READ ME: Dataset for genomic resources for the Neotropical tree genus *Cedrela* (Meliaceae) and its relatives

Kristen N. Finch, B.A.; F. Andrew Jones, Ph.D.; Richard Cronn, Ph.D.

Initial Submission: 8/22/2018; Updated: 11/29/2018 & 1/22/2019

Background

Tree species in the genus *Cedrela* are threatened by timber overexploitation across the Neotropics. Genetic identification of processed timber can be used to supplement wood anatomy to assist in the taxonomic and source validation of protected species and populations of *Cedrela*. However, few genetic resources exist that enable both species and source identification of *Cedrela* timber products. We developed several genomic resources including a leaf transcriptome, chloroplast genome, and diagnostic single nucleotide polymorphisms (SNPs) that may assist the classification of *Cedrela* specimens to species and geographic origin. In this supplementary directory, we share the assembled transcriptome reference, hybridization capture probe sequences, and the chloroplast genome reference for a single specimen, CEOD_NYBG (*C. odorata* from the living collection at the New York Botanical Garden). We also include: chloroplast genome sequences for each of the 43 specimens screened in our diversity panel (as separate files and as a combined file, aligned and unaligned), the VCF file containing SNPs for species and origin prediction for *Cedrela*, and data sets to replicate our statistical analysis using R. By sharing these resources, we hope to enable future research on this widespread Neotropical tree genus.

The Dataset

This dataset is available as supplement for a research article submitted in August 2018 and published in January 2019. Information about the study, study system, and premise for this dataset can be found the research article:

Finch, K.N., Jones, F.A. and Cronn, R.C., 2019. Genomic resources for the Neotropical tree genus *Cedrela* (Meliaceae) and its relatives. BMC Genomics, 20(1), p.58.

DOI https://doi.org/10.1186/s12864-018-5382-6

Read the open access article here

Usage

The generation of these resources was supported by the United States Agency for International Development (19318814Y0010-140001), the United States Forest Service International Programs, and the United States Forest Service Pacific Northwest Research Station.

About

This document was prepared by Kristen N. Finch to explain files that are included as external supplement the article.

Questions about this document should be directed to Kristen N. Finch using the contact email provided under this ORCID ID: https://orcid.org/0000-0003-2098-7546.

- READ_ME.html will open a rendered version of this file in a web browser.
- READ_ME.Rmd R markdown file that can be opened in R, RStudio, or in a text editor.
- READ_ME.pdf rendered PDF version of this document.

All files are contained in a single directory called, "Finch_etal_2018_supplement" within which the reader will find the following files or subdirectories. I will explain each subdirectory in some detail and the files they contain.

chloroplast_assemblies/

This subdirectory contains files pretaining to our *de novo* chloroplast genome assembly for CEOD_NYBG or the *C. odorata* from the living collection New York Botanical Garden. This subdirectory also contains sequence files from the reference guided genome assembly for each of the specimens in the diversity panel. Files in this directory can be opened text editors with the exception of Microsoft word. Notepad++ for PC users and Text Editor or Text Wrangler for Mac users can be used to view and edit the .fasta, .result, .gff, and .py files. Another option in BioEdit for PC users which had many tools for editing .fasta files via a GUI. The .result files are best views in FigTree or Mesquite (see main text for citations). The .gff file can be opened in Microsoft Excel.

File	Description
CEOD_NYBG_chloroplast_reference.fasta	*de novo* chloroplast genome sequence for CEOD_NYBG
CEOD_NYBG_chloroplast_RASTannotation.gff	annotation result from the RAST server
NOVOPlasty_config_file.txt	configuration file used for NOVOPlasty chloroplast assembly of CEOD_NYBG
NOVOPlasty_seed_seq_rbcl.fasta	*C. odorata* chloroplast *rbcl* sequence used to seed the NOVOPlasty chloroplast assembly of CEOD_NYBG
files ending *_chloroplast_refg.fasta	reference guided chloroplast genome sequence for each specimen in the diversity panel
chloroplast_RAxML_bestTree.result	RAxML tree file output for the best tree
chloroplast_RAxML_bootstrap.result	RAxML bootstrap output file (contains 1000 trees)
chloroplast_alignment.fasta	multifasta containing aligned chloroplast genomes used for this article
chloroplast_unaligned.fasta	multifasta containing chloroplast genomes used for this article
percent_ns.py	custom python script for estimating the percent of ambiguous nucleotides (depth <2 reads)

R_analysis/

In this subdirectory, the reader will find R code necessary to replicate our analysis with the files provided. .R and .Rmd files can be opened in R, Rstudio, or a text editor.

File	Description
R_analysis.R	R script for this analysis. Script should run from top to bottom if described versions are maintained
R_analysis.Rmd	R markdown file for this analysis. This file can be opened in R, RStudio, or in a text editor.
R_analysis.html	rendered version of the R markdown file that will open in a web browser.

R_analysis/subdirectories

covstats_1Msubset_gene_targets/

Contains files relevant for assessing target capture efficiency/bias. These were used to provide summary statistics in the main text results section and to generate main text Fig. 3 and Fig. S2 in Additional File 1. See the R_analysis.Rmd file for descriptions of the variables contained within the coverage statistics files. All files can be opened in Microsoft Excel. While performing the analysis in R with these files, users should frequently examine generated data frames with the function head(). The View() function can also be used, but viewing large datasets can cause R to crash.

File	Description
files ending *_1Msubset_gene_targets_covstats.txt	one coverage statsitics file per individual in the diversity panel generated by mapping a subset of 1 million reads to the gene target sequences with bbmap.sh

covstats_allreads_gene_targets/

Contains files relevant for assessing sequence yield and on-target yield (see Glossary of terms in R_analysis.html). These were used to provide summary statistics in the main text results section, main text Table 2, and to generate main text Fig. 2. See the R_analysis.Rmd file for descriptions of the variables contained within the coverage statistics files. All files can be opened in Microsoft Excel. While performing the analysis in R with these files, users should frequently examine generated data frames with the function head(). The View() function can also be used, but viewing large datasets can cause R to crash.

File	Description
files ending *_allreads_gene_targets_covstats.txt	one coverage statsitics file per individual in the diversity panel generated by mapping all generated reads to gene target sequences with bbmap.sh
gene_targets_gene_families.txt	gene families associated with each of the gene targets used for probe design
yield_reads_generated.txt	sequence read counts generated with grep -c (see R_analysis.html)

covstats_chloroplast/

Contains files relevant for assessing chloroplast reference-guided genome assembly for the specimens in the diversity panel. These statistics contributed to main text Table 2. See the R_analysis.Rmd file for descriptions of the variables contained within the coverage statistics files. All files can be opened in Microsoft Excel. While performing the analysis in R with these files, users should frequently examine generated data frames with the function head(). The View() function can also be used, but viewing large datasets can cause R to crash.

File	Description
files ending *_cp_covstats.txt	one coverage statsitics file per individual in the diversity panel generated by mapping all generated reads to the de novo chloroplast genome assembly for CEOD_NYBG with bbmap.sh
chloroplast_percent_n_result.txt	the result output from percent_ns.py

map/

Contains files relevant for generating main text Fig. 1 or the sample map. All files can be opened in Microsoft Excel.

File	Description
TM_WORLD_BORDERS-0.3.dbf	support file for map
TM_WORLD_BORDERS-0.3.prj	support file for map
TM_WORLD_BORDERS-0.3.shp	shapefile used to generate map
TM_WORLD_BORDERS-0.3.shx	support file for map
TM_WORLD_BORDERS_SIMPL-0.3.dbf	support file for map
TM_WORLD_BORDERS_SIMPL-0.3.prj	support file for map
TM_WORLD_BORDERS_SIMPL-0.3.shp	support file for map
TM_WORLD_BORDERS_SIMPL-0.3.shx	support file for map
country_labels.csv	data frame with coordinates for country labels

File	Description
sample_points.csv	data frame with coordinates for samples and sample labels

nuclear_vcf_snps/

Contains files relevant for assessing the amount and quality of information available after target capture for the *Cedrela* specimens using the gene targets as the reference. These files were used to generate main text Fig. 4. Files ending .frq, .imiss, .lqual, .fst are tab-delimited plain text files. All files can be opened in Microsoft Excel. While performing the analysis in R with these files, users should frequently examine generated data frames with the function head().

File	Description
cedrela_nuclear.vcf.gz	another copy of the unfiltered nuclear vcf generated from *Cedrela* specimens in the diversity panel; not actually used in the R analysis
cedrela_nuclear_noindels_biall_0miss.frq	allele frequency estimates for all SNPs in the VCF
cedrela_nuclear_noindels_biall_0miss.imiss	individual missingness estimates for each specimen
cedrela_nuclear_noindels_biall_0miss.lqual	site quality estimates for all SNPs in the VCF
cedrela_nuclear_noindels_biall_0miss_global.weir.fst	Weir-Cockerham's FST estimates for all SNPs in the VCF with species as populations
gene_targets_gene_families.txt	gene families associated with each of the gene targets used for probe design

probe_design/

CEOD_NYBG_cov_stats_DNAreads_to_transcriptome.txt

Coverage statsitics file from bbmap.sh generated by mapping DNA reads from CEOD_NYBG to the transcriptome of CEOD_NYBG. We used this file to identify putative low-copy sequences that would be suitable for hybridization probe design for target capture. This file was used to generate Fig. S1 in Additional File 1. See the R_analysis.Rmd file for descriptions of the variables contained within the coverage statistics files. This file can be opened in Microsoft Excel.

transcriptome/

This subdirectory contains files pretaining to our *de novo* transcriptome assembly for CEOD_NYBG, the transcripts we selected as gene targets for hybridization probe design, and the probe sequences provided by MYBaits. Files in this directory can be opened text editors with the exception of Microsoft word. Notepad++ for PC users and Text Editor or Text Wrangler for Mac users can be used to view and edit the .fasta (aka .fas) and .txt files. Another option in BioEdit for PC users which had many tools for editing .fasta files via a GUI. Note here .fasta.masked is simply a .fasta file and should be handled as such. The .zip files here can be decompressed an opened in a text editor or Excel, but I advise against opening these files in any program because of their size. Examining these files is not explicitly necessary, but if you wish to, I would read them into R with read.table() to explore them.

File	Description
CEOD_NYBG_transcriptome_v1.fasta	*de novo* transcriptome assembly for CEOD_NYBG
CEOD_NYBG_transcriptome_v1.fasta.masked	transcriptome output from RepeatMasker
MYBaits_probe_seqs_19740.fas	multifasta of hybridization probe sequences
TRAPID_CEOD_NYBG_transcriptome_GOs.zip	gene ontology terms associated with the transcriptome
TRAPID_CEOD_NYBG_transcriptome_gene_families.txt	gene families associated with the transcriptome
TRAPID_CEOD_NYBG_transcriptome_proteins.zip	protein domains associated with the transcriptome
TRAPID_CEOD_NYBG_transcriptome_results.pdf	summary of TRAPID/PLAZA2.5 results for the transcriptome

File	Description
gene_targets.fasta	multifasta of transcripts we selected as gene targets for hybridization probe design

cedrela_nuclear.vcf.gz

UNFILTERED Variant Call Format file generated with reads from target capture of the *Cedrela* specimens in the diversity panel. The gene target sequences (transcripts from the *de novo* transcriptome assembly for CEOD_NYBG that we selected as gene targets for hybridization probe design) were used as reference for alignment. See R_analysis/R_analysis.html for tips for filtering with vcftools. This file can be decompressed and opened in a text editor, but using R or vcftools for filtering in recommended due to its size. See package vcfR if interested in processing vcfs without vcftools.

Update Log

Original submission: 8/22/2018.

Updated:

11/29/2018 - minor adjustments to code: legends moved into plots, color palette changed for colorblindness accessibility.

1/22/2019 - information about the research article DOI added to the "About" Section. R session information added below to ensure users can repeat analysis.

```
sessionInfo()
```

```
## R version 3.5.1 (2018-07-02)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS  10.14.2
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_3.5.1  backports_1.1.2 magrittr_1.5    rprojroot_1.3-2
##  [5] tools_3.5.1     htmltools_0.3.6 yaml_2.2.0      Rcpp_0.12.19
##  [9] stringi_1.2.4   rmarkdown_1.10  knitr_1.20      stringr_1.3.1
## [13] digest_0.6.18   evaluate_0.12
```