README.md

README

Data for: Genotyping-by-sequencing and ecological niche modeling illuminate phylogeography, admixture, and Pleistocene range dynamics in quaking aspen (*Populus tremuloides*)

Justin C. Bagley ✉️

Department of Biology, Virginia Commonwealth University, Richmond, VA

Departamento de Zoologia, Universidade de Brasília, Brasília, DF, Brazil

Neander M. Heming

Departamento de Zoologia, Universidade de Brasília, Brasília, DF, Brazil

Eliécer E. Gutiérrez

Universidade Federal de Santa Maria, Santa Maria, RS, Brazil

LICENSE

CITATION

If you use scripts from this accession as part of your published research, we request that you cite the dataset as follows (also see DOI information below):

- Bagley, J.C., Heming, N.M., & Gutiérrez, E. E. (2018). Data for: Genotyping-by-sequencing and ecological niche modeling illuminate phylogeography, admixture, and Pleistocene range dynamics in quaking aspen (*Populus tremuloides*). *Mendeley Data*, v1, available at: http://dx.doi.org/10.17632/jhkhvdgyfy.1.

Alternatively, please provide the following link to this software/data accession in your manuscript:

- http://dx.doi.org/10.17632/jhkhvdgyfy.1

DOI

The DOI for this accession is as follows: doi:10.17632/jhkhvdgyfy.1. The CITATION section above illustrates how to cite this code using the DOI.

INTRODUCTION

In support of the manuscript by Bagley et al. (in review) on quaking aspen phylogeography and ecological niche modeling (ENM), this accession dataset provides 1) the laboratory protocol used to extract DNA from aspen leaf tissues (modified from Strauss Lab); 2) code used to conduct two independent runs of the TASSEL-GBSv2 SNP discovery pipeline (Glaubitz et al. 2014) on our final (combined) genotyping-by-sequencing (GBS) dataset; 3) resulting SNP variant files from TASSEL-GBSv2 and final filtered variant call data files used during our genomic analyses; and 4) unfiltered vs filtered species occurrence data files and computer code used during our ENM analyses of our focal taxon, *Populus tremuloides*. A file tree list of the contents of this accession is given in text format under CONTENTS below.

Users interested in GBS pipelines can see how we ran TASSEL-GBSv2, including changes to the default parameters and ways that calls to the Burrows-Wheeler alignment tool bwa (Li & Durbin 2009) were incorporated into the workflow. Anyone with facility in population genetics and analysis of current population genomic data will be able to quickly use the final SNP dataset, e.g. to check SNPs or experiment with different filtering strategies, or use the SNP variant files to conduct population genetic analyses in `R` packages mentioned in the Materials and Methods section of the paper (Bagley et al. in review) or other software.

In this README, we list the files and analysis scripts contained within this accession, we briefly describe the genomic data files provided, and we briefly explain how ENM Rscripts herein were strung together in a pipeline workflow suitable for UNIX-like environments with recent R and MaxEnt installs.

CONTENTS

Scripts and other files contained in this accession.

Text representation of the current directory file tree structure of the accession:

```
/
|
|-Aspen_DNA_Extraction_Protocol.pdf
|
|- SNP_Discovery_Pipeline
|   |-- final_run
|   |   |-- final_TASSEL-GBSv2_pipeline.sh
|   |   |-- Mock-Strauss_key_file.txt
|   |
|   |- noTReps_run
|      |-- noTReps_TASSEL-GBSv2_pipeline.sh
|      |-- Mock-Strauss_key_file.txt
|
|- SNP_VCF_Data_Files
|   |-- ref_3_noTReps_finalProductionSNPs.vcf.gz
|   |-- ref_1_2_finalProductionSNPs.vcf.gz
|   |-- AspenSNPs.33.5K_n183.vcf.gz
|
|- Occurrence_Data_Files
|   |-- Ptrem_Merged_records_spThin.filtered.csv
|   |-- Ptrem_Merged_records__NOT-filtered.csv
|
|- ENM_Results
|   |-- species_RESULTS
|   |   |-- totalArea.aspen.csv
|   |   |-- sel.mdls.Aspen.csv
|   |   |-- OmRate.aspen.csv
|   |   |-- FracPredArea.aspen.csv
|   |   |-- var.Contribution.Aspen.csv
|   |   |-- var.PermImportanceAspen.csv
|   |
|   |- cluster_RESULTS
|      |-- totalArea.MCPea.csv
|      |-- FracPredArea.MCPea.csv
|      |-- OmRate.MCPea.csv
```

```
|      |-- var.Contribution.Aspen.cltr1.MCPea.csv
|      |-- var.Contribution.Aspen.cltr2.MCPea.csv
|      |-- var.Contribution.Aspen.cltr3.MCPea.csv
|      |-- var.PermImportanceAspen.cltr1.MCPea.csv
|      |-- var.PermImportanceAspen.cltr2.MCPea.csv
|      |-- var.PermImportanceAspen.cltr3.MCPea.csv
|      |-- sel.mdls.Aspen.cltr3.MCPea.csv
|      |-- sel.mdls.Aspen.cltr2.MCPea.csv
|      |-- sel.mdls.Aspen.cltr1.MCPea.csv
|
|- Calibration_Areas
|   |-- species_MCP
|   |   |-- Aspen.occ.poly.dbf
|   |   |-- Aspen.occ.poly.prj
|   |   |-- Aspen.occ.poly.shp
|   |   |-- Aspen.occ.poly.shx
|   |
|   |- species_MCP_Calib
|   |   |-- Aspen.bffr.dbf
|   |   |-- Aspen.bffr.prj
|   |   |-- Aspen.bffr.shp
|   |   |-- Aspen.bffr.shx
|   |
|   |- species_MCcP
|   |   |-- Aspen.concave.occ.poly.dbf
|   |   |-- Aspen.concave.occ.poly.prj
|   |   |-- Aspen.concave.occ.poly.shp
|   |   |-- Aspen.concave.occ.poly.shx
|   |
|   |- cluster_MCPea_Calib
|   |   |-- Aspen.cltr2.MCPea.dbf
|   |   |-- Aspen.cltr2.MCPea.prj
|   |   |-- Aspen.cltr2.MCPea.shp
|   |   |-- Aspen.cltr2.MCPea.shx
|   |   |-- Aspen.cltr3.MCPea.dbf
|   |   |-- Aspen.cltr3.MCPea.prj
|   |   |-- Aspen.cltr3.MCPea.shp
|   |   |-- Aspen.cltr3.MCPea.shx
|   |   |-- Aspen.cltr1.MCPea.dbf
|   |   |-- Aspen.cltr1.MCPea.prj
|   |   |-- Aspen.cltr1.MCPea.shp
|   |   |-- Aspen.cltr1.MCPea.shx
|   |
|   |- cluster_MCP
|      |-- Aspen.cltr1.occ.poly.dbf
|      |-- Aspen.cltr1.occ.poly.prj
|      |-- Aspen.cltr1.occ.poly.shp
|      |-- Aspen.cltr1.occ.poly.shx
|      |-- Aspen.cltr2.occ.poly.dbf
|      |-- Aspen.cltr2.occ.poly.prj
|      |-- Aspen.cltr2.occ.poly.shp
|      |-- Aspen.cltr2.occ.poly.shx
|      |-- Aspen.cltr3.occ.poly.dbf
|      |-- Aspen.cltr3.occ.poly.prj
|      |-- Aspen.cltr3.occ.poly.shp
|      |-- Aspen.cltr3.occ.poly.shx
|
|- ENM_Rscripts
      |-- script2_cluster_dataPrep-ENMAnalysis.R
      |-- script1B_dataPrep-ENMAnalysis.R
      |-- script3_manuscriptFigures.R
      |-- script1A_climScenarios2Grd.R
```

TASSEL-GBSv2 SNP DISCOVERY PIPELINE CODE

Within the SNP_Discovery_Pipeline folder of the accession, there are two subdirectories corresponding to the two independent runs of the pipeline discussed in the main text (other runs were conducted varying the different parameters available at different steps of the pipeline, but are not presented; JCB, unpublished results). The 'final' folder contains reusults of the final reference assembly-based run from which production SNPs were filtered and used in our final population genomic and phylogenomic analyses presented in the paper. The 'noTReps' folder corresponds to the no-technical-replicates run mentioned in the main text and Appendix S1 of the Supporting Information. Each folder contains the shell script to run the pipeline, as well as the key file for TASSEL-GBSv2. As noted in the pipeline documentation, the key file is the "file listing barcodes distinguishing the samples (REQUIRED)", and thus links barcodes to sample names and other metadata. The pipelines for the different runs are largely the same; the main difference between them is the key files, with the noTReps key file omitting barcodes and IDs corresponding to technical replicates from Schilling et al.'s (2014) GBS dataset (lanes).

The raw data were too large (>100GB) to be included in this accession due to space limitations. However, the data are available through means listed in the Data Accessibility section of the main text.

Other than installing dependencies (TASSEL-GBSv2 and bwa, etc.), using a LINUX supercomputer, and following the official documentation for the pipeline, all that is required to replicate our runs of this pipeline is to format the raw data in '*.fastq.gz' or '*.fastq.txt.gz' format described in the documentation aand run the pipeline. Our final raw data files were named by lanes/plates described in the key files, and were moved to the fastq/ subdirectory in each run folder on our Linux-based supercomputing cluster at the Virginia Commonwealth University (CHiPC) facility. The final gzipped raw fastq files had the following names and sizes:

```
$ ## Run within final run folder:
$ cd fastq/
$ ls *.gz
./AXXXXXXXXX_1_fastq.txt.gz
./AXXXXXXXXX_2_fastq.txt.gz
./AH353KBBXX_8_fastq.txt.gz
$
$ ll
total 63221504
-rw-rw-r-- 1 jcbagley jcbagley 32896418130 Jul 22 11:16 AH353KBBXX_8_fastq.txt.gz
-rw------- 1 jcbagley jcbagley 18436171873 Jul 22 11:18 AXXXXXXXXX_1_fastq.txt.gz
-rw------- 1 jcbagley jcbagley 13406144516 Jul 22 11:17 AXXXXXXXXX_2_fastq.txt.gz
```

SNP VARIANT FILE OVERVIEW

After running the TASSEL-GBSv2 pipeline as described in the text and the SNP Discovery Pipeline section above, we obtained variant call format (VCF) files 'ref_1_2_finalProductionSNPs.vcf.gz' (final run) and 'ref_3_noTReps_finalProductionSNPs.vcf.gz' (noTReps run) containing the raw production SNP calls output from the pipeline's ProductionSNPCallerPluginV2 plugin. We then filtered these files using vcftools and filtering parameters discussed in the Materials and Methods section of the main text. The resulting filtered VCF file from the final run was then used as the starting point for subsequent analyses. Essentially all of our genetic analyses in `R` were baed on reading the filtered VCF files directly into the `R` environment and manipulating them there to create other classes of data objects used by various `R` software packages such as `hierfstat` (Goudet 2005) and `adegenet` (Jombart & Ahmed 2011), discussed in the main text. Therefore, we also include a gzipped version of our final filtered VCF file, 'AspenSNPs.33.5K_n183.vcf.gz', in this accession, in the 'SNP_VCF_Data_Files' folder.

ENM PIPELINE OVERVIEW

General Structure and Occurrence Data Files

As described in the main text, the ENM analyses essentially focused on running `ENMeval` (Muscarella et al. 2014) and `MaxEnt` (Phillips et al. 2006) from within the `R` environment, using controls and automation functions available in the wrapper software package `ENMwizard` (Heming et al. 2018). For more information and the latest release of `ENMwizard`, see the corresponding GitHub repository. Our ENM analysis pipeline started from an excellent, large set of unfiltered *P. tremuloides* occurrence records (>100,000 records), which we used the `spThin` package (Aiello-Lammens et al. 2015) to filter down to ~14,000 records in the final 'filtered' dataset. Both the unfiltered and filtered occurrence records files are included in this accession. As shown in the file tree above, the unfiltered vs filtered occurrences files are included within the 'Occurrence_Data_Files' subdirectory and are named 'Ptrem_Merged_records__NOT-filtered.csv' and 'Ptrem_Merged_records_spThin.filtered.csv', respectively.

Pipeline Files

Subsequent ENM analyses were run in `R` v3.5.0 (available here) on MacBook computers running macOS High Sierra using the four Rscripts above, which were prepared by NMH and EEG and modified by JCB, as follows:

- script1A_climScenarios2Grd.R
- script1B_dataPrep-ENMAnalysis.R
- script2_cluster_dataPrep-ENMAnalysis.R
- script3_manuscriptFigures.R

These Rscripts take as input various GIS files (links provided within the script; descriptions in the main text, e.g. Table 1) and the *P. tremuloides* occurrence data, and together they output all of the ENM model selection results, model parameters, calibration area shapefiles, maps of model output and projections across different climatic scenarios, and other files in the ENM folders of this accession. Moreover, they are meant to be run in numerically increasing order--first the 'script1A*' file, then 'script1B*' file, and so on, to the final 'script3*' file. Interested users can run them in this order on the data files to replicate our full ENM analysis.

Within this accession, we have organized the output from these Rscripts into two subdirectories, the 'ENM_Results' and 'Calibration_Areas' folders. The Calibration_Areas folder contains shapefiles for all calibration areas (MCP, MCcP) discussed in the text and Appendix S1. The ENM_Results folder contains compilations of results files for analyses conducted 1) at the whole-species level, within a 'species_RESULTS' subdirectory, and 2) at the level of intraspecific genetic clusters within *P. tremuloides*, in the 'cluster_RESULTS' subdirectory. It is *these subdirectories* which contain all of the original results data files included together in different spreadsheet tabs of the Data S2 Excel (*.xlsx) of the Supporting Information. Moreover, the first spreadsheet within the Data S2 file contains a key and notes on the contents of all of these files, as well as the `R` software packages and functions that generated them.

REFERENCES

- Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, 38, 541–545.
- Bagley, J. C., Heming, N. M., Gutiérrez, E. E., Devisetty, U. K., Mock, K. E., Eckert, A. J., & Strauss, S. H. (in review). Genotyping-by-sequencing and ecological niche modeling illuminate phylogeography, admixture, and Pleistocene range dynamics in quaking aspen (*Populus tremuloides*). *Molecular Ecology*.
- Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Resources*, 5(1), 184–186.
- Heming, N. M., Dambros, C., & Gutiérrez, E. E. (2018). ENMwizard: AIC model averaging and other advanced techniques in Ecological Niche Modeling made easy. R package version 0.1.7. Accessed at https://github.com/HemingNM/ENMwizard.
- Jombart, T., & Ahmed, I. (2011). Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27, 3070–3071.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760.
- Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M., & Anderson, R. P. (2014).

ENMeval: an R package for conducting spatially independent evaluations and estimating optimal model complexity for MAXENT ecological niche models. *Methods in Ecology and Evolution*, 5, 1198–1205.

- Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190, 231–259.