Genome assembly of the milky mangrove *Excoecaria agallocha*

Hong Kong Biodiversity Genomics Consortium

Project Coordinator and Co-Principal Investigators: Jerome H.L. Hui[1], Ting Fung Chan[2], Leo L. Chan[3], Siu Gin Cheung[4], Chi Chiu Cheang[5,6], James K.H. Fang[7], Juan Diego Gaitan-Espitia[8], Stanley C.K. Lau[9], Yik Hei Sung[10,11], Chris K.C. Wong[12], Kevin Y.L. Yip[13,14], Yingying Wei[15]

DNA extraction, library preparation and sequencing: Sean T.S. Law[1], Wai Lok So[1]

Genome assembly and gene model prediction: Wenyan Nong[1]

Genome analysis and quality control: Wenyan Nong[1]

Sample collector and logistics: David T.W. Lau[16], Sean T.S. Law[1], Shing Yip Lee[17], Ho Yin Yip[1]

1. School of Life Sciences, Simon F.S. Li Marine Science Laboratory, State Key Laboratory of Agrobiotechnology, Institute of Environment, Energy and Sustainability, The Chinese University of Hong Kong, Hong Kong, China

2. School of Life Sciences, State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Hong Kong SAR, China

3. State Key Laboratory of Marine Pollution and Department of Biomedical Sciences, City University of Hong Kong, Hong Kong SAR, China

4. State Key Laboratory of Marine Pollution and Department of Chemistry, City University of Hong Kong, Hong Kong SAR, China

5. Department of Science and Environmental Studies, The Education University of Hong Kong, Hong Kong SAR, China

6. EcoEdu PEI, Charlottetown, PE, C1A 4B7, Canada

7. Department of Food Science and Nutrition, Research Institute for Future Food, and State Key Laboratory of Marine Pollution, The Hong Kong Polytechnic University, Hong Kong SAR, China

8. The Swire Institute of Marine Science and School of Biological Sciences, The University of Hong Kong, Hong Kong SAR, China

9. Department of Ocean Science, The Hong Kong University of Science and Technology, Hong Kong SAR, China

10. Science Unit, Lingnan University, Hong Kong SAR, China

11. School of Allied Health Sciences, University of Suffolk, Ipswich, IP4 1QJ, UK

12. Croucher Institute for Environmental Sciences, and Department of Biology, Hong Kong Baptist University, Hong Kong SAR, China

13. Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China

14. Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA.

15. Department of Statistics, The Chinese University of Hong Kong, Hong Kong SAR, China
16. Shiu-Ying Hu Herbarium, School of Life Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China
17. Simon F.S. Li Marine Science Laboratory, The Chinese University of Hong Kong, Hong Kong SAR, China

Correspondence on behalf of the consortium: jeromehui@cuhk.edu.hk

Abstract

The milky mangrove *Excoecaria agallocha* is a latex-secreting mangrove that are distributed in tropical and subtropical regions. While its poisonous latex is regarded as a potential source of phytochemicals for biomedical applications, the genomic resources of *E. agallocha* remains limited. Here, we present a chromosomal level genome of *E. agallocha*, assembled from the combination of PacBio long-read sequencing and Omni-C data. The resulting assembly size is 1,332.45 Mb and has high contiguity and completeness with a scaffold N50 of 58.9 Mb and a BUSCO score of 98.4 %. 73,740 protein-coding genes were also predicted. The milky mangrove genome provides a useful resource for further understanding the biosynthesis of phytochemical compounds in *E. agallocha*.

Introduction

The milky mangrove *Excoecaria agallocha* (Euphorbiaceae)(Figure 1F), also known as blind-your-eye mangrove due to its toxic properties causing blindness when its milky latex in contact with eyes, can be found in the brackish water in tropical mangrove forests. In documented human history, this plant has traditionally been used to treat pains and stings from marine organisms, ulcers, as well as leprosy (Vadlapudi et al 2009; Kaliamurthi & Selvaraj 2016); and is also rich in phytoconstituents and potential source of bioactive compounds, such as polyphenols and terpenoids, for biomedical applications (Raghavanpillai Sabu et al., 2022; Sadeer et al., 2023). *E. agallocha* is dioecious, and contrary to typical mangrove species, it does not exhibit specialized aerial roots for gas exchange (Mondal et al 2016; Kader and Sinha 2022). It has a relatively wide distribution globally, including Australia, Bangladesh, India, and Hong Kong. While it has important ecological values in mangroves, such as being the food sources of jewel bugs, genome of this ecologically important species is also lacked.

Context

To date, a few molecular and genomic studies have been conducted on *E. agallocha*. These include a transcriptomic study on the flower sex determination of this dioecious species (Zhou et al., 2022) and the assembly of its chloroplast genome (Shi et al., 2020). However, the genome of this mangrove species remained missing. Previous studies have reported different karyotypes of *E. agallocha*, including 2n = 108 (Das et al., 1999), 2n = 130 (Das et al., 2011) and 2n = 140 (Sidhu, 1961). Its reported chromosome numbers were remarkably different to other species in the same genus, such as *Excoecaria acerifolia* Didr. 2n = 24 (Perry, 1943).

Here, *E. agallocha* has been selected as one of the species to be sequenced under the Hong Kong Biodiversity Genomics Consortium (a.k.a. EarthBioGenome Project Hong Kong),

formed by researchers from 8 publicly funded universities in Hong Kong, in light to provide a useful resource for further understanding of its biology, ecology, evolution, and to set a foundation to carry out any necessary conservation measures .

Methods

Sample collection

Leaf tissues of a male individual of *E. agallocha* were collected at a mangrove sandy shore at Wu Kai Sha, New Territories, Hong Kong in February 2023. The sample was snap-frozen with liquid nitrogen and stored at -80°C until DNA extraction.

High molecular weight DNA extraction

High molecular weight (HMW) genomic DNA isolation was started from grinding 1 g of leaf tissues with liquid nitrogen and performed using NucleoBond HMW DNA kit (Macherey Nagel Item No. 740160.20) with prior CTAB treatment. In brief, around 0.8 g of sample was digested in 5 mL CTAB (Doyle & Doyle, 1987) with addition of 1% PVP for 1 h. After RNAse A treatment, 1/3 volume (~1.6 mL) of 3M potassium acetate was added for contaminant precipitation, followed by two washes of chloroform:IAA (24:1). The resulting supernatant (~4.2 mL) was topped up to 6 ml by adding H1 buffer from NucleoBond HMW DNA kit and continued with the manufacturer's protocol. The resulting DNA was eluted with 80 µL elution buffer (PacBio Ref. No. 101-633-500) and was subject to quality check using the NanoDrop™ One/OneC Microvolume UV-Vis Spectrophotometer, Qubit® Fluorometer, and overnight pulse-field gel electrophoresis.

Pacbio library preparation and sequencing

Prior to library preparation, DNA shearing was performed. Briefly, a dilution of 5 µg HMW DNA in 120 µL elution buffer was transferred to a g-tube (Covaris Part No. 520079) for 6 passes of centrifugation with 1,990 x *g* for 2 min, followed by DNA purification with SMRTbell® cleanup beads (PacBio Ref. No. 102158-300). 2 µL sheared DNA was used to perform overnight pulse-field gel electrophoresis while the remaining sheared DNA was stored in a 2 mL DNA LoBind® Tube (Eppendorf Cat. No. 022431048) at 4 °C overnight. Subsequently, a SMRTbell library was constructed using the SMRTbell® prep kit 3.0 (PacBio Ref. No. 102-141-700), following the manufacturer's protocol. In brief, the sheared DNA was processed with DNA repair and then each DNA strand was polished at both ends and tailed with an A-overhang, followed by ligation of T-overhand SMRTbell adapters. The SMRTbell library was purified using SMRTbell® cleanup beads and 2 µL of eluted sample was subject to quantity assessment using Qubit® Fluorometer and fragment size examination with overnight pulse-field gel electrophoresis. After that, a nuclease treatment was processed to eliminate non-SMRTbell structures and a final size-selection step with 35% AMPure PB

beads was performed to remove short fragments in the library.

The final library preparation for sequencing was performed with The Sequel® II binding kit 3.2 (PacBio Ref. No. 102-194-100). In brief, the SMRT bell structures were annealed and bound with Sequel II® primer 3.2 and Sequel II® DNA polymerase 2.2, respectively. A final cleanup was processed with SMRTbell® cleanup beads, followed by an addition of serial diluted Sequel II® DNA Internal Control Complex. The library was loaded with the diffusion loading mode at an on-plate concentration of 90 pM. The sequencing was performed on the Pacific Biosciences SEQUEL IIe System running for 30-hour movies with 120 min pre-extension to generate HiFi reads. In total, two SMRT cells were used for the sequencing. Details of the resulting sequencing data are listed in Supplementary Information 1.

Omnic-C library preparation and sequencing

A nuclei isolation procedure was performed from 2 g ground leaf tissues, following the modification of Workman et al. (2018) ((https://www.pacb.com/wp-content/uploads/Procedure-checklist-Isolating-nuclei-from-plant-tissue-using-TissueRuptor-disruption.pdf). The resulting nuclei pellet was used to construct an Omni-C library using the Dovetail® Omni-C® Library Preparation Kit (Dovetail Cat. No. 21005) by following the manufacturer's instructions. In brief, the nuclei pellet was resuspended in 4 mL 1X PBS, followed by crosslinking with formaldehyde and DNA digestion with endonuclease DNase I. The concentration and fragment size of the digested lysate was quantified using Qubit® Fluorometer and TapeStation D5000 HS ScreenTape, respectively. Subsequently, both ends of DNA were polished and ligation of biotinylated bridge adaptors were proceeded at 22°C for 30 min. Proximity ligation between crosslinked DNA fragments was conducted at 22°C for 1 hour, followed by crosslink reversal of DNA and then purification with SPRIselect™ Beads (Beckman Coulter Product No. B23317).

End repair and adapter ligation were conducted using the Dovetail™ Library Module for Illumina (Dovetail Cat. No. 21004). In brief, DNA was tailed with an A-overhang and then ligated with Illumina-compatible adapters at 20 °C for 15 min. The Omni-C library was sheared using USER Enzyme Mix and purified with SPRIselect™ Beads. Afterwards, the DNA fragments were isolated using Streptavidin Beads. The DNA library was amplified with Universal and Index PCR Primers from the Dovetail™ Primer Set for Illumina (Dovetail Cat. No. 25005). A final size selection step was done with SPRIselect™ Beads to retain DNA fragments ranging between 350 bp and 1000 bp only. The concentration and fragment size of the library was validated by Qubit® Fluorometer and TapeStation D5000 HS ScreenTape, respectively. The qualified library was eventually sequenced on an Illumina HiSeq-PE150 platform. Details of the resulting sequencing data are listed in Supplementary Information 1.

Genome assembly and gene model prediction

De novo genome assembly was performed using Hifiasm (Cheng et al., 2021), which was then searched against the NT database using BLAST for the input for BlobTools (v1.1.1) (Laetsch & Blaxter, 2017) to identify and remove any possible contaminations (Supplementary Information 2). Haplotypic duplications were removed using "purge_dups" based on the depth of HiFi reads (Guan et al., 2020). Furthermore, proximity ligation data sequenced from the Omni-C library were employed to scaffold the assembly with YaHS (Zhou et al., 2022).

Gene model prediction was run by funannotate (Palmer & Stajich, 2020) using the following parameters "--repeats2evm --protein_evidence uniprot_sprot.fasta --genemark_mode ES --optimize_augustus --organism other --max_intronlen 350000". The predicted gene models from various prediction sources including GeneMark, high-quality Augustus predictions (HiQ), pasa, Augustus, GlimmerHM and snap were combined and processed with Evidence Modeler to produce the annotation files.

Repeat annotation

Annotation of transposable elements (TEs) were conducted using the Earl Grey TE annotation workflow pipeline (version 1.2, https://github.com/TobyBaril/EarlGrey) (Baril et al., 2022).

Macrosynteny analysis

The longest gene transcripts of gene models of *E. agallocha* and four other chromosome-level genomes from the Euphorbiaceae family, namely *Hevea brasiliensis* (GCF_030052815.1; Cheng et al., 2023), *Manihot esculenta* (GCA_033783085.1; Landi et al., 2023), *Ricinus communis* (GCF_019578655.1; Lu et al., 2022) and *Euphorbia peplus* (GCA_028411795.1; Johnson et al., 2023) were extracted for retrieving orthologous gene pairs through reciprocal BLASTp hits (e-value 1e-5) using diamond (v2.0.13) (Buchfink et al., 2021). Macrosynteny was analysed between *E. agallocha* and the other four genomes using MCScanX (Wang et al., 2012) with default parameters.

Results and discussion

Genome assembly of Excoecaria agallocha

A total of 33.20 Gb of HiFi reads from the whole genome of milky mangrove *Excoecaria agallocha* were generated by PacBio sequencing. After scaffolding with 75 Gb Omni-C data, 86.08% of the sequences were assembled into 18 pseudochromosomes (Figure 1C). The assembled genome size was 1,332.45 Mb, with 1,402 scaffolds and a scaffold N50 of 58.95 Mb. The complete BUSCO value was estimated to be 98.4 % (viridiplantae_odb10) (Figure 1B; Table 1). The GC content was 32.17%. A total of 73,740 protein-coding genes

were predicted, with a mean coding sequence length of 288 bp and a BUSCO score of 82.1% (Figure 1B).

Repeat content analysis revealed that transposable elements (TEs) account for 61.2% of the milky mangrove assembly. While half of which were unclassified TEs (31.94%), the remaining TEs were classified into LTR retransposons (22.47%), DNA transposons (5.86%) and LINE retranspsons (Figure 1D; Table 2). SINE and other TEs (Penelope, Rolling circle, simple repeat, microsatellite) contribute less than 0.01% of the genome respectively.

Conclusion and future perspective

The genome assembly of *E. agallocha* presented in this study is the first genomic resource for this mangrove species, which provides a valuable resource for further investigation in the biosynthesis of phytochemical compounds in its milky latex and for the understanding of biology and evolution in genome architecture in the Euphorbiaceae family.

Data validation and quality control

During HMW DNA extraction and Pacbio library preparation, quality control of the sample or library was assessed with the NanoDrop™ One/OneC Microvolume UV-Vis Spectrophotometer, Qubit® Fluorometer, and overnight pulse-field gel electrophoresis. The Omni-C library was validated by Qubit® Fluorometer and TapeStation D5000 HS ScreenTape.

For the genome assembly, BlobTools (v1.1.1) (Laetsch & Blaxter, 2017) was used to identify and remove any possible contaminations (Supplementary Information 2). Benchmarking Universal Single-Copy Orthologs (BUSCO, v5.5.0) (Manni et al., 2021) was run with a collection of single-copy orthologs dataset for the Viridiplantae (Viridiplantae Odb10) to validate the completeness of the genome assembly and gene annotation (Table 1; Supplementary Information 3).

Disclaimer

The genomic data generated in this study was not assessed for the potential level of polyploidy.

Data availability

The raw reads generated in this study were deposited in the NCBI database under the SRA accession SRR24631716 and SRR26908863. The genome, genome annotation and repeat annotation files were made publicly available in Figshare (https://figshare.com/s/53162fa238f429c7fdae).

Authors' contribution

JHLH, TFC, LLC, SGC, CCC, JKHF, JDG, SCKL, YHS, CKCW, KYLY and YW conceived and supervised the study; STSL, DTWL and JSYL collected the samples; STSL and WLS carried out DNA extraction, library preparation and genome sequencing; HYY arranged the logistics of samples; WN performed genome assembly and gene model prediction.

Competing interest

The authors declare that they do not have competing interests.

Funding

This work was funded and supported by the Hong Kong Research Grant Council Collaborative Research Fund (C4015-20EF), CUHK Strategic Seed Funding for Collaborative Research Scheme (3133356) and CUHK Group Research Scheme (3110154).

References

1. Baril T, Imrie RM, Hayward A. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. bioRxiv. 2022. https://doi.org/10.1101/2022.06.30.498289
2. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature methods. 2021;18(4):366-8.
3. 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):170-5.
4. Cheng H, Song X, Hu Y, Wu T, Yang Q, An Z, Feng S, Deng Z, Wu W, Zeng X, Tu M. Chromosome□level wild *Hevea brasiliensis* genome provides new tools for genomic□assisted breeding and valuable loci to elevate rubber yield. Plant Biotechnology Journal. 2023;21(5):1058-72.
5. Das AB, Basak UC, Das P. Genetic erosion of wetland biodiversity in Bhitarkanica forest of Orissa, India. Biologia, Bratislava. 1999;54(4):415-22.
6. Das AB, Jena S, Pradhan C, Chand PK. Genetic variability among male populations of a minor mangrove *Excoecaria agallocha* L. as evident by chromosome morphology and DNA markers. The Nucleus. 2011;54:39-47.
7. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical bulletin. 1987.
8. Ellison J, Koedam NE, Wang Y, Primavera J, JinEong O, Yong JW, Nam VN. *Excoecaria agallocha*. The IUCN Red List of Threatened Species 2010: e. T178842A7623953.
9. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36(9):2896-8.
10. Johnson AR, Yue Y, Carey SB, Park SJ, Kruse LH, Bao A, Pasha A, Harkess A, Provart

NJ, Moghe GD, Frank MH. Chromosome-level genome assembly of Euphorbia peplus, a model system for plant latex, reveals that relative lack of Ty3 transposons contributed to its small genome size. Genome Biology and Evolution. 2023;15(3):evad018.

11. Kong W, Wang Y, Zhang S, Yu J, Zhang X. Recent Advances in assembly of plant complex genomes. Genomics, Proteomics & Bioinformatics. 2023;21(3):427-439.
12. Laetsch DR, Blaxter ML. BlobTools: Interrogation of genome assemblies. F1000Research. 2017;6(1287):1287.
13. Landi M, Shah T, Falquet L, Niazi A, Stavolone L, Bongcam-Rudloff E, Gisel A. Haplotype-resolved genome of heterozygous African cassava cultivar TMEB117 (*Manihot esculenta*). Scientific Data. 2023;10(1):887.
14. Liu J, Shi C, Shi CC, Li W, Zhang QJ, Zhang Y, Li K, Lu HF, Shi C, Zhu ST, Xiao ZY. The chromosome-based rubber tree genome provides new insights into spurge genome evolution and rubber biosynthesis. Molecular plant. 2020;13(2):336-50.
15. Lu J, Pan C, Fan W, Liu W, Zhao H, Li D, Wang S, Hu L, He B, Qian K, Qin R. A Chromosome-level Genome Assembly of Wild Castor Provides New Insights into its Adaptive Evolution in Tropical Desert. Genomics, proteomics & bioinformatics. 2022;20(1):42-59.
16. Luo J, Ren W, Cai G, Huang L, Shen X, Li N, Nie C, Li Y, Wang N. The chromosome-scale genome sequence of *Triadica sebifera* provides insight into fatty acids and anthocyanin biosynthesis. Communications Biology. 2022 Aug 4;5(1):786.
17. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Molecular biology and evolution. 2021;38(10):4647-54.
18. Mondal S, Ghosh D, Ramakrishna K. A complete profile on blind-your-eye mangrove *Excoecaria agallocha* L.(Euphorbiaceae): Ethnobotany, phytochemistry, and pharmacological aspects. Pharmacognosy reviews. 2016;10(20):123.
19. Palmer JM, Stajich J. Funannotate v1. 8.1: Eukaryotic genome annotation. Zenodo https://doi. org/10.5281/zenodo. 2020;4054262.
20. Perry BA. Chromosome number and phylogenetic relationships in the Euphorbiaceae. American Journal of Botany. 1943:527-43.
21. Raghavanpillai Sabu K, Sugathan S, Idhayadhulla A, Woldemariam M, Aklilu A, Biresaw G, Tsegaye B, Manilal A. Antibacterial, antifungal, and cytotoxic activity of *Excoecaria agallocha* leaf extract. Journal of Experimental Pharmacology. 2022:17-26.
22. Sadeer NB, Zengin G, Mahomoodally MF. Biotechnological applications of mangrove plants and their isolated compounds in medicine-a mechanistic overview. Critical reviews in biotechnology. 2023;43(3):393-414.
23. Shi C, Han K, Li L, Seim I, Lee SM, Xu X, Yang H, Fan G, Liu X. Complete chloroplast

genomes of 14 mangroves: phylogenetic and comparative genomic analyses. BioMed Research International. 2020;2020.

24. Sidhu SS. Vegetation of East Godavari. InProc. 48th Indian Sci. Congress 1961 (Vol. 3, pp. 358-9).
25. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, Kissinger JC. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic acids research. 2012;40(7):e49.
26. Wöhner TW, Emeriewen OF, Wittenberg AH, Schneiders H, Vrijenhoek I, Halász J, Hrotkó K, Hoff KJ, Gabriel L, Lempe J, Keilwagen J. The draft chromosome-level genome assembly of tetraploid ground cherry (*Prunus fruticosa* Pall.) from long reads. Genomics. 2021;113(6):4173-83.
27. Workman R, Timp W, Fedak R, Kilburn D, Hao S, Liu K. High molecular weight DNA extraction from recalcitrant plant species for third generation sequencing. *Protocol Exchange* (published online 18 June 2018. https://protocolexchange.researchsquare.com/article/nprot-6785/v1).
28. Zhou Y, Hao L, Huang L, Tang X, Zhuo D, Wang LY, Zhang Y. De novo transcriptome assembly reveals characteristics of flower sex determination of *Excoecaria agallocha*. Annals of Forest Science. 2022;79(1):1-2.
29. Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 2023;39(1):btac808.

Table 1. Genome statistic and sequencing information

Table 2. Summary of transposable element annotation

Figure 1. Genomic information of *Excoecaria agallocha*. **A)** Picture of a female *Excoecaria agallocha*; **B)** Genome statistics; **C)** Omni-C contact map of the assembly; **D)** Pie chart and repeat landscape plot of repetitive elements in the assembled genome.

Figure 2. Macrosynteny dot plot comparing the genome of *Exoecaria agallocha* with *Ricinus communis* **(A)**, *Euphorbia peplus* **(B)**, *Hevea brasiliensis* **(C)**, *and Manihot esculenta* **(D)**.

Supplementary Information 1. Summary of genomic sequencing data.

Supplementary Information 2. Genome assembly QC and contaminant/cobiont detection.

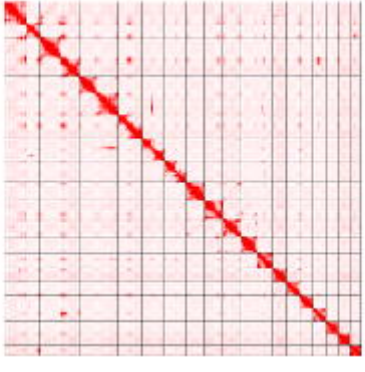
Supplementary Information 3. Information of 18 pseudochromosomes and BUSCO result.

A)

B)

	Fenneropenaeus aguilerichi
Total length (bp)	1,332,447,307
Number	1,403
N_count	0.05%
N50	58,048,051
N50n	8
BUSCO (Genome)	98.40%
Gene models	74,730
Protein coding gene	72,740
BUSCO (Proteome)	92.10%
Repetitive elements	61.24%

C)



D)

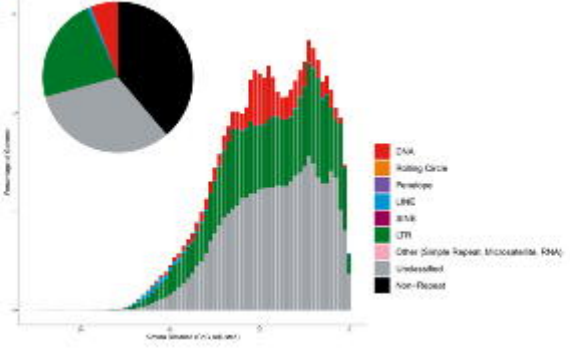


Figure 1