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 28, 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 (Ekblom and Galindo, 2011). Many popular software packages used by researchers for analysis of comparative resequencing data (see Excoffier and Heckel, 2006) were not designed to handle these novel data types or efficiently analyze the large volumes of data now being generated. Despite the decreasing cost of sequencing, researchers must allocate finite resources and balance the depth of sequencing and the breadth of a sample. This poses a challenge for population genetics analysis, which generally requires accurate polymorphism calls in a broad sample. This is especially true of studies that span multiple subpopulations, as this balance must be met for each partition (Felsenstein, 2006; Pluzhnikov and Donnelly, 1996). While the experimental design challenges with molecular population genetic studies have existed for at least a decade, high throughput sequencing brings the technical challenges of 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 (Danecek et al., 2011; Garrigan, 2013; Hutter et al., 2006; Purcell et al., 2007). However, few of these tools offer the flexibility to handle data with the characteristics of high throughput sequence data. Korneliussen et al. (Korneliussen et al., 2014) 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 (Durand et al., 2011), site frequency spectrum estimation (Nielsen et al., 2012), and calculation

Should be just 2014

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

of neutrality test statistics (Korneliussen et al., 2013). ANGSD works directly with the alignment formats produced from standard high throughput sequence analysis pipelines, which removes the need for the user to transform the data into a niche format. One of the most important features of ANGSD is that analyses are integrated over per-site genotype likelihoods, rather than only on pre-determined variable sites. This permits ANGSD to calculate common population genetic descriptive statistics on low-coverage sequencing data, and handle missingness due to variation in coverage.

Here we present ANGSD-wrapper, a user-friendly interface to stable versions of ANGSD. ANGSD-wrapper takes the form of a set of configuration files and 'wrapper' scripts (Figure S1) that streamline the execution of multi-step pipelines required for data analysis in ANGSD. ANGSD-wrapper also eases the configuration of ANGSD-related programs such as ngsPopGen (Fumagali), ngsF(Vieira et al., 2013), and ngsAdmix (Skotte et al., 2013). Additionally, the wrapper scripts are written against "frozen" versions of ANGSD and supporting tools, for consistency of analysis. 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 scripting language. 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 (Gentzsch, 2001), Slurm (Jette et al., 2002) or TORQUE (Staples, 2006). A Python installation (van Rossum) is required for some light, dynamic pre-processing of the data. An installation of the statistical software R (R Core Team, 2014) 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` (Chang et al., 2015), `APE` (Paradis et al., 2004), `Lattice` (Sarkar, 2008), `Hmisc` (Jr et al., 2015), `data.table` (Dowle et al., 2015), `DT` (Xie, 2015), and `shinythemes` (Chang, 2015), as well as `genomeIntervals` (Gagneur et al., 2015) from Bioconductor (Huber et al., 2015) for the visualization package. These dependent packages are installed automatically upon first run of visualization interface.

change of language

has anyone tested AW on a cluster?

ANGSD-wrapper is divided into scripts associated with analytical approaches implemented in ANGSD and associated software. ANGSD-wrapper provides a common configuration file, `Common_Config`, 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

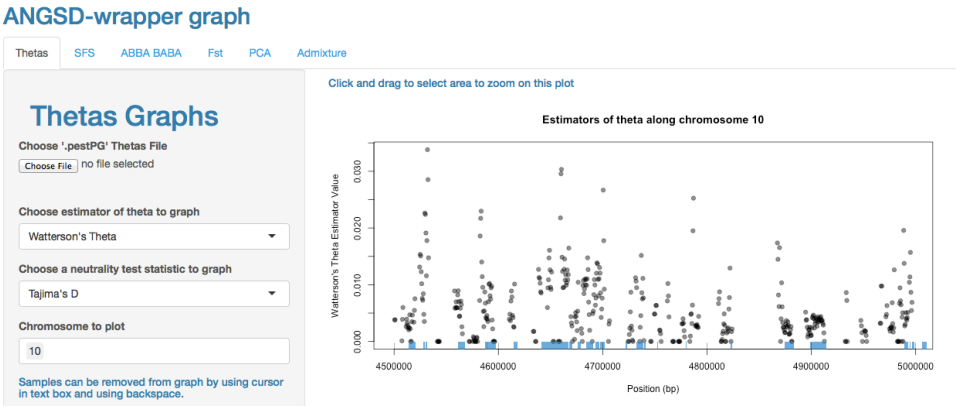

Figure 1: Visualization of Watterson's θ estimated by ANGSD across a 1.5Mb region of chromosome 10 in *Zea mays* ssp. *mays* using ANGSD-wrapper. Darker colors indicate a higher density of points. Blue boxes indicate gene annotations provided by a GFF file.

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 is started from within ANGSD-wrapper and launched locally from a standard 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.

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 10Mb region on chromosome 10 (Chia et al., 2012). 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.

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*. We show that the maize SFS is skewed towards more intermediate-frequency variants (Figure 2A-B, Tajima's D of 0.2 and 0.0085 in maize and teosinte, respectively), likely a result of the bottleneck associated with maize domestication (Beissinger et al., 2015; Eyre-Walker et al., 1998). Using the "Thetas" method we find further evidence of the domestication bottleneck, with mean levels of pairwise nucleotide diversity in this region in maize $\approx 25\%$ lower than in teosinte (0.0061 and 0.0082, respectively; 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 Hufford et al. (2012). There are no genes in this region that have been identified as potential domestication candidates, consistent with the lack of extended regions of high F_{ST} in our analysis (Figure S3).

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 (Skotte et al., 2013), we identify structure within domesticated maize separating three high-latitude temperate landraces from the other tropical accessions (Figure 3). This analysis is a supervised method that assigns proportions of the each individual's genome to a

what are vertical lines in pi density plot? need to be consistent about saying "nucleotide diversity" vs "pairwise diversity" vs "pairwise differences" etc. I vote for "pairwise nucleotide diversity"

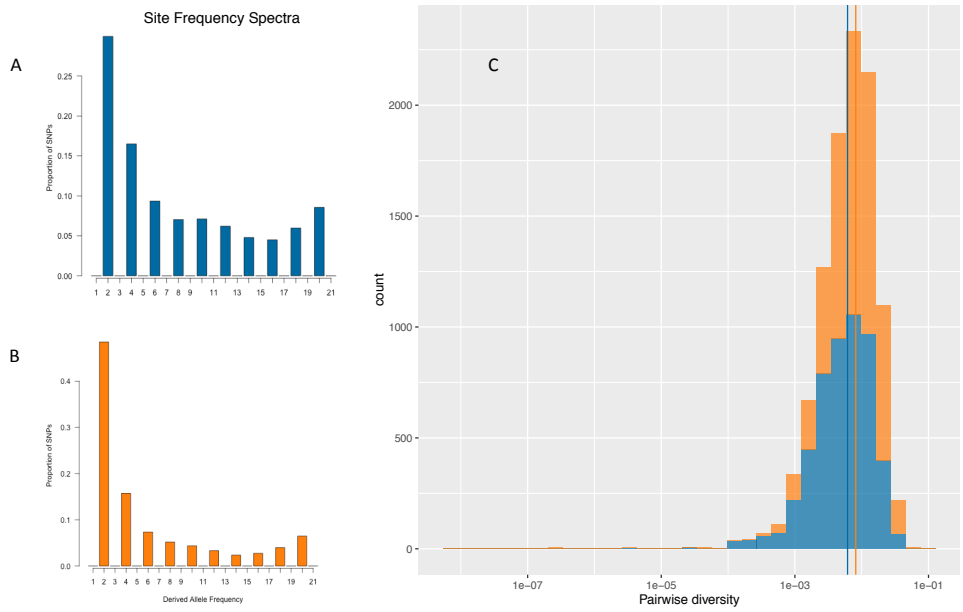

Figure 2: Summary statistics for *Zea mays*. Derived site frequency spectra for A. maize and B. teosinte. C. distribution of pairwise differences for maize (blue) and teosinte (orange). Pairwise diversity results are visualized separately from the interactive graphics and colors were added to A and B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

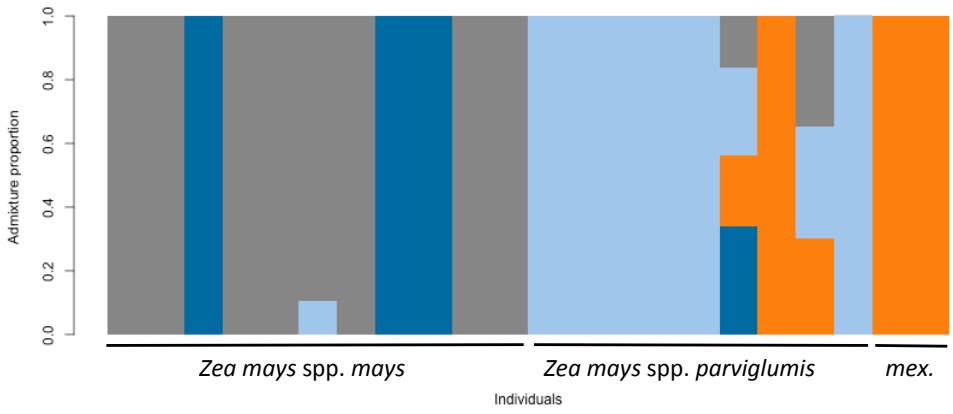

Figure 3: Admixture analysis for *Zea mays* ssp. *mays*, *Zea mays* ssp. *parviglumis*, and *Zea mays* ssp. *mexicana* (*mex*) with colors representing the K=4 source populations. Individuals are shown in the same order as Table S1.

putative source population, represented by the colors. We ran this analysis over a range of number of ancestral populations (K= 2, 3, 4; See Figure S2) and chose the value with the maximum likelihood while keeping the number of parameters as low as possible. In maize, we see that most individuals come from one source population (dark grey or dark blue), which is evidence for no admixture between these lowland maize samples and ssp. *mexicana*, and is consistent with an independent analysis from SNP genotyping (Hufford et al., 2013). *Zea mays* ssp. *mexicana* clusters into its own group (orange), along with a single accession of ssp. *parviglumis* collected from a region in which many teosinte populations appear to be the result of admixture between the two subspecies (Fang et al., 2012). 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.

Conclusions

ANGSD-wrapper provides an easy-to-use interface that simplifies many population genetic analyses implemented in ANGSD (Korneliussen et al., 2014) 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

Beissinger, T. M., Wang, L., Crosby, K., Durvasula, A., Hufford, M. B., and Ross-Ibarra, J. (2015). Recent demography drives changes in linked selection across the maize genome. *bioRxiv*.

Chang, W. (2015). *shinythemes: Themes for Shiny*. R package version 1.0.1.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2015). *shiny: Web Application Framework for R*. R package version 0.11.1.

Chia, J.-M., Song, C., Bradbury, P. J., Costich, D., de Leon, N., Doebley, J., Elshire, R. J., Gaut, B., Geller, L., Glaubitz, J. C., et al. (2012). Maize hapmap2 identifies extant variation from a genome in flux. *Nature Genetics*, 44(7):803–807.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., et al. (2011). The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158.

Dowle, M., Srinivasan, A., Short, T., with contributions from R Saporta, S. L., and Antonyan, E. (2015). *data.table: Extension of Data.frame*. R package version 1.9.6.

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8):2239–2252.

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

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

Eyre-Walker, A., Gaut, R. L., Hilton, H., Feldman, D. L., and Gaut, B. S. (1998). Investigation of the bottleneck leading to the domestication of maize. *Proceedings of the National Academy of Sciences*, 95(8):4441–4446.

Fang, Z., Pyhäjärvi, T., Weber, A. L., Dawe, R. K., Glaubitz, J. C., González, J. d. J. S., Ross-Ibarra, C., Doebley, J., Morrell, P. L., and Ross-Ibarra, J. (2012). Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics*, 191(3):883–894.

Felsenstein, J. (2006). Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Molecular biology and evolution*, 23(3):691–700.

Fumagali, M. ngspopgen. https://github.com/mfumagalli/ngsPopGen. Accessed: 2016-01-18.

Gagneur, J., Toedling, J., Bourgon, R., and Delhomme, N. (2015). *genomeIntervals: Operations on genomic intervals*. R package version 1.22.1.

Garrigan, D. (2013). Popbam: tools for evolutionary analysis of short read sequence alignments. *Evolutionary Bioinformatics Online*, 9:343.

Gentzsch, W. (2001). 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. IEEE Computer Society.

Huber, W., Carey, J., V., Gentleman, R., Anders, S., Carlson, M., Carvalho, S., B., Bravo, C., H., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, D., K., Irizarry, A., R., Lawrence, M., Love, I., M., MacDonald, J., Obenchain, V., Ole's, K., A., Pag'es, H., Reyes, A., Shannon, P., Smyth, K., G., Tenenbaum, D., Waldron, L., Morgan, and M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121.

Hufford, M. B., Lubinksy, P., Pyhäjärvi, T., Devengenzo, M. T., Ellstrand, N. C., and Ross-Ibarra, J. (2013). The genomic signature of crop-wild introgression in maize. *PLoS Genetics*.

Hufford, M. B., Xu, X., Van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R. A., Elshire, R. J., Glaubitz, J. C., Guill, K. E., Kaeppler, S. M., et al. (2012). Comparative population genomics of maize domestication and improvement. *Nature Genetics*, 44(7):808–811.

Hutter, S., Vilella, A. J., and Rozas, J. (2006). Genome-wide dna polymorphism analyses using variscan. *BMC Bioinformatics*, 7(1):409.

Jette, M. A., Yoo, A. B., and Grondona, M. (2002). 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.

Jr, F. E. H., with contributions from Charles Dupont, and many others. (2015). *Hmisc: Harrell Miscellaneous*. R package version 3.17-1.

Korneliussen, T., Moltke, I., Albrechtsen, A., and Nielsen, R. (2013). Calculation of tajima's d and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, 14(1):289.

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). Angsd: analysis of next generation sequencing data. *BMC Bioinformatics*, 15(1):356.

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

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

Pluzhnikov, A. and Donnelly, P. (1996). Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics*, 144(3):1247–1262.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York. ISBN 978-0-387-75968-5.

Skotte, L., Korneliussen, T. S., and Albrechtsen, A. (2013). Estimating Individual Admixture Proportions from Next Generation Sequencing Data. *Genetics*.

Staples, G. (2006). Torque resource manager. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, SC '06, New York, NY, USA. ACM.

van Rossum, G. Python. https://www.python.org. Accessed: 2016-01-27.

Vieira, F. G., Fumagalli, M., Albrechtsen, A., and Nielsen, R. (2013). Estimating inbreeding coefficients from ngs data: Impact on genotype calling and allele frequency estimation. *Genome Research*, 23(11):1852–1861.

Xie, Y. (2015). *DT: A Wrapper of the JavaScript Library 'DataTables'*. R package version 0.1.

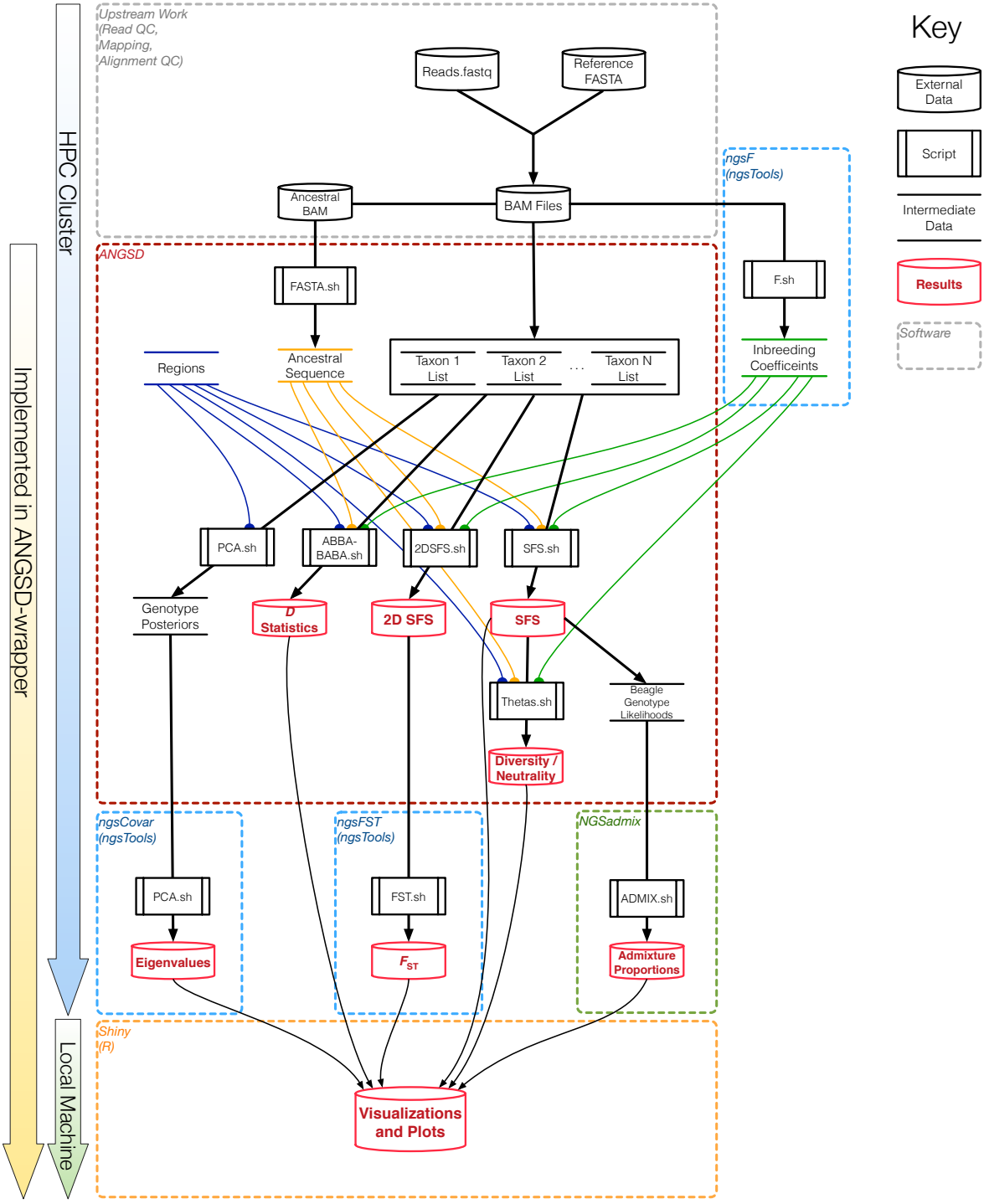

Figure S1: Workflow diagram for all methods available in ANGSD-wrapper.

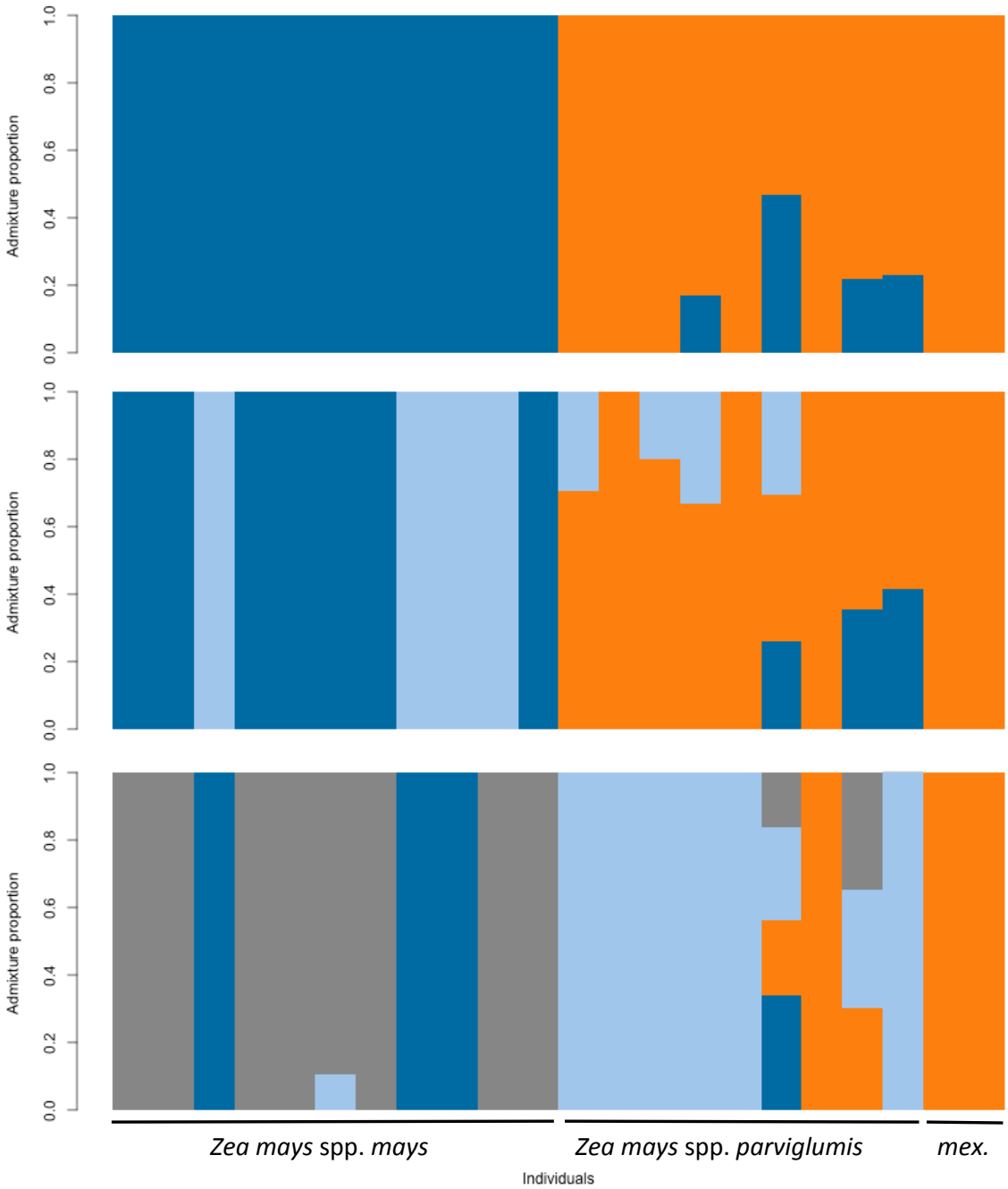

Figure S2: Admixture analysis for K=2 (top), K=3 (middle), and K=4 (bottom).

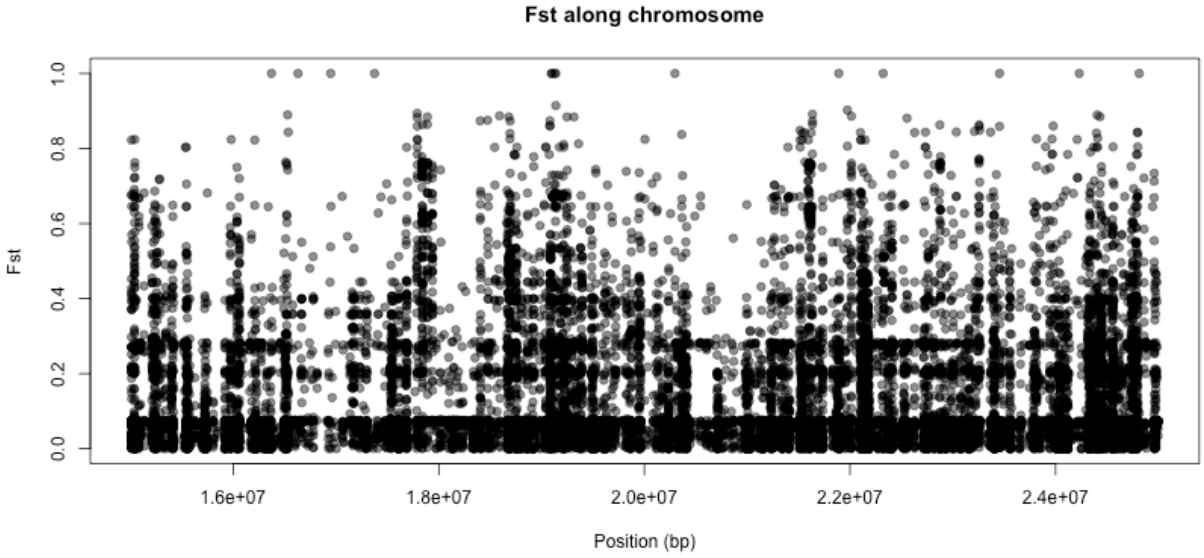

Figure S3: F_{ST} values plotted against base pair position on chromosome 10 of maize. F_{ST} is calculated between the maize and teosinte samples.

Sample	**Mean Depth**	**Species**
BKN009	7.00194	*Zea mays* spp. *mays*
BKN011	6.9777	*Zea mays* spp. *mays*
BKN014	6.78829	*Zea mays* spp. *mays*
BKN015	3.739	*Zea mays* spp. *mays*
BKN019	3.71944	*Zea mays* spp. *mays*
BKN022	3.71268	*Zea mays* spp. *mays*
BKN025	3.53799	*Zea mays* spp. *mays*
BKN026	3.91798	*Zea mays* spp. *mays*
BKN027	7.05576	*Zea mays* spp. *mays*
BKN033	3.85336	*Zea mays* spp. *mays*
BKN035	3.74839	*Zea mays* spp. *mays*
TIL01	3.60133	*Zea mays* spp. *parviglumis*
TIL03	4.07518	*Zea mays* spp. *parviglumis*
TIL04	6.09163	*Zea mays* spp. *parviglumis*
TIL07	5.11419	*Zea mays* spp. *parviglumis*
TIL09	5.29004	*Zea mays* spp. *parviglumis*
TIL11	3.15477	*Zea mays* spp. *parviglumis*
TIL15	6.87873	*Zea mays* spp. *parviglumis*
TIL16	2.67186	*Zea mays* spp. *parviglumis*
TIL17	2.61892	*Zea mays* spp. *parviglumis*
TIL08	6.09453	*Zea mays* ssp. *mexicana*
TIL25	13.1566	*Zea mays* ssp. *mexicana*
TDD39103	8.62	*Tripsacum dactyloides*

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