ANGSD-wrapper: utilities for analyzing next generation sequencing data

Arun Durvasula[1,†], Paul J. Hoffman[2,†], Tyler V. Kent[1], Chaochih Liu[2], Thomas J. Y. Kono[2], Peter L. Morrell[2] and Jeffrey Ross-Ibarra[1,3,*]

[1]Department of Plant Sciences, University of California, Davis, CA 95616
[2]Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108
[3]Center for Population Biology and Genome Center, University of California, Davis, CA 95616
[†]These authors contributed equally.
[*]email: rossibarra@ucdavis.edu

January 18, 2016

Abstract

High throughput sequencing has changed many aspects of population genetics, molecular ecology, and related fields, affecting both experimental design and data analysis. The software package ANGSD allows users to perform a number of population genetic analyses on high-throughput sequencing data. The package is specifically designed to produce more accurate results for samples with low sequencing depth and makes use of full genome data while handling a wide array of sampling and experimental designs. Here we present ANGSD-wrapper, a user-friendly interface for running ANGSD and visualizing results. ANGSD-wrapper includes a number of 'wrapper' scripts that facilitate configuration and execution of multi-step analyses. ANGSD-wrapper also provides interactive graphing of ANGSD results to enhance data exploration. We demonstrate the usefulness of ANGSD-wrapper by analyzing resequencing data from populations of wild and domesticated *Zea*. ANGSD-wrapper is freely available from https://github.com/mojaveazure/angsd-wrapper.

Introduction

High throughput sequencing has revolutionized evolutionary genetics, allowing researchers to quickly assay large numbers of individuals or survey fine-scale patterns of variation across the genome. Application of these approaches has led to changes in both experimental design and data analysis [1]. Many popular software packages used by researchers for analysis of comparative resequencing data [see 2] were not designed to handle these novel data types or efficiently analyze the large volumes of data now being generated. In particular, short read sequencing has brought new challenges, including highly variable coverage, missing data, and high per-nucleotide error rates.

A number of tools have recently been published to estimate population genetic descriptive statistics using high throughput sequencing data [3, 4, 5, 6], but the majority of these either make limiting assumptions about the data (e.g., all sites have been sequenced, all genomes are haploid, sequencing is to sufficient depth, all individuals are outcrossing) or are specialized tools offering a narrow set of analysis options. Korneliussen et al. [7] recently published the software package ANGSD, which enables users to flexibly perform a large number of common population genetic analyses, including estimation of diversity statistics, admixture analysis including Patterson's D statistic [8], site frequency spectrum estimation [9], and calculation of neutrality test statistics [10]. One of the most important features of ANGSD is that most analyses are performed directly on genotype likelihoods, freeing users from the requirement of calling variants or genotypes and permitting analysis of low-coverage data or sequences with large amounts of missing data.

Here we present ANGSD-wrapper, a user-friendly interface to ANGSD. ANGSD-wrapper takes the form of a set of configuration files and 'wrapper' scripts (Figure S1) that streamline the execution of

Method	Calculations	Interactive Graphing
SFS	Site frequency spectrum	Yes
2DSFS	Joint site frequency spectrum, F_{ST}	Yes
ABBA/BABA	Patterson's D statistic	Yes
Ancestral	Extract ancestral sequence from BAM file	No
Genotypes	Genotype likelihood estimations	No
PCA	Principal component analysis	Yes
Thetas	Diversity statistics (θ_w, θ_π, Fu and Li's θ, Fay's θ) and neutrality tests (Tajima's D, Fu and Li's D, Fu and Li's F, Fay and Wu's H, Zeng's E)	Yes
Inbreeding	Calculate per-individual inbreeding coefficients with ngsF	No
Admixture	Perform admixture analysis	Yes

Table 1: Table of methods implemented in ANGSD-wrapper

multi-step pipelines required for data analysis in ANGSD. ANGSD-wrapper also eases the configuration of ANGSD-related programs such as ngsPopGen [**?**], ngsF[11], and ngsAdmix [12]. Because the large volume of data associated with high throughput sequence analysis is often difficult to visualize, ANGSD-wrapper also provides a suite of interactive visualization tools to plot results and explore patterns at multiple scales. We demonstrate some of the analyses possible using ANGSD-wrapper when applied to low-coverage whole-genome data from domesticated maize and two related wild teosinte subspecies. ANGSD-wrapper is freely available from https://github.com/mojaveazure/angsd-wrapper.

Methods

ANGSD-wrapper is a set of configuration files and scripts written primarily in the Bash UNIX shell. The scripts can be run either on a standalone computer with a UNIX terminal, or on computing clusters where they can be submitted to a queuing system such as SGE [13], Slurm [14] or TORQUE [15]. An installation of the statistical software R [16] is required to make use of the visualization tools incorporated in ANGSD-wrapper. The visualization portion of ANGSD-wrapper also requires installation of the R packages shiny [17], genomeIntervals [18], and ape [19].

ANGSD-wrapper is divided into scripts associated with analytical approaches implemented in ANGSD and associated software. ANGSD-wrapper provides a common configuration file, "common.conf," which holds variables that are likely to remain constant across analyses, including identifiers for chromosomal regions and the paths to project directories. In ANGSD-wrapper, each method is self-contained in a shell script which uses information from the common configuration file and a method-specific configuration file. Each analysis is run using a simple command:

```
$ angsd-wrapper <method> <configuration_file>
```

Analyses supported by ANGSD-wrapper are shown in Table 1. A detailed flowchart of each of these workflows is shown in Figure S1, and additional details, documentation, a tutorial, and a wiki can be found on the GitHub page: https://github.com/mojaveazure/angsd-wrapper/wiki.

The visualization software included with ANGSD-wrapper is contained within it's own directory called "shinyGraphing." This application must be started in R and can be accessed locally from a web browser. This software provides a graphical user interface (GUI) to quickly and interactively plot results obtained from ANGSD-wrapper. Each tab in the GUI contains plots for different ANGSD methods.

In order to use the plotting software, the user navigates to the desired tab and uploads the appropriate results file. The Shiny server automatically parses ANGSD output files and creates the resulting plot(s) (Figure 1), which can be saved using the browser's built in image saving capabilities.

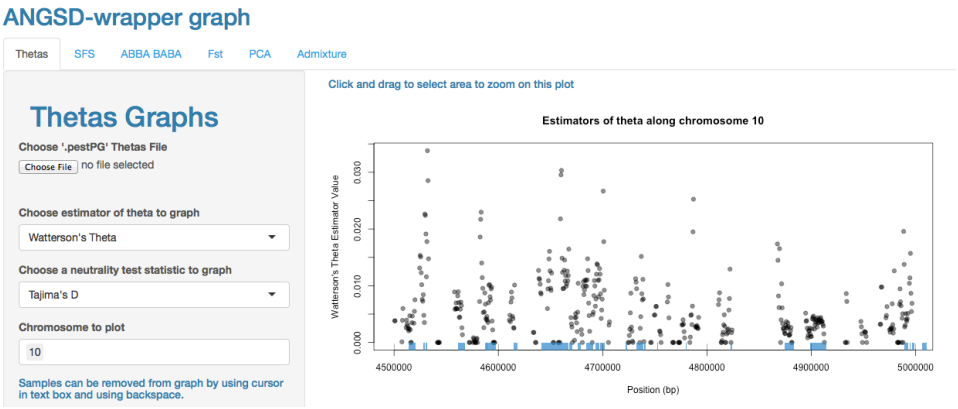

Figure 1: A visualization of Watterson's θ estimated by ANGSD across a 1.5 megabase region of chromosome 10 in *Zea mays* spp. *mays* using ANGSD-wrapper. Blue boxes indicate genic regions provided by a GFF annotation.

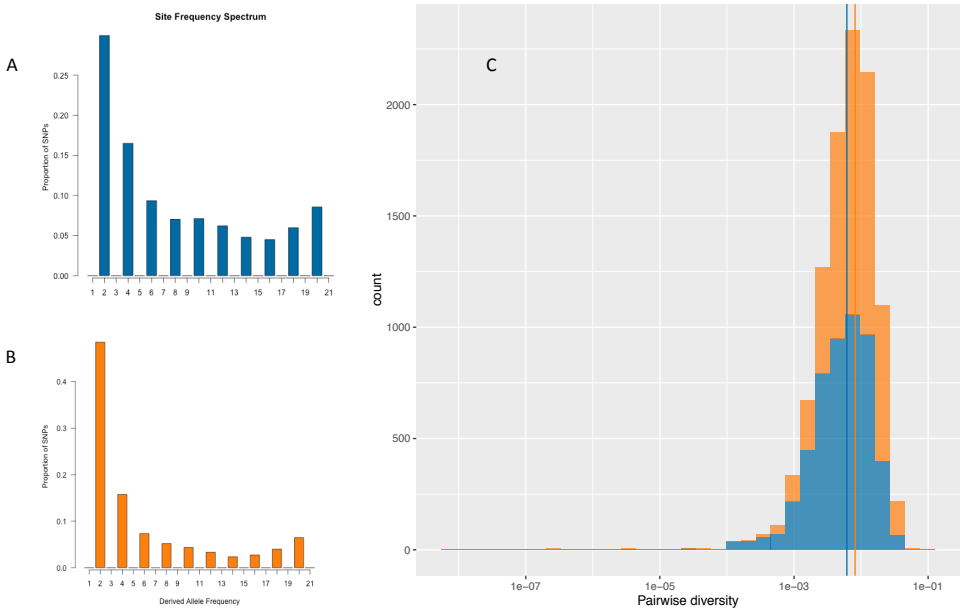

Figure 2: Summary statistics for *Zea mays*. Site frequency spectra for A. maize and B. teosinte. C. distribution of pairwise differences for maize (blue) and teosinte (orange). Pairwise diversity results are visualized separately from the interactive graphics and colors were added to A and B using a custom script.

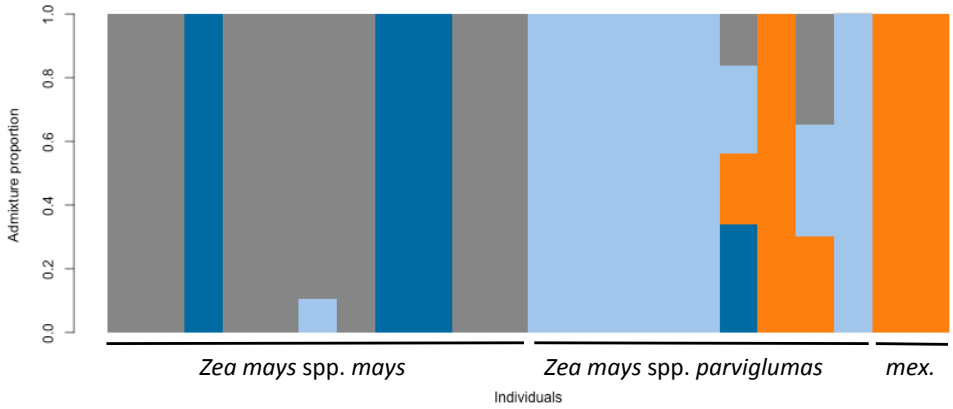

Figure 3: Admixture analysis for *Zea mays* spp. *mays*, *Zea mays* spp. *parviglumas*, and *Zea mays* spp. mexicana (mex) with K=4 source populations. Individuals are shown in same order as Table S1.

As a demonstration of analyses in ANGSD-wrapper, we explore patterns of diversity in a single genomic region of domesticated maize and wild teosinte. We used a subset of the resequenced samples from the Maize HapMap2 project (Table S1) and calculated summary statistics using a 10 megabase region on chromosome 10 [20]. The data are available at https://figshare.com/articles/Example_Data_tar_bz2/2063442. In the following we refer to methods listed in Table 1 when describing analyses.

please add Tripsacum to this table and add species identity to the table

We first use the SFS method to estimate the site frequency spectrum (SFS) of both maize and its wild progenitor *Zea mays* ssp. *parviglumis*, assuming an inbreeding coefficient of $F = 1$ for these highly inbred samples. The SFS and diversity statistics were calculated using ancestral states inferred in the "Ancestral Sequence" method from a single resequenced genome of *Tripsacum dactyloides*. Consistent with previous results [21], we find mean levels of nucleotide diversity in this region to be **X** , and show that the maize SFS is skewed towards more intermediate-frequency variants (Figure 2A-B, Tajima's D of **X** and **X** in maize and teosinte, respectively), likely a result of the bottleneck associated with maize domestication [22]. We find further evidence of the effect of domestication using the Thetas method, revealing lower overall levels of diversity in maize (Figure 2C). Using the 2DSFS method, which includes an F_{ST} calculation, we find a mean F_{ST} in this region of 0.116, nearly identical to the genome-wide value of 0.11 reported in [21]. There are no loci in this region that have previously been identified as showing evidence of selection during domestication, and these data also do not find any large consecutive regions of high F_{ST} .

Finally, we include two samples of the related wild teosinte *Zea mays* ssp. mexicana to assess evidence for admixture. Using the Admixture method, which implements an estimate of admixture proportion from genotype likelihoods [12], we identify structure within domesticated maize separating three high-latitude temperate landraces from the other tropical accessions (Figure 3). We find no evidence of admixture between these lowland maize samples and ssp. mexicana, consistent with an independent analysis from SNP genotyping [23]. *Zea mays* ssp. mexicana clusters into its own group, along with a single accession of ssp. parviglumis collected from region in which many teosinte populations appear to be the result of admixture between the two subspecies [24]. A single ssp. parviglumis accession from the Northern extent of the range does not appear to be well-classified with these data, likely due to our relatively limited geographic and genomic sampling.

parviglumis spelled wrong in figure legend

Conclusions

ANGSD-wrapper provides an easy-to-use interface that simplifies many population genetic analyses implemented in ANGSD [7] and permits the exploration of genome-scale results through interactive visualization. ANGSD-wrapper is under active development to incorporate updates to the ANGSD software package.

Acknowledgements

We acknowledge funding support from the US National Science Foundation (IOS-1238014 to JRI and IOS-1339393 to PLM), from a University of Minnesota Doctoral Dissertation Fellowship supporting TJYK, and from the University of California Davis Plant Sciences Department. We thank members of the Ross-Ibarra and Morrell Labs for discussion and software testing. We thank the authors of ANGSD and related programs for answering questions, particularly Matteo Fumagalli and Filipe Vieira. Finally, we would like to thank Felix Andrews for statistical advice, although we did not follow it.

References

[1] Robert Ekblom and Juan Galindo. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1):1–15, 2011.

[2] Laurent Excoffier and Gerald Heckel. Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics*, 7(10):745–758, 2006.

[3] Daniel Garrigan. Popbam: tools for evolutionary analysis of short read sequence alignments. *Evolutionary bioinformatics online*, 9:343, 2013.

[4] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[5] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[6] Stephan Hutter, Albert J Vilella, and Julio Rozas. Genome-wide dna polymorphism analyses using variscan. *BMC bioinformatics*, 7(1):409, 2006.

[7] Thorfinn S Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. Angsd: analysis of next generation sequencing data. *BMC bioinformatics*, 15(1):356, 2014.

[8] Eric Y Durand, Nick Patterson, David Reich, and Montgomery Slatkin. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8):2239–2252, August 2011.

[9] R. Nielsen, T. Korneliussen, A. Albrechtsen, Y. Li, and J. Wang. SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. *PLoS ONE*, 7(7):e37558, 2012.

[10] Thorfinn Korneliussen, Ida Moltke, Anders Albrechtsen, and Rasmus Nielsen. Calculation of tajima's d and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, 14(1):289, 2013.

[11] Filipe G Vieira, Matteo Fumagalli, Anders Albrechtsen, and Rasmus Nielsen. Estimating inbreeding coefficients from ngs data: Impact on genotype calling and allele frequency estimation. *Genome research*, 23(11):1852–1861, 2013.

[12] L. Skotte, T. S. Korneliussen, and A. Albrechtsen. Estimating Individual Admixture Proportions from Next Generation Sequencing Data. *Genetics*, Sep 2013.

[13] W. Gentzsch. Sun grid engine: Towards creating a compute power grid. In *Proceedings of the 1st International Symposium on Cluster Computing and the Grid*, CCGRID '01, pages 35–, Washington, DC, USA, 2001. IEEE Computer Society.

[14] Morris A. Jette, Andy B. Yoo, and Mark Grondona. Slurm: Simple linux utility for resource management. In *In Lecture Notes in Computer Science: Proceedings of Job Scheduling Strategies for Parallel Processing (JSSPP) 2003*, pages 44–60. Springer-Verlag, 2002.

[15] Garrick Staples. Torque resource manager. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, SC '06, New York, NY, USA, 2006. ACM.

[16] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[17] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2015. R package version 0.11.1.

[18] Julien Gagneur, Joern Toedling, Richard Bourgon, and Nicolas Delhomme. *genomeIntervals: Operations on genomic intervals*, 2015. R package version 1.22.1.

Sample	Mean Depth
BKN009	7.00194
BKN011	6.9777
BKN014	6.78829
BKN015	3.739
BKN019	3.71944
BKN022	3.71268
BKN025	3.53799
BKN026	3.91798
BKN027	7.05576
BKN033	3.85336
BKN035	3.74839
TIL01	3.60133
TIL03	4.07518
TIL04	6.09163
TIL07	5.11419
TIL09	5.29004
TIL11	3.15477
TIL15	6.87873
TIL16	2.67186
TIL17	2.61892
TIL08	6.09453
TIL25	13.1566

Table S1: Table of samples used in analysis with mean depth over the region 15000000-25000000 on chromosome 10

[19] E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

[20] Jer-Ming Chia, Chi Song, Peter J Bradbury, Denise Costich, Natalia de Leon, John Doebley, Robert J Elshire, Brandon Gaut, Laura Geller, Jeffrey C Glaubitz, et al. Maize hapmap2 identifies extant variation from a genome in flux. *Nature genetics*, 44(7):803–807, 2012.

[21] Matthew B Hufford, Xun Xu, Joost Van Heerwaarden, Tanja Pyhäjärvi, Jer-Ming Chia, Reed A Cartwright, Robert J Elshire, Jeffrey C Glaubitz, Kate E Guill, Shawn M Kaeppler, et al. Comparative population genomics of maize domestication and improvement. *Nature genetics*, 44(7):808–811, 2012.

[22] Timothy Mathes Beissinger, Li Wang, Kate Crosby, Arun Durvasula, Matthew B Hufford, and Jeff Ross-Ibarra. Recent demography drives changes in linked selection across the maize genome. *bioRxiv*, 2015.

[23] Matthew B Hufford, Pesach Lubinksy, Tanja Pyhäjärvi, Michael T Devengenzo, Norman C Ellstrand, and Jeffrey Ross-Ibarra. The genomic signature of crop-wild introgression in maize. *PLoS genetics*, 2013.

[24] Zhou Fang, Tanja Pyhäjärvi, Allison L Weber, R Kelly Dawe, Jeffrey C Glaubitz, José de Jesus Sánchez González, Claudia Ross-Ibarra, John Doebley, Peter L Morrell, and Jeffrey Ross-Ibarra. Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics*, 191(3):883–894, 2012.

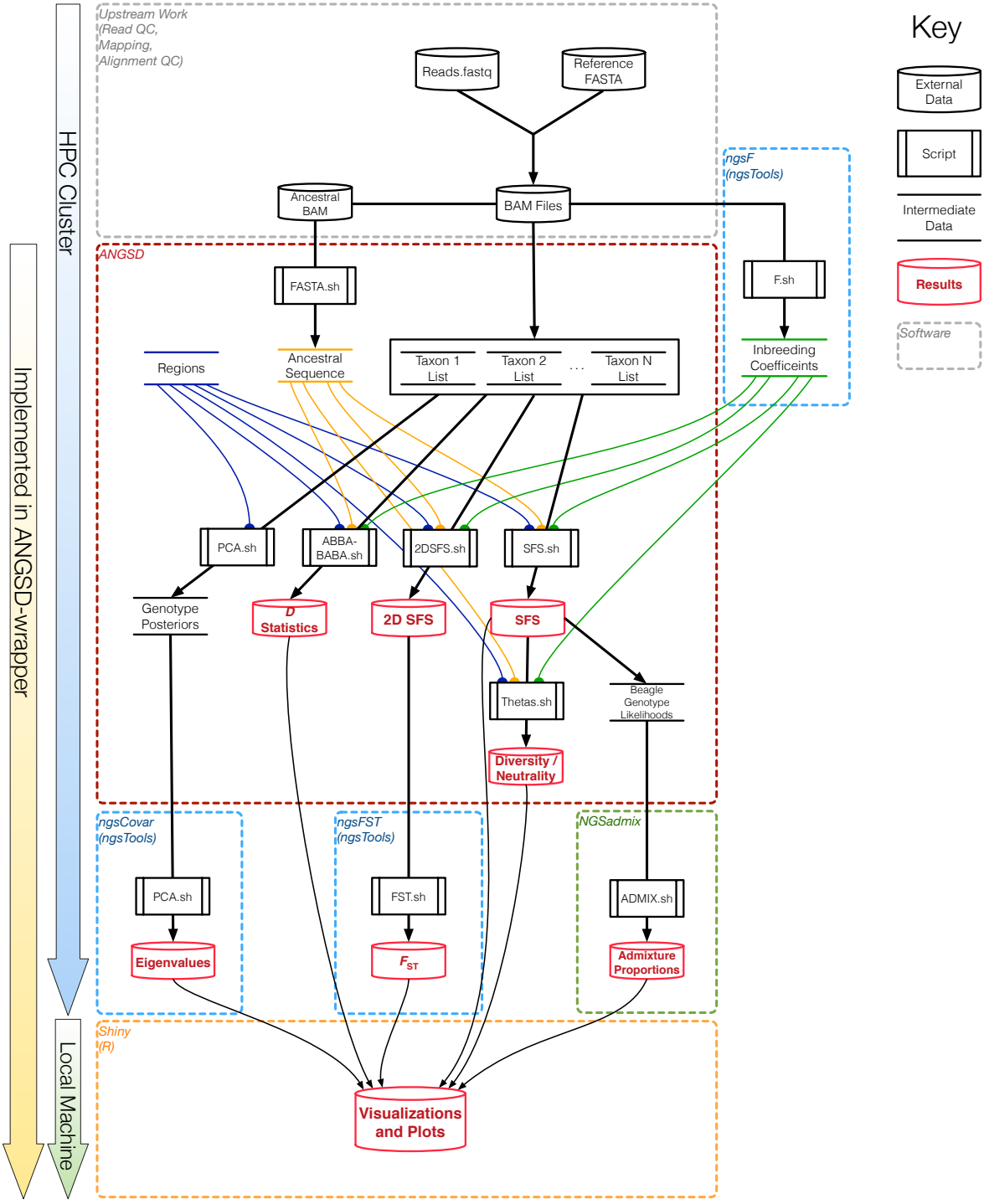

Figure S1: Workflow diagram for all methods available in ANGSD-wrapper.

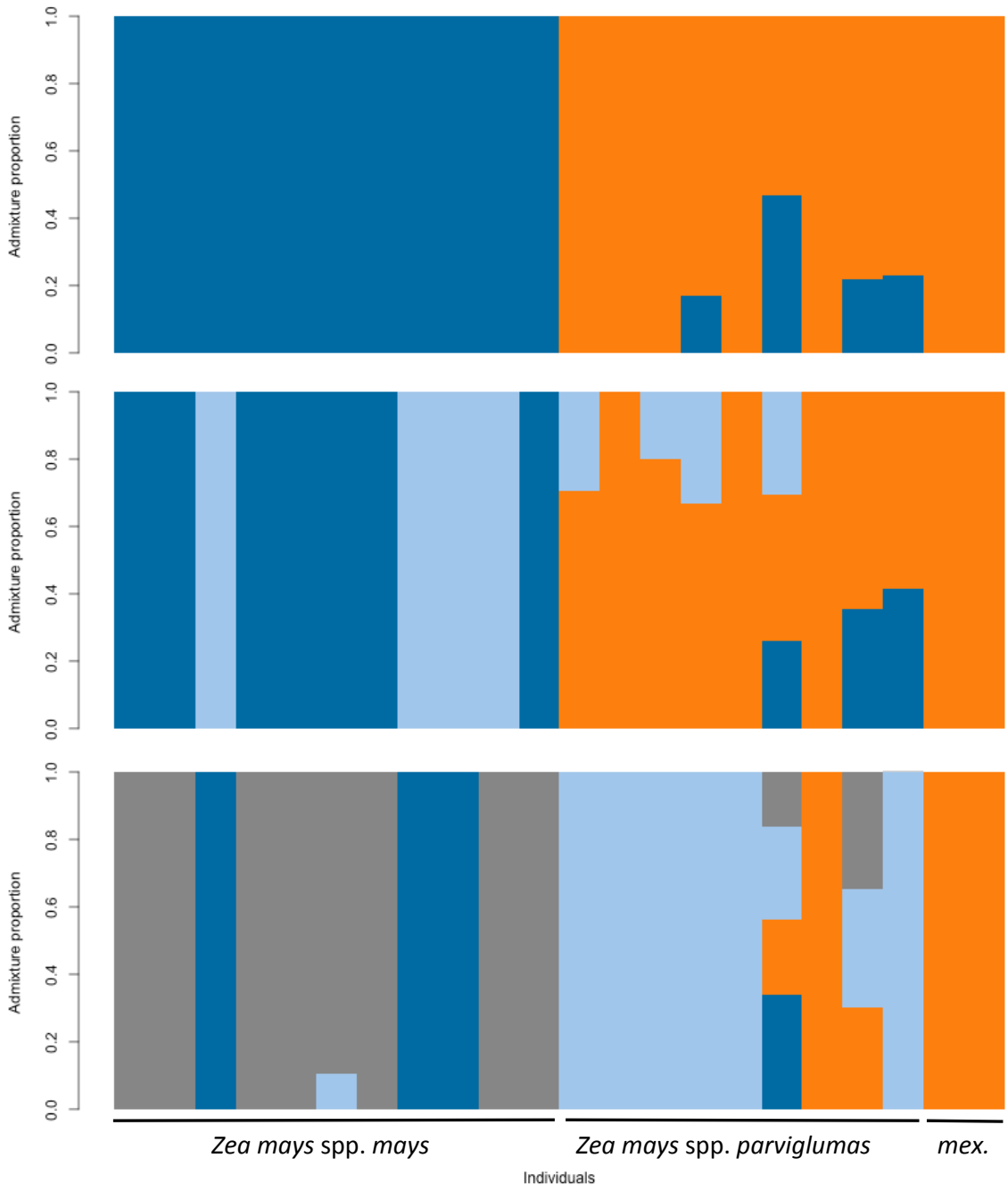

Figure S2: Admixture analysis for K=2 (top), K=3 (middle), and K=4 (bottom).