

An efficient repeat masking library for the genomic data of coconut and related trees

Shri Hari Prasad, Deepu Mathew*, Ravisankar Valsalan

Bioinformatics Centre, Kerala Agricultural University, Thrissur – 680 656, India

*Correspondence author: Deepu Mathew, e-mail: deepu.mathew@kau.in

Abstract

Even though repeat masking using custom designed libraries significantly improves the genome annotation and gene prediction, such libraries for palm trees are yet to be designed and made accessible to the researchers. In this study, a repeat library was designed and validated for use in coconut and related palm genomes. Coconut genome with chromosome-level assembly was used to design independent libraries for tall and dwarf ecotypes, which were subsequently merged. Efficiency of the combined *de novo* library in genome annotation and gene prediction was assessed in comparison with the conventional libraries (Dfam+RepBase), using RepeatMasker. *De novo* library had 76.3 % efficiency in coconut genomes compared to 3.51 % in custom libraries and number of genes predicted was reduced from an average of 193,099 to 31,022. In date palm, oil palm and sago palm also, combined library gave higher repeat masking and reduced the number of genes predicted. The *de novo* library can be accessed at <http://www.kau.in/repeat-libraries>.

Keywords: *Elaeis guineensis*, Genome sequencing, NGS, *Areca-ceae*, *Cocos nucifera*

Introduction

Coconut is a very important oil nut in the tropical regions of the world, with 63.6 million tons produced from a global area of 11.3 million ha (Sairam and Jayasekhar, 2018). Similarly, the palm crops dates and oil palm yield 10.0 and 416.3 million tons yield from 1.3 and 28.9 million ha, respectively (FAO, 2023). Though these tree crops support the livelihoods of millions of people around the globe, sufficient genomic research is yet to be done. The genomes of these trees are reported recently and much effort have to be diverted to elucidate their genome organization.

Identification of repetitive sequences is an imperative step in the *de novo* assembly and annotation of the newly sequenced genomes. Even after being an essential step in EST sequence assembly, gene finding and probe design, repeat libraries are available only for few prominent crops. Conventional alternative is masking against the recognized repeats from model organisms, available in transposable/repeat element libraries such as Dfam and RepBase Update. The common tool for identification, classification and masking of repeated regions, including low-complexity sequences as well as interspersed repetitions, is RepeatMasker (<http://repeatmasker.org>), which matches the input genome sequence with libraries such as Dfam and RepBase Update (<http://www.girinst.org/>).

Masking with generic repeat libraries will be less successful compared to that with species-specific libraries (Malde and Jonassen, 2008). Use of TE libraries of model organisms may lead to the missing of important species-specific repeat families, generating erroneous output during homology searches and misrepresenting the evolutionary paths (Kazazian, 2004). Many a times such libraries appear shorter in terms of the update frequency compared to the exponential sequencing programmes. The need for such libraries for newly sequenced genomes is not addressed by RepeatMasker. RepeatModeler (<http://www.repeatmasker.org/RepeatModeler/>) creates reference repeat libraries for the eukaryotic genomes. This study details the development and validation of a repeat library for the genomes of coconut and related palms.

Material and Methods

Genome sequences

The database survey has identified three assemblies of coconut, belonging to 'Hainan Tall' (GCA 008124465.1), 'Catigan Green Dwarf' (GCA 006176705.1) and 'Chowghat Green Dwarf' (GCA 003604295.1).

Gene finding and repeat masking

The processes were carried out using DELL PowerEdge R740xd Rack Server with 2x Intel Xenon silver 4116 processor (@ 2.10GHz × 48), having a hard disk capacity of 4.7 TB, 128 GB RAM, running on CentOS 7. Initially, gene finding was done using Augustus which is a eukaryotic gene prediction tool based on the Hidden Markov Model incorporating a series of established methodologies and sub-models (Stanke and Waack, 2003). Since the predicted number of genes was considerably more than the reported, efficiency of repeat masking was checked using RepeatMasker employing the conventional repeat libraries, Dfam and RepBase Update. Since the repeat masking realized was negligible (~3.51 %), the development of a crop specific library was attempted.

De novo repeat library development

RepeatModeler was used to prepare an exhaustive repeat library for coconut, through an integrative method. The *de novo* repeat identification by RepeatModeler is facilitated by two different discovery methods, RepeatScout (Price et al., 2005) and RECON (Bao and Eddy 2002), which is followed by consensus generation and classification. Modelling was performed on the three coconut genomes to get repeat libraries (consensi.fa.) and repeat classification was carried out in these libraries using RepeatClassifier, a part of the RepeatModeler. The commands followed were:

```
#Create a Database for RepeatModeler, choose your search
engine and input file.
./BuildDatabase -name -engine -dir
#Run RepeatModeler,select the database, choose your search
engine, the number of threads
./RepeatModeler -database -engine -pa -LTRStruct
#if classification step is not carried out automatically, run
RepeatClassifier
./RepeatClassifier -consensi -stockholm -engine -repeatmas-
ker_dir
#to merge the libraries
cat <file_1> <file_2> >outfile_name
#to run RepeatMasker
./RepeatMasker -species name of specie -s -lib -dir out_file path
path_to genomefile.fna
```

Initially, using the classified consensus sequences from tall and dwarf ecotype genomes, repeat masking was attempted in the corresponding genomes. Subsequently, the resultant repeat library (consensi.fa.classified) files of the genomes were merged. The non-redundant library thus generated was used to mask repeats in the genomes using RepeatMasker with the *-s* (sensitive) option, which allowed the masking of low complexity DNA sequences, simple repeats and transposable elements. Likewise, the repetitive elements in a genome can be masked by soft-masking or hard-masking, whereas the latter was followed in this study. In soft-masking, the repetitive elements are masked in lowercase, while rest of the genome is retained in uppercase. While in hard-masking, the repetitive elements are masked with the letter 'N' or 'X'. The masked genome thus prepared was used for gene prediction using Augustus using the combined *de novo* repeat library.

Validation of the repeat library

To check the suitability in other palms, genome assemblies of date palm (PDK50-GCA_000181215.3), oil palm (EO8-GCA_000441515.1) and sago palm (GCA_017589505.1) were retrieved and subjected to repeat masking in RepeatMasker using the *de novo* developed tall, dwarf as well as the combined consensus sequence libraries, with same parameters mentioned above.

Results

Initial attempts for gene finding using Augustus from the 'Hainan Tall', 'Chowghat Green Dwarf' and 'Catigan Green Dwarf' genomes had predicted 214,050, 142,443 and 222,803 genes, respectively. These numbers were higher than 28039, 51953 and 34958 genes reported in these cultivars. Analysis with RepeatMasker employing the conventional Dfam+RepBase library has identified and masked only 4.60, 1.04 and 4.91% of the repeats in these genomes (Table S1). Using RepeatModeler, separate repeat libraries were developed for the tall and dwarf coconut ecotypes and by integrating these, a combined, exhaustive repeat library for coconut was developed. The *de novo* libraries are made freely available to the researchers at <http://www.kau.in/repeat-libraries>.

Using this combined library, 80.52, 66.18 and 82.19 % repetitive elements were identified in 'Hainan Tall', 'Chowghat Green Dwarf' and 'Catigan Green Dwarf' coconut genome assemblies, respectively (Fig. S1). Prediction from the masked genomes yielded 31997, 29952, and 31119 genes, which were comparable with the number of genes reported previously (Fig. 1).

Of the 80.52 % repeats identified in 'Hainan Tall', 62.89 % belonged to LTR elements (Table S2). In 'Catigan Green Dwarf' and 'Chowghat Green Dwarf', 82.19 and 66.19 % of genomes were repetitive elements, with LTR elements representing its 64.93 and 46.89 %, respectively. (Tables S2, S3). With the combined *de novo* library, masking was enhanced by 1.52, 1.13 and 1.02 % compared to the individual libraries in 'Hainan Tall', 'Chowghat Green Dwarf' and 'Catigan Green Dwarf', respectively (Table S3).

In date palm, oil palm and sago palm, the new combined library has masked 34.21, 42.48 and 16.59 % of the repetitive elements (Fig. S2), which was 2.83, 3.46 and 1.73 % higher compared to the mean masking by the individual tall and dwarf libraries, respectively. Gene prediction using the combined library has given 78890, 47076 and 87755 genes, respectively (Tables S4-S6).

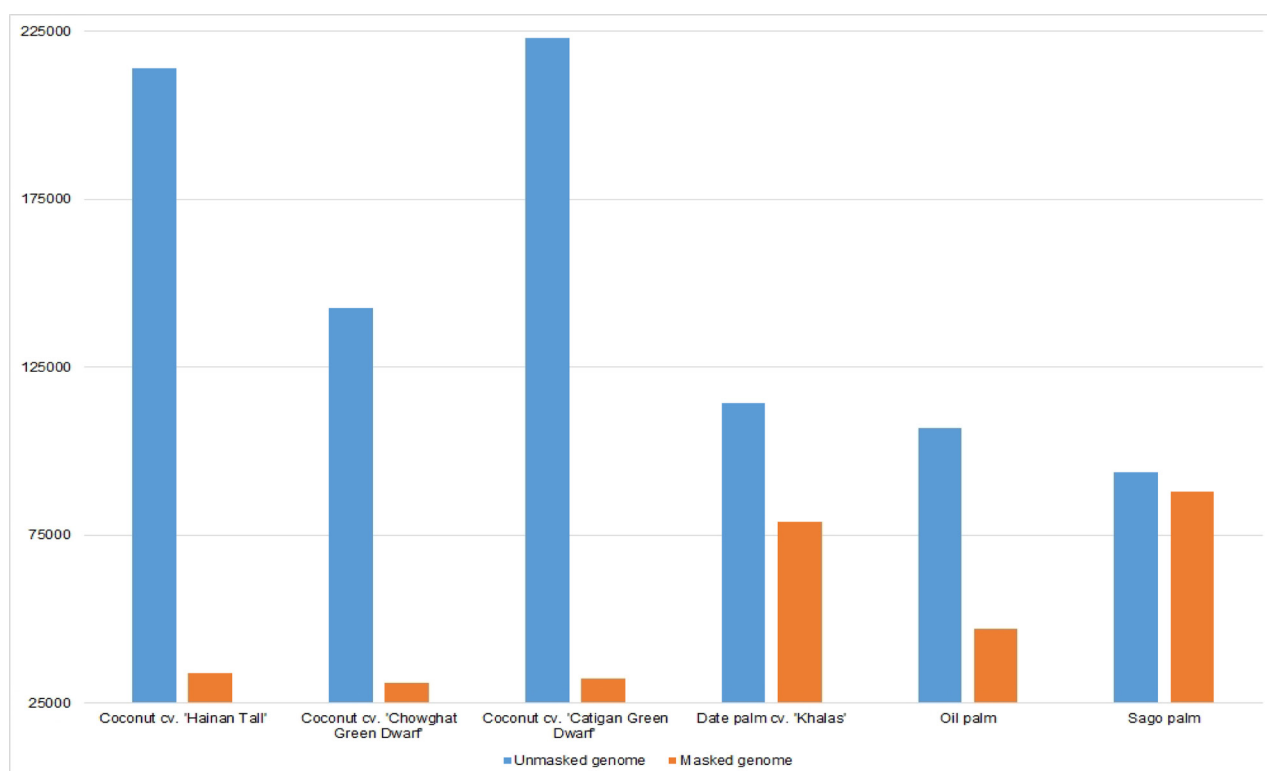


Fig. 1

Gene prediction from the masked and unmasked genomes of coconut, oil palm and sago palm

Discussion

Using RepBase library, Mondal et al. (2018) were able to mask only 19.89 % of the repeats in *Oryza coarctata* genome compared to 36.15 % obtained with the *de novo* library developed using RepeatModeler (Bansal et al., 2020). Similarly, in *Oryza granulate*, custom library has masked 61.98 % repeats (Shi et al., 2020). In this study, masking obtained with the conventional libraries was negligible, necessitating the development of a comprehensive library for coconut genomes. In coconut, Xiao et al. (2017) had identified and masked 74.48 % of repetitive elements in the genome assembly of 'Hainan Tall'. Similarly, Lantican et al. (2019) had identified 78.33 % repetitive elements in the genome assembly of coconut cv. 'Catigan Green Dwarf', in which 60.26 % accounted for LTR elements. Compared to their results, our analysis showed 82.19 % overall masking, with 64.93 % LTR elements, which shows the better efficiency of the repeat library designed. The genome assembly of 'Chowghat Green Dwarf' (Mulyar et al., 2020) had reported 77.29 % repeat masking, of which 58.85 % were LTRs. These results are comparable to the present results of 66.18 % overall masking, comprising 46.89 % LTRs. However, none of these libraries have been released in public domain.

Since the extent of masking in other palm genomes is not reported, direct comparison of masking efficiency was not possible and hence gene prediction results were used. In oil palm,

Singh et al. (2013) identified 158,946 genes compared to 47,076 in this study. Number of genes in date palm reported by Al-Dous et al. (2011) was slightly lower than what we got, recommending crop-wise customisation of library. The draft genome of sago palm is yet to be published.

Conclusions

The *de novo* combined library was effective in masking the repeats in genome assemblies of coconut, oil palm, date palm and sago palm. In coconut assembly, upto 82.19 % repeats were masked compared to an average of 3.51 % by conventional libraries and correspondingly reduced the number of genes predicted. In oil palm, it gave 42.48 % masking and number of genes predicted reduced from 106,886 to 47,076. In date palm, it showed 34.21 % repeat masking, proving efficiency across palm genomes.

References

- Al-Dous EK, George B, Al-Mahmoud ME et al (2011) De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nature Biotechnology* 29(6):521-7. <https://doi.org/10.1038/nbt.1860>

- Bansal J, Gupta K, Rajkumar MS, Garg R, Jain M (2020) Draft genome and transcriptome analyses of halophyte rice *Oryza coarctata* provide resources for salinity and submergence stress response factors. *Physiologiae Plantarum* 173(4):1309-1322. <https://doi.org/10.1111/ppl.13284>
- Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research* 12(8):1269-76. <https://doi.org/10.1101/gr.88502>
- FAO (2023) FAOSTAT, Food and Agriculture Organization of the United Nations, Rome, Italy, available at <https://www.fao.org/faostat/en/#data/QCL>
- Kazazian HH (2004) Mobile elements: drivers of genome evolution. *Science* 303(5664):1626-32. <https://doi.org/10.1126/science.1089670>
- Lantican DV, Strickler SR, Canama AO et al (2019) De novo genome sequence assembly of dwarf coconut (*Cocos nucifera* L.'Catigan Green Dwarf') provides insights into genomic variation between coconut types and related palm species. *G3: Genes Genomes Genetics* 9(8):2377-93. <https://doi.org/10.1534/g3.119.400215>
- Mondal TK, Rawal HC, Chowrasia S et al (2018) Draft genome sequence of first monocot-halophytic species *Oryza coarctata* reveals stress-specific genes. *Scientific Reports* 8:13698. <https://doi.org/10.1038/s41598-018-31518-y>
- Muliyar RK, Chowdappa P, Behera SK et al (2020) Assembly and annotation of the nuclear and organellar genomes of a dwarf coconut (Chowghat Green Dwarf) possessing enhanced disease resistance. *Omics* 24(12):726-42. <https://doi.org/10.1089/omi.2020.0147>
- Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinformatics* 21(Suppl-1):i351-i358. <https://doi.org/10.1093/bioinformatics/bti1018>
- Sairam CV, Jayasekhar S (2018) World coconut economy: Sectoral issues, markets and trade. In: Nampoothiri KUK, Krishnakumar V, Thampan P, Nair MA (Eds) *The Coconut Palm (Cocos nucifera L.) - Research and Development Perspectives*, eBook ISBN 978-981-13-2754-4, Springer, Singapore, pp. 801-820. https://doi.org/10.1007/978-981-13-2754-4_17
- Shi C, Li W, Zhang QJ et al (2020) The draft genome sequence of an upland wild rice species, *Oryza granulata*. *Scientific Data* 7:131. <https://doi.org/10.1038/s41597-020-0470-2>
- Singh R, Ong-Abdullah M, Low ET et al (2013) Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature* 500(7462):335-9. <https://doi.org/10.1038/nature12309>
- Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19(Suppl-2):ii215-ii225. <https://doi.org/10.1093/bioinformatics/btg1080>
- Xiao Y, Xu P, Fan H et al (2017) The genome draft of coconut (*Cocos nucifera*). *Gigascience* 6(11):gix095. <https://doi.org/10.1093/gigascience/gix095>