*plants*

MDPI

*Article*

# A Pipeline NanoTRF as a New Tool for *De Novo* Satellite DNA Identification in the Raw Nanopore Sequencing Reads of Plant Genomes

**Ilya Kirov [1,2,*], Elizaveta Kolganova [1], Maxim Dudnikov [1,2], Olga Yu. Yurkevich [3], Alexandra V. Amosova [3] and Olga V. Muravenko [3,*]**

[1] All-Russia Research Institute of Agricultural Biotechnology, Timiryazevskaya Str. 42, Moscow 127550, Russia
[2] Moscow Institute of Physics and Technology, Dolgoprudny 141701, Russia
[3] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 119991, Russia
* Correspondence:kirovez@gmail.com (I.K.); olgmur1@yandex.ru (O.V.M.)

**Abstract:** High-copy tandemly organized repeats (TRs), or satellite DNA, is an important but still enigmatic component of eukaryotic genomes. TRs comprise arrays of multi-copy and highly similar tandem repeats, which makes the elucidation of TRs a very challenging task. Oxford Nanopore sequencing data provide a valuable source of information on TR organization at the single molecule level. However, bioinformatics tools for *de novo* identification of TRs in raw Nanopore data have not been reported so far. We developed NanoTRF, a new python pipeline for TR repeat identification, characterization and consensus monomer sequence assembly. This new pipeline requires only a raw Nanopore read file from low-depth (<1×) genome sequencing. The program generates an informative html report and figures on TR genome abundance, monomer sequence and monomer length. In addition, NanoTRF performs annotation of transposable elements (TEs) sequences within or near satDNA arrays, and the information can be used to elucidate how TR–TE co-evolve in the genome. Moreover, we validated by FISH that the NanoTRF report is useful for the evaluation of TR chromosome organization—clustered or dispersed. Our findings showed that NanoTRF is a robust method for the *de novo* identification of satellite repeats in raw Nanopore data without prior read assembly. The obtained sequences can be used in many downstream analyses including genome assembly assistance and gap estimation, chromosome mapping and cytogenetic marker development.

**Keywords:** satellite DNA; Nanopore sequencing; genome; tandem repeats; pipeline



**Citation:** Kirov, I.; Kolganova, E.; Dudnikov, M.; Yurkevich, O.Y.; Amosova, A.V.; Muravenko, O.V. A Pipeline NanoTRF as a New Tool for *De Novo* Satellite DNA Identification in the Raw Nanopore Sequencing Reads of Plant Genomes. *Plants* **2022**, *11*, 2103. https://doi.org/10.3390/plants11162103

Received: 30 June 2022
Accepted: 11 August 2022
Published: 12 August 2022



## 1. Introduction

Satellite DNA (satDNA) consists of multi-copy tandemly organized repeats (TRs) and can comprise a substantial portion of most eukaryotic genomes [1]. Although satDNA was discovered more than 60 years ago, it is still an enigmatic part of eukaryotic genomes, and its current function in the cell is still a matter of debate [2]. It was shown that TRs play multiple important roles in a number of biological processes including cell division, gene expression regulation and genome architecture [1,3–5]. The well-known examples of functionally important TRs are centromeric repeats, which have been identified and elucidated in a number of species [4,6]. Centromere repeats are bound with the centromere histone H3 (CENH3) and involved in centromere location establishment as well as proper chromosome segregation during cell division [6]. TRs have been an essential part of all plant genome studies to date, and they have been shown to be involved in plant genome evolution and speciation [7–9].

Since TRs form long arrays in the genome, they can be relatively easy visualized by fluorescent in situ hybridization (FISH) [10–16]. This property of TRs has been broadly exploited in plant cytogenetic studies to develop chromosome markers, which are useful tools to trace individual chromosomes during cell division, and also to study chromosome

evolution and rearrangements [9,11–13,17–22]. Detection of some TRs by FISH can be achieved even without a denaturation step and long hybridization, allowing for a quick (few hours) chromosome identification [13,19–21]. TR-based FISH karyotyping has been established for many plant species including model organisms and crops [11–13,16,19–22]. FISH-mapped centromeric TRs are also a valuable resource for the integration of chromosome maps and genome assembly, as well as for the validation of chromosome-level genome assembly [10].

The discovery of new TRs has for a long time been based on 'wet' lab methods, including density centrifugation [23] and genomic DNA restriction [24–26]. Although these methods played an important role in the initial elucidation of TR composition and chromosome localization, they are technically challenging and do not provide information on all the TRs in a genome. An alternative group of methods exploits genome assembly data for the TR search; these methods are based on a representative of string-matching algorithms, e.g., Tandem Repeat Finder [27], nucleotide autocorrelation functions [15] or Fourier transformation [28,29]. The results of all these computational methods strongly depend on the quality and contiguity of the genome assembly. However, repetitive sequences are often significantly underrepresented or collapsed in the genome assembly, and therefore, assembly-based methods for TR identification usually underestimate the TR copy number in the genome [10,30,31]. The real breakthrough in satellite DNA studies was the development of methods based on the analysis of next-generation sequencing data [32,33]. These methods performed similarity-based read clustering, *de novo* repeat family identification and annotation, as well as repeat sequence assembly. Moreover, deep characterization of repeatome composition and evolution have been carried out in a number of species [9,10,13,34–38]. However, TR identification based on short-read NGS data lacks information about the organization of repeats at the genomic scale. At the same time, the genome context of TR location is important for understanding the origin and evolution of new TR families [14].

The introduction of Oxford Nanopore Technology (ONP, Oxford, UK) sequencing has revolutionized the field of genomics, enabling high-throughput long-read sequencing at a low price with the use of portable sequencing machines (MinIONs). In the context of tandem repeat research, raw ONP reads provide a backbone for sequencing long arrays of tandem repeats and studying the genomic context of TR organization [14,39,40]. In addition, ONP data can also be used to decipher the epigenetic profile of tandem repeats. Currently, several algorithms have been proposed to identify TRs in individual ONP reads, such as TideHunter [41] and NCRF [42]. However, *de novo* identification of novel high-copy TRs from raw ONP data obtained by low-depth genome sequencing and their classification into families, as well as the estimation of total genome abundancy, are still not straightforward. Moreover, this type of the data is rapidly accumulated in databases. In this study, we present NanoTRF (https://github.com/Kirovez/NanoTRF, accessed on 3 June 2022), a computational pipeline for the *de novo* identification, quantification and consensus assembly of high-copy TRs in raw and low-depth Nanopore sequencing data.

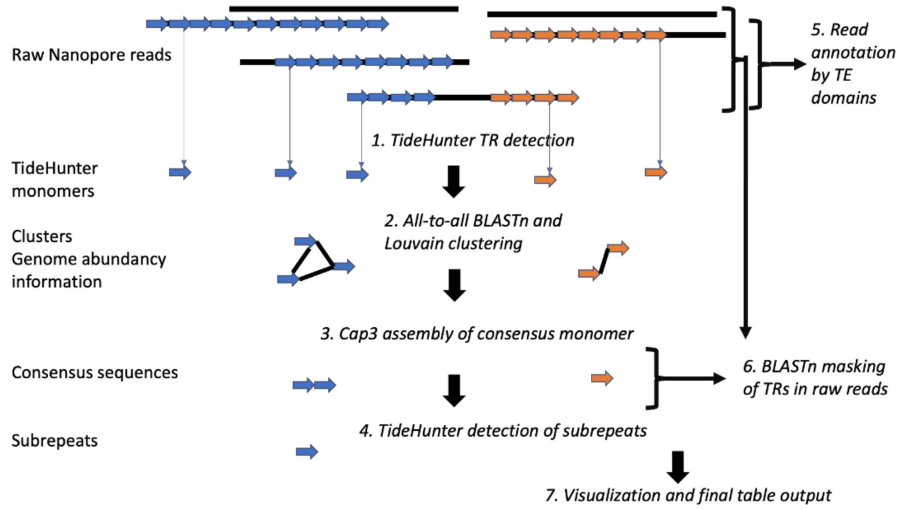
## 2. Results

### *2.1. Description of NanoTRF*

The key aim of this pipeline is the identification of high-copy tandem repeats (TRs) and the reconstruction of their consensus sequences. The only input required by NanoTRF is the raw Nanopore read file in fastq or fasta format from low-depth (it was tested on 0.1–1×) genome coverage sequencing. NanoTRF includes several steps (Figure 1): (1) TR detection in individual raw Nanopore reads by TideHunter software [41]; (2) an all-to-all similarity search between identified single-read TR sequences using BLASTn [43] and clustering of single-read TRs followed by community detection using Louvain heuristics and TR genome abundancy calculations; (3) TR consensus monomer assembly by cap3 [44] for each cluster; (4) TideHunter analysis of individual consensus and detection of subrepeats; (5) a similarity search of nanopore reads carrying TRs from each cluster with TE protein domains;

(6) conversion of BLASTn consensus monomer sequences to raw reads for calculation of the percentage of reads similar to the TRs; (7) Draw cluster layout and read annotation (pie chart and histogram of read coverage by TRs and read annotation by TE domains) figures and writing the final report table. NanoTRF reports several files: (i) a fasta file possessing consensus contig sequences of TRs assembled from monomers in individual clusters by cap3; (ii) a summary table containing per-cluster information and (iii) a html report with general information about TRs, graph layouts, figures, read coverage by TRs and read annotation by TE domains.

A




B

| Cluster Name | CLuster | Genome abundancy, % | Genome abundancy corrected, % | Per Read Coverage By Cluster, % | Per Read Coverage By Cluster Pie, % | Min monomer size assembled by Cap3 | Max monomer size assembled by Cap3 | Subrepeats length | Annotation | Per read annotation |
|---|---|---|---|---|---|---|---|---|---|---|
| clust1 | | 0.5258 | 0.629 | | | 333 | 2104 | 336 - 377 | Reads: 74 (6.38%) : Class_II_EnSpm/CACTA_CACTA-TPase:17 | |
| clust29 | | 0.4191 | 0.494 | | | 329 | 4430 | 301 - 388 | Reads: 75 (7.07%) : Class_II_EnSpm/CACTA_CACTA-TPase:29 | |
| clust9 | | 0.3728 | 0.365 | | | 409 | 2867 | 361 - 390 | Reads: 32 (4.49%) : Class_I_Tekay_Ty3-RH:9 | |
| clust0 | | 0.3119 | 0.319 | | | 394 | 4456 | 376 - 405 | Reads: 31 (4.71%) : Class_I_Tekay_Ty3-RT:12 | |
| clust11 | | 0.2551 | 0.222 | | | 165 | 3206 | 154 - 172 | Reads: 34 (5.11%) : Class_I_Tekay_Ty3-INT:9 | |

**Figure 1.** Schematic representation of the NanoTRF pipeline and the report. (**A**) The scheme showing the seven main steps in the NanoTRF pipeline: (1) TR identification in individual ONP reads by TideHunter; (2) all-to-all similarity search between TRs and clustering of highly similar TRs; (3) consensus monomer assembly by cap3; (4) detection of subrepeats in the contigs; (5) ONP read annotation by the TE protein database; (6) calculation of ONP read coverage by TRs and (7) final report generation. Monomers of distinct TRs are colored in orange and blue. (**B**) A screenshot of the output table from the html file generated by NanoTRF.

### *2.2. TR Abundancy Calculated from Long (NanoTRF) and Short (TAREAN) Reads Are Well Correlated*

To validate the NanoTRF pipeline, we performed Nanopore sequencing of the genome of *Deschampsia antarctica* E. Desv. (Poaceae), a species with a well-characterized satellitome composition [9,38]. In total, we obtained 1,452,313 reads with N50 1305 bp and a total number of bases of ~4 Gb, representing about 0.8× of coverage of the *D. antarctica* genome. In total, 43 highly abundant TR clusters (genome abundancy 0.52–0.01%) with a median monomer length of 390 bp (42–2192 bp) were identified by NanoTRF (Supplementary File S1). Based on the genome abundancy of each TR, we estimated that 3.5% of the *D. antarctica* genome is occupied by satellite DNA.

We also compared the NanoTRF results with the results from the TAREAN software [33], which uses Illumina reads. For this, we performed identification of TRs by TAREAN using publicly available Illumina NGS reads for *D. antarctica* [37]. The comparison of TRs found by NanoTRF and TAREAN revealed that 93% (15 of 16 high-confident TRs) of the TRs identified by TAREAN were also detected by NanoTRF. Based on the genome abundancy of each TR found by TAREAN, we calculated that 2.44% of the *D. antarctica* genome is occupied by satellite DNA. This value is in good accordance with the NanoTRF results (3.5%). We further compared the genome abundancy, and the monomer length calculated for TRs found by NanoTRF and TAREAN resulted in a good correlation (correlation coefficients were 0.87 and 0.95, respectively ($p$-value < $1.1 \times 10^{-5}$; Figure 2A,B). Similarly, alignment of the monomer sequences assembled by NanoTRF and TAREAN revealed that most of the sequences had >93% similarity.

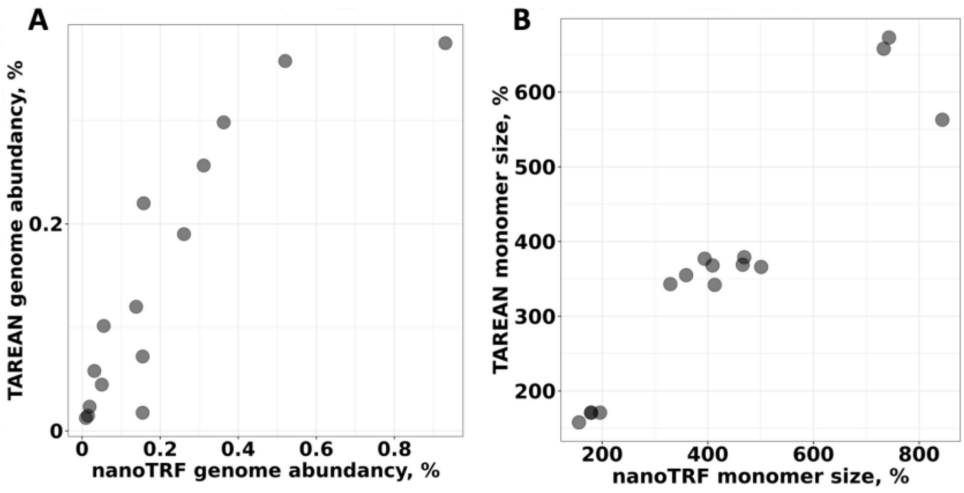



**Figure 2.** Comparison of the results of the NanoTRF and TAREAN identification of TRs in the *D. antarctica* genome based on the (**A**) genome abundancy and (**B**) monomer length. Individual TRs are represented as black dots. X-axis and Y-axis show the results from NanoTRF and TAREAN, respectively.

Thus, our results demonstrated that NanoTRF is a robust tool for the *de novo* identification, sequence assembly and genome abundancy prediction of high-copy tandem repeats from raw ONP data.

### *2.3. NanoTRF Data for Clustered and Dispersed TRs*

On the genomic level, TRs can be dispersed over chromosomes or can generate long arrays. TRs that are organized in megabase-sized arrays are usually a part of centromeres, subtelomeres or heterochromatin regions of plant chromosomes. The ability to distinguish between dispersed and clustered TRs is an important task.

To achieve this, NanoTRF provides data on the percentage of the sequence of individual Nanopore reads that shows similarity to NanoTRF clusters. This data is presented as a histogram and a pie chart in the NanoTRF html report, and it can be useful for the

estimation of the relative TR array size. To validate this, we compared the read coverage using three TRs of *D. antarctic*—Da322, Da97 and Da238—having dispersed, dispersed and clustered and clustered chromosome organization, respectively (Figure 3). We found that the majority of reads in the clusters of all three repeats were covered by a TR of >90% (Figure 3). However, the percentage of the reads covered by less than 90% of their sequence (Figure 2) by the corresponding TR was 2 and 1.5-times higher for the dispersed (Da322) and dispersed + clustered (Da97) TRs, respectively, compared to the TR with a clustered chromosome organization (Da238).

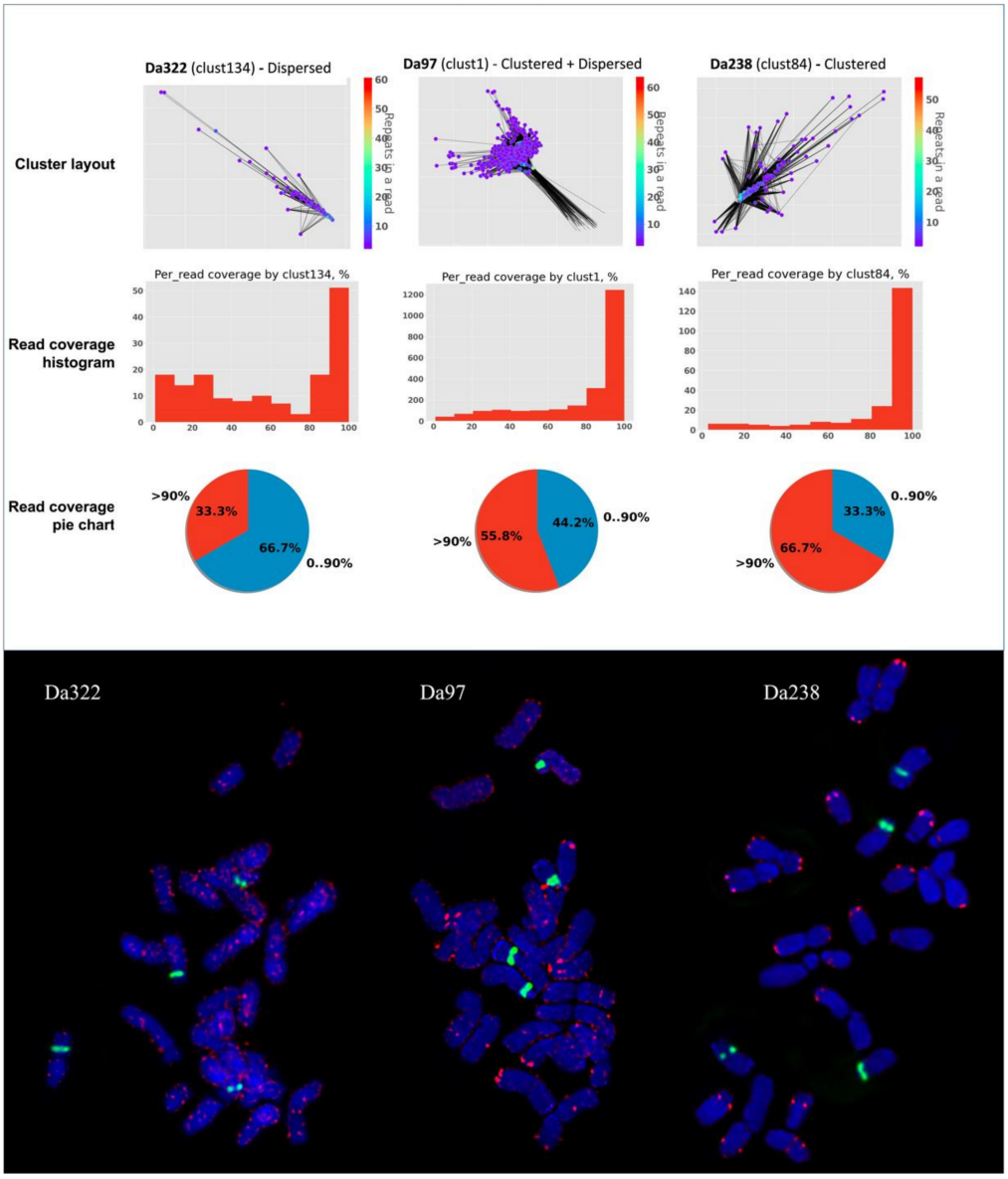



**Figure 3.** Comparison of the cluster layouts and read coverage data of three TRs with different FISH patterns on chromosomes of *Deschampsia antarctica*: Da322 (dispersed), Da97 (dispersed and clustered) and Da238 (clustered) probes. Bottom picture shows the results of FISH experiments with labeled individual TRs (Da322, Da97 and Da238; red fluorescence signals) and 45S rDNA (green fluorescence signals). Chromosomes are stained by DAPI (blue fluorescence signals).

These results showed that the read coverage data provided by NanoTRF can be used to estimate the chromosome organization of TRs and also to select TRs that are suitable for molecular cytogenetic studies as chromosome FISH markers.

### *2.4. Cluster Annotation Showed an Association between TR and Transposable Elements*

Previous reports on some plant species, e.g., *Lathyrus sativus* L. (Fabaceae), have suggested that TR origin and evolution can be tightly connected with the amplification of certain TE families, including Ogre elements [14]. NanoTRF performs automatic annotation of reads in each cluster with the use of the TE protein domain database. This information can be useful for dissecting TR–TE coevolution events.

To assess this, we performed a NanoTRF analysis of a subset (total read length—500 Mb) of ultra-long reads of *L. sativus* that were previously used for tracing TR origins using Ogre elements. Using the similarity search, we identified the NanoTRF clusters corresponding to Fab TRs of *L. sativus*. Among them, FabTR2 was shown to generate long-arrays in the genome that were occasionally disturbed by Ogre insertions [14]. Similarly, the NanoTRF cluster of FabTR2 (clust1) had only 15.8% of its reads partially (<90%) covered by this TR—a feature corresponding to clustered TRs (see above). Additionally, 5% of the reads in this cluster showed similarity to domains of TEs, including Ogre elements, corroborating with occasional insertions of these elements into FabTR2 arrays. Opposite to FabTR2, another TR—FabTR58—was shown to be mostly part of Ogre element copies [14]. Indeed, the analysis of the corresponding NanoTRF cluster (clust59) also showed that this repeat does not generate long arrays in the genome, as follows from the very low number of reads (3%) with >90% coverage by the TR (Figure 4A). Besides this, 25% of the reads possessed a similarity to Ogre element protein domains, which is five-times higher than for the FabTR2 cluster (Figure 4B)—supporting the previous conclusion of the frequent co-occurrence of FabTR58 and Ogre elements in the genome.

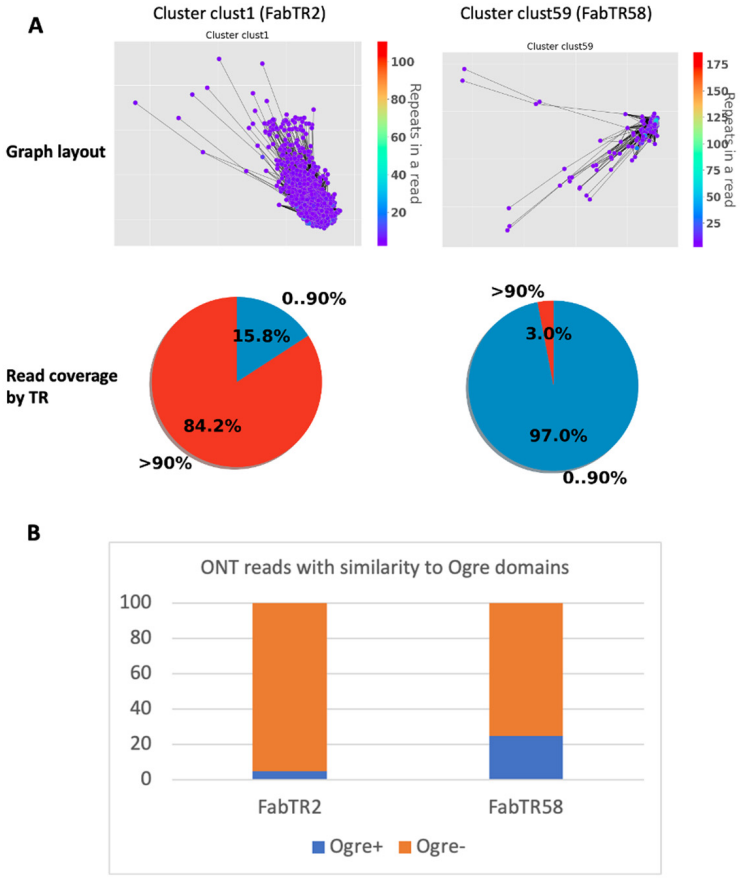



**Figure 4.** NanoTRF results for the analysis of the FabTR2 and FabTR58 repeats. (**A**) Graph layouts and pie chart showing the percentage of reads with different TR occupancies for clusters clust1 (FabTR2) and clust59 (FabTR58). (**B**) Bar plot of the percentage of reads in the two clusters (clust1 and clust59) possessing similarity to Ogre TE domains.

These results demonstrate that read annotation by TE domains and read coverage by TRs provided by NanoTRF are useful for the elucidation of TR–TE co-location in the genome.

## 3. Discussion

Satellite DNA is an important component of many eukaryotic genomes including human and plants. In plant genomes, satellite DNA is presented in a large number of families and with wide diversity, and it plays an essential role in organizing the structural integrity and functioning of genomes [1,4,5]. TRs exhibit remarkable diversity in monomer size, genome abundancy, chromosome localization and sequence composition, even between closely related species [3,13,16,45,46]. During the last decade, the number of discovered satellite tandem repeats (TRs) has rapidly accumulated because of new tools developed for *de novo* TR identification using short reads from NGS sequencing [32,33]. However, progress in the understanding of the genomic organization of TRs is lagging behind because of challenges in TR assembly and their frequent underestimation in sequenced genomes [30]. At the same time, the long reads obtained by the Oxford Nanopore (ONP) method are useful for the sequencing of long arrays of repetitive DNA [40]. The number of available datasets with ONP long reads is rapidly growing as they can be generated in a conventional laboratory using a portable (e.g., MinION) device and easy library preparation protocols. Theoretically, the length of the ONP reads is only limited by the size of the isolated DNA fragments, making ONP reads an attractive tool for the study of the genomic organization of TRs [14,39,40]. However, the identification of satellite tandem repeats in raw ONP data obtained by low-depth genome sequencing has been hampered by the absence of the user-friendly tools. To fill in this gap, we developed NanoTRF, a simple, robust and easy-to-use pipeline for the direct identification, quantification and analysis of TRs using raw, low-depth ONP data.

The advantages of NanoTRF are the following: (1) it requires neither prior knowledge about TR sequences nor genome assembly, and therefore it can be used for a broad range of species; (2) even low coverage ONP genomic data (>0.1×) is suitable for NanoTRF; (3) NanoTRF calculates the genome abundance for each TR and the results are comparable with TAREAN [33]; (4) NanoTRF provides additional information on read occupancy by TRs and, as we have shown in the present study, this data can be used for TR filtering according to their genome organization—clustered or dispersed; (5) the read coverage data, together with the read annotation by TE domains, is useful for the elucidation of TR–TE coevolution and genome organization and (6) the html report provides user-friendly access to the NanoTRF results.

Recent studies have utilized ONP reads to elucidate TR genome organization features in plant genomes [10,14,47,48]. The very long size of the ONP reads open the door for the investigation of TRs in individual genomic arrays. Although ONP reads are error-prone, it still possible to detect individual TR arrays in raw reads without contig assembly. This is crucial for satellitome studies as the assembly step may introduce artefacts and significantly shorten the TR array [10]. Previous studies [14] have shown that ONP reads can also provide insights into TR organization on the chromosome level. There can be differences in the chromosome-scale organization of TRs, with some TRs forming long arrays (clustered TRs) while others form short clusters dispersed throughout the genome (dispersed TRs). Not all tandemly organized sequences generate long arrays in the genome [13]. At the same time, clustered TRs are attractive items for molecular cytogenetics and also FISH-based karyotyping. NanoTRF provides information on the portion of ONP reads with similarity to distinct TRs in order to distinguish clustered and dispersed TRs. Using FISH, we demonstrated that NanoTRF reports are indeed useful for separating TRs based on genomic organization. In the case of clustered TRs, this information is useful when selecting TRs for FISH-based chromosome marker design. Another valuable feature in NanoTRF is the TE annotation step in the TR-carrying ONP reads. Modern similarity search algorithms (e.g., DIAMOND [49]) and TE protein database [50] make it possible to detect TE sequences

in raw ONP reads with up to 15–20% errors. A previous study demonstrated that this approach can be used to decode TE–TR coevolution events [14]. By using NanoTRF, we have demonstrated that TE–TR coevolution events can now be detected automatically.

In conclusion, NanoTRF introduces a new modern toolbox for TR analysis using Nanopore raw genomic sequencing data in plants and other eukaryotes.

## 4. Materials and Methods

### *4.1. Overview of NanoTRF*

NanoTRF is written in Python 3 and depends on several libraries including matplotlib (https://matplotlib.org/, accessed on 3 June 2022), biopython [51], networkx [52] and python-louvain (https://github.com/taynaud/python-louvain, accessed on 3 July 2022). NanoTRF includes several steps (Figure 1) and modules. The first module begins with the identification of single-read TRs ('srTRs') in individual ONT reads using TideHunter [49] with default parameters. After this, all obtained srTR sequences are subjected to all-to-all similarity comparison by BLASTn [50] with the following optional parameters: -word_size (default = 24), -max_hsps = 1 and –evalue (default = 2). Third-clustering module works using the Louvain clustering method, which groups similar srTRs into communities and attempts to merge similar nodes from similar communities, further building a new network with node communities (srTR clusters). In the next stage, clusters consisting of less than 6 srTRs are removed from further analysis. Next, the module launches the Cap3 software with the additional parameters (-h 100 -n -2 -m 3 -p 80 -s 600) to perform assembly of consensus TR repeats from the srTRs of a cluster. The genome abundancy for each cluster is calculated as follows: $(\sum_{i=0}^{i=n} nrep * len * 100)/\mathrm{TRL}$, where n—number of monomer sequences in a cluster, nrep—number of repeats of the monomer occurs in the read, len—monomer length and TRL—the total length of all input reads. The individual clusters are drawn by networkx [48] and matplotlib libraries. Then, the corrected genome abundancy is estimated by blastn (-word_size 28) masking of raw ONT reads using assembled TRs. The corrected genome abundancy is calculated as the sum of the masked base pairs of the ONT reads by individual TRs divided by the total number of raw ONT reads. The obtained information is also used to calculate the percentage of the ONT reads covered by TRs. Then, the raw reads for which the TRs were clustered are subjected to annotation by DIAMOND [51] and the RExDB [52] database of transposon protein domain sequences.

### *4.2. Plant Material and DNA Isolation*

The seeds of *D. antarctica* (KEW-0521613, St. Georgia, Falkland Is.) were obtained from the Seed Conservation Department of Kew Royal Botanic Gardens (Kew, UK). Genomic DNA was isolated from green leaves using a Genomic DNA Purification kit (Thermo Fisher Scientific, Waltham, MA, USA). The concentration of DNA was assessed with a Quibit 4 fluorometer (Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA).

### *4.3. Library Preparation and Nanopore Sequencing*

The library for the ONP sequencing was prepared using an SQK-LSK109 Ligation Sequencing Kit (Oxford Nanopore Technology, Oxford, UK) for 1D genomic DNA sequencing. A MinION (ONT) instrument with an R9.4.1 flow-cell (ONT) was used for sequencing. Sequencing was controlled by MinKNOW v18.07.2 (Oxford Nanopore) and stopped after 48 h. Basecalling was carried out by Guppy software v 4.0.11 (Oxford Nanopore Technology, UK).

### *4.4. Search of TRs in Illumina Data by TAREAN Software*

Publicly available Illumina reads of *D. antarctica* [37] were downloaded from NCBI (accession number—SRR1158316). Quality checks and adapter trimming were performed by the FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, accessed on 1 May 2022) and Trimmomatic [53] tools, respectively. TAREAN [33] software was run as a part of the RepeatExplorer pipeline [32].

### *4.5. Fluorescence in Situ Hybridization*

In the FISH assays, we used two wheat DNA probes: pTa71 enclosing 18S-5.8S-26S (45S) rDNA and pTa794 containing 5S rDNA [54]. These DNA probes were labelled directly with fluorochromes Aqua 431 dUTP or Red 580 dUTP (ENZO Life Sciences, NY, USA) by nick translation according to the manufacturers' protocols. Additionally, oligonucleotide probes Da322, Da97 and Da238 were designed based on the satDNA sequences of Da322, Da97 and Da238 (Table 1). These probes were produced and labelled directly with Cy3-dUTP in *Syntol* (Moscow, Russia). FISH procedures were performed as described previously [9].

**Table 1.** Sequences of the TRs of *Deschampsia antarctica* used for generation of the oligonucleotide FISH probes.

| Tandem Repeat/Genome Proportion, % | Length, bp | Sequence |
|---|---|---|
| Da 97/0.21 | 342 | CCCACGGGCTAGGGTTTCGCTGGAAAAGTACCGCCGGAGCGCCG<br>GAATCCCACGAAAACTTGCGTGTGGCCCTAGCATGCATGCACAA<br>GTGTGGTGGAAGGTTCCTAGATGCAATACGTAGCTCCCGGGTGCG<br>ATCCTGTTCGCGCGCATGCGATAACACTTAGAAAACTGCTGGACC<br>TCTGGGAGGAATCTCCCGCTACGGGTCAAACGGAGGGCAACCGG<br>TGGAATCATGGGCCCAACCTTGGTTTTCCATGTAGATATGCCCTA<br>AACAAACCCAAACCAACAAAAAAGTACATTGGTCAACCCTCGTA<br>CGGAGAATGCTAGGGGGCTAGACCTGCGGGT |
| Da 238/0.042 | 379 | GCCTAACACCCTATCGTAGACACCCATGGGTTGGGGCGCAGTGC<br>ACGTAATACTATACGGATCCAGCGTTCCATCGAATTTTGAGTTTTT<br>ACTGCAGAAACTTCCATTTTTCCTAGACTTGTGAGCACTTTTTGAG<br>GCCCTAAAAAGGCTTTTTTGGGGTCGAGATGGTCCGCACGCGTGC<br>TGGGGTGTGTGCACGTATGTAAAATCATCCGGATTGCAAAAATTA<br>GAAGTCCTTTTATC<br>CTAGTTCTCCGAGATCTTTCTAACGCCTTCGAAACCGCCTCAATC<br>GGAGCTCGTTCTCATTCGCGTCGTTAGTATTAACAAAGTTCCTCCG<br>TACGATGATCCTTTGCTTTCAACGGTCACCGTTTCTTCTCAGGCGTGA |
| Da 322/0.013 | 342 | GGTCTAGGGTTTCCCCGGATACAGACCACCGGAGCGTCGGAATC<br>GCTGGAAAACTTGCATGTGTCCCTAACATATGTGTACAAGTGTGA<br>TGTAAGGTTGGTAGATGGCATATCTAGGTCCCAGGCGTGACGCTG<br>TTCGCAGACATGGGCTAACACTTGGTAAAATCCTGGATCTGTATG<br>TGGAAACTCCCGCTACGGGTCAACCGGAGCCTATTTTATGGTAAA<br>GTAGGCCCAACCTCTGCTTTCCATGTACATATGTCCTAAACAAAC<br>CAAAACAAGGAAAAAACTCCATTGGTAAACCCTCGTACGGAGA<br>AAGCTATAGGGGTAGATCTGCGGGGTCCCC |

## 5. Conclusions

Long reads provide indispensable information on repetitive DNA organization. However, the exploiting of raw ONP data obtained by low-depth genome sequencing for high-copy tandem repeat discovery has been hampered by the absence of user-friendly tools. In the present study, we developed NanoTRF, a pipeline for TR repeat identification, characterization and consensus monomer sequence assembly. NanoTRF only requires a raw Nanopore read file from low-depth (<1×) genome sequencing. An informative html report and figures on TR genome abundance, monomer sequence and monomer length, as well as annotation of transposable elements (TEs) sequences within or near TR arrays is generated. The obtained TR sequences can be used in many downstream analyses including genome assembly assistance and gap estimation, chromosome mapping and cytogenetic marker development. We believe that NanoTRF will significantly accelerate the progress in satellitome research as it opens the way for rapid TR identification, placing individual TRs into their genomic context.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/plants11162103/s1, File S1: *D. antarctica* satellitome composition unraveled by NanoTRF.

**Author Contributions:** Conceptualization, I.K. and O.V.M.; methodology, I.K. and O.V.M.; software, I.K. and E.K.; validation, I.K., O.V.M.; formal analysis, I.K., E.K., M.D., O.Y.Y. and A.V.A.; investigation, I.K., E.K., M.D., O.Y.Y., A.V.A. and O.V.M.; writing—original draft preparation, I.K., E.K., O.Y.Y., A.V.A. and O.V.M.; writing—review and editing, I.K. and O.V.M.; visualization, I.K., E.K., M.D., O.Y.Y., A.V.A. and O.V.M.; supervision I.K. and O.V.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Russian Science Foundation (project No. 22-26-00222).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The Nanopore data produced for this study are available in Sequence Read Archive (SRA) NCBI under Bioproject Accession PRJNA708177.

**Acknowledgments:** The authors acknowledge Jiri Macas (Biology Centre ASCR, Ceske Budejovice, Czech Republic) for their valuable comments and suggestions on this study.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

1. Garrido-Ramos, M.A. Satellite DNA: An evolving topic. *Genes* **2017**, *8*, 230.
2. Shatskikh, A.S.; Kotov, A.A.; Adashev, V.E.; Bazylev, S.S.; Olenina, L.V. Functional Significance of Satellite DNAs: Insights from Drosophila. *Front. Cell Dev. Biol.* **2020**, *8*, 312. [CrossRef]
3. Plohl, M.; Meštrović, N.; Mravinac, B. Satellite DNA evolution. *Genome Dyn.* **2012**, *7*, 126–152.
4. Plohl, M.; Meštrović, N.; Mravinac, B. Centromere identity from the DNA point of view. *Chromosoma* **2014**, *123*, 313–325.
5. Hartley, G.; O'Neill, R.J. Centromere Repeats: Hidden Gems of the Genome. *Genes* **2019**, *10*, 223.
6. Talbert, P.B.; Henikoff, S. What Makes a Centromere? *Exp. Cell Res.* **2020**, *389*, 111895.
7. Ferree, P.M.; Barbash, D.A. Species-Specific Heterochromatin Prevents Mitotic Chromosome Segregation to Cause Hybrid Lethality in Drosophila. *PLoS Biol.* **2009**, *7*, e1000234. [CrossRef]
8. Nadachowska-Brzyska, K.; Burri, R.; Olason, P.I.; Kawakami, T.; Smeds, L.; Ellegren, H. Demographic Divergence History of Pied Flycatcher and Collared Flycatcher Inferred from Whole-Genome Re-Sequencing Data. *PLoS Genet.* **2013**, *9*, e1003942. [CrossRef]
9. Amosova, A.V.; Yurkevich, O.Y.; Bolsheva, N.L.; Samatadze, T.E.; Zoshchuk, S.A.; Muravenko, O.V. Repeatome Analyses and Satellite DNA Chromosome Patterns in *Deschampsia sukatschewii*, *D. cespitosa*, and *D. antarctica* (Poaceae). *Genes* **2022**, *13*, 762. [CrossRef]
10. Saint-Oyant, L.H.; Ruttink, T.; Hamama, L.; Kirov, I.; Lakhwani, D.; Zhou, N.-N.; Bourke, P.; Daccord, N.; Leus, L.; Schulz, D. A High-Quality Genome Sequence of Rosa Chinensis to Elucidate Ornamental Traits. *Nat. Plants* **2018**, *4*, 473–484.
11. Divashuk, M.G.; Alexandrov, O.S.; Razumova, O.V.; Kirov, I.V.; Karlov, G.I. Molecular Cytogenetic Characterization of the Dioecious Cannabis Sativa with an XY Chromosome Sex Determination System. *PLoS ONE* **2014**, *9*, e85118. [CrossRef]
12. Kirov, I.; Gilyok, M.; Knyazev, A.; Fesenko, I. Pilot Satellitome Analysis of the Model Plant, Physcomitrella patens, Revealed a Transcribed and High-Copy IGS Related Tandem Repeat. *Comp. Cytogenet.* **2018**, *12*, 493.
13. Kirov, I.V.; Kiseleva, A.V.; Laere, K.V.; Roy, N.V.; Khrustaleva, L.I. Tandem Repeats of Allium Fistulosum Associated with Major Chromosomal Landmarks. *Mol. Genet. Genom.* **2017**, *292*, 453–464.
14. Vondrak, T.; Robledillo, L.Á.; Novák, P.; Koblížková, A.; Neumann, P.; Macas, J. Characterization of Repeat Arrays in Ultra-Long Nanopore Reads Reveals Frequent Origin of Satellite DNA from Retrotransposon-Derived Tandem Repeats. *Plant J.* **2020**, *101*, 484–500. [CrossRef]
15. Macas, J.; Navrátilová, A.; Koblížková, A. Sequence Homogenization and Chromosomal Localization of VicTR-B Satellites Differ between Closely Related Vicia Species. *Chromosoma* **2006**, *115*, 437–447.
16. Amosova, A.V.; Ghukasyan, L.; Yurkevich, O.Y.; Bolsheva, N.L.; Samatadze, T.E.; Zoshchuk, S.A.; Muravenko, O.V. Cytogenomics of Deschampsia P. Beauv. (Poaceae) Species Based on Sequence Analyses and FISH Mapping of CON/COM Satellite DNA Families. *Plants* **2021**, *10*, 1105. [CrossRef]
17. Hobza, R.; Lengerova, M.; Svoboda, J.; Kubekova, H.; Kejnovsky, E.; Vyskot. An Accumulation of Tandem DNA Repeats on the Y Chromosome in Silene Latifolia during Early Stages of Sex Chromosome Evolution. *Chromosoma* **2006**, *115*, 376.
18. Kato, A.; Vega, J.M.; Han, F.; Lamb, J.C.; Birchler, J.A. Advances in Plant Chromosome Identification and Cytogenetic Techniques. *Curr. Opin. Plant Biol.* **2005**, *8*, 148–154.

19. Tang, S.; Tang, Z.; Qiu, L.; Yang, Z.; Li, G.; Lang, T.; Zhu, W.; Zhang, J.; Fu, S. Developing New Oligo Probes to Distinguish Specific Chromosomal Segments and the A, B, D Genomes of Wheat (*Triticum aestivum* L.) Using ND-FISH. *Front. Plant Sci.* **2018**, *9*, 1104.
20. Xi, W.; Tang, S.; Du, H.; Luo, J.; Tang, Z.; Fu, S. ND-FISH-Positive Oligonucleotide Probes for Detecting Specific Segments of Rye (*Secale cereale* L.) Chromosomes and New Tandem Repeats in Rye. *Crop J.* **2020**, *8*, 171–181.
21. Xiao, Z.; Tang, S.; Qiu, L.; Tang, Z.; Fu, S. Oligonucleotides and ND-FISH Displaying Different Arrangements of Tandem Repeats and Identification of Dasypyrum Villosum Chromosomes in Wheat Backgrounds. *Molecules* **2017**, *22*, 973.
22. Zhu, M.; Du, P.; Zhuang, L.; Chu, C.; Zhao, H.; Qi, Z. A Simple and Efficient Non-Denaturing FISH Method for Maize Chromosome Differentiation Using Single-Strand Oligonucleotide Probes. *Genome* **2017**, *60*, 657–664.
23. Kit, S. Equilibrium Sedimentation in Density Gradients of DNA Preparations from Animal Tissues. *J. Mol. Biol.* **1961**, *3*, 711-IN2.
24. Alix, K.; Baurens, F.-C.; Paulet, F.; Glaszmann, J.-C.; D'Hont, A. Isolation and Characterization of a Satellite DNA Family in the Saccharum Complex. *Genome* **1998**, *41*, 854–864.
25. Waye, J.S.; Willard, H.F. Human Beta Satellite DNA: Genomic Organization and Sequence Definition of a Class of Highly Repetitive Tandem DNA. *Proc. Natl. Acad. Sci. USA* **1989**, *86*, 6250–6254.
26. Divashuk, M.; Alexandrov, O.; Kroupin, P.Y.; Karlov, G. Molecular Cytogenetic Mapping of Humulus Lupulus Sex Chromosomes. *Cytogenet. Genome Res.* **2011**, *134*, 213–219.
27. Benson, G. Tandem Repeats Finder: A Program to Analyze DNA Sequences. *Nucleic Acids Res.* **1999**, *27*, 573–580.
28. Sharma, D.; Issac, B.; Raghava, G.P.; Ramaswamy, R. Spectral Repeat Finder (SRF): Identification of Repetitive Sequences Using Fourier Transformation. *Bioinformatics* **2004**, *20*, 1405–1412. [CrossRef]
29. Yadav, Y.; Sharma, S.N.; Shakya, D.K. Detection of Tandem Repeats in DNA Sequences Using Short-Time Ramanujan Fourier Transform. In *Transactions on Computational Biology and Bioinformatics*; IEEE/ACM: New York, NY, USA, 2021; pp. 1583–1591. [CrossRef]
30. Peona, V.; Weissensteiner, M.H.; Suh, A. How Complete Are "Complete" Genome Assemblies?—An Avian Perspective. *Mol. Ecol. Resour.* **2018**, *18*, 1188–1195.
31. Tørresen, O.K.; Star, B.; Mier, P.; Andrade-Navarro, M.A.; Bateman, A.; Jarnot, P.; Gruca, A.; Grynberg, M.; Kajava, A.V.; Promponas, V.J.; et al. Tandem Repeats Lead to Sequence Assembly Errors and Impose Multi-Level Challenges for Genome and Protein Databases. *Nucleic Acids Res.* **2019**, *47*, 10994–11006. [CrossRef]
32. Novak, P.; Neumann, P.; Pech, J.; Steinhaisl, J.; Macas, J. RepeatExplorer: A Galaxy-Based Web Server for Genome-Wide Characterization of Eukaryotic Repetitive Elements from next-Generation Sequence Reads. *Bioinformatics* **2013**, *29*, 792–793.
33. Novák, P.; Ávila Robledillo, L.; Koblížková, A.; Vrbová, I.; Neumann, P.; Macas, J. TAREAN: A Computational Tool for Identification and Characterization of Satellite DNA from Unassembled Short Reads. *Nucleic Acids Res.* **2017**, *45*, e111. [CrossRef]
34. Lower, S.S.; McGurk, M.P.; Clark, A.G.; Barbash, D.A. Satellite DNA Evolution: Old Ideas, New Approaches. *Curr. Opin. Genet. Dev.* **2018**, *49*, 70–78.
35. Peška, V.; Mandáková, T.; Ihradská, V.; Fajkus, J. Comparative Dissection of Three Giant Genomes: Allium Cepa, Allium Sativum, and Allium Ursinum. *Int. J. Mol. Sci.* **2019**, *20*, 733.
36. Kreplak, J.; Madoui, M.-A.; Cápal, P.; Novák, P.; Labadie, K.; Aubert, G.; Bayer, P.E.; Gali, K.K.; Syme, R.A.; Main, D. A Reference Genome for Pea Provides Insight into Legume Genome Evolution. *Nat. Genet.* **2019**, *51*, 1411–1422.
37. González, M.L.; Chiapella, J.O.; Urdampilleta, J.D. Characterization of Some Satellite DNA Families in *Deschampsia antarctica* (Poaceae). *Polar Biol.* **2018**, *41*, 457–468.
38. González, M.L.; Chiapella, J.; Topalian, J.; Urdampilleta, J.D. Genomic Differentiation of *Deschampsia antarctica* and *D. cespitosa* (Poaceae) Based on Satellite DNA. *Bot. J. Linn. Soc.* **2020**, *194*, 326–341. [CrossRef]
39. Dvorkina, T.; Bzikadze, A.V.; Pevzner, P.A. The String Decomposition Problem and Its Applications to Centromere Analysis and Assembly. *Bioinformatics* **2020**, *36*, i93–i101. [CrossRef]
40. Miga, K.H.; Koren, S.; Rhie, A.; Vollger, M.R.; Gershman, A.; Bzikadze, A.; Brooks, S.; Howe, E.; Porubsky, D.; Logsdon, G.A.; et al. Telomere-to-Telomere Assembly of a Complete Human X Chromosome. *Nature* **2020**, *585*, 79–84. [CrossRef]
41. Gao, Y.; Liu, B.; Wang, Y.; Xing, Y. TideHunter: Efficient and Sensitive Tandem Repeat Detection from Noisy Long-Reads Using Seed-and-Chain. *Bioinformatics* **2019**, *35*, i200–i207. [CrossRef]
42. Harris, R.S.; Cechova, M.; Makova, K.D. Noise-Cancelling Repeat Finder: Uncovering Tandem Repeats in Error-Prone Long-Read Sequencing Data. *Bioinformatics* **2019**, *35*, 4809–4811. [CrossRef]
43. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
44. Huang, X.; Madan, A. CAP3: A DNA Sequence Assembly Program. *Genome Res.* **1999**, *9*, 868–877. [CrossRef]
45. Lee, H.-R.; Zhang, W.; Langdon, T.; Jin, W.; Yan, H.; Cheng, Z.; Jiang, J. Chromatin Immunoprecipitation Cloning Reveals Rapid Evolutionary Patterns of Centromeric DNA in Oryza Species. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 11793–11798.
46. Talbert, P.B.; Kasinathan, S.; Henikoff, S. Simple and Complex Centromeric Satellites in *Drosophila* Sibling Species. *Genetics* **2018**, *208*, 977–990. [CrossRef]
47. Wang, B.; Yang, X.; Jia, Y.; Xu, Y.; Jia, P.; Dang, N.; Wang, S.; Xu, T.; Zhao, X.; Gao, S.; et al. High-quality *Arabidopsis thaliana* genome assembly with nanopore and HiFi long reads. *Genom. Proteom. Bioinform.* **2021**. [CrossRef]
48. Naish, M.; Alonge, M.; Wlodzimierz, P.; Tock, A.J.; Abramson, B.W.; Schmücker, A.; Mandáková, T.; Jamge, B.; Lambing, C.; Kuo, P.; et al. The genetic and epigenetic landscape of the Arabidopsis centromeres. *Science* **2021**, *374*, eabi7489.
49. Buchfink, B.; Xie, C.; Huson, D.H. Fast and Sensitive Protein Alignment Using DIAMOND. *Nat. Methods* **2015**, *12*, 59–60. [CrossRef]

50. Neumann, P.; Novák, P.; Hoštáková, N.; Macas, J. Systematic Survey of Plant LTR-Retrotransposons Elucidates Phylogenetic Relationships of Their Polyprotein Domains and Provides a Reference for Element Classification. *Mob. DNA (UK)* **2019**, *10*, 1. [CrossRef]
51. Cock, P.J.A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423. [CrossRef]
52. Hagberg, A.; Swart, P.; Chult, D.S. Exploring Network Structure, Dynamics, and Function Using NetworkX. In Proceedings of the 7th Python in Science Conference, Pasadena, CA, USA, 1 January 2008.
53. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics (Oxf. Engl.)* **2014**, *30*, 2114–2120. [CrossRef]
54. Gerlach, W.L.; Bedbrook, J.R. Cloning and Characterization of Ribosomal RNA Genes from Wheat and Barley. *Nucleic Acids Res.* **1979**, *7*, 1869–1885. [CrossRef]