forests

MDPI

Article

Genome Survey Sequencing of *Betula platyphylla*

Sui Wang [1,2], Su Chen [1], Caixia Liu [1], Yi Liu [1], Xiyang Zhao [1], Chuanping Yang [1,*] and Guan-Zheng Qu [1,*]

[1] State Key Laboratory of Tree Genetics and Breeding, Northeast Forestry University, 26 Hexing Road, Harbin 150040, China; wangsui.ws@163.com (S.W.); chensunefu@163.com (S.C.); liuyi1992521@163.com (Y.L.); zhaoxyphd@163.com (X.Z.); yangchuanping_ycp@163.com (C.Y.); quguanzheng@yahoo.com (G.Q.)

[2] Key Laboratory of Soybean Biology in Chinese Ministry of Education, Northeast Agricultural University, 600 Changjiang Street, Harbin 150030, China

* Correspondence: quguanzheng@yahoo.com (G.-Z.Q.); yangchuanping_ycp@163.com (C.Y.); Tel.: +86-0451-8219-2695

Received: 13 August 2019; Accepted: 17 September 2019; Published: 20 September 2019

Abstract: Research Highlights: A rigorous genome survey helped us to estimate the genomic characteristics, remove the DNA contamination, and determine the sequencing scheme of *Betula platyphylla*. Background and Objectives: *B. platyphylla* is a common tree species in northern China that has high economic and medicinal value. However, there is a lack of complete genomic information for this species, which severely constrains the progress of relevant research. The objective of this study was to survey the genome of *B. platyphylla* and determine the large-scale sequencing scheme of this species. Materials and Methods: Next-generation sequencing was used to survey the genome. The genome size, heterozygosity rate, and repetitive sequences were estimated by k-mer analysis. After preliminary genome assembly, sequence contamination was identified and filtered by sequence alignment. Finally, we obtained sterilized plantlets of *B. platyphylla* by plant tissue culture, which can be used for third-generation sequencing. Results: We estimated the genome size to be 432.9 Mb and the heterozygosity rate to be 1.22%, with repetitive sequences accounting for 62.2%. Bacterial contamination was observed in the leaves taken from the field, and most of the contaminants may be from the genus *Mycobacterium*. A total of 249,784 simple sequence repeat (SSR) loci were also identified in the *B. platyphylla* genome. Among the SSRs, only 11,326 can be used as candidates to distinguish the three *Betula* species. Conclusions: The *B. platyphylla* genome is complex and highly heterozygous and repetitive. Higher-depth third-generation sequencing may yield better assembly results. Sterilized plantlets can be used for sequencing to avoid contamination.

Keywords: *Betula platyphylla*; genome survey sequencing; sequence contamination; SSR

1. Introduction

Betula platyphylla, commonly called white birch or Asian white birch, is a broadleaved deciduous hardwood tree species belonging to the genus *Betula* in the family Betulaceae (Figure 1). This tree can be found in temperate or subarctic places of Asia [1]. This tree species is best grown in medium to wet, well-drained, sandy, or rocky loams in full sun to part shade but it is also tolerant to salt, cold, light, drought and flooding, adapts well to numerous types of soil, and can endure harsh conditions [2]. The most typical characteristic of *B. platyphylla* is papery hoary bark and a long, horizontal lenticel [3]. *B. platyphylla* diploid cells have 28 chromosomes (the somatic number, 2n) [4]. This tree species is monoecious, but the germination rate of its self-bred progeny is often notably low. In northern China, *B. platyphylla* is an important afforestation and economic tree species. Recent studies have indicated that this tree species also has high medicinal value. The bark of this tree contains numerous

triterpenoids, such as betulin, betulinic acid, lupeol, oleanolic acid, and ursolic acid [5,6]. These triterpenoids have antibacterial, antimalarial, anti-inflammatory, anthelmintic, and antioxidant properties. These compounds can also serve as inhibitors with strong potential for preventing and treating HIV and cancers [7–9].

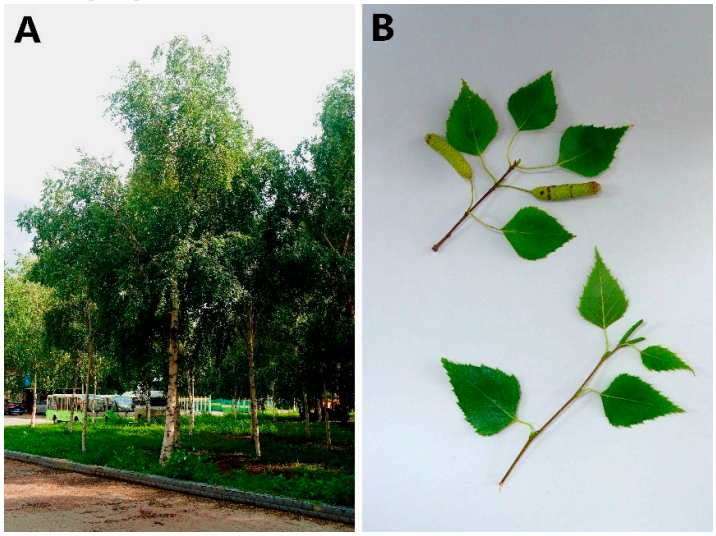

Figure 1. Adult tree (**A**), leaves and inflorescences (**B**) of *Betula platyphylla* on the Northeast Forestry University campus (45°43′21″N, 126°38′30″E).

The State Key Laboratory of Tree Genetics and Breeding has been engaged in *B. platyphylla* research for more than 30 years. A series of studies has been carried out on *B. platyphylla*, involving growth and development, environmental stress, and metabolic networks, for example. We also selected some plus trees (the best individuals in a group) and superior families (the best families in a larger group) through conventional breeding [6,10,11]. Despite the onset of today's genome era, there remains a lack of complete genomic information for *B. platyphylla*, which severely constrains the progress of relevant research. We plan in the future to sequence the whole genome of *B. platyphylla*, which will lay a foundation for molecular and breeding research in this tree species. To understand the genome information and determine the sequencing scheme, before large-scale sequencing, we conducted a genome survey of *B. platyphylla* using next-generation sequencing technology.

2. Materials and Methods

2.1. Plant Materials and Sequencing

Tender leaves were collected from an adult *B. platyphylla* plus tree located on the Northeast Forestry University campus (45°43′21″N, 126°38′30″E). After cleaning and disinfection with 70% alcohol, leaves were stored in liquid nitrogen. Total genomic DNA was extracted from tender leaves using the CTAB (Cetyl Trimethyl Ammonium Bromide) method [12]. The DNA were electrophoresed on a 1% agarose gel to assess quality and then sent to BGI (Beijing Genomics Institute, Shenzhen, Guangdong, China) in dry ice. After using the NanoDropTM 2000 Spectrophotometer (Thermo Scientific, Waltham, MA, USA) to determine the 260/280 and 260/230 ratios of the DNA, three paired-end (insert sizes = 200 , 500, and 800 bp) Illumina libraries were prepared and sequenced on the HiSeq 2000 platform (Illumina, San Diego, CA, USA) with the paired-end 100 bp.

2.2. Data Cleaning

To obtain high-quality and vector/adaptor-free reads, raw paired-end reads were filtered using the NGSQC Toolkit v 2.3.3 (cut-off read length for HQ = 70%, cut-off quality score = 20, trim reads from 5′ = 3, trim reads from 3′ = 7) [13]. Then, FastQC (v 0.11.5) was used to check the quality of the clean reads.

2.3. Genome Size Estimation by k-mer Analysis

The Genomic Character Estimator (GCE, v 1.0) was used to estimate the genome size, repeat structure, and heterozygous rate of *B. platyphylla* [14]. The program kmer_freq_hash was used to count *k*-mer frequency. We used all the clean reads to survey the genome and set $k = 17$ in *k*-mer analysis. The *k*-mer distribution was drawn using Excel 2016.

2.4. Preliminary Genome Assembly and GC Content Analysis

First, we utilized the Edena assembler [15] with default parameters to preliminarily assemble all the clean reads into contigs. Next, a Perl script was written to calculate the GC content of each contig. Then, we extracted the sequencing depth of each contig from the Edena result. Finally, the densCols function in R was used to plot the GC depth distribution of the contigs.

2.5. Identification and Filtration of Genomic Pollutants

We extracted the contigs with a GC content not less than 60% from the GC depth distribution. All the selected contigs were blasted against the NCBI (National Center for Biotechnology Information) NT (Nucleotide collection) database (updated on 2018/12/4). We set max_taret_seqs=5, and all the other parameters were set to default. A Perl script written by us was used to calculate the hit species distribution. We extracted all the microorganism sequences from the result and built a pollution database. All the contigs assembled by Edena were blasted against the pollution database. The blast cut-off was set to max_target_seqs=1, evalue≤1e-5, and the query coverage ≥50%. After removing contamination from the contigs, we plotted the GC depth distribution again to confirm that most of the pollution was removed.

2.6. Scaffold Construction, Gap Closing, and Sequence Polishing

We further extended and scaffolded preassembled contigs. Contigs longer than 1000 bp were selected. All the clean paired-end reads were mapped to the filtered contigs using bowtie2 (v 2.3.5) [16]. After removing duplication, all SAM (Sequence Alignment/Map) files were converted to TAB (a tab-delimited file containing information about the positions of the reads on the contigs) files. According to the information in the TAB files, we used SSPACE (v 3.0) [17] to assess the order, distance and orientation of contigs and combine them into scaffolds. Then, GapCloser (v 1.12) was used to close most of the gaps. Finally, Pilon [18] was used to polish the GapCloser results and obtain relatively accurate assembly sequences. The quality of the completeness of the assembly was assessed using the BUSCO (Benchmarking Universal Single-Copy Orthologs) v3 [19] method based on a benchmark of 1,440 conserved plant genes.

2.7. SSR Identification and Analysis

Simple sequence repeats (SSRs) were detected using the Perl script MISA (MIcroSAtellite identification tool) by setting the minimum number of repeats to 10, 5, 4, 3, 3, and 3 for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides, respectively. We also downloaded two other published birch genomes, *B. pendula* (https://genomevolution.org/CoGe/GenomeInfo.pl?gid=35079) and *B. nana* (http://birchgenome.org/data) [20,21]. CandiSSR [22] was used to identify polymorphic SSRs (PolySSRs) and to automatically design primer pairs for each identified PolySSR in the three *Betula* species.

2.8. Acquisition and Culture of Tissue Culture Seedlings of B. platyphylla

In April 2016, we selected some tender leaves from a healthy *B. platyphylla* plus tree in the experimental base of the State Key Laboratory of Tree Genetics and Breeding (Northeast Forestry University). After disinfection with 70% alcohol and 2% sodium hypochlorite solution, leaves were cut into 1 × 1 cm pieces and placed on dedifferentiation medium. When the callus was produced, we transferred it to the redifferentiation medium and waited for bud formation. When the height of

seedlings was approximately 1 cm, they were transferred to rooting medium. All medium formulations are shown in Table 1. The environmental conditions of tissue culture were a temperature of 24 °C, light intensity of 3000 Lux, 16 h photoperiod, and a relative humidity of 60%.

Table 1. Formula of tissue culture medium for *B. platyphylla*.

Medium	Basic Medium	Plant Hormone (mg/L)	Sucrose (g/L)	Agar (g/L)	pH	Others (g/L)
Dedifferentiation medium	WPM	NAA 0.02 6-BA 1.0	30	6.0	5.8~6.0	--
Redifferentiation medium	WPM	NAA 0.02 6-BA 1.0 GA3 0.5	30	6.0	5.8~6.0	--
Rooting medium	1/2 MS	IBA 0.4 NAA 0.02	20	6.5	5.8~6.0	Activated carbon 2.0

WPM: Woody plant medium.

MS: Murashige & Skoog Basal Medium.

2.9. Sequencing and GC Content Analysis of Tissue Culture Plantlets

Several *B. platyphylla* leaves from tissue culture seedlings were taken, and DNA was extracted according to the method described above. A small insert size library of 300 bp was constructed and then sequenced on the HiSeq X platform (Illumina, San Diego, CA, USA) with the paired-end 150 bp. A total of 18.8 Gb of data was obtained. After data cleaning, the Edena assembler was also used to build contigs. We filtered out contigs less than 1500 bp in length and redrew the GC depth distribution. At the same time, in order to further confirm whether there was contamination in the sequence, all the filtered contigs were blasted against the NCBI NT database with max_target_seqs=1. Finally, we calculated the hit species distribution again.

3. Results

3.1. Sequencing and Data Cleaning

Through quality inspection, the extracted DNA was qualified and we used it to build libraries. After removing adapters and primers, approximately 29.3 Gb of data were obtained. Because of the poor quality of the front and tail of reads, we trimmed and filtered them appropriately. After data cleaning, we obtained approximately 23.3 Gb of clean reads. The filtration rate was 20.5%. Table 2 shows the basic statistics for the sequencing data.

Table 2. Basic statistics for the sequencing reads.

Library	200 bp	500 bp	800 bp	300 bp*
Raw reads length (bp)	100	100	100	150
Raw data (Gb)	14.5	12.0	5.0	23.5
Raw Q20 (%)	96.9	94.4	86.5	94.8
Clean reads length	90	90	90	135
Clean data (Gb)	12.3	9.6	3.2	18.4
Clean Q20 (%)	99.1	98.6	97.4	97.7

* The last column (300 bp library) is the results of the following sequencing of tissue culture plantlet.

3.2. Genome Size Estimation by k-mer Analysis

Based on published genomes of *Betula* genus plants and *C*-values from the Plant DNA *C*-values Database (http://data.kew.org/cvalues/, C Mean = 0.63), we estimated that there is little possibility that the genome size of *B. platyphylla* is larger than 1 Gb. Therefore, we set $k = 17$ for *k*-mer analysis. We detected 20,620,036,537 17-mer sequences out of a total of 501,475,979 different 17-mers. Figure 2 shows the 17-mer frequency distribution; two peaks were available. The peak at ~50× depth was

recognized as a main peak and the peak at half depth (~25×) of the main peak was recognized as a heterozygous peak. On the right side of the main peak, there was also a distinct trail at approximately 100× depth, resulting from repeat sequences. According to the 17-mer curve, we preliminarily estimated that *B. platyphylla* has a highly heterozygous diploid genome. After running the GCE program in hybrid mode, the main peak was calculated at 46× depth and the heterozygote at 23× depth. It meant that the average sequencing depth of the genome was 46×. As a result, we estimated the genome size to be 432.9 Mb and the heterozygosity rate to be 1.22%; the repetitive sequences accounted for 62.2% of the whole genome.

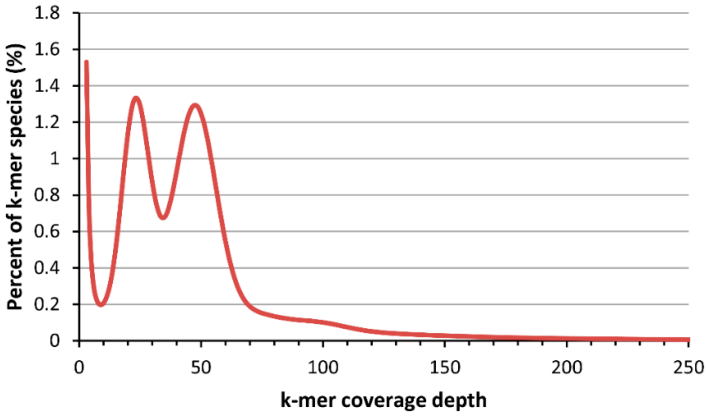

Figure 2. Distribution of 17-mer frequency of *B. platyphylla*.

3.3. Preliminary Genome Assembly and GC Content Analysis

With the help of Edena, 1,434,832 contigs were preliminarily assembled, totaling 649 Mb. The N50 contig was 704 bp, and the longest contig was 39,604 bp. The overall GC content of the whole contigs was 35.5%. We filtered contigs less than 1000 bp in length and finally obtained 246 Mb of assembled sequences consisting of 120,934 contigs. Next, we analyzed the GC content and sequencing depth distribution of these contigs. Because the average coverage of each contig was given in the Edena assembly results, we only needed to calculate the GC content of each contig. As shown in Figure 3, most of the contigs were concentrated in the 20%–50% GC content and 20×–50× average depth area. In this area, contigs had two gravity centers, which were located at an average depth of approximately 25× and 50×, respectively. Combined with the *k*-mer analysis, we inferred that the two gravity centers correspond to the heterozygosity and main peak, respectively. In areas over 80× depth, there were also a small number of contigs, which may be repetitive sequences in the genome. It should be noted that in the lower right corner of the figure, there was also an abnormal contigs distribution area. This area was obviously a stray from the group, its GC content is over 60% and its average depth is low. We speculated that it may be a result of bacterial contamination.

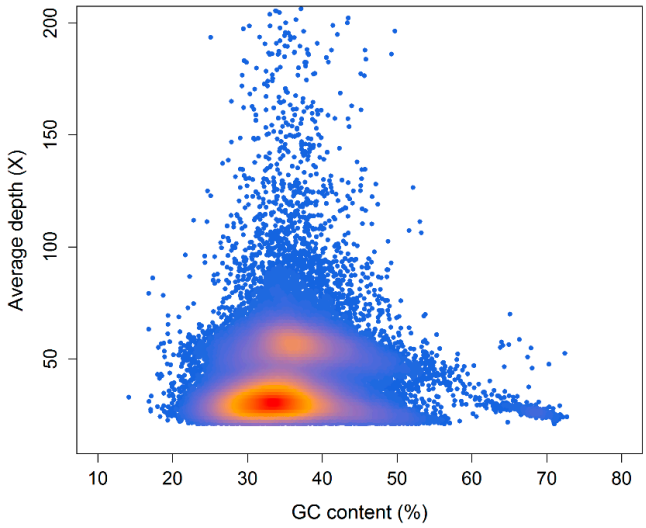

Figure 3. GC depth distribution of *B. platyphylla* genome (before filtration). The X-axis is the GC content and the Y-axis is the average depth. Each dot represents a contig. Color from red to orange and then to blue indicates dot density from high to low.

3.4. Identification and Filtration of Genomic Pollutants

To confirm our inference, contigs with a GC content over 60% were extracted. We obtained 332 contigs, totaling 5.6 Mb. The blast results showed that majority of these contigs aligned with bacteria, most of them belonging to the genus *Mycobacterium* (Table 3). Therefore, we speculated that they were fragments of bacterial genome. Of course, a small proportion of contigs hit some relatives of *B. platyphylla*, such as *Juglans regia*, *Cajanus cajan,* and *Theobroma caca*. These contigs may be fragments with a high GC content in the *B. platyphylla* genome. To remove as much contamination as possible, we selected all prokaryotic genomes from the blast results and built a pollution database. All contigs assembled by Edena were mapped to the database. After removing all contaminated contigs, we redrew a GC depth distribution figure (Figure 4). There was no obvious abnormal dot distribution area in the figure.

Table 3. Species distribution of contigs with high GC content.

Species Name	Hit Number	Species Name	Hit Number
Mycobacterium chimaera	229	*Cutibacterium avidum*	1
Mycobacterium intracellulare	187	*Mycobacterium gilvum*	1
Mycobacterium marseillense	186	*Mycobacterium vanbaalenii*	1
Mycobacterium yongonense	106	*Malus x*	1
Mycobacterium indicus	76	*Mycobacterium dioxanotrophicus*	1
Mycobacterium avium	26	*Kocuria palustris*	1
Mycobacterium colombiense	12	*Acidiphilium multivorum*	1
Mycobacterium abscessus	4	*Gluconobacter oxydans*	1
Juglans regia	4	*Mycobacterium ulcerans*	1
Mycobacterium marinum	4	*Mycobacterium sinense*	1
Roseomonas gilardii	2	*Theobroma cacao*	1
Mycobacterium thermoresistibile	2	*Mycobacterium lepraemurium*	1
Xanthobacter autotrophicus	1	*Prunus avium*	1
Cajanus cajan	1	*Paracoccus yeei*	1
Mycobacterium liflandii	1	*Glycine max*	1
Mycobacterium rhodesiae	1	*Prunus persica*	1
Acidiphilium cryptum	1	*Herrania umbratica*	1
Rhodococcus jostii	1	*Vitis vinifera*	1

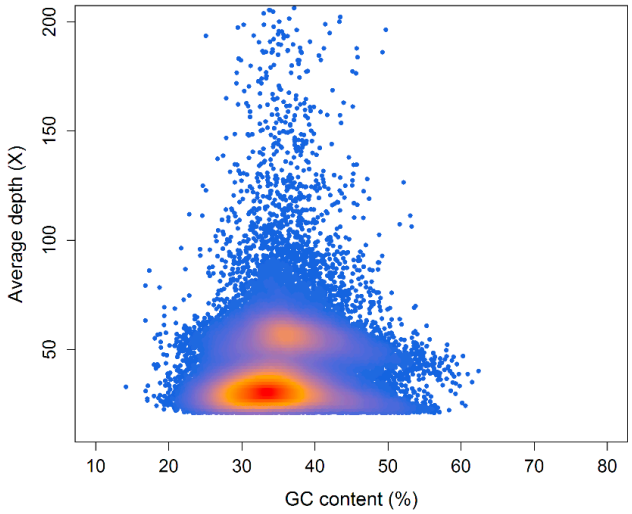

Figure 4. GC depth distribution of the *B. platyphylla* genome (after filtration).

3.5. Scaffold Construction, Gap Closure, and Sequence Polishing

After filtering, we selected contigs longer than 1000 bp and further improved the genome to the scaffold level. Through strict sequence alignment, the insert size of each library was confirmed. With the help of SSPACE, 120,575 contigs were connected to 79,580 scaffolds. After closing gaps and polishing sequences, 79,580 scaffolds were finally assembled, totaling 250.1 Mb. The N50 scaffold was 4312 bp and the longest scaffold was 85,240 bp (Table 4). The BUSCO results showed that only 25.1% of the region had complete gene coverage (including 1.9% duplicated ones), 10.8% were fragmented and 64.1% were missing.

Table 4. Statistics of assembled *B. platyphylla* genome sequences.

Item	Scaffold	Contig
No. of sequences	79,580	97,097
Total length (bp)	250,090,936	249,878,124
N50 length (bp)	4312	3081
N90 length (bp)	1336	1244
Max length (bp)	85,240	85,240
GC content (%)	34.8	34.8

3.6. SSR Identification and Analysis

Recently, increasingly numerous theoretical reasons and well-documented examples show that the repetitive sequence of genomic DNA is essential. This sequence not only affects the advanced structure of chromosomes but is also very useful in genome evolution and rearrangement [23,24]. In this study, we only focused on microsatellite sequences. In the *B. platyphylla* assembled scaffolds, a total of 249,784 SSRs were identified (Table S1). Among the SSRs, there were 192,575, 42,276, 11,755, 2337, 568, and 273 mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats, respectively. As expected, the mononucleotide repeat was the predominant type, accounting for 77.1% of all SSRs, followed by the di- (16.9%) and tri- (4.7%) types. In addition, we identified 158 SSR types in total. Figure 5 shows the top 20 frequency of SSR types. Except for mononucleotide and dinucleotide repeat types, some other repeat types, such as AAT/ATT, AAG/CTT, and AAAT/ATTT, also account for a high proportion.

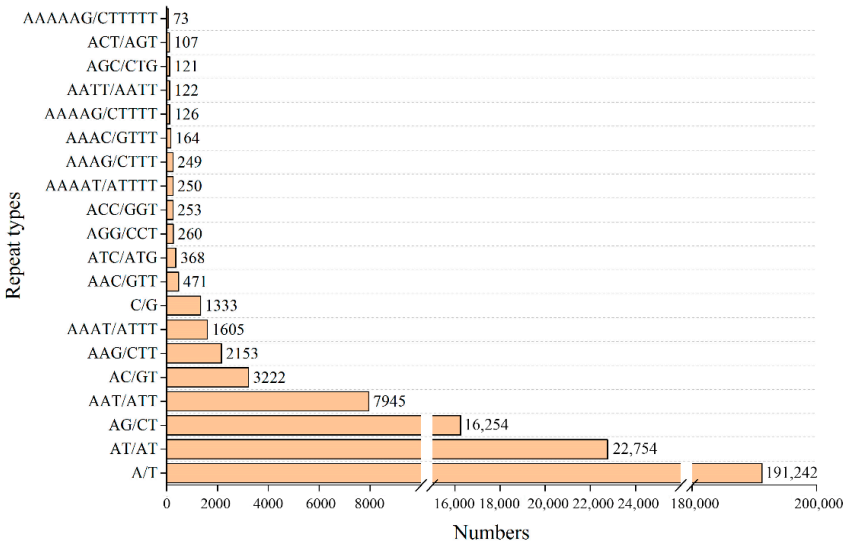

Figure 5. Top 20 frequency of *B. platyphylla* SSR types (considering sequence complementary).

Because SSRs are often useful in developing lineage-specific markers, we used CandiSSR to identify candidate polymorphic SSRs among *Betula* species. According to the genome sequence of *B. pendula*, *B. nana*, and *B. platyphylla*, 11,326 possible SSR loci were found among the three *Betula* species (Table S2), accounting for only 4.5% of all SSRs in *B. platyphylla*. With the help of Primer 3, 10,096 pairs of primers were designed (Table S3).

3.7. Sequencing and GC Content Analysis of Tissue Culture Plantlets

Through tissue culture, we obtained sterilized plantlets of *B. platyphylla*. After filtering, 14.4 Gb of clean reads were left. With the help of Edena, 840,231 contigs were assembled, totaling 454.3 Mb. Among the contigs, 79,852 were longer than 1000 bp (149.9 Mb). Figure 6 shows that the points in the GC depth distribution of *B. platyphylla* tissue culture plantlets are more concentrated than before. In addition, there is no obvious abnormal accumulation area, especially in the high GC content area. Table 5 further shows the annotation information for all filtered contigs aligned to the NCBI NT database (Top 20 species of hit number). All the species in Table 5 are plants, and most of them have a closer genetic relationship with *B. platyphylla*. The top species on the list is *Juglans regia*, which belongs to Fagales, as well as *B. platyphylla*. Its genome sequence was published in 2016 [25]. These results showed that there was no obvious bacterial contamination in the sequence.

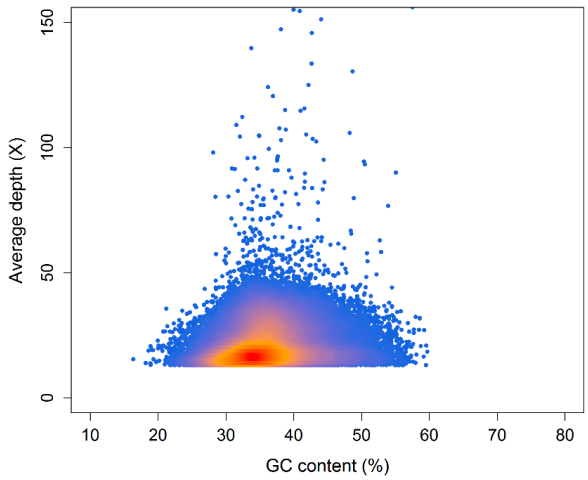

Figure 6. GC depth distribution of *B. platyphylla* tissue culture plantlets.

Table 5. Species distribution of contigs of *B. platyphylla* tissue culture plantlets.

Species name	Hit Number	Align Length (bp)	Species Name	Hit Number	Align Length (bp)
Juglans regia	8726	5,851,287	*Ziziphus jujuba*	89	51,320
Vitis vinifera	638	658,331	*Prunus persica*	89	70,721
Theobroma cacao	332	194,640	*Betula nana*	84	57,369
Betula pendula	263	667,040	*Malus x*	66	27,903
Corylus avellana	217	211,681	*Hevea brasiliensis*	65	36,133
Betula platyphylla	197	64,844	*Fragaria vesca*	55	26,960
Betula luminifera	191	98,488	*Prunus mume*	55	33,181
Quercus robur	162	110,642	*Pyrus x*	54	23,152
Prunus avium	141	77,758	*Glycine max*	54	29,659
Populus trichocarpa	135	83,113	*Citrus clementina*	43	29,869

4. Discussion

A genome survey is an important step before large-scale genome sequencing. It gives a preliminary understanding of the genomic characteristics of the species to be sequenced, including genome size, repeat structure, and heterozygous rate. Before carrying out the third-generation sequencing of the *B. platyphylla* genome, we surveyed the genome by next-generation sequencing. The high heterozygosity rate led to larger assembly than actual, which interfered with the coverage estimation, so we did not calculate sequencing coverage in this study. As expected, the quality of our pre-assembly genome was poor and only 25.1% of the region had complete gene coverage. A large number of genome regions were filtered or lost. Such genome sequence was clearly insufficient for many studies, but it was enough for genome survey and SSR identification. We determined that the genome size of *B. platyphylla* is approximately 430 Mb. This is not very large for a woody plant, but its heterozygosity rate is as high as 1.22%, and 62.2% of the sequence was repetitive, which represents the complex genome. For sequencing species with high heterozygosity, there were three main methods in the past. The first method is to reduce the heterozygosity rate by continuous self-crossing or inbreeding. This approach is very suitable for self-compatible species with short reproductive cycles, but more difficult for annual trees. The genome sequencing of *B. pendula*, a relative species of *B. platyphylla*, was accomplished by this strategy. After long-term breeding, the *B pendula* individual used for reference genome sequencing can be induced to flower within one year and has a lower heterozygosity rate [20]. This approach is extremely beneficial for sequencing, but unfortunately *B. platyphylla* lacks such varieties. The long breeding cycle and self-incompatibility make this method difficult to achieve. The second is to use the haploid breeding technique. This approach can fundamentally solve the impact of heterozygosity, but it is not helpful to repeat sequences. This method was adopted in the recently published rose genome sequencing [26,27]. These researchers used anther culture to generate a homozygote for sequencing and obtained one of the most complete plant genomes to date (contig N50 is as high as 24 Mb). However, the technology is also full of challenges and requires a long cycle for *B. platyphylla*. The third is an assembly strategy. A series of heterozygous genome assembly software, such as Hapsembler [28], HapCompass [29], Platanus [30], and dipSPAdes [31], was developed. These researchers solved some problems to some extent, but it was still far from enough. Currently, the third-generation sequencing and some new haplotype assembly tools make it possible [32–35]. With the decreasing price of third-generation sequencing, we decided to use it to sequence *B. platyphylla* and then assemble a haplotype genome.

Identification and filtration of genomic pollutants is another important part of the genome survey. Although field sampling inevitably involves contamination, it is more evident in the DNA of *B. platyphylla*. From the preliminary filtered assembly results, the contigs of pollutants account for only 0.3% (359 in 120,934) of the total number, but the length of pollutant contigs accounts for 2.4% (6.1 in 257.8 Mb) of the total assembled length. This finding indicates that the length of bacterial contigs is longer in the assembly results. In fact, 94 of the longest top 100 contigs of the preliminary filtered assembly genome are from bacteria. According to the annotation results, most of the pollutants in leaves belong to the genus *Mycobacterium*. Because *Mycobacterium* sp. are commonly found in soil, dust, and water [36], the bacterial contamination in the DNA may come from plant surfaces, but some studies have also shown that they may be endophytes in plants [37–39]. Some

evidence suggests that endophytes may have a positive effect on plants [40,41], but *Mycobacterium* sp. may cause growth retardation [42]. Based on the current evidence, we cannot determine the source of the pollutants, but we believe that if they are not removed, they will inevitably affect genome assembly, annotation, and subsequent analysis. Although most pollution sequences can be removed by alignment, this is clearly not the best solution. Bacterial DNA will take up some of the sequencing resources and affect assembly results. Cultivating sterilized plantlets of *B. platyphylla* for genome sequencing can fundamentally solve this problem and provide materials for downstream molecular biology experiments. The contamination in the sequencing reads must have an impact on the survey results, but we do not know how great that impact is. We attempted to remove the contaminated sequences from reads by sequence alignment and then repeat the genome survey. The filtered results predicted the genome size to be 429.5 Mb, the heterozygosity rate to be 1.17%, and the repetitive sequences as 62.0% of the whole genome. All the estimated results were not significantly different from before. Therefore, the contamination did not cause excessive interference to the genome survey of *B. platyphylla*. We estimated that this result may be due to the low proportion of the contaminating sequences. The removal of these contaminated sequences led to a decrease in total number of *k*-mer, especially in the vicinity of the heterozygous peak. As a result, the heterozygosity rate decreased a little more, while the size and repetition rate of the genome decreased slightly. But overall, the changes were not great.

Microsatellites, or simple sequence repeats (SSRs), are among the most informative and multipurpose molecular markers. In the past, SSRs recognition was not easy, and the specificity of PCR primers was poor. The advent of high-throughput sequencing has facilitated the development of SSRs across the genome, while being quick, efficient, and cost-effective, even in non-model plant populations with limited or no background genetic information [43]. Today, SSR identification has become a part of routine genome survey analysis [44–46]. The reads used for survey analysis were preliminarily assembled. Although these assembly results are generally poor, they are enough for most SSR analyses. From the results of *B. platyphylla* SSR analysis, it is highly similar to *B. alnoides*, another species of Betulaceae [47]. At present, genome-wide identification of SSR usually yields a large number of loci. However, clearly, not all SSRs can be used as markers. Traditional experimental screening of the SSR polymorphic status and their subsequent applicability to genetic studies is extremely labor-intensive and time-consuming. In this study, CandiSSR was used to screen the preliminary SSR loci based on the draft genome sequence of the species to be tested, and candidate polymorphic SSR loci were obtained. Obviously, the closer the species are, the fewer the candidate SSR loci are. Finally, only 4.5% of all SSRs in *B. platyphylla* were filtered, which demonstrates the importance of virtual filtering in SSR identification and analysis.

5. Conclusions

In this study, we surveyed the *B. platyphylla* genome to understand the genomic characteristics of this species. Its genome is relatively complex. We estimated the genome size to be 432.9 Mb and the heterozygosity rate to be 1.22%, with repetitive sequences accounting for 62.2%. Field sampling of the *B. platyphylla* individual involved bacterial contamination and we cultivated sterilized plantlets to solve this problem fundamentally. A total of 249,784 SSR loci were also identified in the *B. platyphylla* genome. Among the SSRs, only 11,326 can be used as candidates to distinguish the three *Betula* species. This study may provide a fundamental resource for future large-scale sequencing and molecular breeding studies of *B. platyphylla*.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1

Author Contributions: S.W. designed the experiments, analyzed the data and wrote the manuscript. C.L. and Y.L. prepared the sample. S.C., X.Z., C.Y. and G.Q. revised the manuscript. All authors approved the final manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (No.31770712). The funding bodies had no role in the design of the study, collection, analysis, or interpretation of data or in the writing of the manuscript.

Acknowledgments: We sincerely thank Xiaoying Na for her help with the tissue culture of *Betula platyphylla*.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, Z.X. *Dendrology*, 2nd ed.; China Forestry Publishing House: Beijing, China, 2008.
2. Mijiti, M.; Zhang, Y.M.; Zhang, C.R.; Wang, Y.C. Physiological and molecular responses of *Betula platyphylla* Suk to salt stress. *Trees* **2017**, *31*, 1653–1665.
3. MobileReference: *The Illustrated Encyclopedia of Trees and Shrubs: An Essential Guide to Trees and Shrubs of the World*; MobileReference: Boston, MA, USA, 2008.
4. Wei, Z.G.; Zhang, K.X.; Yang, C.P.; Liu, G.F.; Liu, G.J.; Lian, L.; Zhang, H.G. Genetic linkage maps of *Betula platyphylla* Suk. based on ISSR and AFLP markers. *Plant Mol. Biol. Report.* **2010**, *28*, 169.
5. Krasutsky, P.A. Birch bark research and development. *Nat. Prod. Rep.* **2006**, *23*, 919–942.
6. Wang, S.; Zhao, H.; Jiang, J.; Liu, G.; Yang, C. Analysis of three types of triterpenoids in tetraploid white birches (*Betula platyphylla* Suk.) and selection of plus trees. *J. For. Res.* **2015**, *26*, 623–633.
7. Yogeeswari, P.; Sriram, D. Betulinic acid and its derivatives: A review on their biological properties. *Curr. Med. Chem.* **2005**, *12*, 657–666.
8. Fu, J.Y.; Qian, L.B.; Zhu, L.G.; Liang, H.T.; Tan, Y.N.; Lu, H.T.; Lu, J.F.; Wang, H.P.; Xia, Q. Betulinic acid ameliorates endothelium-dependent relaxation in L-NAME-induced hypertensive rats by reducing oxidative stress. *Eur. J. Pharm. Sci.* **2011**, *44*, 385–391.
9. Ríos, J.L.; Manez, S. New pharmacological opportunities for betulinic acid. *Planta Med.* **2018**, *84*, 8–19.
10. Liang, D.Y.; Zhang, X.X.; Wang, C.; Wang, X.W.; Li, K.L.; Liu, G.F.; Zhao, X.Y.; Qu, G.Z. Evaluation of *Betula platyphylla* Families Based on Growth and Wood Property Traits. *For. Sci.* **2018**, *64*, 663–670.
11. Zhao, X.Y.; Bian, X.Y.; Liu, M.R.; Li, Z.X.; Li, Y.; Zheng, M.; Teng, W.H.; Jiang, J.; Liu, G.F. Analysis of genetic effects on a complete diallel cross test of *Betula platyphylla*. *Euphytica* **2014**, *200*, 221–229.
12. Porebski, S.; Bailey, L.G.; Baum, B.R. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Report.* **1997**, *15*, 8–15.
13. Patel, R.K.; Jain, M. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS ONE* **2012**, *7*, e30619.
14. Liu, B.H.; Shi, Y.J.; Yuan, J.Y.; Hu, X.S.; Zhang, H.; Li, N.; Li, Z.Y.; Chen, Y.X.; Mu, D.S.; Fan, W. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. *arXiv* **2013**, arXiv:1308.2012.
15. Hernandez, D.; François, P.; Farinelli, L.; Osterås, M.; Schrenzel, J. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Res.* **2008**, *18*, 802–809.
16. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357.
17. Boetzer, M.; Henkel, C.V.; Jansen, H.J.; Butler, D.; Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **2011**, *27*, 578–579.
18. Walker, B.J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C.A.; Zeng, Q.D.; Wortman, J.; Young, S.K.; et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **2014**, *9*, e112963.
19. Robert, M.W.; Mathieu, S.; Felipe, A.S.; Mose, M.; Panagiotis, I.; Guennadi, K.; Evgenia, V.K.; Evgeny, M.Z. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **2017**, *35*, 543–548.
20. Salojarvi, J.; Smolander, O.P.; Nieminen, K.; Rajaraman, S.; Safronov, O.; Safdari, P.; Lamminmaki, A.; Immanen, J.; Lan, T.Y.; Tanskanen, J.; et al. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nat. Genet.* **2017**, *49*, 904–912.
21. Wang, N.; Thomson, M.; Bodles, W.J.; Crawford, R.M.; Hunt, H.V.; Featherstone, A.W.; Pellicer, J.; Buggs, R.J. Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Mol. Ecol.* **2013**, *22*, 3098–3111.
22. Xia, E.X.; Yao, Q.Y.; Zhang, H.B.; Jiang, J.J.; Zhang, L.P.; Gao, L.Z. CandiSSR: An Efficient Pipeline used for Identifying Candidate Polymorphic SSRs Based on Multiple Assembled Sequences. *Front. Plant Sci.* **2016**, *6*, 1171.
23. Cournac, A.; Koszul, R.; Mozziconacci, J. The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic Acids Res.* **2015**, *44*, 245–255.

24. Shapiro, J.A.; Von Sternberg, R. Why repetitive DNA is essential to genome function. *Biol. Rev.* **2005**, *80*, 227–250.
25. Martinez-Garcia, P.J.; Crepeau, M.W.; Puiu, D.; Gonzalez-Ibeas, D.; Whalen, J.; Stevens, K.A.; Paul, R.; Butterfield, T.S.; Britton, M.T.; Reagan, R.L.; et al. The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. *Plant J.* **2016**, *87*, 507–532.
26. Hibrand Saint-Oyant, L.; Ruttink, T.; Hamama, L.; Kirov, I.; Lakhwani, D.; Zhou, N.N.; Bourke, P.M.; Daccord, N.; Leus, L.; Schulz, D.; et al. A high-quality genome sequence of Rosa chinensis to elucidate ornamental traits. *Nat. Plants* **2018**, *4*, 473.
27. Raymond, O.; Gouzy, J.; Just, J.; Badouin, H.; Verdenaud, M.; Lemainque, A.; Vergne, P.; Moja, S.; Choisne, N.; Pont, C.; et al. The Rosa genome provides new insights into the domestication of modern roses. *Nat. Genet.* **2018**, *50*, 772.
28. Donmez, N.; Brudno, M. Hapsembler: An assembler for highly polymorphic genomes. In Proceedings of the International Conference on Research in Computational Molecular Biology, Vancouver, BC, Canada, 28–31 March 2011; pp. 38–52.
29. Aguiar, D.; Istrail, S. HapCompass: A fast cycle basis algorithm for accurate haplotype assembly of sequence data. *J. Comput. Biol.* **2012**, *19*, 577–590.
30. Kajitani, R.; Toshimoto, K.; Noguchi, H.; Toyoda, A.; Ogura, Y.; Okuno, M.; Yabana, M.; Harada, M.; Nagayasu, E.; Maruyama, H.; et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **2014**, *24*, 1384–1395.
31. Safonova, Y.; Bankevich, A.; Pevzner, P.A. dipSPAdes: Assembler for Highly Polymorphic Diploid Genomes. *J. Comput. Biol.* **2015**, *22*, 528–545.
32. Chin, C.S.; Peluso, P.; Sedlazeck, F.J.; Nattestad, M.; Concepcion, G.T.; Clum, A.; Dunn, C.; O'Malley, R.; Figueroa-Balderas, R.; Morales-Cruz, A.; et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **2016**, *13*, 1050.
33. Koren, S.; Rhie, A.; Walenz, B.P.; Dilthey, A.T.; Bickhart, D.M.; Kingan, S.B.; Hiendleder, S.; Williams, J.L.; Smith, T.; Phillippy, A.M. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **2018**, *36*, 1174.
34. Roach, M.J.; Schmidt, S.A.; Borneman, A.R. Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinform.* **2018**, *19*, 460.
35. Huang, S.F.; Kang, M.J.; Xu, A.L. HaploMerger2: Rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics* **2017**, *33*, 2577–2579.
36. Miskoff, J.A.; Chaudhri, M. Mycobacterium Chimaera: A Rare Presentation. *Cureus* **2018**, *10*, e2750.
37. Quambusch, M.; Pirttila, A.M.; Tejesvi, M.V.; Winkelmann, T.; Bartsch, M. Endophytic bacteria in plant tissue culture: Differences between easy- and difficult-to-propagate *Prunus avium* genotypes. *Tree Physiol.* **2014**, *34*, 524–533.
38. Koskimaki, J.J.; Hankala, E.; Suorsa, M.; Nylund, S.; Pirttila, A.M. Mycobacteria are hidden endophytes in the shoots of rock plant [*Pogonatherum paniceum* (Lam.) Hack.](Poaceae). *Environ. Microbiol. Rep.* **2010**, *2*, 619–624.
39. Taber, R.A.; Thielen, M.A.; Falkinham J.O., III; Smith, R.H. *Mycobacterium scrofulaceum*: A bacterial contaminant in plant tissue culture. *Plant Sci.* **1991**, *78*, 231–236.
40. Goh, C.H.; Vallejos, D.F.V.; Nicotra, A.B.; Mathesius, U. The impact of beneficial plant-associated microbes on plant phenotypic plasticity. *J. Chem. Ecol.* **2013**, *39*, 826–839.
41. Ulrich, K.; Stauber, T.; Ewald, D. *Paenibacillus*—A predominant endophytic bacterium colonising tissue cultures of woody plants. *Plant Cell Tissue Organ Cult.* **2008**, *93*, 347–351.
42. Laukkanen, H.; Soini, H.; Kontunen-Soppela, S.; Hohtola, A.; Viljanen, M. A mycobacterium isolated from tissue cultures of mature *Pinus sylvestris* interferes with growth of Scots pine seedlings. *Tree Physiol.* **2000**, *20*, 915–920.
43. Taheri, S.; Lee, A.T.; Yusop, M.R.; Hanafi, M.M.; Sahebi, M.; Azizi, P.; Shamshiri, R.R. Mining and Development of Novel SSR Markers Using Next Generation Sequencing (NGS) Data in Plants. *Molecules* **2018**, *23*, 399.
44. Zhou, X.J.; Dong, Y.; Zhao, J.J.; Huang, L.; Ren, X.P.; Chen, Y.N.; Huang, S.M.; Liao, B.S.; Lei, Y.; Yan, L.Y.; et al. Genomic survey sequencing for development and validation of single-locus SSR markers in peanut (*Arachis hypogaea* L.). *BMC Genom.* **2016**, *17*, 420.

45. An, J.Y.; Yin, M.Q.; Zhang, Q.; Gong, D.T.; Jia, X.W.; Guan, Y.J.; Hu, J. Genome Survey Sequencing of *Luffa Cylindrica* L. and Microsatellite High Resolution Melting (SSR-HRM) Analysis for Genetic Relationship of *Luffa* Genotypes. *Int. J. Mol. Sci.* **2017**, *18*, 1942.
46. Li, G.Q.; Song, L.X.; Jin, C.Q.; Li, M.; Gong, S.P.; Wang, Y.F. Genome survey and SSR analysis of *Apocynum venetum*. *Biosci. Rep.* **2019**, *39*, BSR20190146.
47. Tan, J.; Guo, J.J.; Yin, M.Y.; Wang, H.; Dong, W.P.; Zeng, J.; Zhou, S.L. Next Generation Sequencing-Based Molecular Marker Development: A Case Study in *Betula Alnoides*. *Molecules* **2018**, *23*, 2963.