Scientific Data

https://doi.org/10.1038/s41597-026-06554-3

Article in Press

The chromosome-scale genome assembly, annotation of *Bischofia polycarpa* (H. Lév.) Airy Shaw, Phyllanthaceae

Received: 21 November 2024

Accepted: 29 December 2025

Cite this article as: Xin, G., Wang, G., Liu, B. *et al.* The chromosome-scale genome assembly, annotation of *Bischofia polycarpa* (H. Lév.) Airy Shaw, Phyllanthaceae. *Sci Data* (2026). https://doi.org/10.1038/s41597-026-06554-3

Guiliang Xin, Gang Wang, Bobin Liu, Daizhen Zhang, Boping Tang, Chuanyuan Deng & Lie Wang

The chromosome-scale genome assembly, annotation of *Bischofia polycarpa* (H. Lév.) Airy Shaw, Phyllanthaceae

Guiliang Xin[1], Gang Wang[1], Bobin Liu[1], Daizhen Zhang[1], Boping Tang[1], Chuanyuan Deng[2] & Lie Wang[3]*

[1]Jiangsu Key Laboratory for Bioresources of Saline Soils, Yancheng Teachers University, Yancheng 224007, China

[2]College of Landscape Architecture and Art, Fujian Agriculture and Forestry University, Fuzhou 350002, Fujian, China.

[3]Art School, Hunan University of Information Technology, Changsha 410151, Hunan, China.

*Correspondence: Lie Wang (wanglie8610051138@outlook.com)

Bischofia polycarpa (2n = 68), belonging to Phyllanthaceae family, is a native deciduous tree with naturally distribution ranging from southern Qinling Mountains and Huaihe River basin to the northern regions of Fujian and Guangdong, China. It holds significant horticultural, ornamental, and medicinal value and serves as a crucial winter food resource for wild birds. Herein, we report a de novo genome assembly for *B*. *polycarpa*, utilizing a combination of PacBio HiFi Reads and Hi-C data. In total, the genome size reaches 585.68 Mb with a contig N50 of 12.62 Mb, and 99.06% (580.18 Mb) of the assembly successfully anchored on 34 chromosomes. The genome comprises approximately 62.77 % repetitive sequences and 32,554 protein-coding genes, of which 96.15% could be functionally annotated. The BUSCO analysis reveals a genome completeness of 95.42% (n=1,540), including 1,499 (92.87%) single-copy BUSCOs and 41 (2.54%) duplicated BUSCOs. This high-quality genome of the Phyllanthaceae enriches our understanding of the genetic underpinnings of plant reproductive ecology.

Background & Summary

The Phyllanthaceae family, previously classified within the Euphorbiaceae as the subfamily Phyllanthoideae, is distinguished by having two ovules per locule, contrasting with the single ovule per locule found in Euphorbiaceae[1]. Recent molecular phylogenetic analyses have restricted Euphorbiaceae to uniovulate subfamilies, prompting the separation of Phyllanthoideae, now recognized as Phyllanthaceae, within the order Malpighiales[2].

Phyllanthaceae is a predominantly tropical family of shrubs and treelets and rarely of herbs and trees, encompassing 59 genera and over 2000 species. Despite their significant ornamental, ecological, edible, and

medicinal value[3-6], the paucity of genomic data has impeded molecular research, germplasm exploration, and genetic improvement of essential traits. To date, six species, *Phyllanthus cochinchinensis*[7], *Sauropus androgynus*[8], *Phyllanthus emblica and Sauropus spatulifolius*[9], *Baccaurea ramiflora*[10], and *Flueggea virosa*[11], have been genomically characterized.

The genus *Bischofia*, within Phyllanthaceae, includes *Bischofia polycarpa* and *Bischofia javanica*. *B. polycarpa*, a deciduous tree exceeding 15 meters, is renowned for its medicinal and ecological value, as well as its striking autumn foliage. Its fruits, used in wine and oil production, are rich in linolenic, linoleic, and oleic acids, which are crucial in preventing cardiovascular ailments, hypertension, hyperlipidemia, and high cholesterol[12]. Furthermore, these persistent fruits provide a vital food resource for animals (especially the wild-birds) from autumn through winter. Current research on *B. polycarpa* mainly focuses on its phytochemical and biological components, along with its pharmacological applications. However, the lack of reference genome sequences and limited functional genomics studies constrain our understanding of its molecular mechanisms, posing challenges for further research advancement.

In this study, we present a de novo chromosome-level genome assembly of *B*. *polycarpa* using PacBio HiFi sequencing combined with Hi-C technology. The assembled genome size of *B*. *polycarpa* was 585.68 Mb, with a contig N50 of 12.62 Mb. A total of 580.18 Mb (99.06%) of the assembled sequences were anchored to 34 chromosomes with a complete BUSCO score of 95.42%. A total of 32,554 hypothetical proteins were identified through genome annotation and prediction.. This high-quality genome offers invaluable insights, not only enhancing the study of phylogenetic relationships and genetic diversity but also facilitating breeding systems, comparative genetics, and genomic research within Phyllanthaceae.

Methods and results

Sample collection and genomic DNA extraction. For the de novo genome assembly, young leaves were meticulously gathered from individual potted plant clones with the same genetic background in Yangchen, China (37.38°N, 120.20°E; Fig. 1). The collected specimens underwent a sterilization process involving 70% ethyl alcohol followed by a rinse with distilled water. Subsequently, these samples were harvested, swiftly frozen in liquid nitrogen, and preserved at -80°C until DNA extraction. Total genomic DNA was meticulously extracted from young leaves using the conventional cetyltrimethylammonium bromide (CTAB) method[13]. Quantitative PCR (Q-PCR) was conducted to achieve precise library concentration and confirm sufficient library density. DNA quantity and quality were meticulously gauged using Nanodrop (NANODROP2000, Thermo Fisher

Scientific, USA) and Qubit QubitTM3Flurometer (Life Technologies, CA, USA).

Figure 1 goes here

llumina sequencing. For the Illumina short-reads sequencing process, qualifed gDNA was randomly fragmented using the Covaris LE220-plus Focused-ultrasonicator (Covaris, Woburn, MA, USA). Subsequently, the paired-end Illumina sequencing library with an insert size of 350 bp was generated using TruSeq Nano DNA HT sample preparation kit (Illumina) using the following procedures: (a) library construction begins with genomic DNA that is subsequently fragmented; (b) Blunt-end fragments are created; (c) fragments are narrowly size selected with sample purification beads; (d) A-base is added; (e) dual-index adapters are ligated to the fragments and final product is ready for cluster generation; (f) ligated product is amplified and ready for cluster generation. Next, the qualified library sequenced on an Illumina Novaseq 6000 platform (Biomarker Technologies Co., Ltd., Beijing, China) with 150 bp paired-end reads. Raw data were filtered by the fastp v0.20.0 (https://github.com/OpenGene/fastp)[14] using the following strategies: (a) filtered reads with adapters; (b) trimmed reads off low-quality bases at the 5′-and 3′-ends; (c) filtered reads with unknown bases more than 10%; (d) filtered duplicated reads due to polymerase chain reaction amplification; and (e) filtered reads with low-quality bases (Q $\leqslant$ 5) more than 50%. Finally, yielding 110.15 Gb of high-fidelity clean data. These data were subsequently employed for genome size estimation and heterozygosity analysis (Table 1).

Genome survey analysis. The short-reads from illumina platform were quality filtered by fastp[14]using the parameters is -q 10 -u 50 -y -g -Y 10 -e 20 -l 100 -b 150 -B 150. Finally, a total of 110.15 Gb of clean data from the Illumina platform was used for the genome survey. Genome size, heterozygosity and repeat content were estimated from the *k*-mer frequency distribution using jellyfish v2.1.4[15] and Genomescope v2.0[16]. The 21-mer analyses yielded an estimated genome size of 549.22 Mb, with heterozygosity of 0.76% and repeat content of 52.57%, GC content of 34.96% (Fig. 2; Table 1).

Figure 2 goes here

Karyotype analysis. *B. polycarpa* seeds were germinated at room temperature. Once the roots extended to a length of 1.5-2 cm, the root tips were exposed to nitrous oxide for 2.5 hours[17]. The cells were then treated with 8-hydroxyquinoline to stop mitosis at metaphase step, resulting in a substantial number of cells at this stage. The results revealed a total of 34 detected chromosomes, indicating a diploid plant with a karyotype of 2n = 2x = 68 (Fig. 3a-c).

Hi-C sequencing. For Hi-C library construction, fresh leaf tissues (4g) were cross-linked by vacuum infiltration

with 3% formaldehyde at 4°C for 30 min, followed by quenching with 0.375 M glycine for 5 min. After cell lysis, endogenous nucleases were inactivated using 0.3% SDS. The chromatin DNA was digested with 100 U of MboI (New England Biolabs, Ipswich, MA, USA), labeled with biotin-14-dCTP (Invitrogen; Thermo Fisher Scientific, Waltham, MA, USA), and ligated with 50 U of T4 DNA ligase (New England Biolabs). Cross-links were reversed, and ligated DNA was purified using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. The purified DNA (4ug) was sheared into 300-700 bp fragments by Covaris S220 (Covaris Inc, USA), subjected to blunt-end repair, A-tailing, and adapter ligation, and then enriched through biotin-streptavidin pulldown and PCR amplification. After library construction, the Qubit 3.0 and GX platform were used to measure the library concentration and insert size (Insert Size), respectively, and the Q-PCR method was applied to accurately quantify the effective concentration of the library to ensure its quality. The final Hi-C libraries were quantified and sequenced on an Illumina NovaSeq 6000 platform (Illumina, San Diego, CA, USA) based on Sequencing By Synthesis (SBS) in paired-end 150 bp mode at Biomarker Technologies Co., Ltd. (Beijing, China), generating 59.45 Gb of high-quality data, equivalent to approximately 101.51× coverage of the genome. HiC-Pro v2.10.0[18] was used to assess the Hi-C libraries quality.

Figure 3 goes here

Property	**Min**	**Max**
Homozygous (aa)	99.21%	99.26%
Heterozygous (ab)	0.74%	0.79%
Genome Haploid Length	546,553,461 bp	549,224,509 bp
Genome Repeat Length	287,304,616 bp	288,708,695 bp
Genome Unique Length	259,248,845 bp	260,515,814 bp
Model Fit	48.61%	92.07%
Read Error Rate	0.29%	0.29%

Table 1. Genome evaluation of the *B. polycarpa.*

PacBio sequencing and genome assembly. For the PacBio sequencing process, genomic DNA was sheared into ~ 15 kb fragments by Megaruptor®°2. The SMRTbell library was constructed using the SMRTbell Express Template Prep kit 2.0 (Pacifc Biosciences, Menlo Park, CA, USA). The qualified library was sequenced on the

PacBio Sequel II platform (Pacifc Biosciences, Menlo Park, CA, USA) using the circular consensus sequencing (CCS) mode. Briefly, 10 μg of the sheared DNA was carried into the first enzymatic reaction to remove single-stranded overhangs followed by treatment with repair enzymes to repair any damage that may be present on the DNA backbone.After DNA damage repair, ends of the double-stranded fragments were polished and subsequently tailed with an A-overhang at the 3'end. Ligation with T-overhang SMRTbell adapters was performed at 20°C for 60 minutes. Following ligation, the SMRTbell library was digested by exonuclease and purified with 0.45X AMPure PB beads. The size distribution and concentration of the library were assessed using the FEMTO Pulse automated pulsed-field capillary electrophoresis instrument (Agilent Technologies, Wilmington, DE) and the Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). Following library characterization, 3 μg was subjected to a size selection step using the Sage ELF system (Sage Science, Beverly, MA) to collect SMRTbells 15 -18kb. After size selection, the library was purified with 1X AMPure PB beads. Library size and quantity were assessed using the FEMTO Pulse and the Qubit dsDNA HS reagents Assay kit. Sequencing primer and Sequel II DNA Polymerase were annealed and bound, respectively, to the final SMRTbell library. The library was loaded at an on-plate concentration of 55 pM using diffusion loading. SMRT sequencing was performed using a single 8M SMRT Cell on the Sequel II System with Sequel II Sequencing Kit, 1800-minute movies by Pacific Biosciences (USA). After quality control and filtering, the clean dataset comprised 18.06 Gb from 614.38Mb long-reads (Table 1), accounting for ~31.59 × of the genome. The PacBio HiFi reads were initially assembled into contigs using hifiasm v0.16.1-r37526[19] with default parameters. This process yielded an assembly of 585.68 Mb and a CtgN50 length of 12.62 Mb (Fig 3b; Table 2). Notably, the size of the assembled genome is slightly exceeded our initial estimates (~549.22 Mb Mb).

The completeness of the genome assembly was evaluated using Benchmarking Universal Single-Copy Orthologues (BUSCO v5.2.2) against the OrthoDB database, achieving a completeness score of 95.42%. (Table 2)[20]. The accuracy of the draft assembly was further assessed by mapping all short reads to the genome assembly utilizing the BWA-MEM v0.7.17-r1188[21], resulting in a mapping rate and genome coverage of 98.79% and 99.96%, respectively. And using SAMtools with the -M parameter to filter out multiply mapped reads. The refined uniquely mapped reads rate is 95.38% (Table 2).

For chromosome construction, Lachesis[22] was used for preliminary processing of Hi-C data, clustering, sequencing and orientation of contigs, thereby generating preliminary chromosomal assembly results (saved as a .assembly file). Import the assembly results generated by Lachesis (.assembly file) as input into 3D-DNA

pipeline (version v201008) for further optimization. The visualization module of 3D-DNA generates .hic and .assembly files compatible with Juicebox Assembly Tools (JBAT)[23]. These files are imported into JBAT for manual inspection and adjustment to resolve potential misassemblies or misorientations. After manual adjustment in JBAT, the FINAL assembly result (.FINAL.fasta file) is regenerated using the finalize module of the 3D-DNA pipeline (v201008), ultimately producing a chromosome-level genome sequence. Ultimately, approximately 580.18 Mb of genome sequence was anchored to 34 chromosomes, representing 99.06% of the assembled genome size (Fig. 3d, e). These results collectively affirm the high completeness and reliability of our *B*. *polycarpa* genome assembly.

Genome assembly	
Total length of assembly (Mb)	585.68
Number of contigs	162
Contig N50 (Mb)	12.62
chromosome number	34
Scaffold N50 (Mb)	17.18
Number of gaps	60
Anchor rate (%)	99.06
Mapped lllumina reads (%)	90.79
BUSCO (%)	95.42
Mapping rate	98.79
Genome coverage	99.96
Uniquely mapped reads rate	95.38
GC content (%)	36.79
K-mer completeness	74
Genome annotation	
Total length of repeats (Mb)	350.63 Mb
Repeats percentage of assembly (%)	60.43
Number of protein-coding genes	32,554
Average gene length (bp)	3383.16
Average intron length (bp)	1772.05

Average exon length (bp)	1611.12

Table 2. Statistics of the *B. polycarpa* genome assembly and annotation.

Transcriptome sequencing. Four different tissue samples, including leaf, stem, flower, and root, were collected from the same adult plant and used for RNA extraction using RNAplant Plus Reagent according to the manufacturers' instructions (Tiangen, Beijing, China). To ensure optimal library quality, the Qubit® 4.0 Fluorometer (Life Technologies, CA, USA) and Agilent 2100 systems (Agilent Technologies, CA, USA) were employed to assess cDNA concentration and insert size.

Only RNA with good quality could move on to following procedures. Then, qualified RNA were processed for library construction, and the procedures are described as follow: (1) mRNA was isolated by Oligo (dT)-attached magnetic beads. (2) mRNA was then randomly fragmented in fragmentation buffer. (3) First-strand cDNA was synthesized with fragmented mRNA as template and random hexamers as primers, followed by second-strand synthesis with addition of PCR buffer, dNTPs, RNase H and DNA polymerase I. Purification of cDNA was processed with AMPure XP beads. (4) Double-strand cDNA was subjected to end repair. Adenosine was added to the end and ligated to adapters. AMPure XP beads were applied here to select fragments within certain size range. (5) cDNA library was obtained by certain rounds of PCR on cDNA fragments generated from step 4. Sequencing libraries were generated using NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (#E7530L, NEB, USA) following the manufacturer's recommendations and then sequenced on the Illumina NovaSeq 6000 platform for sequencing with read length of PE150, yielding a total of 36.69 M Reads (10.97 Gb) of clean data , with a minimum yield of 10.97 Gb clean data per sample. The percentage of Q30 bases in each sample surpassed 94.40%.

Genome annotation. We used transcriptome data generated in this study to guide gene prediction of assembled genomes. The RNA-Seq transcripts were combined into gene assemblies using PASA v2.3.3[24] that was customized to recognise an additional donor splice site (GA). TransDecoder v5.2.0[25] (https://github.com/TransDecoder/TransDecoder/) was used to predict open reading frames on the PASA assembled transcripts. Complete proteins (CDS with both start and stop codons) predicted by TransDecoder that had valid genome coordinates and more than one exon were retained for further analysis.

The quality of genome annotations among the different gene prediction methods was evaluated using three primary metrics: (1) the mono-exonic (single-exon) and multi-exonic (multiple-exon) ratio, (2) conserved single-copy orthologs queried from the predicted gene models using BUSCO (embryophyta database version 10), and

(3) gene prediction assessment with EnTAP version 0.10.8[26] using a 70% reciprocal functional annotation approach with NCBI's RefSeq Plant and UniProt databases.

Annotation of repetitive sequences. Transposon element (TE) and tandem repeat were annotated by the following workflows. TE were identified by a combination of homology-based and de novo approaches. Initially, a custom de novo repeat library for the genome was created using RepeatModeler2 v2.0.1[27] (http://www.repeatmasker.org/RepeatModeler/), which can automatically deploys two de novo repeat finding programs, including RECON (v1.0.8)[28] and RepeatScout (v1.0.6)[29], RepeatClassifier (v4.12)[30] was employed to classify the predicted results using Dfam (v3.5)[31] known database. Then full-length long terminal repeat retrotransposons (fl-LTR-RTs) were identified using LTRharvest (v1.5.10)[32] and LTR_FINDER (v1.07)[33]. High-quality intact fl-LTR-RTs and a non-redundant LTR library were then refined LTR_retriever v2.9.0[34]. Non-redundant species-specific TE library was constructed by combining the *de novo* TE sequences library above with the known Dfam (v3.5) database. Final TE sequences in the *B. polycarpa* genome were identified and classified by homology search against the library using RepeatMasker v4.1.2[35]. Tandem repeats were annotated by Tandem Repeats Finder(TRF 409)[36] and the MIcroSAtellite identification tool (MISA v2.1)[37] (Table 3).

Type	Number	Length (bp)	Percentage in genome (%)
ClassI:Retroelement	268,246	331,664,537	56.63
ClassI/DIRS	2	127	0
ClassI/LINE	6,801	1,360,154	0.23
ClassI/LTR/Caulimovirus	319	422,447	0.07
ClassI/LTR/Copia	49,853	67,092,021	11.46
ClassI/LTR/ERV	3,090	213,722	0.04
ClassI/LTR/Gypsy	69,233	116,077,402	19.82
ClassI/LTR/Ngaro	489	33,316	0.01
ClassI/LTR/Pao	112	7,083	0
ClassI/LTR/Unknown	136,782	146,297,169	24.98
ClassI/SINE	1,565	161,096	0.03
ClassII:DNA transposon	148,790	35,991,896	6.15
ClassII/CACTA	4,757	273,410	0.05
ClassII/Crypton	91	3,918	0
ClassII/Dada	764	36,975	0.01
ClassII/Ginger	191	8,589	0
ClassII/Helitron	24,853	7,760,570	1.33

ClassII/IS3EU	250	13,072	0
ClassII/Kolobok	954	62,137	0.01
ClassII/Maverick	133	9,611	0
ClassII/Merlin	562	29,510	0.01
ClassII/Mutator	1,468	496,829	0.08
ClassII/P	435	32,379	0.01
ClassII/PIF-Harbinger	1,929	114,072	0.02
ClassII/PiggyBac	263	13,005	0
ClassII/Sola	2	90	0
ClassII/Tc1-Mariner	450	27,107	0
ClassII/Unknown	105,810	25,248,239	4.31
ClassII/Zisupton	1,761	1,624,571	0.28
ClassII/hAT	4,117	237,812	0.04
Unknown	23	1,277	0
Total	417,059	367,657,710	62.77

Table 3. Statistics of repeat sequences in the *B. polycarpa* genome.

Annotation of Protein coding gene. To annotate protein-coding genes, we employed a synthesis of three methodologies: de novo prediction, homology search, and transcript-based assembly. De novo gene models were generated using two ab initio gene-prediction tools, Augustus (v3.1.0)[38] and SNAP (2006-07-28)[39]. For the homolog-based prediction, GeMoMa (v1.7)[40] was utilized with reference gene model from species such as *Euphorbia peplus*, *Hevea brasiliensis*, *Mercurialis annua*, *Manihot esculenta*, *Oryza sativa*, *Ricinus communis*, *Speranskia yunnanensis*, *Triadica sebifera*, and *Vernicia fordii*. For the transcript-based prediction, RNA-sequencing data were mapped to the reference genome using Hisat (v2.1.0)[41] and assembled by Stringtie (v 2.1.4)[42]. GeneMarkS-T (v5.1)[43] were used to predict genes based on the assembled transcripts. The PASA (v2.4.1)[24] software was used to the gene structure annotation [44], and full-length genes were detected by comparing them with reference protein sequences using PASA. Specifically, PASA employs BLAT and GSNAP[45] to align transcript sequences to the genome, enabling structural annotation of genes. The AUGUSTUS v.3.4.0[38] was trained using the full-length gene set for five rounds of optimization. Subsequently, the MAKER2 pipeline (v. 2.31.9)[46] was used for annotation based on ab initio predictions, transcript annotations, and homologous protein evidence. Given the relatively low accuracy of the MAKER2 annotation process, further integration of the MAKER2 and PASA annotations was performed using EVidenceModeler

(EVM, v. 1.1.1)[47] to generate consistent gene annotations. To avoid introducing TE-coding regions, TEsorter (v. 1.4.1)[48] was used to identify the TE protein domains in the genome, and EVM was used to shield them. In addition, PASA was used to optimize the EVM annotation by adding UTR and alternative splicing, and the gene annotations with abnormal coding frames (containing internal termination codons or ambiguous bases, lacking start or termination codons) and those that were too short (<50 aa) were removed. A total of 32,554 protein-coding genes, averaging 3,383.16 bp in length, were predicted in the 585.68 Mb assembled genome (Table 4; Fig. 3f).

Species	***Hevea brasiliensis***	***Oryza sativa***	***Speranskia yunnanensis***	***Mercurialis annua***	***Ricinus communis***
GeneNum	44146	28466	25467	24982	20357
Genelen (bp)	1.73E+08	1.15E+08	72563487	91684367	83066674
AveGenlen (bp)	3907.92	4052.58	2849.31	3670.02	4080.5
ExonLen (bp)	50327414	65452894	30428266	47359610	44641771
AveExonLen (bp)	1140.02	2299.34	1194.81	1895.75	2192.94
ExonNum	226568	179353	133571	150085	142268
AveExonNum	5.13	6.3	5.24	6.01	6.99
CDSLen (bp)	50327414	38272296	30428266	34112085	27634484
AveCDSlen (bp)	1140.02	1344.49	1194.81	1365.47	1357.49
CDSNum	226568	143856	133571	129925	114854
AveCDSNum	5.13	5.05	5.24	5.2	5.64
IntronLen (bp)	1.22E+08	49907785	42135221	44324757	38424903
AveIntronLen (bp)	2767.9	1753.24	1654.5	1774.27	1887.55
IntronNum	182422	150887	108104	125103	121911
AveIntronnum	4.13	5.3	4.24	5.01	5.99
Species	***Manihot esculenta***	***Triadica sebifera***	***Vernicia fordii***	***Euphorbia peplus***	***Bischofia polycarpa***
GeneNum	29523	32240	46827	25473	32554
Genelen (bp)	1.42E+08	2.00E+08	1.51E+08	69154581	1.10E+08
AveGenlen (bp)	4795.75	6191.32	3219.38	2714.82	3383.16
ExonLen (bp)	64216289	57757475	54694694	31148182	52448283
AveExonLen (bp)	2175.13	1791.48	1168.02	1222.79	1611.12
ExonNum	203349	185683	181048	127274	171202
AveExonNum	6.89	5.76	3.87	5	5.26
CDSLen (bp)	40321581	41190649	43230189	31149232	43357386
AveCDSlen (bp)	1365.77	1277.63	923.19	1222.83	1331.86
CDSNum	160571	179348	176687	127275	166233

AveCDSNum	5.44	5.56	3.77	5	5.11
IntronLen (bp)	77368771	1.42E+08	96059314	38006399	57687261
AveIntronLen (bp)	2620.63	4399.83	2051.37	1492.03	1772.05
IntronNum	173826	153443	134221	101801	138648
AveIntronnum	5.89	4.76	2.87	4	4.26

Table 4. Comparison of the protein coding gene annotations.

Functional annotation of protein-coding genes. Gene functions annotation were deduced based on the best alignments with databases such as NCBI Non-Redundant (NR), EggNOG[49], KOG, TrEMBL[50], InterPro, and Swiss-Prot, using diamond blastp (diamond v0.9.29.130)[51] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[52] with an E-value threshold of 1E-3. Protein domainszz were annotated via InterProScan (v5.34-73.0)[53] utilizing InterPro protein databases, and motifs and domains were identified with PFAM databases[54]. Gene Ontology (GO) IDs were assigned based on TrEMBL, InterPro and EggNOG. Approximately 31,302 (about 96.15 %) of the predicted protein-coding genes of 32,554 were functionally annotated with known genes, conserved domains, and GO terms. (Table 5).

Functional annotation	***B. polycarpa***
SwissProt	70.94%
Nr	95.19%
KEGG	74.78%
Pfam	81.32%
eggNOG	82.53%
TrEMBL	95.90%
GO	80.02%
KOG	58.00%
Total	96.15%

Table 5. Statistical analyses of the gene functional annotations of *B. polycarpa* genome.

Annotation of non-coding RNA genes. tRNAscan-SE v2.0.7[55] was used to identify tRNA genes, which serve as adaptor bridging mRNA genetic code with amino acids in proteins. For rRNA identification, barrnap (v0.9)[56] was employed. snoRNAs, guiding chemical modifications of other RNAs, and miRNAs and snRNAs were identified using Infernal (v1.1)[57] against the Rfam (v14.5) database[58]. In total, 626 tRNAs, 875 rRNAs, 78 miRNAs, 123 snRNA and 120 snoRNA were predicted.

Pseudogene prediction Pseudogenes, resembling functional genes but lacking biological function due to

mutations like insertions and deletions, were identified using GenBlastA (v1.0.4)[59] after masking predicted functional genes (Table 6). Candidates were analyzed for premature stop and frame-shift mutations with GeneWise v2.4.1[60].

Pseudogene	Statistic
Total_Number	183
Total_len	1,030,096
Average_Len	5628.94

Table 6. Results of pseudogene prediction

Data Records

The complete genome sequence data, encompassing paired-end short reads, HiFi reads, Hi-C interaction reads, and genome files were deposited in the Genome Warehouse in National Genomics Data Center[61, 62] with the accession number PRJCA031760. The finalized chromosome assembly has been archived with accession number GWHFHGA00000000.1. The RNA-seq data from various tissues are accessible under the same BioProject with accession number PRJCA031768. And the assembly and annotation data were deposited in NCBI GenBank under accession number GCA_053574235.1[63]. RNA-seq data from various tissues are accessible under the BioProject (PRJNA1365770) with accession number SRR36186603[64]. The genomic read data was deposited in NCBI GenBank with accession numbers SRR36589530[65]. The genome annotation files were available in the Figshare databas[66].

Technical Validation

The genome assembly's quality was evaluated through several dimensions:

Genome Completeness: Assessed using CEGMA[67] (integrated with tblastn, genewise, and geneid) and BUSCO (v5.2.2)[20], the analysis revealed that 95.42% of the BUSCO genes were complete in the final assembly, with 92.87% existing as single-copy genes, 2.54% being duplicated, and 0.87% fragmented. We also evaluated the genome assembly quality using the Merqury (v. v1.3) software[68], employing second-generation DNA sequencing data for the assessment. The results exhibit a completeness rate exceeding 74%. To evaluate the assembly integrity and the uniformity of sequencing coverage, the short sequences were aligned to the assembled genome using the BWA software.

Mapping and Coverage: Short reads from the library of *B. polycarpa* were aligned to the assembled genome

using BWA v0.7.17-r1188, achieving a mapping rate of 98.79% and coverage of 99.96%.

Integrity and Coverage: Minimap2[69] was employed to align HiFi reads with the assembled genome, resulting in an impressive mapping rate of 99.57% (Fig. 4a).

Hi-C Interaction Heatmap: The comprehensive genome-wide Hi-C interaction heat map demonstrated superior assembly quality (Fig. 2C). The organization of interaction contacts in and around the chromosomal regions, as illustrated by the Hi-C heatmap, further corroborates the assembly's precision.

Furthermore, the LTR Assembly Index (LAI) was computed using LTR_retriever v2.9.0[70], yielding a score of 12.37, indicative of a reference-quality genome (Fig. 4b).

Figure 4 goes here

Data availability

The finalized chromosome assembly were deposited in NCBI GenBank under BioProject (PRJNA1267844) with accession number GCA_053574235.1. RNA-seq data from various tissues are accessible under the BioProject (PRJNA1365770) with accession numbers SRR36186603. The genome annotation files (GFF3, GTF, FASTA) were available in the Figshare databas. All datasets are publicly available without restriction.

Code availability

All sofware and pipelines were executed in strict accordance with the manuals and protocols provided by the published bioinformatics tools. No custom programming or coding was used.

References

1. Webster GL: **Synopsis of the genera and suprageneric taxa of Euphorbiaceae**. *Annals of the Missouri Botanical Garden* 1994, **81**(1):33-144.
2. GROUP TAP: **An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II**. *Botanical Journal of the Linnean Society* 2003, **141**:399-436.
3. Kawakita A, Kato M: **Diversity of Phyllanthaceae plants**. *Obligate pollination mutualism* 2017:81-115.
4. Mazumdar AB, Chattopadhyay S: **Sequencing, de novo assembly, functional annotation and analysis of Phyllanthus amarus leaf transcriptome using the Illumina Platform**. *Frontiers in Plant Science* 2016, **6**:1199.
5. Ahmad B, Nabia H, Abdur Rauf, Shumaila Bashir, Huang Linfang, Mujeeb-ur Rehman, Mohammad S. Mubarak, Md. Sahab Uddin, Saud Bawazeer, Mohammad Ali Shariati *et al*: **Phyllanthus emblica: a comprehensive review of its therapeutic benefits**. *South African Journal of Botany* 2021, **138**(1):278-310.
6. Rani NZA, Lam KW, Jalil J, Mohamad HF, Ali MSM, Husain K: **Mechanistic studies of the antiallergic activity of Phyllanthus amarus Schum. & Thonn. and its compounds**. *Molecules* 2021, **26**(3):695.
7. Zhang WT, Xu S, Gu Y, Jiao M, Mei Y, Wang J: **The first high-quality chromosome-level genome assembly of Phyllanthaceae (Phyllanthus cochinchinensis) provides insights into flavonoid biosynthesis**. *Planta* 2022, **256**(6):109.
8. Xia FG, Li Bin, Song Kangkang, Yankun W, Zhuangwei H, Haozhen L, Xiaohua Z, Fangping L, Long Y: **Polyploid genome assembly provides insights into morphological development and ascorbic acid accumulation of Sauropus androgynus**. *International journal of molecular sciences* 2024, **25**(1):300.
9. Li F, Hou Z, Xu S, Han D, Li B, Hu H, Liu J, Cai S, Gan Z, Gu Y *et al*: **Haplotype-resolved genomes of octoploid species in Phyllanthaceae family reveal a critical role for polyploidization and hybridization in speciation**. *The Plant Journal* 2024, **119**(1):348-363.
10. Huang J, Chen J, Shi M, Zheng J, Chen M, Wu L, Zhu H, Zheng Y, Wu Q, Wu F: **Genome assembly provides**

insights into the genome evolution of Baccaurea ramiflora Lour. *Scientific Reports* 2024, **14**(1):4867.
11. Chen B-Z, Yang Z-J, Wang W-B, Hao T-T, Yu P-B, Dong Y, Yu W-B: **Chromosome-level genome assembly and annotation of Flueggea virosa (Phyllanthaceae)**. *Scientific Data* 2024, **11**(1):875.
12. Wannamethee SG, Barbara JJ, Lennon L, Papacosta O, Whincup PH, Hingorani AD: **Serum conjugated linoleic acid and risk of incident heart failure in older men: the British Regional heart study**. *Journal of the American Heart Association* 2018, **7**:e006653.
13. Allen GC, Flores-Vergara M, Krasnyanski K, Thompson W: **A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide**. *Nature protocols* 2006, **1**(5):2320-2325.
14. Chen S, Zhou Y, Chen Y, Gu J: **fastp: an ultra-fast all-in-one FASTQ preprocessor**. *Bioinformatics* 2018, **34**(17):i884-i890.
15. Marçais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers**. *Bioinformatics* 2011, **27**(6):764-770.
16. Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes**. *Nature communications* 2020, **11**(1):1432.
17. He Z, Zhang W, Luo X, Huan J: **Five Fabaceae Karyotype and Phylogenetic Relationship Analysis Based on Oligo-FISH for 5S rDNA and (AG3T3)3**. In: *Genes.* vol. 13; 2022.
18. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, Heard E, Dekker J, Barillot E: **HiC-Pro: an optimized and flexible pipeline for Hi-C data processing**. *Genome biology* 2015, **16**:1-11.
19. Cheng H, Concepcion GT, Feng X, Zhang H, Li H: **Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm**. *Nature Methods* 2021, **18**(2):170-175.
20. Simão FA, Waterhouse RM, Panagiotis I, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics* 2015, **31**(19):3210-3212.
21. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
22. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J: **Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions**. *Nature biotechnology* 2013, **31**(12):1119-1125.
23. Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL: **Juicebox.js provides a Cloud-Based Visualization System for Hi-C Data**. *Cell Systems* 2018, **6**(2):256-258.e251.
24. Haas BJ, Delcher AL, Mount SM, Wortman JR, Jr RKS, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD *et al*: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies**. *Nucleic acids research* 2003, **31**(19):5654-5666.
25. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M *et al*: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis**. *Nature Protocols* 2013, **8**(8):1494-1512.
26. Hart AJ, Ginzburg S, Xu M, Fisher CR, Rahmatpour N, Mitton JB, Paul R, Wegrzyn JL: **EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes**. *Molecular ecology resources* 2020, **20**(2):591-604.
27. Flynn JM, Hubley R, Rosen J, Clark AG, Smit AF: **RepeatModeler2 for automated genomic discovery of transposable element families**. *PNAS* 2020, **117**(17):9451-9457.
28. Bao Z, Eddy SR: **Automated de novo identification of repeat sequence families in sequenced genomes**. *Genome Research* 2002, **12**(8):1269-1276.
29. Price AL, Jones NC, Pevzner PA: **De novo identification of repeat families in large genomes**. *Bioinformatics* 2005, **21**(suppl_1):i351-i358.
30. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AFJPotNAoS: **RepeatModeler2 for automated genomic discovery of transposable element families**. 2020, **117**(17):9451-9457.
31. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ: **The Dfam database of repetitive DNA families**. *Nucleic Acids Research* 2016, **44**(D1):D81-D89.
32. Ellinghaus D, Kurtz S, Willhoeft U: **LTRharvest , an efficient and flexible software for de novo detection of LTR retrotransposons**. *BMC Bioinformatics* 2008, **9**:18.
33. Xu Z, Wang H: **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons**. *Nucleic Acids Research* 2007, **35**:W265-W268.
34. Ou S, Jiang N: **LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons** *Plant Physiology* 2018, **176**(2):1410-1422.
35. Tarailo-Graovac M, Chen NS: **Using RepeatMasker to identify repetitive elements in genomic sequences**.

Current protocols in bioinformatics 2009, **25**(1):4-10.
36. Benson G: **Tandem repeats finder: a program to analyze DNA sequences**. *Nucleic Acids Research* 1999, **27**(2):573-580.
37. Beier S, Thiel T, Thomas M, Scholz U, Mascher M: **MISA-web: a web server for microsatellite prediction**. *Bioinformatics* 2017, **33**(16):2583-2585.
38. Stanke M, Diekhans M, Baertsch R, Haussler D: **Using native and syntenically mapped cDNA alignments to improve de novo gene finding**. *Bioinformatics* 2008, **24**(5):637-644.
39. Korf I: **Gene finding in novel genomes**. *BMC Bioinformatics* 2004, **14**(5):59.
40. Jens K, Michael W, Erickson JL, Schattat MH, Jan G, Frank H: **Using intron position conservation for homology-based gene prediction**. *Nucleic acids research* 2016, **44**(9):e89-e89.
41. Kim D, Langmead B, Salzberg SL: **HISAT: A fast spliced aligner with low memory requirements**. *Nature methods* 2015, **12**(4):357-360.
42. Pertea M, Pertea GM, M AC, Cheng CT, Mendell JT, Salzberg SL: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads**. *Nature biotechnology* 2015, **33**(3):290-295.
43. Tang S, A Lomsadze, Borodovsky M: **Identification of protein coding regions in RNA transcripts**. *Nucleic Acids Research* 2015, **43**(12):e78.
44. Grabherr MG, Haas BJ, Yassour M, Levin JZ, others: **Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data**. *Nature Biotechnology* 2013, **29**:644.
45. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ: **GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality**. *Statistical genomics: methods and protocols* 2016:283-334.
46. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell MJGr: **MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes**. 2008, **18**(1):188-196.
47. Haas BJ, Salzberg SL, Zhu W, Mihaela P: **Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments**. *Genome biology* 2008, **9**:1-22.
48. Zhang R-G, Li G-Y, Wang X-L, Dainat J, Wang Z-X, Ou S, Ma Y: **TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes**. *Horticulture Research* 2022, **9**:uhac017.
49. Huerta-Cepas J, Szklarczyk D, Davide H, Hernández-Plaza A, K FS, Helen C, R MD, Ivica L, Thomas R, Juhl JL: **eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses**. *Nucleic acids research* 2019, **47**(D1):D309-D314.
50. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I *et al*: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**. *Nucleic Acids Research* 2003, **31**(1):365-370.
51. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND**. *Nature Methods* 2015, **12**(12):59-60.
52. Kanehisa M, Sato Y, Kawashima M, Mao T: **KEGG as a reference resource for gene and protein annotation**. *Nucleic Acids Research* 2015, **44**(D1):D457-D462.
53. Jones P, Binns D, Chang HY, Fraser M, Li W, Anulla CM, William HM, Maslen J, Mitchell A, Nuka G: **InterProScan 5: genome-scale protein function classification**. *Bioinformatics* 2014, **30**(9):1236-1240.
54. Finn RD, Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R. and Eddy, S.R.: **Pfam: clans, web tools and services**. 2006, **34**(Database issue):D247-251.
55. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence**. *Nucleic acids research* 1997, **25**(5):955-964.
56. Torkel L: **A novel method for predicting ribosomal RNA genes in prokaryotic genomes**. *Lund University* 2017.
57. Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches**. *Bioinformatics* 2013, **29**(22):2933-2935.
58. Griffiths-Jones S, Simon M, Mhairi M, Ajay K, R ES, Alex B: **Rfam: annotating non-coding RNAs in complete genomes**. *Nucleic acids research* 2005, **33**(Database issue):D121-124.
59. She R, Chu JSC, K.Wang, Pei J, Chen N: **genBlastA: Enabling BLAST to identify homologous gene sequences**. *Genome research* 2009, **19**(1):143-149.
60. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise**. *Frontiers in Plant Science* 2004, **14**(5):988.
61. Partners C-NMa: **Database resources of the national genomics data center, China national center for bioinformation in 2024.** *Nucleic acids research* 2024, **52**(D1):D18-D32.

62. Chen M, Ma Y, Wu S, Zheng X, Kang H, Sang J, Xu X, Hao L, Li Z, Gong Z *et al*: **Genome Warehouse: a public repository housing genome-scale data**. *Genomics Proteomics Bioinformatics* 2021, **19**(4):584-589.
63. **NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCA_053574235.1 (2025).**
64. **NCBI GenBank https://identifiers.org/ncbi/insdc.sra:SRR36186603 (2025).**
65. **NCBI GenBank https://identifiers.org/ncbi/insdc.sra:SRR36589530 (2025).**
66. Xin G, Wang G, Liu B, Zhang D, Boping Tang, Deng C, Wang L: **The chromosome-scale genome assembly, annotation of *Bischofia polycarpa* (Levl.) Airy Shaw, Phyllanthaceae.** *Figshare https://doi.org/106084/m9figshare27458694* (2025).
67. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes**. *Bioinformatics* 2007, **23**(9):1061-1067.
68. Rhie A, Walenz BP, Koren S, Phillippy AMJGb: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies**. 2020, **21**:1-27.
69. Li H: **Minimap2: pairwise alignment for nucleotide sequences**. *Bioinformatics* 2018, **31**(18):3094-3100.
70. Ou SJ, Jiang N: **LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons**. *Plant Physiology* 2018, **176**(2):1410-1422.

Author contributions

L. Wang and B.B conceived and designed the study, C.Y. revised the manuscript. G.L. prepared the materials. G.L. and C.Y. analyzed the data and wrote the manuscript. G. Wang, D.Z., B.P., and B.B. edited and improved the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgements

This work was supported by the Program for Young Talents of Science and Technology in Universities of Yancheng Teachers University (grant number: 206670157, and 204670012); Hunan Provincial Natural Science Foundation of China (grant number: 2024JJ5295), and General Project of Philosophy and Social Sciences in Hunan Province (grant number: 22YBA306); Key Scientific Research Projects of Hunan Provincial Education Department (grant number: 24A0751). Thanks to Professor Chenglang Pan from Minjiang University for providing the photographs of *Bischofia*.

Figure legends

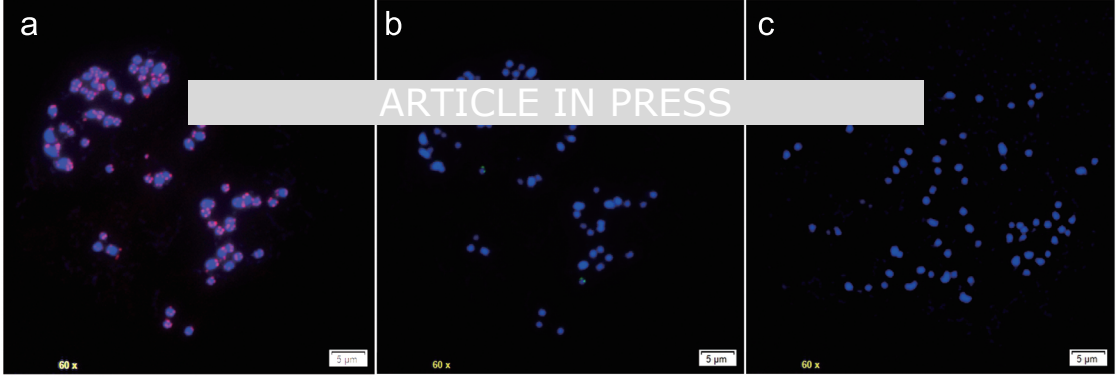
Fig. 1 The morphology of *B. polycarpa* (Levl.) Airy Shaw. (**a**) Inflorescence of female, twigs and leaves. (**c**) Inflorescence of female flower (FF). (**d**) Persistent maturing fruits and developing fruits. (**b**) Male flowers. (**e~f**) The developing male flowers. (**g**) The blooming male flowers and the anther dehiscence.

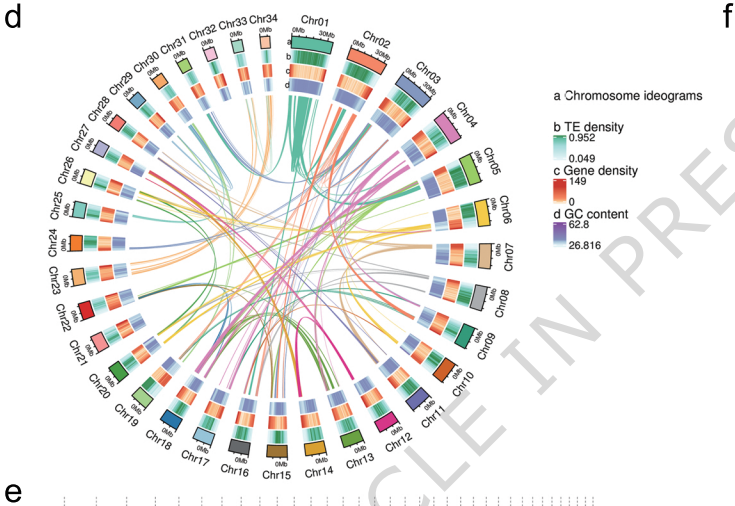
Fig. 2 Genome survey at 21 K-mer of *B. polycarpa* estimated using GenomeScope version 2.0. A *K*-mer distribution map with $k = 21$ was constructed utilizing a 350 bp library data to assess genome size, repeat ratio, and heterozygosity. The primary software employed includes Jellyfish v2.1.4 (parameter: -h 10000000000) and Genomescope v2.0, (parameter: -k 21 -p 2 -m 100000)[16]. The vertical dotted lines indicate the peaks

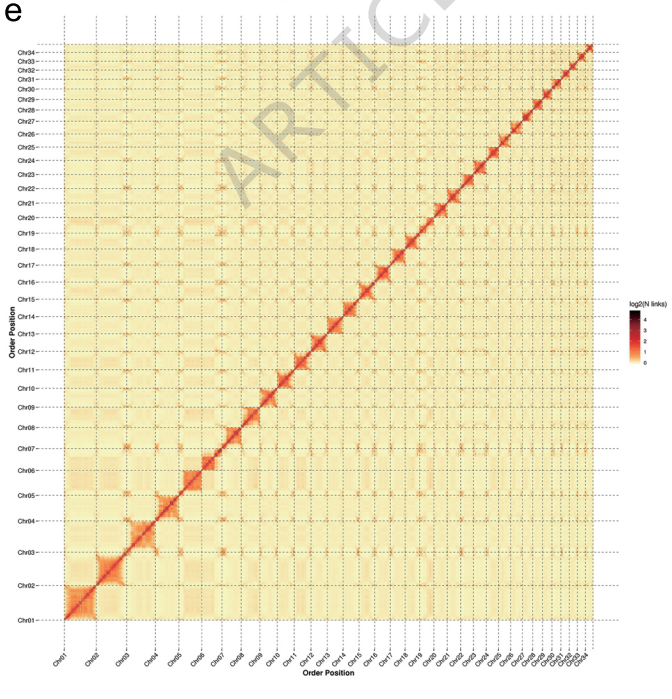
corresponding to the heterozygous, homozygous, and duplicated sequences.

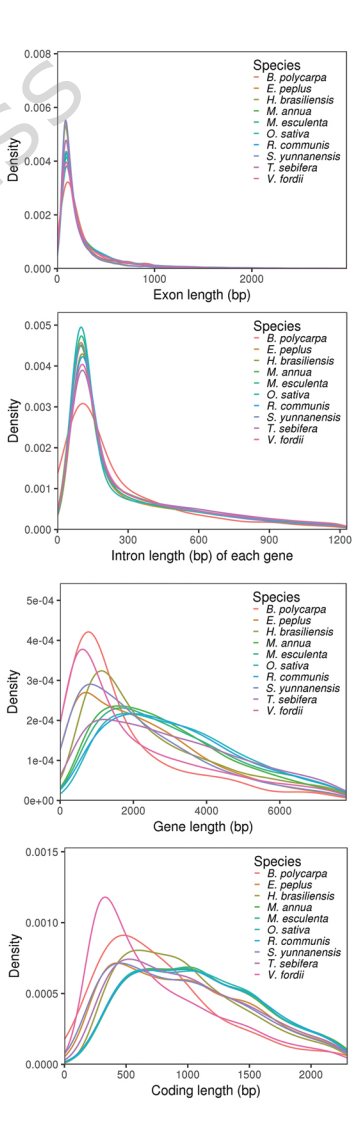
Fig. 3 Karyotype analysis and genome size of *B. polycarpa.* (a) Fluorescence in situ hybridization (FISH) using telomere-specific probes revealed telomeric signals at chromosome ends, confirming a diploid chromosome number of 2n=2x=68. **(b)** FISH analysis with a 5S rDNA repeat sequence probe demonstrated two distinct hybridization signals, supporting its diploid status (2n=2x=68). (**c**) Chromosome counting via DAPI counterstaining validated the total chromosome number as 68. (**d**) Distribution of genomic features of *B. polycarpa*. Tracks 'a-d' represent tandem repeat density, LTR Gypsy density, LTR Copia density, TE density, GC content, and gene density, respectively. (**e**) The Hi-C interaction heatmap at chromosome-level. The heatmap indicates that the intra-chromosome interactions (blocks on the diagonal line) are stronger compared to the inter-chromosome interactions. (**f**) Comparisons of the predicted gene models between the *B. polycarpa* genome and other species genomes, including gene length, coding length, extron length, and intron length.

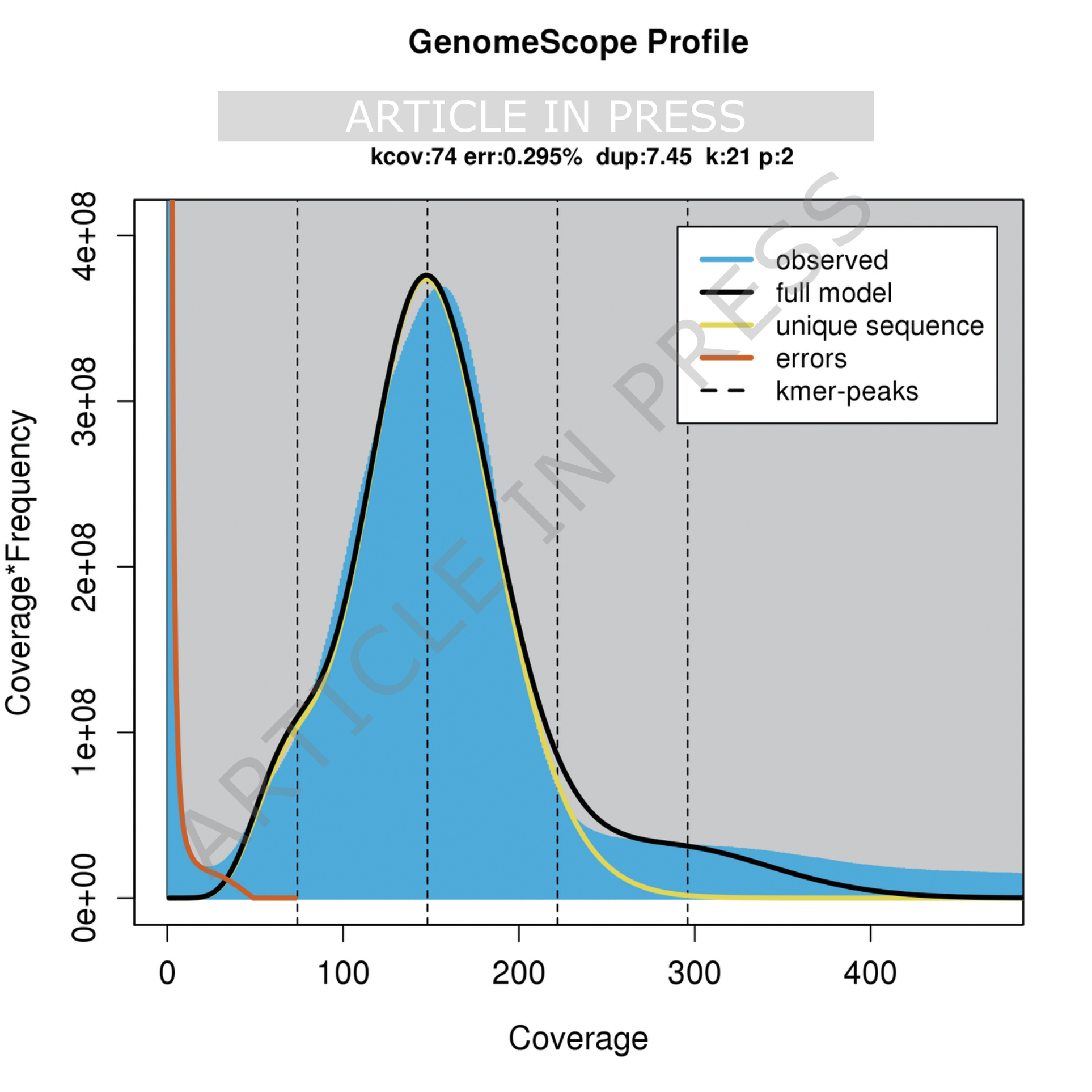
Fig. 4 Quality assessment of the assembled *B. polycarpa* genome. (**a**) The distribution map of reads under varied depths. The horizontal axis represents sequencing depth, while the vertical axis indicates coverage. The distribution closely approximates a Poisson distribution, suggesting high assembly quality. (**b**) The temporal dynamics of long terminal repeat (LTR) retrotransposon bursts in *B. polycarpa* and other representative species.



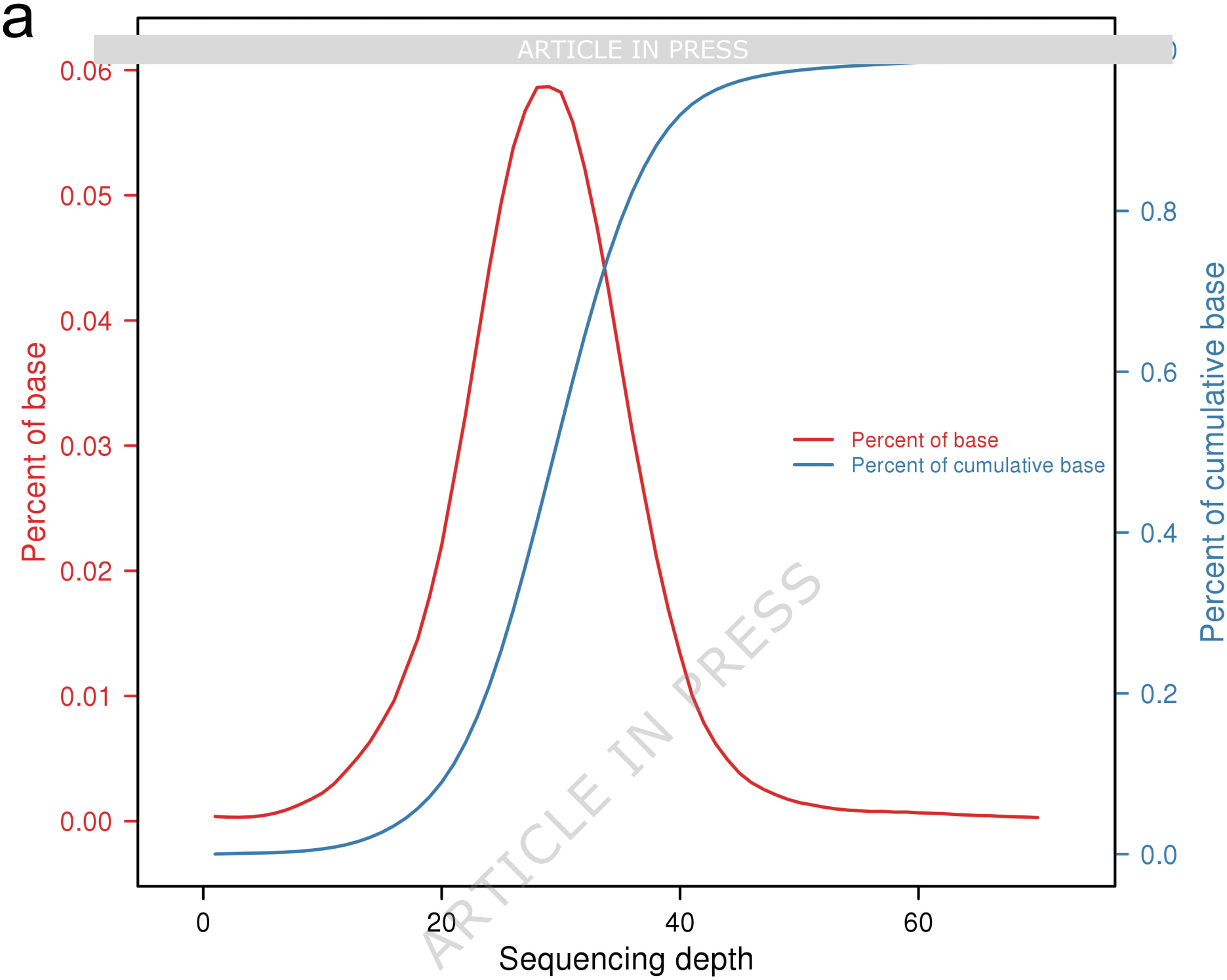







GenomeScope Profile
ARTICLE IN PRESS
kcov:74 err:0.295% dup:7.45 k:21 p:2
observed
full model
unique sequence
errors
kmer-peaks
Coverage*Frequency
0e+00
1e+08
2e+08
3e+08
4e+08
0
100
200
300
400
Coverage
ARTICLE IN PRESS

a



b