04%, and CG/GC for only 0.06% (Figure 5b). Among the trinucleotide repeat motifs, AAT/ATT and AAG/CTT accounted for 43.80% and 24.23% respectively, and CCG/CGG accounted for only 1.05% (Figure 5c). Comprehensive analyses of the characteristics of all SSR repeat types revealed that the SSR loci in the rambutan genome were rich in A and T, with a strong base preference. Figure 6 shows the SSR motifs categorized by their unit sizes and the number of repeats. Dinucleotide and trinucleotide repeats were much more abundant than the other four types. The variations in SSR repeat numbers decreased with increasing motif length.

Discussion

Reliable data for genome size is critical for many fields of research, including taxonomy and evolution. This information is also essential for gene cloning and genome sequencing projects. Flow cytometry has been used as a standard method to predict genome size. However, recently, next generation sequencing-based K-mer methods have been successfully and widely used to estimate genome size. In this study, the genome size of rambutan (*N. lappaceum* cv ‘BR7’) was estimated to be 386.95 Mb by a K-mer analysis. Taking differences among cultivars into account, this estimate is generally consistent with those determined by flow cytometry for *N. lappaceum* cv. ‘Gula Gatu’ (343 Mb) and *N. lappaceum* cv. ‘Rongrien’ (333 Mb) (VanBuren *et al.* 2011). The genome size of rambutan is similar to that of pulasan (*N. ramboutan-ake*) (356 Mb), which is another tropical fruit tree and a species that is closely related to rambutan in the genus *Nephelium*. However, the size of the rambutan genome is substantially smaller than those of other *Sapindaceae* species such as lychee (554 Mb), longan (471.88 Mb) (Lin *et al.* 2017), longli (450 ~ 678 Mb) (VanBuren *et al.* 2011), and yellowhorn (439.97 Mb) (Liang *et al.* 2019), as well as *Acer yangbiense* (640 Mb) (Yang *et al.* 2019). The difference in genome size among these species indicates significant divergence. Compared with other sequenced tropical fruit tree genomes, the rambutan genome is smaller than those of banana (523 Mb) (D'Hont *et al.* 2012), kiwifruit (758 Mb) (Huang *et al.* 2013), pineapple (524.07 Mb) (Redwan *et al.* 2016), coconut (2150-2420 Mb) (Xiao *et al.* 2017; Lantican *et al.* 2019), durian (738 Mb) (The *et al.* 2017), and similar to those of papaya (372 Mb) (Ming et al. 2008), sweet orange (367 Mb) (Xu *et al.* 2013), and mulberry(357 Mb) (He *et al.* 2013). In general, a small genome size implies that genetic analyses and whole genome sequencing will be relatively simple and inexpensive.

However, the genome of *N. lappaceum* L. is complex with an estimated heterozygosity rate of 0.91% and 45.72% repetitive sequences. High heterozygosity in fruit trees may be due to self-incompatibility, artificial grafting, or asexual reproduction. The estimated heterozygosity rate in the whole genome of *N. lappaceum* (0.91%) is higher than those of kiwifruit (0.536%) (Huang *et al.* 2013), papaya (0.06%) (Ming *et al.* 2008), and *Acer yangbiense* (Yang *et al.* 2019), lower than that of durian (1.14%) (Lantican *et al.* 2019), and similar to that of longan (0.88%) (Lin *et al.* 2017). Analyzing the GC content not only evaluates the randomization of genome sequencing but also provides important clues about gene density, gene expression, replication time, recombination, and evolutionary relationships (Du *et al.* 2010). Analyses of the genome survey data revealed that the GC content in the rambutan genome is 33.52%. While analyzing the GC-depth distribution, we detected a *Methylobacterium* endophyte contaminant in the rambutan sample. *Methylobacterium* spp. are widely distributed in the natural environment, mainly in soil, leaf surfaces, and other plant tissues. During genome-wide sequencing, we can remove contigs derived from bacteria from total contigs with the homologous alignment method. Another approach would be to use rambutan samples obtained from aseptic tissue culture, which would fundamentally avoid bacterial contamination, improve sequencing accuracy, and reduce sequencing costs. Based on the survey of the rambutan genome, we suggest to use a combination of Illumina and PacBio sequencing technology with Hi-C technology and resequencing technology for subsequent large-scale sequencing. The sequencing depth needs to be increased appropriately, to >30 Gb and about 70x depth.

Genomic SSR markers are reliable, highly polymorphic, often multi-allelic, and easy to amplify. Consequently, they have been widely used to analyze genetic diversity and to construct genetic maps. However, the lack of genomic resources available for rambutan has restricted the use of microsatellite markers. To date, no genome-wide SSR markers have been developed and published for rambutan. In this study, 71,291 SSR loci without mono-nucleotide repeats were identified from the genome survey data. Among the di-nucleotide repeat motifs, AT/AT accounted for 67.38% and was the most abundant type, followed by AG/CT, accounting for 24.04%. This was consistent with the result that AAT/ATT was the most abundant type of tri-nucleotide motifs, and A/T-rich motifs were abundant in the rambutan genome. The abundance of A/T motifs in the genome has also been reported for grape (Jaillon *et al.* 2007) and mulberry (He *et al.* 2013).

Conclusion

This study contributes the following new information: (1) A genome survey was performed using the Illumina platform, and a K-mer analysis indicated that rambutan has a genome of about 386.95 Mb in size with a heterozygosity rate of 0.91%. (2) From the initial assembled rambutan genome, 5,636 SSRs with various types of motif repeats were detected, and 4,749 genomic SSR markers were identified. (3) Fifty novel SSR markers were randomly selected and validated by screening representative cultivars. The SSR-based phylogenetic analysis indicated that 69 rambutan accessions could be divided into five groups. This new genomic information will be useful for designing strategies for whole genome sequencing, and the new SSR markers will facilitate studies on genetics and marker-assisted breeding in rambutan.

**Acknowledgments**

We thank Jennifer Smith, PhD, from Liwen Bianji, Edanz Group China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

**Author’s contributions**

All authors made substantial contributions to the conception, design, and drafting of this work as individual experts in their fields. In particular,L.X.E. carried out the experimental work, statistically analyzed the data, and wrote the manuscript. L.X.E. and Z.Z.X. designed the experiment and edited the manuscript. G.H.M., G.Y., D.J.L., and M.J.H. collected samples and assisted with data entry.

References

Sukmandari, NS; GK Dash, WHW Jusof, M Hanafi (2017). A review on *Nephelium lappaceum* L. *Res J Pharm Technol* 10: 2819-2827 <http://rjptonline.org/HTMLPaper.aspx?Journal=Research+Journal+of+Pharmacy+and+Technology;PID=2017-10-8-68>

Gusman J.A., PJ Tsai (2015). Extraction of antioxidant compounds from rambutan (*Nephelium lappaceum* L.) peel as agricultural waste in Taiwan. *J Trop Crop Sci* 2: 10-16 <https://doi.org/10.29244/jtcs.2.2.10-16>

Palanisamy U, CH Ming, T Masilamani, T Subramaniam Thavamanithevi, LL Teng, AK Radhakrishnan (2008). Rind of the rambutan, *Nephelium lappaceum* L., a potential source of natural antioxidants. *Food Chem* 109: 54-63 <https://doi.org/10.1016/j.foodchem.2007.12.018>

Nont T, T Aphiwat, R Nuansri (2008). Antioxidant and antibacterial activities of *Nephelium lappaceum* L. extracts. *Food Sci Tech* 41: 2029-2035 <http://www.thaiscience.info/Article%20for%20ThaiScience/Article/5/10016424.pdf>

Palanisamy U, LT Ling, T Manaharan, D Appleton (2011). Rapid isolation of geraniin from *Nephelium lappaceum* rind waste and its anti-hyperglycemic activity. *Food Chem J* 127, 21-27. <https://doi.org/10.1016/j.foodchem.2010.12.070>

Ampthill B (2010). Sapindaceous fruits: botany and horticulture. *Hortic Rev* 51: 143 <https://a.glgoo.top/books?hl=zh-CN&lr=&id=fFj92fyIIhkC&oi=fnd&pg=PA143&dq=Sapindaceae+fruits:+botany+and+horticulture&ots=duV3z3m7vK&sig=GAGqi-iqQYqMLxsVgFVAgY_DrX4#v=onepage&q&f=false>

Wei X, LH Wang, YX Zhang, XQ Qi, XL Wang, X Ding, J Zhang, XR Zhang (2014). Development of simple sequence repeat (SSR) markers of sesame (*Sesamum indicum*) from a genome survey. *Molecules*, 19: 5150-5162 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6270694/>

Hou SY, ZX Sun, BL Hu, DM Xu, B Wu, B Zhang, XC Wang, YH Han, LJ Zhang, ZJ Qiao, et al (2016**)**. Genetic diversity of buckwheat cultivars (*Fagopyrum tartaricum* Gaertn.) assessed with SSR markers developed from genome survey sequences. *Plant Mol Biol Rep* 34: 233-241 [https://link.springer.com/article/10.1007%2Fs11105-015-0907-5](https://link.springer.com/article/10.1007/s11105-015-0907-5)

An J, M Yin, Q Zhang, D Gong, X Jia, Y Guan, J Hu (2017). Genome survey sequencing of *Luffa Cylindrica* L. and microsatellite high resolution melting (SSR-HRM) analysis for genetic relationship of luffa genotypes. *Inter J Mol Sci* 18: 1942  <https://doi.org/10.3390/ijms18091942>

Porebski S, LG Bailey, BR Baum (1997). Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol Biol Rep* 15: 8–15 [https://xs.scihub.ltd/https://doi.org/10.1007/BF02772108](https://xs.scihub.ltd/https%3A//doi.org/10.1007/BF02772108)

Li RQ, W Fan, G Tian, HM Zhu, L He, J Cai, QF Huang, QL Cai, YQ Bai, et al (2010). The sequence and de novo assembly of the giant panda genome. *Nature* 463: 311–317 <https://www.nature.xilesou.top/articles/nature08696>

VanBuren R, JG Li, F Zee, JH Zhu, CM Liu, AK Arumuganathan, R Ming (2011). Longli is not a hybrid of longan and lychee as revealed by genome size analysis and trichome morphology. *Tropical Plant Biol* 4: 228-236 <https://link.springer.com/article/10.1007/s12042-011-9084-3>

Lin YL, JM Min, RL Lai, ZY Wu, YK Chen, LL Yu, CZ Cheng, YC Jin, QL Tian, QF Liu, et al (2017). Genome-wide sequencing of longan (*Dimocarpus longan* Lour.) provides insights into molecular basis of its polyphenol-rich characteristics. *Gigascience* 6: 1-14 [http://oar.icrisat.org/10118/1/Genome-wide%20sequencing%20of%20longan%20(Dimocarpus.pdf](http://oar.icrisat.org/10118/1/Genome-wide%20sequencing%20of%20longan%20%28Dimocarpus.pdf)

Liang Q, HY Li, SK Li, FL Yuan, JF Sun, QC Duan, QY Li, R Zhang, YL Sang, N Wang, XW Hou, KQ Yang, JN Liu, L Yang, X Hou (2019). The genome assembly and annotation of yellowhorn (*Xanthoceras sorbifolium* Bunge). *GigaScience* 8: giz071 <https://doi.org/10.1093/gigascience/giz071>

Yang J, HM Wariss, L Tao, RG Zhang, QZ Yun, P Hollingsworth, ZL Dao, GF Luo, HJ Guo, YP Ma, WB Sun (2019). De novo genome assembly of the endangered *Acer yangbiense*, a plant species with extremely small populations endemic to Yunnan Province, China. *GigaScience* 8: giz085 <https://doi.org/10.1093/gigascience/giz103>

D'Hont A, F Denoeud, JM Aury, F-C Baurens, F Carreel, O Garsmeur, B Noel, S Bocs, G Droc, M Rouard, et al (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 488: 213-217 <https://www.nature.com/articles/nature11241.pdf>

Huang S, J Ding, D Deng, J Ding, DJ Deng, W Tang, HH Sun, DY Liu, L Zhang, XL Niu (2013). Draft genome of the kiwifruit Actinidia chinensis. *Nat Commun* 4: 2640 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4089393/>

Redwan RM, A Saidin, SV Kumar (2016). The draft genome of MD-2 pineapple using hybrid error correction of long reads. *DNA Research* 23: 427-439 <https://doi.org/10.1093/dnares/dsw026>

Xiao Y, PW Xu, HK Fan, L Baudouin, W Xia, S Bocs, JY Xu, Q Li, AP Guo, LX Zhou, et al (2017). The genome draft of coconut (*Cocos nucifera*). GigaScience 6: 1-11 <https://doi.org/10.1093/gigascience/gix095>

Lantican DV, SR Strickler, AO Canama, RR Gardoce, LA Mueller, HF Galvez (2019). De novo genome sequence assembly of dwarf coconut (*Cocos nucifera* L. ‘Catigan Green Dwarf’) provides insights into genomic variation between Coconut types and related palmspecies. *G3: Genes, Genomes, Genetics* 9: 2377-2393 <https://doi.org/10.1534/g3.119.400215>

Teh BT, K Lim, CH Yong, CCY Ng, SR Rao, V Rajasegaran, WK Lim, CK Qng, K Chan, VKY Cheng, et al (2017). The draft genome of tropical fruit durian (*Durio zibethinus*). *Nature* *Genetics* 49: 1633-1641 [https://xs.scihub.ltd/https://doi.org/10.1038/ng.3972](https://xs.scihub.ltd/https%3A//doi.org/10.1038/ng.3972)

Ming R, SB Hou, Y Feng, Q Yu, A Dionne-Laporte, JH Saw, P Senin, W Wang, BV Ly, KL Lewis, et al (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452: 991-996 <https://www.nature.com/articles/nature06856>

Xu Q, LL Chen, XA Ruan, DJ Chen, AD Zhu, CL Chen, D Bertrand, WB Jiao, BH Hao, MP Lyon, et al (2013). The draft genome of sweet orange (*Citrus sinensis*). *Nat Genet* *45*: 59–66 [https://xs.scihub.ltd/https://doi.org/10.1038/ng.2472](https://xs.scihub.ltd/https%3A//doi.org/10.1038/ng.2472)

He NJ, C Zhang, XW Qi, SC Zhao, Y Tao, GJ Yang, TH Lee, XY Wang, QL Cai, D Li, et al (2013). Draft genome sequence of the mulberry tree *Morus notabilis*. *Nat Commun* 4: 2445 <https://www.nature.com/articles/ncomms3445>

Du HL, HF Hu, YH Meng, WH Zheng, F Ling, JF Wang, XQ Zhang, QH Nie, XN Wang (2010). The correlation coefficient of GC content of the genome-wide genes is positively correlated with animal evolutionary relationships. *FEBS Lett* *584*: 3990-3994 <https://doi.org/10.1016/j.febslet.2010.08.003>

Jaillon O, JM Aury, B Noel, A Policriti, C Clepet, A Casagrande, N Choisne, S Aubourg, N Vitulo, C Jubin, et al (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449: 463-467. <https://doi.org/10.1038/nature06148>

**Figure Captions**

****

**Fig. 1:** Characteristics of rambutan fruit.

****

**Fig. 2:** Analysis of K-mer (K=17) distribution based on whole genome data for *Nephelium lappaceum* L. X-axis is depth (X) and the Y-axis is proportion that represents the frequency at that depth.



B

A

**Fig. 3:** Distribution of contigs. A: Contig coverage depth and length; B: contig coverage depth and number.



**Fig. 4:** GC content and sequencing depth. X-axis represents GC content; y-axis represents average depth. Each dot represents a contig. Right: distribution of sequencing depth; top: distribution of GC content. Red part represents high density of dots in scatter plot.

****

c

b

a

**Fig. 5:** Distribution of 71,291 SSRs in rambutan genome based on repeat type. a, Frequency of different SSR repeat types; b, frequency of different dinucleotide SSR motifs; c, frequency of different trinucleotide SSR motifs.

****

Fig. 6: Distribution and frequency of SSR repeats of various lengths. X-axis, SSR repeat numbers; y-axis, frequency of SSR types.

**Table Captions**

Table 1: Sequencing data statistics

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Library | Insert size | Raw Reads | Raw Base (bp) | Effective Rate (%) | Clean Base (bp) | Error Rate (%) | Q20 (%) | Q30 (%) | GC Content (%) |
| DES02312-S  | 350 | 180,235,212 | 54,070,563,600 | 99.99 | 54,065,451,600 | 0.03 | 95.90 | 90.70 | 33.52 |

Table 2: Statistics of assembled genome sequence for *Nephelium lappaceum* L.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Assembly | Total length (bp) | Total number | Max length (bp) | N50 length (bp) | N90 length (bp) |
| Contig | 356,250,469 | 987,898  | 68,350 | 923 | 126 |
| Scaffold | 360,534,817 | 723,955 | 142,353 | 1,929 | 147 |

Table 3: Species showing homology to contigs with high GC contents

|  |  |
| --- | --- |
| Species | Contig number |
| *Methylobacterium radiotolerans* | 2406 |
| *Methylobacterium extorquens* | 1384 |
| *Methylobacterium nodulans* | 258 |
| *Methylobacterium populi* | 112 |
| *Starkeya novella* | 55 |
| *Sporisorium reilianum* | 31 |