Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Draft genome and SSR data mining of Typhonium flagelliforme, an anti-cancer medicinal plant

Devit Purwoko
National Research and Innovation Agency
Siti Zulaeha
National Research and Innovation Agency
Gemilang Rahmadara
National Research and Innovation Agency
Suparjo Suparjo
Research Organization for Health, BRIN. Science and Technology Area of Sukarno
Teuku Tajuddin
National Research and Innovation Agency
Syahnada Jaya Syaifullah
National Research and Innovation Agency
Ani Kurniawati
IPB University
Willy Bayuardi Suwarno
IPB University
Sobir Sobir
sobir@apps.ipb.ac.id

IPB University

Short Report

Keywords: Assembly, Genomics, Illumina, Microsatellite, Rodent tubber, SSR data mining

Posted Date: August 12th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7296811/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Genetic Resources and Crop Evolution on December 15th, 2025. See the published version at https://doi.org/10.1007/s10722-025-02661-z.

Abstract

Typhonium flagelliforme, a medicinal plant endemic to Indonesia and belonging to the Araceae family, has garnered significant attention due to its potential anticancer properties. Given its therapeutic relevance, this species represents a promising genetic resource for future plant breeding initiatives. In the present study, whole genome sequencing (WGS) of *T. flagelliforme* was performed using the Illumina NextSeq 2000 platform. Sequencing was conducted with a paired-end 150 bp (PE150) approach, yielding approximately 112 GB of raw data. The estimated genome size was 714.70 Mb, with an assembly contig N50 of 3,971 bp and a BUSCO completeness score of 76.08%. Also, we identified 64.41% repetitive DNA from the genome assembly, in which retroelements occupied 21.40% of the total genome. This first *T. flagelliforme* genome is expected to contribute to a better understanding of its genetics for molecular breeding programs, development of medicinal plant-based biotechnology, and sustainable conservation of rodent tubber germplasm.

Introduction

Indonesia is a megabiodiversity country rich in medicinal plants with great potential to be developed as raw materials for the pharmaceutical industry, phytotherapy, and biotechnology. One of the herbal plants that has long been known to have medicinal properties is the rodent tubber (*Typhonium flagelliforme* Lodd.), a member of the Araceae family. Various studies have shown that the rodent tubber has the potential as an anticancer, antioxidant, antimicrobial, and anti-inflammatory agent, as well as other important pharmacological activities (Mohan et al. 2011; Farida et al. 2014; Mirgane et al. 2020; Septaningsih et al. 2021).

Although the rodent tubber has high bioactive potential, the development of its cultivation and widespread utilization still faces various challenges. One of the main obstacles is the limited genetic information of this plant, which hinders the breeding of superior varieties and the development of derivative products. Previous studies have mostly been in the form of secondary metabolite bioprospecting or biological activity tests, while genomic data for this plant is still very minimal or even not yet comprehensively available.

Next-generation sequencing (NGS) technologies have revolutionized genomics by enabling rapid generation of high-throughput sequencing data (Satam et al. 2023; Panahi et al. 2024; Patwekar et al. 2025). With the advancement of molecular technology, especially Next Generation Sequencing (NGS), more sophisticated and efficient approaches are now available to explore the genetic information of non-model organisms. One of the latest technologies, Whole Genome Sequencing (WGS), allows comprehensive mapping of the entire genome of a species, including previously inaccessible genetic regions such as introns, intergenic regions, and regulatory elements. Sequencing and assembling the genomes of large eukaryotic organisms is still a challenge (Collins 2018; Liao et al. 2019; Blaxter et al. 2022). Genome assembly is a foundational task in genomics, reliant on computational assemblers that vary in algorithmic design and performance under different conditions (Alhakami et al. 2017; Mochizuki

et al. 2023). Optimizing assembler selection remains critical for accurate genome reconstruction. The process of assembling the genome of the rodent tubber was carried out de novo considering the lack of complete information on the genome structure, such as the length, location, and DNA composition of this species.

Using NGS, Yin et al. (2021) assembled the complete genome of *Colocasia esculenta* followed by Pan et al. (2024) in revealing InDel-SSR associated with genes which regulate leaf development. Several members of the Araceae family have had their complete genomes successfully sequenced in recent years. Notable examples include *Metroxylon sagu* (Purwoko et al. 2019), *Amorphophallus titanum* (Frisse et al. 2022), *Amorphophallus konjac* (Li et al. 2023), and *Amorphophallus albus* (Duan el al. 2025), which provide valuable genomic resources for comparative and evolutionary studies within the family. Despite growing genomic resources in Araceae, rodent tuber remains underexplored. Here, we report the first genome sequence of *Typhonium flagelliforme*, a medicinal plant endemic to Indonesia and belonging to the Araceae family, has garnered significant attention due to its potential anticancer properties. The application of WGS to the rodent tubber is expected to provide fundamental information on genome structure, gene annotation, biosynthesis of secondary metabolite compounds, and identification of key genes involved in plant defense mechanisms and production of bioactive compounds. These data are important as a basis for molecular breeding programs, development of medicinal plant-based biotechnology, and sustainable conservation of rodent tubber germplasm.

Materials and methods

Sample collection and DNA extraction

We collected leaves of *Typhonium flagelliforme* from Laboratory for Biotechnology BRIN. Science and Technology Area of BJ Habibie, Serpong, South Tangerang, Banten, Indonesia. Genomic DNA (gDNA) was extracted from whole leaves using Genomic DNA extraction using Quick-DNA Magbead Plus Kit (Zymoresearch, D4082) following the manufacturer's protocol. Initial quantification and purity were performed using Nanodrop 2000 (Thermo Scientific), DNA visualization using 1% TBE agarose gel electrophoresis, quantification accuracy using Qubit dsDNA HS Assay Kits (Thermo Scientific) using Qubit™ Flex Fluorometer, DNA integrity quality check using 4150 TapeStation (Agilent).

DNA sequencing and data quality control

Total extracted gDNA was used as input for library preparation. gDNA was fragmented using an enzymatic method to match the expected insert size. The fragmented DNA was ligated with Illumina-compatible adapters (forward adapter: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA, reverse adapter: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT) with a unique index for each sample. Paired-end libraries (PE) generated from gDNA were then quality checked (QC) using the xGen DNA Library Prep EZ UNI Kit (IDT, 10009822) according to manufacturer's specifications and validated using a 4150 TapeStation (Agilent) and Qubit™ Flex Fluorometer. Libraries that passed QC were sequenced using Illumina NextSeq 2000 for a read length of 300 cycles (PE150) at PT. Genetika Science Indonesia.

Raw reads obtained in fastq format were further subjected to quality control (QC) steps to ensure data quality before further analysis. QC was performed through initial quality checks using FastQC (version 0.12.1) to evaluate parameters of quality distribution per base, quality distribution per read, GC content, adapter contamination and overrepresented sequences. The resulting reads were cleaned (trimmed) using fastp software for advanced quality control, adapter trimming, quality filtering and per-reads quality trimming (Chen et al 2018). The quality of the reads was re-evaluated with fastQC and summarized using multiQC.

Genome Assembly

The resulting high-quality reads were subsequently assembled with SPAdes v3.15.5 (Prjibelski et al. 2020). Data transformation was performed using samtools v1.19.2 (Danecek et al. 2021). Mapping was performed using bwa (-mem) v0.7.17-r1188 (Li, 2013). The quality of the assembled sequences was determined using Quast v5.0.2 (Gurevich et al. 2013) and Qualimap v2.3 (Okonechnikov et al. 2016). The quality and accuracy of the genome assembly were evaluated through two complementary approaches. First, the filtered paired-end Illumina reads were realigned to the assembled contigs using Bowtie2 v2.4.2 (Langmead and Salzberg, 2012), and alignment statistics were analyzed with SamTools v1.7 (Li et al. 2009) to identify potential assembly errors. Second, assembly completeness and gene content coverage were assessed using the BUSCO pipeline (Simão et al. 2015) by comparing against a lineage-specific dataset, enabling evaluation of conserved single-copy ortholog representation within the assembled genome. To assess the contiguity of the genome assembly, the N50 value was calculated, representing the minimum contig or scaffold length needed to cover 50% of the total genome. Assembly quality metrics, including quality value (QV), k-mer error rate, and k-mer completeness, were determined using Merqury v1.3 (Rhie et al. 2020) with a k-mer size of 21.

Genome annotation and SSR data mining

Gene prediction was carried out using the BRAKER pipeline (Hoff et al. 2016), which integrates evidence-based and ab initio approaches for accurate gene model identification. Functional annotation of predicted proteins was performed through BLASTp searches against the UniRef90 (Suzek et al. 2015) and CAZy (Cantarel et al. 2009) databases to identify homologous sequences and assign putative functions, including carbohydrate-active enzymes. Genomic feature visualization was generated using Circos (Krzywinski et al. 2009), providing a circular representation of genome structure and annotation data. To verify taxonomic assignment, BLASTn was used to align genome scaffolds against the NCBI non-redundant nucleotide (nt) database. Finally, genome completeness was assessed using BUSCO (Benchmarking Universal Single-Copy Orthologs) (Simão et al. 2015), allowing evaluation of the representation of conserved orthologous genes.

Repetitive elements within the *Typhonium flagelliforme* genome were identified through a combination of de novo and homology-based approaches. For the de novo prediction, RepeatModeler v2.0.5 (Flynn et al. 2020) was employed with default parameters to construct a species-specific repeat library. This custom library was then used in RepeatMasker v4.1.5 (Tempel, 2012) for repeat annotation. In parallel,

homology-based detection was conducted using RepeatMasker, supported by the Repbase v4.0.7 database (Bao et al. 2015) and RMBlast v2.2.27 (Korf, 2004) as the search engine. The outputs from both approaches were subsequently merged, and a comprehensive summary of repetitive elements was generated using RepeatMasker. Microsatellite sequences (SSRs) were detected across the genome using the MISA tool (http://pgrc.ipk-gatersleben.de/misa/) as described by Beier et al. (2017). The search parameters were configured to identify repeat motifs with a minimum threshold of 12 repeat units for mononucleotides, 6 for dinucleotides, 5 for trinucleotides, and 4 repeat units for tetra-, penta-, and hexanucleotide motifs. These criteria ensure comprehensive detection of both common and rare microsatellite motifs, contributing to the effective characterization of SSR distribution within the genome.

Results

Sequence Data Quality and Filtering Outcomes

A total of 400,390,368 raw reads were generated prior to quality filtering, with an average read length of 151 bp, yielding approximately 60.46 billion nucleotide bases (Table 1). Following the filtering process, 375,783,156 high-quality reads were retained, representing 93.86% of the initial dataset. The average read length slightly decreased to 144 bp, resulting in a total of 54.07 billion high-quality bases. Significant improvements were observed in the quality parameters. The proportion of Q20 bases (indicating an error probability of ≤ 1%) increased from 95.39–97.19%, while Q30 bases (indicating an error probability of ≤ 0.1%) improved from 90.25–92.69%. These metrics indicate that the quality control step effectively enhanced the proportion of high-confidence sequences within the dataset. A slight decrease in GC content was noted, declining from 42.90% in the raw dataset to 42.54% after filtering. This change is likely attributed to the removal of reads originating from genomic regions with extreme GC content. Nevertheless, the reduction is marginal and does not suggest any significant bias or distortion in the overall genomic composition.

Table 1
Short read sequencing statistical data of *T. flagelliforme*

	Before filtering	**After filtering**
Total reads	400,390,368	375,783,156
Mean length (bp)	151	144
Total bases	60,458,946,000	54,074,738,000
Q20 bases	57,672,652,000 **(95.39%)**	52,553,582,000 **(97.19%)**
Q30 bases	54,562,643,000 **(90.25%)**	50,123,213,000 **(92.69%)**
GC content	42.90%	42.54%

Genome Assembly Summary

The genome assembly of *T. flagelliforme* yielded a total of 235,460 contigs, with an aggregate length of 714,696,677 bp (Table 2). Notably, all contigs measured ≥ 1,000 bp, as indicated by the identical values for total contigs and contigs ≥ 1,000 bp, both amounting to 235,460. The longest contig reached 218,692 bp, demonstrating the assembler's capacity to generate long, continuous sequences. The N50 value was calculated at 3,971 bp, with an L50 of 46,390 contigs, indicating that 50% of the assembly is represented by contigs of at least this length. In addition, N90 and L90 values were 1,355 bp and 173,985 contigs, respectively. The area under the N-curve (auN) was 7,583.7, providing a robust summary of contig length distribution. The total sequence length was 714,696,913 bp (≥ 0 bp), closely matching the length of contigs ≥ 1,000 bp, suggesting only a negligible number of ultra-short sequences. The GC content of the assembled genome was determined to be 41.03%, consistent with typical eukaryotic genome composition.

Table 2
BUSCO assessment of *T. flagelliforme*

Parameter	Number
Assembly Statistics	
# contigs	235,460
# contigs (> = 0 bp)	235,461
# contigs (> = 1000 bp)	235,460
Largest contig	218,692
Total length	714,696,677
Total length (> = 0 bp)	714,696,913
Total length (> = 1000 bp)	714,696,677
N50	3,971
N90	1,355
auN	7,583.70
L50	46,390
L90	173,985
GC (%)	41.03
Genome Completeness	
Complete BUSCOs (C)	194 (76.08%)
Complete and single-copy BUSCOs (S)	154 (60.39%)
Complete and duplicated BUSCOs (D)	40 (15.69%)
Fragmented BUSCOs (F)	45 (17.65%)
Missing BUSCOs (M)	16 (6.27%)
Total BUSCO groups searched	255
Merqury	
Quality Value (QV)	38.17
k-mer error rate	1.55449e-04
k-mer completeness (%)	60.731

The completeness of the assembled genome was evaluated using the Benchmarking Universal Single-Copy Orthologs (BUSCO) approach (Simão et al. 2015), which provides a quantitative measure of gene

space completeness based on evolutionarily conserved orthologous genes. A total of 255 BUSCO groups were searched using a plant-specific lineage dataset (Table 2). The results revealed that 194 BUSCOs (76.08%) were identified as complete, indicating the presence of most essential gene components within the assembly. Among these, 154 BUSCOs (60.39%) were detected as complete and single-copy, while 40 BUSCOs (15.69%) were categorized as complete and duplicated, suggesting the existence of some gene duplications, which is a common feature in plant genomes due to polyploidy or segmental duplications. In addition, 45 BUSCOs (17.65%) were classified as fragmented, reflecting the presence of partial gene sequences that may be the result of fragmented assembly or incomplete gene prediction. Only 16 BUSCOs (6.27%) were deemed missing, indicating a small fraction of conserved genes not captured in the current assembly. Quality value (QV) and k-mer completeness evaluated by Merqury, showed a QV of 38.17 and completeness of 60.73%. These results collectively demonstrate that the genome assembly exhibits a high level of completeness, with over three-quarters of essential gene content represented, thereby providing a solid foundation for downstream gene prediction and functional annotation.

Figure 1 shows the k-mer multiplicity spectrum generated using Merqury to assess the quality and completeness of the *T. flagelliforme* genome assembly based on Illumina short-read data. The x-axis represents the k-mer multiplicity (how many times a k-mer appears in the read dataset), and the y-axis shows the total count of k-mers at each multiplicity level. The gray-shaded region on the left corresponds to unique k-mers found only in the raw reads but not present in the assembly, commonly attributed to sequencing errors. The red-shaded region represents k-mers occurring once in the assembly, while blue, green, purple, and orange areas indicate k-mers appearing 2, 3, 4, and more than 4 times, respectively. The spectrum exhibits a unimodal peak, characteristic of a diploid genome, with most valid k-mers peaking at a multiplicity around 6–7. This indicates a good representation of homozygous regions, with a tail suggesting the presence of repetitive or multi-copy regions. The relatively small gray area indicates a low proportion of missing k-mers, reflecting a high level of assembly completeness and low sequencing error. Additionally, the high overlap between k-mers in the reads and assembly supports strong base-level accuracy (QV) of the assembled genome.

Genome annotation and SSR data mining

The genome-wide analysis revealed that interspersed repeats occupy approximately 487,814,671 base pairs, accounting for 64.41% of the entire genome sequence (Table 3). Among these, retroelements were the most prevalent class, comprising 227,861 elements and covering 162,055,080 bp, which corresponds to 21.4% of the genome. Within retroelements, long terminal repeat (LTR) elements dominated, with 174,741 instances occupying 135,777,481 bp (17.93%). Among the LTRs, the Gypsy/DIRS1 superfamily was most abundant, contributing 10.74% of the genome (93,400 elements), followed by Ty1/Copia elements (79,920 elements; 7.08%). Retroviral elements were detected in relatively low numbers (964 elements, 0.05%). Long interspersed nuclear elements (LINEs) were the second most abundant retroelement group, comprising 52,394 elements and spanning 26,108,712 bp (3.45%). The most dominant LINE family was L1/CIN4, with 41,167 elements (2.87%), while RTE/Bov-B

elements contributed 0.58%. No elements were detected from the CRE/SLACS, L2/CR1/Rex, R1/LOA/Jockey, or R2/R4/NeSL LINE subgroups. Short interspersed nuclear elements (SINEs) were rare, consisting of only 726 elements (0.02%), while Penelope elements were minimally present with 180 instances (0.01%). DNA transposons were also present but in smaller proportions than retroelements, with 33,428 elements occupying 21,669,896 bp (2.86%). The hobo-Activator and MULE-MuDR families were the most represented, contributing 0.80% and 1.43% of the genome, respectively. Tourist/Harbinger elements accounted for 0.14%, whereas other families such as En-Spm, PiggyBac, and Transib-related elements were absent. Rolling-circle elements were identified in 559 copies, occupying 469,727 bp (0.06%). A substantial fraction of the genome was composed of unclassified elements, totaling 944,678 sequences and occupying 304,049,562 base pairs, which corresponds to 40.15% of the entire assembly. This high proportion likely reflects the presence of highly diverse or lineage-specific repetitive elements that are not yet represented in existing repeat annotation databases. Additionally, simple sequence repeats (SSRs) and low complexity regions were identified. Simple repeats accounted for 0.90% of the genome (159,299 elements), while low complexity sequences comprised 0.16%. Minor fractions of the genome were occupied by small RNAs (0.01%) and satellite sequences (0.04%).

Table 3
De novo identification of sequence repeats in the genome of *T. flagelliforme*

	number of elements*	length occupied (bp)	percentage of sequence (%)
Retroelements	227861	162055080	21.4
SINEs:	726	168887	0.02
Penelope	180	40133	0.01
LINEs:	52394	26108712	3.45
CRE/SLACS	0	0	0
L2/CR1/Rex	0	0	0
R1/LOA/Jockey	0	0	0
R2/R4/NeSL	0	0	0
RTE/Bov-B	11227	4357489	0.58
L1/CIN4	41167	21751223	2.87
LTR elements:	174741	135777481	17.93
BEL/Pao	0	0	0
Ty1/Copia	79920	53654757	7.08
Gypsy/DIRS1	93400	81326359	10.74
Retroviral	964	415639	0.05
DNA transposons	33428	21669896	2.86
hobo-Activator	12410	6065926	0.8
Tc1-IS630-Pogo	1198	366356	0.05
En-Spm	0	0	0
MULE-MuDR	10303	10832871	1.43
PiggyBac	0	0	0
Tourist/Harbinger	1600	1070282	0.14
Other (Mirage, P-element, Transib)	0	0	0
Rolling-circles	559	469727	0.06
Unclassified:	944678	304049562	40.15
Total interspersed repeats:		487814671	64.41

	number of elements*	length occupied (bp)	percentage of sequence (%)
Small RNA:	359	40911	0.01
Satellites:	1759	302417	0.04
Simple repeats:	159299	6822633	0.9
Low complexity:	22316	1202268	0.16

A total of 97,631 genomic simple sequence repeats (gSSRs) were identified from the analyzed genome assembly. Among these, dinucleotide repeats were the most abundant type, accounting for 60.51% (59,076 SSRs) of the total SSRs detected. This was followed by mononucleotide repeats, which constituted 28.83% (28,146 SSRs). Trinucleotide motifs were present at a moderate level with 8,876 SSRs (9.09%), while tetranucleotide, pentanucleotide, and hexanucleotide motifs were observed at much lower frequencies, comprising 1.16% (1,129 SSRs), 0.24% (233 SSRs), and 0.18% (171 SSRs), respectively (Fig. 2).

Dinucleotide repeats emerged as the most prevalent class of simple sequence repeats (SSRs) in the genome assembly, with a total of 59,076 motifs identified. Among these, the AG/CT motif was the most frequent, comprising 23,173 occurrences, followed by the AT/AT motif with 20,934 instances (Fig. 3). The AC/GT motif was also relatively common, with 14,844 repeats identified. In contrast, the CG/CG motif was extremely rare, detected only 125 times, accounting for less than 0.25% of all dinucleotide SSRs.

The analysis revealed 8,876 trinucleotide SSRs within the genome assembly. Of these, the AAG/CTT motif was the most dominant, occurring 3,121 times and representing a notable share of the total trinucleotide SSRs (Fig. 4). This was followed by CCG/CGG with 1,074 repeats, AAT/ATT with 1,127 repeats, and ACT/ATG with 619 repeats. Other relatively frequent motifs included AGG/CCT (682), AGT/ATC (678), AGC/CGT (399), ACG/CTG (459), AAC/GTT (477), and ACC/GGT (240). These data reflect a moderate level of diversity among trinucleotide SSRs, with a notable bias toward A/T-rich motifs.

Discussion

The draft genome assembly of *Typhonium flagelliforme* presented in this study provides a foundational resource for future genetic and genomic research in this underexplored medicinal plant species. The assembly quality, supported by multiple evaluation metrics such as N50, BUSCO completeness, and k-mer spectrum analysis, reflects a high level of sequence integrity and representation. The predominance of long contigs and the low proportion of missing k-mers, as revealed by Merqury analysis, indicate minimal sequencing errors and strong coverage across the genome.

The results reflect a high-quality preliminary assembly, with the majority of contigs exceeding 1,000 base pairs an essential benchmark, as shorter contigs are typically less informative and may impede downstream genomic analyses (Seitz and Nieselt 2017). The presence of a maximum contig length exceeding 218 kb suggests successful recovery of extended genomic regions that may encompass complete gene structures, including regulatory elements and intergenic sequences (Gurevich et al. 2013). Although chromosome-level contiguity has not yet been achieved, this outcome is notable for a de novo assembly of a plant genome, which often presents challenges due to its complexity and repeat content. The moderate N50 and high L50 values reflect a degree of fragmentation in the assembly, indicating that future efforts could aim to improve contiguity through scaffolding technologies or long-read sequencing. The auN metric, which integrates contig lengths more comprehensively than N50, further supports the observation of moderate contiguity while providing a robust measure for comparing assembly quality across tools and datasets (Bradnam et al. 2013). The GC content of 41.03% falls within the normal range for plant genomes and suggests an absence of major biases or contamination. GC content remains a key factor in assessing sequencing and assembly performance, given its influence on DNA stability, sequencing efficiency, and potential representation biases (Benjamini and Speed 2012). Any significant deviation from expected GC content might indicate technical artifacts or foreign sequence contamination, which was not evident in this dataset.

In this study, a whole-genome shotgun sequencing strategy was employed for *T. flagelliforme*, utilizing exclusively paired-end read libraries for de novo assembly. This approach yielded scaffold N50 values in the moderate range, a result largely attributable to the absence of mate-pair libraries, which are known to enhance scaffold length and assembly contiguity. Paired-end libraries alone are often insufficient to resolve complex repetitive regions in plant genomes, which typically have a high proportion of repetitive sequences (Liao et al. 2023). Previous studies have demonstrated that the incorporation of mate-pair libraries can significantly improve assembly metrics, with N50 values increasing by one to two orders of magnitude (Belova et al. 2013).

Despite this limitation, the scaffold N50 in our *T. flagelliforme* assembly exceeds 20 kb, comparable to those reported for *Cannabis sativa* (Wei et al. 2024) and *Tapiscia sinensis* (Zhao et al. 2020). Moreover, the scaffold N50 achieved here is consistent with those obtained in other plant genomes assembled without mate-pair data (Belova et al. 2013), suggesting that a substantial portion of the non-repetitive genomic regions has been successfully captured. To improve assembly quality in applications such as comparative genomics, whole genome duplication analysis, or evolutionary studies in rodent tubers some future efforts may benefit from the integration of additional long-range information, such as mate-pair libraries, optical or physical maps, or cytogenetic data. Nevertheless, the current assembly provides a valuable resource for characterizing key genomic features of *T. flagelliforme*, including its repeat landscape and gene content.

The k-mer spectrum analysis using Merqury provides critical insights into the quality and completeness of the *T. flagelliforme* genome assembly. The unimodal peak observed at a k-mer multiplicity of approximately 6–7 is characteristic of a diploid genome and indicates a high representation of

homozygous regions. Such patterns are typical of well-assembled diploid plant genomes and support the structural integrity of the current assembly (Rhie et al. 2020; Zimin et al. 2017). A minimal, gray-shaded region, representing k-mers present in the raw reads but absent in the assembly, suggests a low sequencing error rate and high completeness of the assembly. This is critical, as an overrepresentation of such "error" k-mers would indicate substantial sequence loss or misassembly (Vurture et al. 2017). The high degree of overlap between k-mers in the raw reads and those in the assembly confirms successful incorporation of most of the genomic information and supports a high base-level accuracy (QV) (Rhie et al. 2020; Jayakumar and Sakakibara, 2022). The small fraction of high-frequency k-mers (multiplicity > 4), typically corresponding to repetitive or multi-copy elements, suggests that repetitive regions were adequately resolved. Though short-read data can pose limitations in assembling highly repetitive regions (Michael and VanBuren, 2020), the smooth tail and narrow error peak observed here indicate that the assembler managed such sequences effectively. Merqury's k-mer-based approach is particularly advantageous for non-model plant species like *T. flagelliforme*, where reference genomes are not available. It allows robust, reference-free assessment of genome assembly metrics including completeness, base accuracy, and phasing quality (Rhie et al. 2020; Lee et al. 2023). This analysis confirms that the assembly is of sufficient quality to serve as a foundation for downstream applications such as gene annotation, SSR marker mining, and comparative genomics across Araceae.

The high proportion of interspersed repeats (64.41%) in this genome is characteristic of many plant genomes, especially those with large genome sizes and complex evolutionary histories. The predominance of retrotransposons, particularly the Gypsy and Copia LTR families, is a common feature in higher plants and has been linked to genome expansion and adaptation (Ou et al. 2019; Su et al. 2021). Gypsy elements, often found near centromeric regions, have been shown to play a role in chromatin structure and genome stability, whereas Copia elements are more frequently associated with euchromatic and gene-rich regions. The considerable presence of LINEs, especially L1/CIN4 elements, supports findings from other angiosperms, where these elements contribute to structural variations and influence gene expression through insertional mutagenesis (Makarevitch et al. 2021). The near absence of certain LINE families (e.g. L2/CR1/Rex and R1/LOA/Jockey) may reflect lineage-specific loss or silencing mechanisms in this species. The identification of DNA transposons, particularly from the hobo-Activator and MULE-MuDR families, is consistent with their widespread occurrence in plant genomes, where they are often implicated in gene duplication and regulatory evolution (Chuong et al. 2023). The large fraction of unclassified repeats (40.15%) highlights the potential presence of novel or species-specific transposable elements that are not captured in current reference databases, underscoring the need for continued annotation improvement and repeat library curation (Navarro-Muñoz et al. 2020).

The microsatellite density in the *T. flagelliforme* genome, calculated at 345,202 SSR/Mb, offers a useful metric for assessing the abundance of these repetitive elements within the genome. This frequency is particularly informative when compared across species, as variation in SSR density may reflect underlying differences in genome organization, mutation rates, or levels of genetic diversity among Typhonium species (Srivastava et al. 2019; Fischer et al. 2017). The comparatively lower number of SSRs observed in *T. flagelliforme* may be attributed to species-specific evolutionary processes, such as

selective constraints or the effects of genetic drift, which could have contributed to the gradual loss or suppression of these repetitive elements over time (Bagshaw, 2017). Understanding the evolutionary forces and genomic contexts that shape microsatellite abundance will be important for elucidating the genetic diversity, adaptability, and evolutionary history of *T. flagelliforme*. Further comparative studies across Araceae genomes may help clarify these patterns.

The predominance of dinucleotide repeats in the *T. flagelliforme* genome is consistent with findings in other plant species, where AT- and AG-rich motifs frequently occur in intergenic and intronic regions (Chen et al. 2021; Singh et al. 2023). Their abundance is likely driven by replication slippage and relaxed selective pressures in non-coding regions. Among these, the AG/CT and AT/AT motifs were the most common, reinforcing their potential as informative and polymorphic molecular markers (Yadav et al. 2022). Although mononucleotide repeats are also abundant, particularly poly-A/T stretches, they are often excluded from marker development due to sequencing errors and homopolymer-related artifacts (Kumar et al. 2020). In contrast, trinucleotide SSRs, which represented 9.09% of total SSRs, are more stable and frequently located within coding regions. Their preservation of the reading frame makes them suitable for gene-associated marker development (Yadav et al. 2022; Chen et al. 2021). Among trinucleotide motifs, AAG/CTT was the most frequent—consistent with its common presence in untranslated and regulatory regions of plant genomes (Kumar et al. 2020). The high occurrence of CCG/CGG motifs suggests possible roles in gene regulation due to their GC-rich content and localization in coding sequences (Ali et al. 2023). In contrast, motifs such as AAT/ATT and ACT/ATG, while less frequent, are valued for their high mutation rates, making them useful for genetic diversity studies (Singh et al. 2023). The lower frequency of tetra-, penta-, and hexanucleotide repeats aligns with previous SSR profiling studies in plants, which report that longer motif units are less common and likely subject to stronger purifying selection (Basak et al. 2019; Ali et al. 2023). The motif-specific distribution observed in this study supports the utility of di- and trinucleotide SSRs for marker-assisted selection, genetic mapping, and diversity analysis in *T. flagelliforme* and related species.

Conclusion

This study presents a comprehensive draft genome assembly of *Typhonium flagelliforme*, generated through a whole-genome shotgun approach using paired-end reads and SPAdes assembler. The assembly yielded a total length of ~ 714.7 Mb with moderate contiguity (N50 = 3.9 kb), and a GC content of 41.03%, consistent with other eukaryotic plant genomes. Despite the absence of mate-pair libraries, the assembly successfully captured large portions of non-repetitive regions and provides a valuable foundation for further genomic investigation. Repeat analysis revealed that interspersed repeats accounted for 64.41% of the genome, with LTR retrotransposons, especially Gypsy and Copia elements, dominating the repetitive landscape. A substantial number of unclassified elements also suggest the presence of lineage-specific or novel transposable elements. The genome exhibited a microsatellite density of 345,202 SSRs/Mb, with dinucleotide repeats, notably AG/CT and AT/AT motifs, representing the most abundant SSR categories. Trinucleotide SSRs, especially AAG/CTT and CCG/CGG, were also prominent and offer strong potential for gene-associated marker development. These findings suggest

that the assembly sufficiently captures both the genic and intergenic regions, enabling downstream analyses such as gene prediction, marker development, and comparative genomics. Given the limited genomic resources currently available for the genus *Typhonium*, this draft genome provides a valuable platform for understanding species-specific traits and for advancing conservation and breeding efforts.

Declarations

Conflict of interest

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Funding

This research was supported by Join Collaboration Inhouse Biological and Environmental Research Organization BRIN Program with number B-1797/III.5/PR.03.06/6/2024 and B1721/III.5/PR.03.06/7/2025 and RIIM Collaborative Platform Biology Biomolecular Structure Biodiversity Batch I number B-2657/III.5/FR.06.00/6/2024. The authors would like to thank the two anonymous reviewers for their insightful feedback on an earlier draft of this paper.

Author Contribution

Conceptualization, DP, So and TT; methodology, DP and So; software, DP; validation, DP, TT; formal analysis, DP and SZ; investigation, DP; resources, Su and GR; data curation, DP, and SZ; writing—original draft preparation, DP, TT, and So; writing—review and editing, DP, So, TT and SJS; visualization, DP, and SJS; supervision, So, TT, AK and WBS; project administration, DP; funding acquisition, DP. All authors have read and agreed to the published version of the manuscript.

Acknowledgement

Thanks to Applied Botany Research Center and Talent Management BRIN for funding DBR research, laboratories and scholarships.

Data Availability

The whole genome sequencing data was deposited in the Short Read Archive (SRA) database under accession number PRJNA1141733. And will be Release on 2026-09-01 or if the paper has been accepted.

References

1. Alhakami H, Mirebrahim H, Lonardi S (2017) A comparative evaluation of genome assembly reconciliation tools. Genome Biol 18: 93. https://doi.org/10.1186/s13059-017-1213-3
2. Ali F, Hussain A, Khan MA et al (2019) Genome-wide SSR discovery and population structure analysis in chickpea (*Cicer arietinum* L.). Genes 10(9):678. https://doi.org/10.3390/genes10090678
3. Bagshaw ATM (2017) Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes, Genome Biology and Evolution 9(9):2428–2443. https://doi.org/10.1093/gbe/evx164
4. Bao W, Kojima KK, Kohany O (2015) Repbase update, a data base of repetitive elements in eukaryotic genomes. Mob DNA. https://doi.org/10.1186/s13100-015-0041-9
5. Basak M, Uzun B, Yol E (2019) Genetic diversity and population structure of the Mediterranean sesame core collection with use of genome-wide SNPs developed by double digest RAD-Seq. PLoS ONE 14(10): e0223757. https:// doi.org/10.1371/journal.pone.0223757
6. Beier S, Thiel T, Münch T et al (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33:2583–2585. https://doi.org/10.1093/bioinformatics/ btx198
7. Belova T, Zhan B, Wright J, Caccamo M, Asp T, Simková H, Kent M, Bendixen C, Panitz F, Lien S, Doležel J, Olsen OA, Sandve SR (2013) Integration of mate pair sequences to improve shotgun assemblies of flow-sorted chromosome arms of hexaploid wheat. BMC genomics 14:222. https://doi.org/10.1186/1471-2164-14-222
8. Benjamini Y, Speed TP (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic acids research 40(10):e72. https://doi.org/10.1093/nar/gks001
9. Blaxter B, Archibald JM, Childers AK, Coddington JA, Crandall KA, Di Palma F, Durbin R, Edwards SV, Graves JAM, Hackett KJ, Hall N, Jarvis ED, Johnson RN, Karlsson EK, Kress WJ, Kuraku S, Lawniczak MKN, Lindblad-Toh K, Lopez JV, Moran NA, Robinson GE, Ryder OA, Shapiro B, Soltis PS, Warnow T, Zhang G, Lewin HA (2022) Why sequence all eukaryotes? Proc. Natl. Acad. Sci. U.S.A. 119(4):e2115636118. https://doi.org/10.1073/pnas.2115636118 (2022).
10. Bradnam KR, Fass JN, Alexandrov A et al (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2(1):10. https://doi.org/10.1186/2047-217X-2-10
11. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes Database (CAZy): An Expert Resource for Glycogenomics. Nucleic Acids Research, 37(Database issue), D233–D238. https://doi.org/10.1093/nar/gkn663
12. Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics (Oxford, England) 34(17):i884–i890. https://doi.org/10.1093/bioinformatics/bty560
13. Chen Y, Zhang L, Li H, et al (2021) Genome-wide identification and characterization of microsatellites in cultivated peanut (*Arachis hypogaea* L.). BMC Genomics 22:453. https://doi.org/10.1186/s12864-021-07761-z

14. Chuong EB, Elde NC, Feschotte C (2023) Regulatory activities of transposable elements: From conflicts to benefits. Nature Reviews Genetics 24:26–44. https://doi.org/10.1038/s41576-022-00513-7
15. Collins A (2018) The Challenge of Genome Sequence Assembly. The Open Bioinformatics Journal 11:231-239. https://doi.org/10.2174/1875036201811010231
16. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H (2021) Twelve years of SAMtools and BCFtools. GigaScience 10(2):giab008. https://doi.org/10.1093/gigascience/giab008
17. Duan L, Qin J, Zhou G, Shen C and Qin B (2025) Genomic, transcriptomic and metabolomic analyses of *Amorphophallus albus* provides insights into the evolution and resistance to southern blight pathogen. Front. Plant Sci. 15:1518058. doi: 10.3389/fpls.2024.1518058
18. Farida Y, Irpan K, Fithriani L (2014) Antibacterial and antioxidant activity of keladi tikus leaves extract (*Typhonium flagelliforme*) (Lodd) Blume. Procedia Chemistry 13: 209 213. https://doi.org/10.1016/j.proche.2014.12.029
19. Fischer MC, Rellstab C, Leuzinger M, Roumet M, Gugerli F, Shimizu KK, Holderegger R, Widmer A (2017) Estimating genomic diversity and population differentiation - an empirical comparison of microsatellite and SNP variation in Arabidopsis halleri. BMC genomics 18(1):69. https://doi.org/10.1186/s12864-016-3459-7
20. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences 117(17):9451–9457. https://doi.org/10.1073/pnas.1921046117
21. Frisse L, Martinez MA, Pirro S (2022) The Complete Genome Sequence of *Amorphophallus titanum,* the Corpse Flower. Biodiversity genomes 2022:10.56179/001c.37841. https://doi.org/10.56179/001c.37841
22. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics (Oxford, England) 29(8):1072–1075. https://doi.org/10.1093/bioinformatics/btt086
23. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M (2016) BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics (Oxford, England) 32(5):767–769. https://doi.org/10.1093/bioinformatics/btv661
24. Jayakumar V, Sakakibara Y (2022) Comprehensive evaluation of de novo genome assemblies using k-mer-based analysis. BMC Genomics 23:124. https://doi.org/10.1186/s12864-022-08313-9
25. Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 14:5–59. https://doi.org/10.1186/1471-2105-5-59
26. Krzywinski M, Schein J, Birol I et al (2009) Circos: An Information Aesthetic for Comparative Genomics. Genome Research 19(9):1639–1645. https://doi.org/10.1101/gr.092759.109
27. Kumar A, Gahlaut V, Kumar S (2020) Genome-wide analysis and development of SSR markers in wheat for marker-assisted selection. Molecular Biology Reports 47:727–736.

https://doi.org/10.1007/s11033-019-05120-6

28. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/ 10.1038/nmeth.1923
29. Lee H, Baek J, Park J, et al (2023) Benchmarking tools for genome assembly validation using simulated short reads. Briefings in Bioinformatics 24(1):bbac519. https://doi.org/10.1093/bib/bbac519
30. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/ btp352
31. Li, H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv. 1303. https://doi.org/10.48550/arXiv.1303.3997.
32. Li L, Yang M, Wei W, Zhao J, Yu X, Impaprasert R, Wang J, Liu J, Huang F, Srzednicki G, Yu L (2023) Characteristics of *Amorphophallus konjac* as indicated by its genome. Sci Rep 13:22684. https://doi.org/10.1038/s41598-023-49963-9
33. Liao X, Li M, Zou Y, Wu FX, Pan Y, Wang J (2019) Current challenges and solutions of *de novo* assembly. Quant Biol 7:90–109. https://doi.org/10.1007/s40484-019-0166-9
34. Liao X, Zhu W, Zhou J, Li H, Xu X, Zhang B, Gao X (2023) Repetitive DNA sequence detection and its role in the human genome. Communications biology 6(1):954. https://doi.org/10.1038/s42003-023-05322-y
35. Makarevitch I, Waters AJ, Hirsch CD (2021) Transposable elements contribute to stress-responsive gene regulation in plants. Plant Physiology 185(2):400–411. https://doi.org/10.1093/plphys/kiab019
36. Michael TP, VanBuren R (2020) Building near-complete plant genomes. Current Opinion in Plant Biology 54:26–33. https://doi.org/10.1016/j.pbi.2019.12.002
37. Mirgane NA, Chandore A, Shivankar V, Gaikwad Y, Wadhawa GC (2021) Phytochemical study and screening of antioxidant, anti-inflammatory *Typhonium flagelliforme*. Research Journal of Pharmacy and Technology 14: 2686–2690. https://doi.org/10.52711/0974-360X.2021.00474
38. Mochizuki T, Sakamoto M, Tanizawa Y, Nakayama T, Tanifuji G, Kamikawa R, Nakamura Y (2023) A practical assembly guideline for genomes with various levels of heterozygosity. Briefings in Bioinformatics 24(6):bbad337. https://doi.org/10.1093/bib/bbad337
39. Mohan S, Bustamam A, Ibrahim S, Al-Zubairi AS, Aspollah M, Abdullah R, Elhassan MM (2011) In Vitro Ultramorphological Assessment of Apoptosis on CEMss Induced by Linoleic Acid-Rich Fraction from Typhonium flagelliforme Tuber. Evidence-based complementary and alternative medicine : eCAM, 2011, 421894. https://doi.org/10.1093/ecam/neq010
40. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW et al (2020) A computational framework to explore large-scale biosynthetic diversity. Nature Chemical Biology 16:60–68. https://doi.org/10.1038/s41589-019-0400-9
41. Okonechnikov K, Conesa A, García-Alcalde F (2016) Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics (Oxford, England) 32(2):292–294.

https://doi.org/10.1093/bioinformatics/btv566

42. Ou S, Su W, Liao Y, et al (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biology 20:275. https://doi.org/10.1186/s13059-019-1905-y
43. Pan R, Zhu Q, Jia X, Li B, Li Z, Xiao Y, Luo S, Wang S, Shan N, Sun J, Zhou Q, Huang Y (2024) Genome-Wide Development of InDel-SSRs and Association Analysis of Important Agronomic Traits of Taro (*Colocasia esculenta*) in China. Current Issues in Molecular Biology 46(12):13347-13363. https://doi.org/10.3390/cimb46120796
44. Panahi B, Jalaly HM, Hamid R (2024) Using next-generation sequencing approach for discovery and characterization of plant molecular markers.Current Plant Biology 40:100412. https://doi.org/10.1016/j.cpb.2024.100412.
45. Patwekar M, Patwekar F, Badarinath AV, Billah AAM, Gorijavolu V, Krishnan K, Shanmugasundaram P, Prasad PD, Kazi AA (2025) Genomic Sequencing: Techniques, Advancements, and the Path Ahead. *J Bio-X Res.* 8:0046. https://doi.org/10.34133/jbioxresearch.0046
46. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A (2020) Using SPAdes De Novo Assembler. Current protocols in bioinformatics 70(1):e102. https://doi.org/10.1002/cpbi.102
47. Purwoko D, Cartealy IC, Tajuddin T, Dinarti D, Sudarsono S (2019) SSR identification and marker development for sago palm based on NGS genome data. Breeding Science 69(1):1–10. https://doi.org/10.1270/jsbbs.18061.
48. Rhie A, Walenz BP, Koren S, Phillippy AM (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome biology 21(1):245. https://doi.org/10.1186/s13059-020-02134-9
49. Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, Thakare RP, Banday S, Mishra AK, Das G, Malonia SK (2023) Next-Generation Sequencing Technology: Current Trends and Advancements. Biology, 12(7):997. https://doi.org/10.3390/biology12070997
50. Seitz A, Nieselt K (2017) Improving ancient DNA genome assembly. PeerJ 5:e3126. https://doi.org/10.7717/peerj.3126
51. Septaningsih DA, Yunita A, Putra CA, Herawati I, Achmadi SS, Heryanto R, Rafi M (2021) Phenolics profiling and free radical scavenging activity of *Annona muricata*, *Gynura procumbens*, and *Typhonium flagelliforme* leaves extract. Indonesian Journal of Chemistry 21: 1140–1147. https://doi.org/10.22146/ijc.62124
52. Singh P, Sinha P, Tiwari R (2023) In silico mining and validation of genomic SSR markers in rice using whole genome sequencing data. Scientific Reports 13:11212. https://doi.org/10.1038/s41598-023-38357-5
53. Simão FA, Waterhouse RM, Ioannidis P et al (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210 3212. https://doi.org/10.1093/bioinformatics/btv351

54. Srivastava S, Avvaru AK, Sowpati DT et al (2019) Patterns of microsatellite distribution across eukaryotic genomes. BMC Genomics 20:153. https://doi.org/10.1186/s12864-019-5516-5
55. Su W, Gu X, Peterson T, Zhang Z (2021) Genome-wide analysis of LTR-retrotransposons in plants highlights the ongoing evolution of genomic repeats. Molecular Plant 14(6):874–887. https://doi.org/10.1016/j.molp.2021.03.006
56. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, the UniProt Consortium (2015) UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches. Bioinformatics 31(6):926–932. https://doi.org/10.1093/bioinformatics/btu739
57. Tempel S (2012) Using and Understanding RepeatMasker. In: Bigot, Y. (eds) Mobile Genetic Elements. Methods in Molecular Biology, vol 859. Humana Press. https://doi.org/10.1007/978-1-61779-603-6_2
58. Vurture GW, Sedlazeck FJ, Nattestad M et al (2017) Genom eScope Fast reference-free genome profiling from short reads. Bioinformatics. Oxford University Press, Oxford, pp 2202–2204
59. Wei H, Yang Z, Niyitanga S et al (2024) The reference genome of seed hemp (*Cannabis sativa*) provides new insights into fatty acid and vitamin E synthesis. Plant Communications 5(1):100718. https://doi.org/10.1016/j.xplc.2023.100718
60. Yadav RK, Singh A, Bhandawat A (2022) Development of EST-SSR markers and assessment of genetic diversity in medicinal plants. Frontiers in Plant Science 13:871927. https://doi.org/10.3389/fpls.2022.871927
61. Yin J, Jiang L, Wang L, Han X, Guo W, Li C, Zhou Y, Denton M, Zhang P (2021) A high-quality genome of taro (*Colocasia esculenta* (L.) Schott), one of the world's oldest crops. Mol Ecol Resour. 21: 68-77. https://doi.org/10.1111/1755-0998.13239
62. Zhao P, Xin G, Yan F, *et al* (2020) The de novo genome assembly of *Tapiscia sinensis* and the transcriptomic and developmental bases of androdioecy. Hortic Res 7:191. https://doi.org/10.1038/s41438-020-00414-w
63. Zimin, A. V., Puiu, D., Luo, M. C., *et al* (2017) Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Research 27(5):787–792. https://doi.org/10.1101/gr.213405.116

Figures

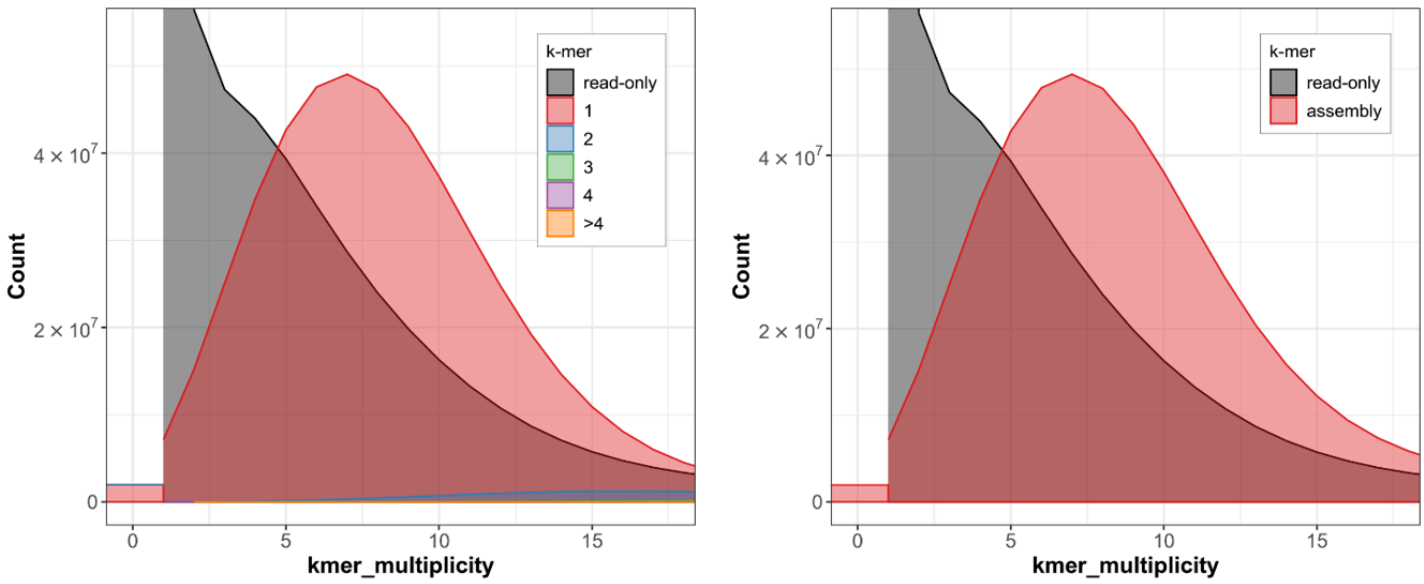


Figure 1

Merqury spectrum plots for *T. flagelliforme* genome assembly. (A) Copy number spectrum plot. (B) Assembly spectrum plot for evaluating K-mer completeness.

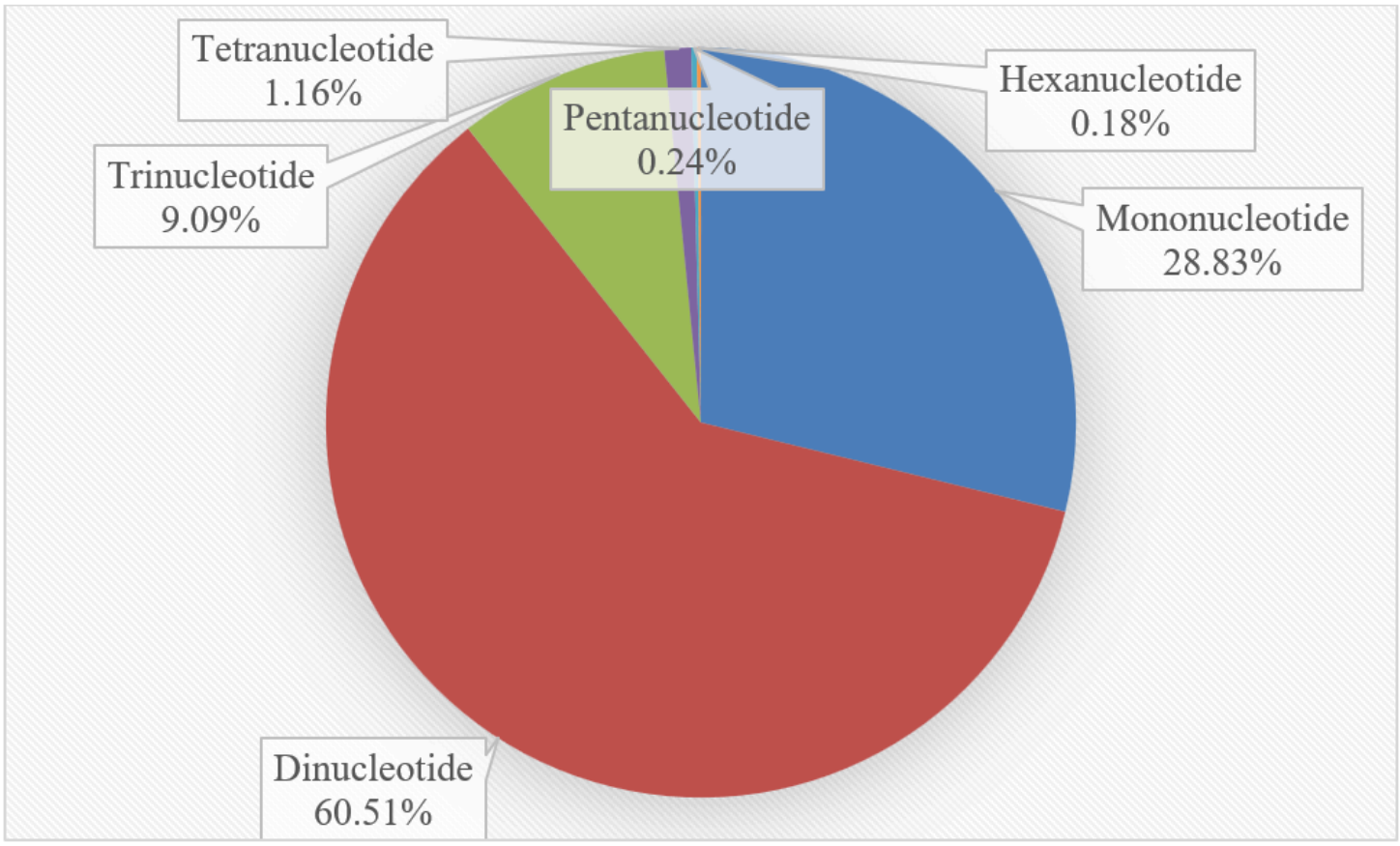


Figure 2

SSR distribution of *T. flagelliforme* genome

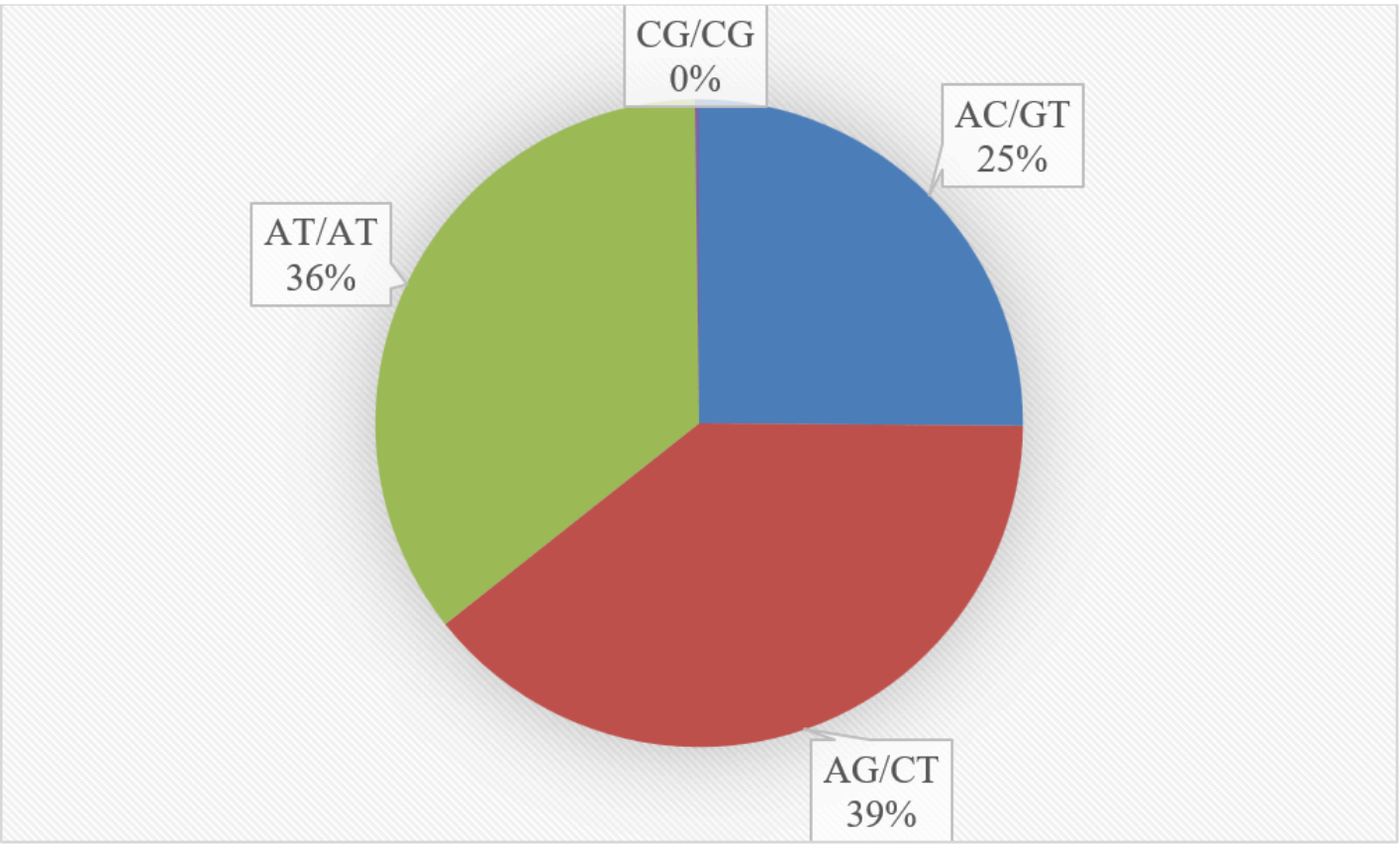


Figure 3

Percentage of different motifs in dinucleotide repeats in *T. flagelliforme* genome

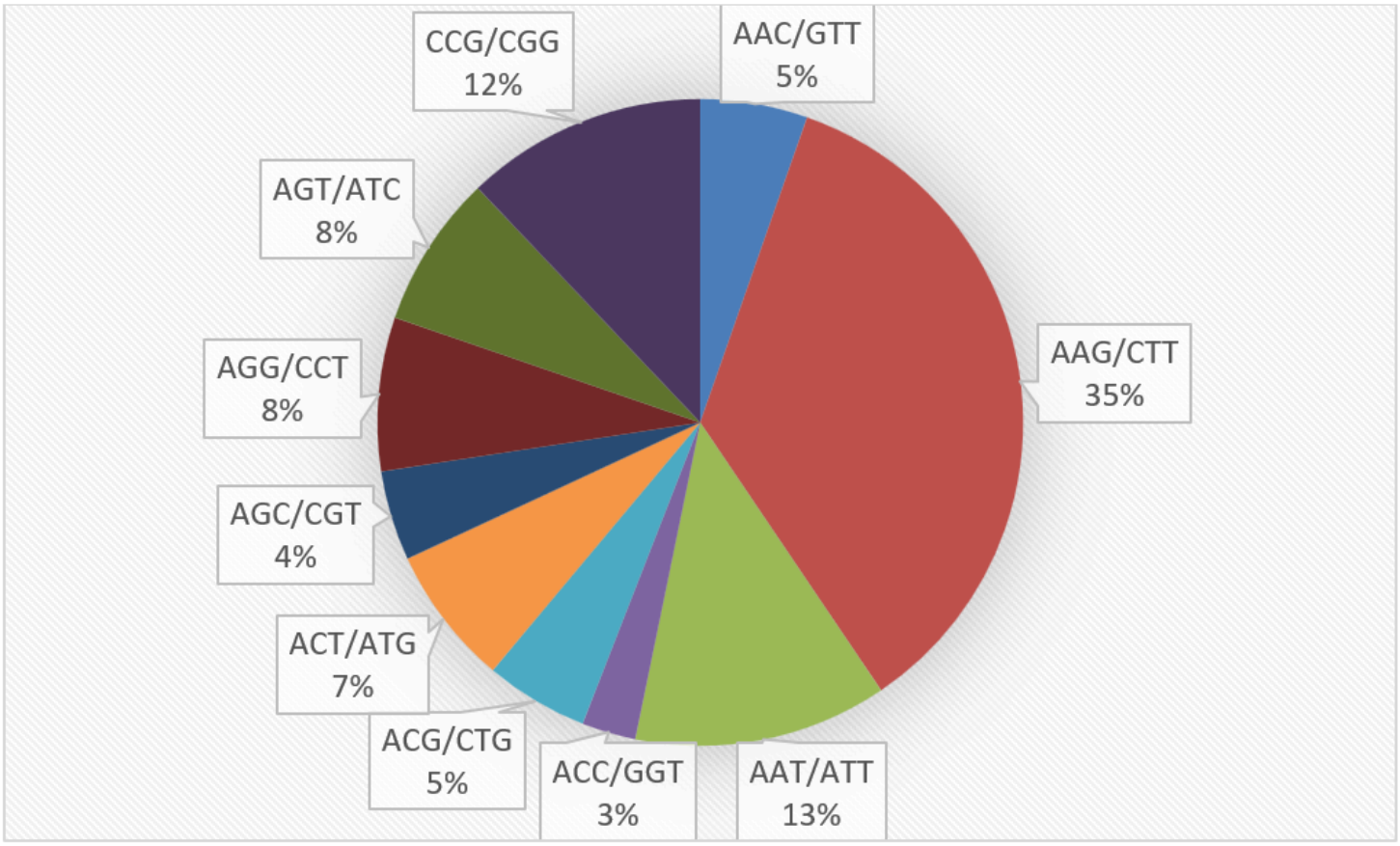


Figure 4

Percentage of different motifs in trinucleotide repeats in *T. flagelliforme* genome