Estimating genetic diversity and population information from short read (ddRAD-seq) type data

Chris Barratt (sDiv / Evolution and Adaptation)
Laura Mendez (Evolution and Adaptation)
Assistant: Dimas Calderon (Evolution and Adaptation)

20-21 June 2022

Objectives

1. Understand best practices on planning and executing a population genomics project based on ddRAD-seq (short read) type data

2. Learn how to download genomic sequence data from the European Nucleotide Archive

3. Familiarise yourself with the Stacks 2 bioinformatics pipeline to process raw reads

4. Be able to generate your own output files after Stacks 2 processing for downstream analyses of population structure, phylogeny and genetic diversity

5. Run analyses to perform (Admixture, sNMF, DAPC, RAxML, F-statistics), understand what you've done and how to interpret the outputs

Schedule

Day 1 (20th June): 09:00 - 15:00

- Brief introductions and research topics of course leaders and course participants (2-3 mins each)
- Introductory lecture on population genomics and ddRAD-seq and which kinds of questions can be answered with these data
- Hands on session for downloading publicly available genomic data (e.g. from the European Nucleotide Archive)
- Stacks 2 lecture (main processing steps)
- Hands on session for each of the major processing steps required for the Stacks 2 pipeline (process_radtags, denovo_map)

Day 2 (21st June): 09:00 - 15:00 (?)

- Hands on sessions to analyse population structure, phylogeny and genetic diversity:
- Admixture (pop structure)
- sNMF (pop structure)
- DAPC (pop structure)
- RAxML (phylogenetic relationships)
- F-stats (FST, nucleotide diversity)

Datasets we will use

We ran a similar course 2019 – the main aim of the course is to get you confident enough to go from raw data to interpreting outputs from analyses. So ask questions!

Myotis escalerai

Leptopelis flavomaculatus

Hyphaene coriacea

What are we doing?

- High-throughput data and ddRAD-seq background – what is it and why use it?
- Sampling strategy prior to RAD-seq type work – how should I plan my sequencing?
- Bioinformatics to generate output files with a HPC cluster using Stacks 2

➢ More specific workshops to follow late 2022 depending on demand (most probably demographic history/local adaptation and Genotype-Environment Association analyses)

Population genetics history

See Wright, Fisher, Haldane in the early-mid 19th century:

- Genetic composition of populations
- Changes in genotypes in response to the four evolutionary forces (drift, natural selection, gene flow and mutation)

Markers have gradually become more numerous and associated methods more sophisticated:

- Allozymes
- AFLPs
- microsatellites
- SNPs (Single Nucleotide Polymorphisms)
- Whole genomes

Genetics vs. Genomics

Genetic data (e.g. nuclear loci, mtDNA, chloroplast DNA)

Genomic data (e.g. WGS, Exon capture, [dd]RAD-seq)

Reduced Library Representation

- Huge increase over time

Supplementary figure S1. Numbers of articles citing the original papers describing each type of RADseq protocol over time. Data (see table) generated using Web of Science. Data for 2015 are extrapolated using numbers of articles cited from January through September 2015. Protocols are arranged by order of first appearance in the literature. CRoPS, complexity reduction of polymorphic sequences; original RAD, original restriction-site associated-DNA; RRL, reduced representation library; GBS, genotyping by sequencing; MSG, multiplexed shotgun genotyping; SBG, sequence-based genotyping; ddRAD, double-digest restriction-site-associated DNA; SLAF-seq, specific-locus amplified fragment sequencing; RestSeq, restriction fragment sequencing.

ddRAD-seq

Positives

- Large data
- Quick data generation
- Pooling of individuals
- High resolution
- Powerful inference
- Cost effective

Negatives

- Laboratory investment (adapters, equipment such as sonicator or Pippin prep)
- Bioinformatic investment – HPC use, also learning how to process data

STUDY DESIGNS

Harnessing the power of RADseq for ecological and evolutionary genomics

Kimberly R. Andrews[1], Jeffrey M. Good[2], Michael R. Miller[3], Gordon Luikart[4] and Paul A. Hohenlohe[5]

Abstract | High-throughput techniques based on restriction site-associated DNA sequencing (RADseq) are enabling the low-cost discovery and genotyping of thousands of genetic markers for any species, including non-model organisms, which is revolutionizing ecological, evolutionary and conservation genetics. Technical differences among these methods lead to important considerations for all steps of genomics studies, from the specific scientific questions that can be addressed, and the costs of library preparation and sequencing, to the types of bias and error inherent in the resulting data. In this Review, we provide a comprehensive discussion of RADseq methods to aid researchers in choosing among the many different approaches and avoiding erroneous scientific conclusions from RADseq data, a problem that has plagued other genetic marker types in the past.

RAD-seq

RAD-seq
Laboratory protocol
Genomic DNA
Restriction enzyme digestion
Ligation of P1 adapter
P1 adapter
Barcode
Restriction site
Pool samples
Shearing and size selection
P2 adapter ligation
Barcoded RAD end
P1
P2
Sheared end
Single end sequencing
Paired end sequencing
Assembly
Assembly
~100 bp contigs
~100 bp + ~400 bp contigs
Single end RAD library
Paired end RAD library
Sequence stacks for bioinformatic processing

Stacks
Bioinformatic processing
Demultiplex, Quality control and clean raw NGS sequence reads
process radtags
Build loci de-novo and call single nucleotide polymorphisms (SNPs) using sequence stacks
ustacks
Assemble catalogue of loci across multiple samples
cstacks
Match homologous reads in each sample to catalogue
sstacks
Create output files for analyses using population maps and specified filtering thresholds
populations
*.phy
*.str
*.tsv
Phylogenetic analysis
Population structure analysis
Genetic diversity estimates
denovo_map.pl

ddRAD seq

Figure 2. Double digest RAD sequencing improves efficiency and robustness while minimizing cost.
(**A**) Traditional Restriction-Site Associated DNA sequencing (RADseq) uses a single restriction enzyme (RE) digest coupled with secondary random fragmentation and broad size selection to generate reduced representation libraries consisting of all genomic regions adjacent to the RE cut site (red segments). (**B**) Double digest RAD sequencing (ddRADseq), by contrast, uses a two enzyme double digest followed by precise size selection that excludes regions flanked by either [a] very close or [b] very distant RE recognition sites, recovering a library consisting of only fragments close to the target size (red segments). Representation in this library is expected to be inversely proportional to deviation from the size-selection target, thus read counts across regions are expected to be correlated between individuals (yellow and green bars).

Sampling design

Genome considerations: (you can't control this…)

- Genome size, GC content?
- Repetitive elements?
- Ploidy?
- Reference genome?

Sampling considerations: (but you can control this…)

- Conceptualize sequence space…
- Number of samples and populations?
- Restriction enzyme(s) selected
- DNA quality

Genome considerations

- Smaller is generally better - gives higher coverage of loci as less restriction sites in a smaller genome, so they are sequenced more often per individual

- Few repetitive elements (paralogs) better, genome duplications pose additional bioinformatic challenges

- Diploid organisms generally easier- most software written to deal with this though there are workarounds (e.g. dDocent)

Sampling considerations

- What are your questions? Pop structure/Phylogeny/Demography – this will dictate how many individuals/populations you need to sequence in your sampling strategy

- RAD-seq uses a rare cutter enzyme (e.g. sbfI, EcoRI, PstI), ddRAD introduces a frequent cutter too (e.g. MspI, MboI) – think about which to use to maximise the information you will get

- DNA quality is **CRUCIAL**! ddRAD in particular needs very high quality DNA, RAD-seq can deal with degraded samples better. You need **ACCURATE** quantification of **ALL** samples prior to any labwork. Samples will compete for sequence reads so try to have a working concentration (x ng/ul) across all samples. Run samples on a gel and also use Qubit to measure e.g. 3x and take average

In the lab…

- Quantify your DNA and samples prior to organising RAD library construction (don't do all together). Bring 96- plates with your DNA aliquots to lab and do protocols there

- Intense concentration needed, but relaxed periods in between (e.g. waiting for PCR / gel steps), so keep alert. One small mistake can ruin hours/days of work!

- Be exceptionally consistent and accurate with pipetting especially at the adapter ligation stages – this will determine if individuals are included in your library when pooling. Also be careful with extracting the gel slices, size selection is very important

Sequencing

- It's wise to run samples on a Bioanalyzer to check the size of the fragments in the final library, sequencing facilities can do this but will charge

- Once your libaries are done you can store them in the fridge/freezer until ready to take to the sequencing facility

- Enjoy that feeling once it's all done and take a break before…

BIOINFORMATICS

"There's your problem...
You've got an extra parenthesis in line 18."

Bioinformatics

- "Bioinformatics" is an extremely broad subject – covers all types of different analyses. For RAD-seq type data the options are usually Stacks, ipyRAD, dDocent, TASSEL, ANGSD to process data

- Downstream analyses will typically use output files from these programs but there are numerous conversion tools if needed (PLINK, PGDSpider)

- Absolutely **NOT** trivial! You need to become familiar with writing code in for linux/unix style environments to do the heavy lifting for you

- Also you need to become with High Performance Computing clusters and how to use them as this will be fundamental to parallelising the workload on multiple processors

Stacks

- Programmed and maintained by Julian Catchen
- Designed specifically for short read [dd]RAD-seq type data
- Written in C++ and Perl
- Well established pipeline which has not changed too much, cited many times
- Good google user group (very useful!)
- Denovo_map.pl, ref_map.pl or modular system (process_radtags, ustacks, cstacks, sstacks, populations) - can rerun any part
- Stacks 2 much improved for paired end data especially, very good software

Workflow

ddRAD-seq
raw sequence
read data

HPC cluster

Stacks 2.61
Perl

- Demultiplex
- Cluster reads
- Build catalogue
- Match catalogue
- Assemble datasets

Output files for
pop structure, genetic
diversity and
phylogenetics

Laura will talk about this more!

process_radtags

- *Myotis escalarai* – Escalara's mouse-eared bat
- We will download data and then process them
- Taking sequenced library and demultiplexig based on individual barcodes (Razgour et al. 2019)

denovo_map

- *Hyphaene coriacea* – a Malagasy palm (Mendez et al. in prep)
- We will run denovo_map on some samples so you can see how to process data in Stacks

Downstream analyses

- *Leptopelis flavomaculatus* – an East African treefrog (Barratt et al. 2018)
- We will look at pop structure, phylogeny and genetic diversity (and FST)

Shell scripts for job submission

- Shell scripts let you send your jobs to the HPC cluster

- Usually one script per individual task, though you can make special scripts which will submit multiple jobs in sequence

- I have a main script header which I edit for each individual job:
  - Job name
  - Notifications (starts, ends, aborts, pauses etc.) and email address where to send them
  - Input/Output directories and where to place job logfile
  - Requirements – processors needed, amount of memory per processor and runtime (used for queueing)

  - Followed by the main lines of code – load programs/environments and relevant instructions

Shell scripts

```bash
#!/bin/bash

#SBATCH --job-name=process_radtags
#SBATCH --mail-user=christopher_david.barratt@uni-leipzig.de
#SBATCH --mail-type=BEGIN,END,FAIL,TIME_LIMIT
#SBATCH --output=/work/barratt/ddRAD-seq_workshop/job_logs/%x-%j.log
#SBATCH --cpus-per-task=1
#SBATCH --mem-per-cpu=8G
#SBATCH --time=12:00:00


## Load modules and activate software


module purge
module load Anaconda3
source activate /gpfs0/global/apps/stacks_2.61

# process_radtags - take raw data from sequencer and output sequences per individual using barcodes and restriction enzyme information
# -p = file path of raw data from sequencer
# -o = output file path of where you want your individual data to go
# -b = barcode identifier (one file showing all barcodes for each individual)
# --inline_inline: barcode is inline with sequence, occurs on single and paired-end read.
# --renz_1 = restriction enzyme used
# --renz_2 = restriction enzyme used
# -c,--clean = clean data, remove any read with an uncalled base.
# -q,--quality = discard reads with low quality scores.
# --filter_illumina = discard reads that have been marked by Illumina's chastity/purity filter as failing.
# -s = set the score limit. If the average score within the sliding window drops below this value, the read is discarded (default 10).i
# -i = input file type, either 'fastq', 'gzfastq' (gzipped fastq), 'bam', or 'bustard' (default: guess, or gzfastq if unable to).
# -y = output type, either 'fastq', 'gzfastq', 'fasta', or 'gzfasta' (default: match input type).
# -t = truncate final read length to this value.


process_radtags -1 /work/barratt/ddRAD-seq_workshop/data/Exercises_1-2/raw/BAT_02_NoIndex_L005_R1_001.fastq.gz -2 /work/barratt/ddRAD-seq_works
```

Stacks pipeline

Demultiplex (all data) – feed in adapter barcodes to get *.fq.gz file per individual (only run once per dataset)

Make “stacks” of common sequences in each individual file – this module is run for each individual sample

Build catalogue library of all stacks from all individuals (run once per dataset)

Match stacks in each individual to the library catalogue – this module is run for each individual sample

Output file generation based on thresholds defined by user and a population map (e.g. only output loci which are present in at least 50% of all samples)

Data structure

Tutorials

Go here: https://cd-barratt.github.io/ddRAD-seq_workshop/

ddRAD-seq_workshop | Home | Tutorials | GitHub

- 00 - Setting up the data and scripts for analysis
- 01 - Downloading data from ENA
- 02 - Demultiplexing sequencing data (process_radtags)
- 03 - Running Stacks 2 (denovo_map)
- 04 - Population structure (Admixture)
- 05 - Population structure (sNMF)
- 06 - Population structure (DAPC)
- 07 - Phylogenetic relationships (RAxML)
- 08 - Genetic Diversity (F-stats)

Estimating … population informatic … RAD-seq) type data

Date & time: June 20-21 20…

Location: German Centre fo… 03, Leipzig. Room: Beehive (Ground floor)

Teachers: Chris Barratt & La…

Course summary | Course website

Schedule

Day 1 (20th June): 09:00 - 15:00

- Brief introductions and research topics of course leaders and course participants (2-3 mins each)
- Introductory lecture on population genomics and ddRAD-seq and which kinds of questions can be answered with these data
- Hands on session for downloading publicly available genomic data (e.g. from the European Nucleotide Archive) - using *Myotis escalerai* data (a European bat species)
- Stacks 2 lecture (main processing steps)
- Hands on session for each of the major processing steps required for the Stacks pipeline (process_radtags, denovo_map) - using *Hyphaene coriacea* data (an African palm)

00 - setup

ddRAD-seq_workshop | Home | Tutorials | GitHub

Setting up the data and scripts for analysis

To get us set up all with files and directory structure that you'll need, let's do the following together in the EVE cluster…

1. Login to the EVE cluster

First we will login to EVE, using either Terminal if you're a mac user, or PuTTy if you're a Windows user. You might need to make sure your VPN is connected to access.

```
# Open your Terminal (Mac) or PuTTy (Windows)
ssh USER@idiv-gateway.ufz.de
# [enter your password]
ssh frontend1
# [enter your password]
```

2. Copy the directory we provide in `/public/` to your own workspace (will take a few mins)

Now we'll copy the directory which is all ready to your own workspace. Ideally it's best practice to have data in a `/data/` directory but this will get quite complicated with all of us, so let's keep it simple and store in your own `/work/$USER/` directories

```
# This line will copy all of the files and directories to your own work space (the ~ symbol )
cp -r /public/ddRAD-seq_workshop/ /work/$USER/
```

01 – Downloading data

ebi.ac.uk/ena/browser/text-search?query=PRJEB29086

EMBL-EBI | Services | Research | Training | About us | EMBL-EBI

ENA European Nucleotide Archive

PRJEB29086 Search

Examples: histone, BN000065

Enter accession View

Examples: Taxon:9606, BN000065, PRJEB402

Home | Submit | Search | Rulespace | About | Support

Text Search

Uses EBI Search to perform a free text search across ENA data. For more detailed usage please refer to the help & documentation section.

Search term: PRJEB29086 Search

Search results for PRJEB29086

- **Read**
  - Experiment (1)
  - Run (5)
- **Study**
  - Study (1)
  - Project (1)

Experiment	
ERX2839233	Illumina HiSeq 2500 paired end sequencing; bat
Run View all 5 results.	
ERR2832483	Illumina HiSeq 2500 paired end sequencing; bat
Study	
ERP111357	Considering adaptive genetic variation in climate change vulnerability assessment reduces species range loss projections
Project	
PRJEB29086	Considering adaptive genetic variation in climate change vulnerability assessment reduces species range loss projections

02 – process_radtags

03 – denovo_map

04 - Admixture

Fast model-based estimation of ancestry in unrelated individuals

David H. Alexander,[1,4] John Novembre,[2] and Kenneth Lange[3]

[1]*Department of Biomathematics, University of California at Los Angeles, Los Angeles, California 90095, USA;* [2]*Department of Ecology and Evolutionary Biology, University of California at Los Angeles, Los Angeles, California 90095, USA;* [3]*Department of Human Genetics and Department of Statistics, University of California at Los Angeles, Los Angeles, California 90095, USA*

Population stratification has long been recognized as a confounding factor in genetic association studies. Estimated ancestries, derived from multi-locus genotype data, can be used to perform a statistical correction for population stratification. One popular technique for estimation of ancestry is the model-based approach embodied by the widely applied program *structure*. Another approach, implemented in the program EIGENSTRAT, relies on Principal Component Analysis rather than model-based estimation and does not directly deliver admixture fractions. EIGENSTRAT has gained in popularity in part owing to its remarkable speed in comparison to *structure*. We present a new algorithm and a program, ADMIXTURE, for model-based estimation of ancestry in unrelated individuals. ADMIXTURE adopts the likelihood model embedded in *structure*. However, ADMIXTURE runs considerably faster, solving problems in minutes that take *structure* hours. In many of our experiments, we have found that ADMIXTURE is almost as fast as EIGENSTRAT. The runtime improvements of ADMIXTURE rely on a fast block relaxation scheme using sequential quadratic programming for block updates, coupled with a novel quasi-Newton acceleration of convergence. Our algorithm also runs faster and with greater accuracy than the implementation of an Expectation-Maximization (EM) algorithm incorporated in the program FRAPPE. Our simulations show that ADMIXTURE's maximum likelihood estimates of the underlying admixture coefficients and ancestral allele frequencies are as accurate as *structure*'s Bayesian estimates. On real-world data sets, ADMIXTURE's estimates are directly comparable to those from *structure* and EIGENSTRAT. Taken together, our results show that ADMIXTURE's computational speed opens up the possibility of using a much larger set of markers in model-based ancestry estimation and that its estimates are suitable for use in correcting for population stratification in association studies.

[Supplemental material is available online at http://www.genome.org. The ADMIXTURE program is freely available at http://www.genetics.ucla.edu/software.]

04 - Admixture

- ML estimation of individual ancestries from SNP datasets
- Uses same statistical model as Structure (Pritchard, Stephens & Donnelly, 2000) – the benchmark for understanding pop structure the past 20 years
- Uses a block relaxation approach for speed, which outperforms the other methods – they simply become inefficient and intractable for large SNP data

04 - Admixture

Cross-validation

The choice of the number of ancestral populations K can prove difficult when the underlying population genetics of a species is poorly understood. STRUCTURE provides a means of estimating the best value of K by computing the *model evidence* for each K from a range of choices. The model evidence is defined as

$$\Pr(N|K) = \int f(N|Q, F, K)\, \pi(Q, F|K)\, dQ\, dF \tag{2}$$

where f represents the data likelihood and π represents a prior density on the parameters. STRUCTURE approximates the integral via Monte Carlo methods. Our optimization framework is not well suited to evaluating this integral. As an alternative, we employ cross-validation. In cross-validation, we aim to identify the best K value as judged by prediction of systematically withheld data points. A similar tactic is also employed by the haplotype analysis program fastPHASE [5] and is inspired by Wold's method for cross-validating PCA models [6].

Our v-fold cross-validation procedure partitions the non-missing genotypes into v roughly equally sized subsets (*folds*). At each of v iterations, the members of one of the folds are masked (temporarily marked as missing) to yield a new data matrix $\tilde{N} = \{\tilde{n}_{ij}\}$. Analysis of the masked data matrix $\tilde{N}$ poses no new challenges. In computing the log-likelihood, score, and observed information matrix of $\tilde{N}$, we simply ignore the entries (i, j) with missing values. Maximization of the log-likelihood readily yields new estimates $\tilde{Q}$ and $\tilde{F}$ for the masked data. We then predict each masked value n_{ij} by $\hat{\mu}_{ij} = 2\sum_k \tilde{q}_{ik}\tilde{f}_{kj}$. Prediction error is estimated by averaging the squares of the deviance residuals for the binomial model [7],

$$d(n_{ij}, \hat{\mu}_{ij}) = n_{ij}\log(n_{ij}/\hat{\mu}_{ij}) + (2 - n_{ij})\log[(2 - n_{ij})/(2 - \hat{\mu}_{ij})], \tag{3}$$

across all masked entries over all folds. Minimizing this estimated prediction error on a grid of K values then suggests the most suitable K.

04 - Admixture

LEA: an R package for Landscape and Ecological Associations studies

LEA contains two main functions **snmf()** and **lfmm()**. For an overview the package, the best introduction in the package vignette (latest release from Github).

- The **snmf()** function produces results similar to the widely-used computer program STRUCTURE. Users can play with a large panel of graphical functions for displaying pie-charts and interpolated admixture coefficients on geographic maps. For a short **introduction on how to run structure-like analyses** and importing files from the structure or tess formats, follow the examples presented in this tutorial. Improved graphical functions are also implemented in the R companion package tess3r.

- The **lfmm()** function implements latent factor mixed models for detecting loci exhibiting association with ecological variables or phenotypic traits. The LEA package also implements genome scans for selection based on ancestral population differentiation tests for detecting signatures of local adaptation when no populations are defined a priori. Here is an example on the use of LFMM to perform a genome-wide ecological association study.

A brief presentation of statistical methods implemented in LEA for the inference of population structure and local adaptation is available as a PDF presentation here.

05 - sNMF

Cross validation scores across k (sNMF)
CV score
0.320
0.325
0.330
0.335
0.340
0.345
0.350
2
4
6
8
10
k clusters

Ancestry
1.0
0.8
0.6
0.4
0.2
0.0
T3416
T3419
T4606
T5621
T5625
T5628
T5629
T5631
T5643
T5648
T5645
T2480
T2841
T3644
T3768
T4075
T4098
T4100
T4101
T4199
T4200
T4534
T4537
T4538
T4540
T4847
T4923
T4960
T4970
T5238
T5239
T3431
T3446
T3447
T3452
T3802
T3809
T4872
T3081
T2979
T3009
T3010
T3074
T3107
T3108
T3157
T3179
T3189
T3190
T3191
T3192
T3193
T3194
T3195
T3196
T3198
T3210
T4536
T4834

06 - DAPC

Hardy–Weinberg proportions for two alleles: the horizontal axis shows the two allele frequencies p and q and the vertical axis shows the expected genotype frequencies. Each line shows one of the three possible genotypes.

- Hardy-Weinberg principle:

- Allele and genotype frequencies will remain constant from generation to generation without external evolutionary influences

- Selection
- Mutation
- Gene flow (migration)
- Genetic drift random changes (e.g. small population sizes)

06 - DAPC

BMC Genomic Data

Home About Articles Submission Guidelines Join The Board

Methodology article | Open Access | Published: 15 October 2010

Discriminant analysis of principal components: a new method for the analysis of genetically structured populations

Thibaut Jombart, Sébastien Devillard & François Balloux

BMC Genetics **11**, Article number: 94 (2010) | Cite this article

98k Accesses | **2589** Citations | **25** Altmetric | Metrics

06 - DAPC

BIC across k
Bayesian Information Criterion
372
374
376
378
380
382
384
2
4
6
8
10
k clusters

DA eigenvalues

Summary of pop structure

- Admixture, sNMF and DAPC analyses suggest k=4
- 1 isolated cluster towards southern Mozambique
- 1 cluster in northern Mozambique
- 1 cluster in southern Tanzania
- 1 large cluster in northern Tanzania and Kenya

07 - RAxML

BIOINFORMATICS APPLICATIONS NOTE

Vol. 30 no. 9 2014, pages 1312–1313
doi:10.1093/bioinformatics/btu033

Phylogenetics

Advance Access publication January 21, 2014

RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies

Alexandros Stamatakis[1,2]

[1]Scientific Computing Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg and [2]Department of Informatics, Institute of Theoretical Informatics, Karlsruhe Institute of Technology, 76128 Karlsruhe, Germany

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Phylogenies are increasingly used in all fields of medical and biological research. Moreover, because of the next-generation sequencing revolution, datasets used for conducting phylogenetic analyses grow at an unprecedented pace. RAxML (Randomized Axelerated Maximum Likelihood) is a popular program for phylogenetic analyses of large datasets under maximum likelihood. Since the last RAxML paper in 2006, it has been continuously maintained and extended to accommodate the increasingly growing input datasets and to serve the needs of the user community.

standard bootstrap search that relies on algorithmic shortcuts and approximations to speed up the search process.

It also offers an option to calculate the so-called SH-like support values (Guindon *et al.*, 2010). I recently implemented a method that allows for computing RELL (Resampling Estimated Log Likelihoods) bootstrap support as described by Minh *et al.* (2013).

Apart from this, RAxML also offers a so-called bootstopping option (Pattengale *et al.*, 2010). When this option is used, RAxML will automatically determine how many bootstrap rep-

07 - RAxML

FigTree v1.4.4 - RAxML_bipartitions.FINAL_bootstrapped.tre
Cartoon
Collapse
Reroot
Rotate
Annotate
Colour
Hilight
Find
Node
Clade
Taxa
Selection Mode
Prev/Next
Layout
Zoom:
Expansion:
Fish Eye:
Root Length:
Curvature:
Align Tip Labels
Current Tree: 1 / 1
Appearance
Trees
Time Scale
Tip Labels
Display: NEW NAME
Colour by: User selection
Font Size: 8
Setup: Colour Font
Format: Decimal
Sig. Digits: 4
Tip Shapes
Node Labels
Node Shapes
Node Bars
Branch Labels
"T5629 – Mozambique, Taratibu"
"T5645 – Mozambique, Taratibu"
"T5643 – Mozambique, Taratibu"
"T5628 – Mozambique, Taratibu"
"T5625 – Mozambique, Taratibu"
"T5631 – Mozambique, Taratibu"
"T5621 – Mozambique, Taratibu"
"T5648 – Mozambique, Taratibu"
"T3419 – Mozambique, Gorongosa"
"T3416 – Mozambique, Gorongosa"
"T4606 – Mozambique, Mount Mabu"
"T3190 – Tanzania, Noto plateau, Lindi"
"T3107 – Tanzania, Litipo FR (edge), Lindi"
"T2979 – Tanzania, Makangala FR, Lindi"
"T3074 – Tanzania, Litipo FR (edge), Lindi"
"T3010 – Tanzania, Makangala FR, Lindi"
"T3009 – Tanzania, Makangala FR, Lindi"
"T3189 – Tanzania, Noto plateau, Lindi"
"T3198 – Tanzania, Noto plateau, Lindi"
"T3108 – Tanzania, Litipo FR (edge), Lindi"
"T3179 – Tanzania, Litipo FR, Lindi"
"T3081 – Tanzania, Litipo FR (edge), Lindi"
"T3196 – Tanzania, Noto plateau, Lindi"
"T3157 – Tanzania, Litipo FR, Lindi"
"T3210 – Tanzania, Noto plateau, Lindi"
"T3191 – Tanzania, Noto plateau, Lindi"
"T3192 – Tanzania, Noto plateau, Lindi"
"T3194 – Tanzania, Noto plateau, Lindi"
"T3193 – Tanzania, Noto plateau, Lindi"
"T3802 – Tanzania, Kibasira swamp, Kilombero valley"
"T3809 – Tanzania, Kibasira swamp, Kilombero valley"
"T3432 – Tanzania, Nyamuete FR, Pwani"
"T3446 – Tanzania, Kiwengoma FR, Pwani"
"T3447 – Tanzania, Kiwengoma FR, Pwani"
"T3431 – Tanzania, Nyamuete FR, Pwani"
"T4540 – Tanzania, Mvuha FR, Uluguru Mountains"
"T4538 – Tanzania, Mkungwe FR, Uluguru Mountains"
"T4847 – Tanzania, Kimboza forest, Uluguru Mountains"
"T4923 – Tanzania, Kimboza forest, Uluguru Mountains"
"T4537 – Tanzania, Milawilia FR, Uluguru Mountains"
"T5238 – Kenya, Shimba Hills lodge"
"T5239 – Kenya, Shimba Hills lodge"
"T2641 – Tanzania, Nguu Mountains"
"T4075 – Tanzania, Zaraninge FR, Pwani"
"T4098 – Tanzania, Gendagenda North FR, Tanga"
"T4101 – Tanzania, Gendagenda North FR, Tanga"
"T4200 – Tanzania, Mabayani bwawa, Tanga"
"T4199 – Tanzania, Mabayani bwawa, Tanga"
"T4970 – Tanzania, Segoma FR, East Usambara"
"T4100 – Tanzania, Gendagenda North FR, Tanga"
"T3644 – Tanzania, Muyuyu FR, Pwani"
"T2480 – Tanzania, Kazimzumbwi FR, Pwani "
"T3768 – Tanzania, Kibasira swamp, Kilombero valley"
0.02

08 – Genetic diversity and FST

# Pop ID	Private	Num_Indv	Var	StdErr	P	Var	StdErr	Obs_Het	Var	StdErr	Obs_Hom	Var	StdErr	Exp_Het	Var	StdErr	Exp_Hom	Var	StdErr	Pi	Var	StdErr	Fis
Kazimzumbwi	0	1.00000	0.00000	0.00000	0.97925	0.00994	0.00114	0.04149	0.03978	0.00227	0.95851	0.03978	0.00227	0.02075	0.00994	0.00114	0.97925	0.00994	0.00114	0.04149	0.03978	0.00227	0.00000
Nguu	0	1.00000	0.00000	0.00000	0.95429	0.02077	0.00138	0.09142	0.08307	0.00275	0.90858	0.08307	0.00275	0.04571	0.02077	0.00138	0.95429	0.02077	0.00138	0.09142	0.08307	0.00275	0.00000
Makangala	7	2.35474	0.22892	0.00469	0.87960	0.03148	0.00174	0.10818	0.04742	0.00213	0.89182	0.04742	0.00213	0.14886	0.04187	0.00200	0.85114	0.04187	0.00200	0.19051	0.06913	0.00258	0.13495
Litipo	17	4.48293	0.99829	0.00920	0.86709	0.02562	0.00147	0.12615	0.03665	0.00176	0.87385	0.03665	0.00176	0.17925	0.03855	0.00181	0.82075	0.03855	0.00181	0.20338	0.04992	0.00206	0.16540
Noto	79	6.72055	1.63839	0.01200	0.85843	0.02435	0.00146	0.11245	0.02510	0.00149	0.88755	0.02510	0.00149	0.19435	0.03663	0.00179	0.80565	0.03663	0.00179	0.21059	0.04308	0.00195	0.24199
Gorongosa	66	1.57317	0.24467	0.00473	0.98287	0.00711	0.00081	0.02739	0.02200	0.00142	0.97261	0.02200	0.00142	0.01945	0.00848	0.00088	0.98055	0.00848	0.00088	0.02959	0.02096	0.00138	0.00329
Nyamuete	0	1.75021	0.18741	0.00394	0.90716	0.02910	0.00155	0.12923	0.07775	0.00253	0.87077	0.07775	0.00253	0.11023	0.03644	0.00174	0.88977	0.03644	0.00174	0.15714	0.07864	0.00255	0.04185
Kiwengoma	2	1.64012	0.23038	0.00439	0.91862	0.02769	0.00152	0.10787	0.07071	0.00243	0.89213	0.07071	0.00243	0.09415	0.03343	0.00167	0.90585	0.03343	0.00167	0.13895	0.07825	0.00256	0.04662
Muyuyu	0	1.00000	0.00000	0.00000	0.95702	0.01965	0.00140	0.08596	0.07858	0.00281	0.91404	0.07858	0.00281	0.04298	0.01965	0.00140	0.95702	0.01965	0.00140	0.08596	0.07858	0.00281	0.00000
Kibasira	5	2.55751	0.24671	0.00463	0.89209	0.02847	0.00157	0.09848	0.04289	0.00193	0.90152	0.04289	0.00193	0.13559	0.03895	0.00184	0.86441	0.03895	0.00184	0.17031	0.06210	0.00233	0.12141
Zaraninge	0	1.00000	0.00000	0.00000	0.95609	0.02003	0.00139	0.08783	0.08012	0.00279	0.91217	0.08012	0.00279	0.04391	0.02003	0.00139	0.95609	0.02003	0.00139	0.08783	0.08012	0.00279	0.00000
Gendagenda	5	2.54421	0.24807	0.00466	0.90735	0.02601	0.00151	0.09734	0.04650	0.00202	0.90266	0.04650	0.00202	0.11611	0.03574	0.00177	0.88389	0.03574	0.00177	0.14636	0.05728	0.00224	0.08240
Mabayani	2	1.65836	0.22494	0.00433	0.92909	0.02449	0.00143	0.09891	0.06538	0.00233	0.90109	0.06538	0.00233	0.08278	0.03009	0.00158	0.91722	0.03009	0.00158	0.12127	0.06912	0.00240	0.03354
Nilo	0	1.00000	0.00000	0.00000	0.98700	0.00633	0.00113	0.02600	0.02533	0.00227	0.97400	0.02533	0.00227	0.01300	0.00633	0.00113	0.98700	0.00633	0.00113	0.02600	0.02533	0.00227	0.00000
Mkungwe	0	1.00000	0.00000	0.00000	0.94352	0.02505	0.00148	0.11296	0.10021	0.00297	0.88704	0.10021	0.00297	0.05648	0.02505	0.00148	0.94352	0.02505	0.00148	0.11296	0.10021	0.00297	0.00000
Mount_Mabu	0	1.00000	0.00000	0.00000	0.98173	0.00880	0.00096	0.03655	0.03521	0.00192	0.96345	0.03521	0.00192	0.01827	0.00880	0.00096	0.98173	0.00880	0.00096	0.03655	0.03521	0.00192	0.00000
Kimboza	1	1.60790	0.23838	0.00445	0.92792	0.02577	0.00146	0.10035	0.06998	0.00241	0.89965	0.06998	0.00241	0.08223	0.03055	0.00159	0.91777	0.03055	0.00159	0.12399	0.07488	0.00249	0.03547
Shimba	11	1.55937	0.24650	0.00460	0.94935	0.01949	0.00129	0.06668	0.04913	0.00205	0.93332	0.04913	0.00205	0.05720	0.02288	0.00140	0.94280	0.02288	0.00140	0.08593	0.05519	0.00218	0.02887
Taratibu	506	6.34113	1.45099	0.01115	0.88510	0.02592	0.00149	0.10616	0.03158	0.00164	0.89384	0.03158	0.00164	0.15156	0.03835	0.00181	0.84844	0.03835	0.00181	0.16535	0.04571	0.00198	0.13523
Milawilia	0	1.00000	0.00000	0.00000	0.97568	0.01157	0.00113	0.04863	0.04627	0.00226	0.95137	0.04627	0.00226	0.02432	0.01157	0.00113	0.97568	0.01157	0.00113	0.04863	0.04627	0.00226	0.00000
Mvuha	0	1.00000	0.00000	0.00000	0.97373	0.01245	0.00120	0.05254	0.04979	0.00240	0.94746	0.04979	0.00240	0.02627	0.01245	0.00120	0.97373	0.01245	0.00120	0.05254	0.04979	0.00240	0.00000
USFR	0	1.00000	0.00000	0.00000	0.98583	0.00689	0.00101	0.02834	0.02754	0.00203	0.97166	0.02754	0.00203	0.01417	0.00689	0.00101	0.98583	0.00689	0.00101	0.02834	0.02754	0.00203	0.00000
Segoma	2	1.30198	0.21081	0.00426	0.94752	0.02149	0.00136	0.07778	0.06374	0.00235	0.92222	0.06374	0.00235	0.05648	0.02355	0.00143	0.94352	0.02355	0.00143	0.09501	0.07135	0.00248	0.02584

Stacks

	Kazimzumbwi	Nguu	Makangala	Litipo	Noto	Gorongosa	Nyamuete	Kiwengoma	Muyuyu	Kibasira	Zaraninge	Gendagenda	Mabayani	Nilo	Mkungwe	Mount_Mabu	Kimboza	Shimba	Taratibu	Milawilia	Mvuha	USFR	Segoma
Kazimzumbwi		0.651907	0.380009	0.245843	0.185239	0.826774	0.406634	0.46134	0.654684	0.29927	0.66	0.342018	0.479256	0.79615	0.619631	0.833682	0.470559	0.604587	0.326466	0.746587	0.736789	0.810533	0.607912
Nguu			0.362694	0.245813	0.188422	0.768773	0.380375	0.425728	0.592215	0.286497	0.5386	0.269778	0.417311	0.679745	0.536655	0.766644	0.417849	0.529602	0.329878	0.645658	0.638634	0.72935	0.523589
Makangala				0.110466	0.0938261	0.434746	0.265176	0.289563	0.348172	0.23379	0.364997	0.306648	0.339679	0.420681	0.345755	0.398659	0.329723	0.371297	0.229157	0.389717	0.387089	0.334244	0.372086
Litipo					0.0731138	0.300475	0.187642	0.20087	0.228359	0.176429	0.249249	0.236882	0.245696	0.277129	0.230545	0.266942	0.23483	0.26014	0.191377	0.258439	0.25522	0.21323	0.260536
Noto						0.23451	0.144221	0.152067	0.169453	0.138688	0.190668	0.196847	0.195949	0.21485	0.177021	0.205833	0.187082	0.202794	0.170957	0.198044	0.194733	0.15124	0.206282
Gorongosa							0.551117	0.593001	0.767856	0.453355	0.771295	0.519967	0.630806	0.856162	0.741021	0.534187	0.641528	0.700392	0.286974	0.833344	0.817072	0.848627	0.73008
Nyamuete								0.247787	0.373081	0.20685	0.389578	0.276753	0.333582	0.454742	0.362535	0.521463	0.316012	0.387806	0.28569	0.418938	0.412273	0.415606	0.380899
Kiwengoma									0.413013	0.222366	0.436921	0.293062	0.364468	0.501476	0.396506	0.570282	0.346847	0.426577	0.29536	0.463642	0.459277	0.466851	0.428211
Muyuyu										0.268489	0.59669	0.327687	0.448297	0.725221	0.562042	0.768005	0.446015	0.544483	0.315035	0.667526	0.656284	0.716524	0.545038
Kibasira											0.298792	0.239943	0.271044	0.346676	0.272004	0.420098	0.253308	0.310499	0.271146	0.316268	0.310553	0.261291	0.310109
Zaraninge												0.280814	0.410198	0.691742	0.551425	0.772658	0.427062	0.527769	0.327692	0.659973	0.649702	0.719101	0.513321
Gendagenda													0.210227	0.309917	0.319603	0.49176	0.280081	0.282164	0.306855	0.360947	0.348694	0.415366	0.239053
Mabayani														0.443088	0.430934	0.611118	0.361015	0.370484	0.319645	0.489186	0.492043	0.554538	0.338305
Nilo															0.675083	0.864421	0.525468	0.563626	0.3436	0.790618	0.797646	0.859102	0.564079
Mkungwe																0.736826	0.346809	0.530152	0.322785	0.576101	0.574257	0.65792	0.527886
Mount_Mabu																	0.617325	0.692999	0.247368	0.830449	0.820858	0.861163	0.716877
Kimboza																		0.439471	0.31958	0.427243	0.422328	0.542962	0.433234
Shimba																			0.326293	0.601226	0.602297	0.6588	0.454051
Taratibu																				0.342791	0.339979	0.315068	0.334298
Milawilia																					0.70432	0.801932	0.613982
Mvuha																						0.782483	0.610214
USFR																							0.66101