ANGSD-wrapper

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, but it handles a wide array of sampling and experimental designs and makes use of full genome data. Here we present ANGSD-wrapper, a user-friendly interface to ANGSD. 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, thus enhancing data exploration. We demonstrate the usefulness of ANGSD-wrapper by analyzing resequencing data from populations of wild and domesticated *Oryza*. ANGSD-wrapper is freely available from https://github.com/ arundurvasula/angsd- wrapper.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1472v1 | CC-BY 4.0 Open Access | rec: 3 Nov 2015, publ: 3 Nov 2015

ANGSD-wrapper

Arun Durvasula[1], Tyler V. Kent[1], Paul J. Hoffman[2], Chaochih Liu[2], Peter L. Morrell[2] , 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

Corresponding Author:

Jeffrey Ross-Ibarra[1,3]

1 Shields Avenue, Department of Plant Sciences, University of California, Davis, CA 95616

Email address: rossibarra@ucdavis.edu

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1472v1 | CC-BY 4.0 Open Access | rec: 3 Nov 2015, publ: 3 Nov 2015

ANGSD-wrapper

Arun Durvasula[1], Tyler V. Kent[1], Paul J. Hoffman[2], 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**

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, but it handles a wide array of sampling and experimental designs and makes use of full genome data. Here we present ANGSD-wrapper, a user-friendly interface to ANGSD. 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, thus enhancing data exploration. We demonstrate the usefulness of ANGSD-wrapper by analyzing resequencing data from populations of wild and domesticated *Oryza*. ANGSD-wrapper is freely available from `https://github.com/arundurvasula/angsd-wrapper`.

Keywords: software, population genetics, genotype likelihood

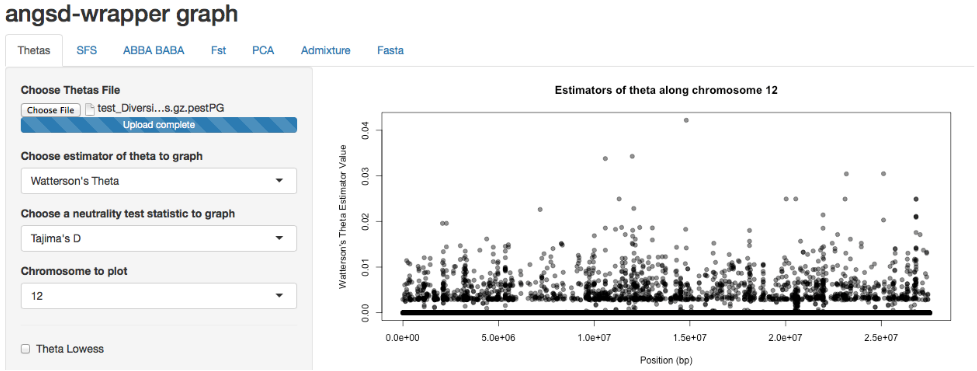
INTRODUCTION

High throughput sequencing has revolutionized evolutionary genetics, allowing researchers to quickly assay large numbers of individuals or survey fine-scale patterns of variation along the genome. Application of these methods has led to changes in both experimental design and data analysis (Ekblom and Galindo, 2011). Many of the popular software packages used by researchers (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. A particular challenge with short read sequencing has been higher per base pair rates of errors and missing data.

A number of tools have recently been published to handle high throughput sequencing data (Garrigan, 2013; Purcell et al., 2007; Danecek et al., 2011; Hutter et al., 2006), 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. (2014) recently published the software package ANGSD, which enables users to flexibly perform a large number of common population genetic analyses, including calculating the ABBA BABA D-statistic, site frequency spectrum estimation (Nielsen et al., 2012), and neutrality test statistics (Korneliussen et al., 2013) 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 S2) that streamline the execution of multi-step pipelines inherent in ANGSD as well as pipelines involving related programs such as ngsPopGen, ngsF(Vieira et al., 2013), and ngsAdmix (Skotte et al., 2013). Because the large volume of data associated with high throughput sequence analysis is often difficult to explore by hand, 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 in an initial

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1472v1 | CC-BY 4.0 Open Access | rec: 3 Nov 2015, publ: 3 Nov 2015



angsd-wrapper graph

Choose Thetas File: test_Diversi...s.gz.pestPG

Upload complete

Choose estimator of theta to graph: Watterson's Theta

Choose a neutrality test statistic to graph: Tajima's D

Chromosome to plot: 12

□ Theta Lowess

Figure 1. A visualization of Watterson's θ estimated by ANGSD across chromosome 12 of *O. glumaepatula* using angsd-wrapper

exploration of low-coverage whole-genome sequence data from populations of the endangered wild rice species *O. glumaepatula* and compare to the domesticated *Oryza sativa*. ANGSD-wrapper is freely available from `https://github.com/arundurvasula/angsd-wrapper`.

IMPLEMENTATION

ANGSD-wrapper is a set of configuration files and scripts written 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 (Gentzsch, 2001), Slurm (Jette et al., 2002) or TORQUE (Staples, 2006). 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), genomeIntervals (Gagneur et al., 2015), and ape (Paradis et al., 2004).

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 the run using a simple bash command:

```
$ bash scripts/<method>.sh scripts/<method>.conf
```

A detailed flowchart of each of these workflows is shown in Figure S2, and additional details, documentation, a tutorial, and a wiki can be found on the GitHub page: `https://github.com/arundurvasula/angsd-wrapper/wiki` or within the scripts directory.

The visualization software included with ANGSD-wrapper is contained within the script directory in a folder called shiny. 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 file of results. 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.

EXAMPLE ANALYSIS

As an example analysis using ANGSD-wrapper, we analyze low-coverage whole-genome sequence data from two populations of the endangered wild rice species *O. glumaepatula* . One of two sampled populations of *O. glumaepatula* occurs in sympatry with domesticated rice *O. indica*. Hybridization between these taxa has been observed Fuchs et al. (2015) and is thought to be a threat to the maintenance of *O. glumaepatula* as a distinct genetic entity Fuchs et al. (2015).

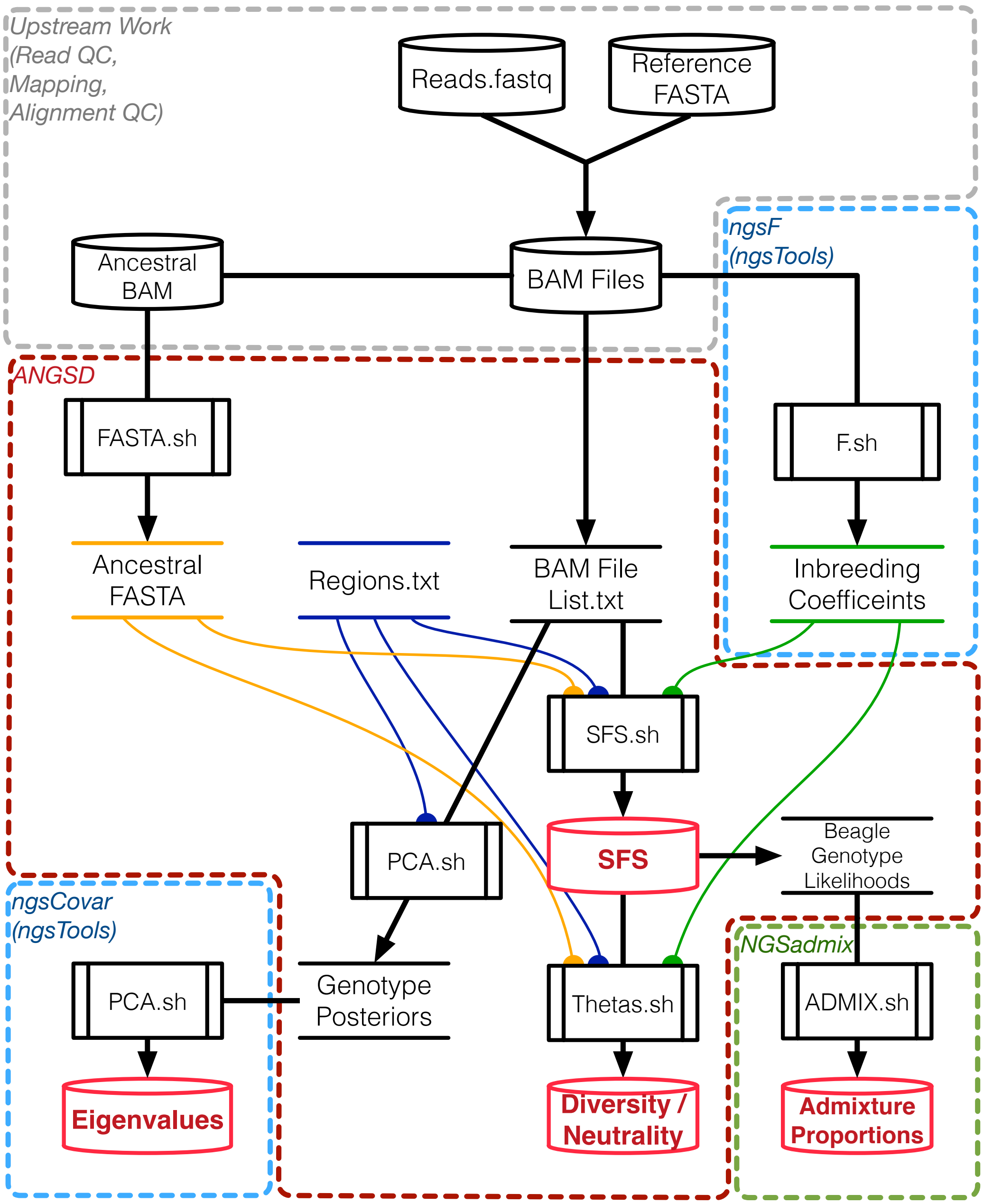


Figure 2. Example analysis workflow diagram.

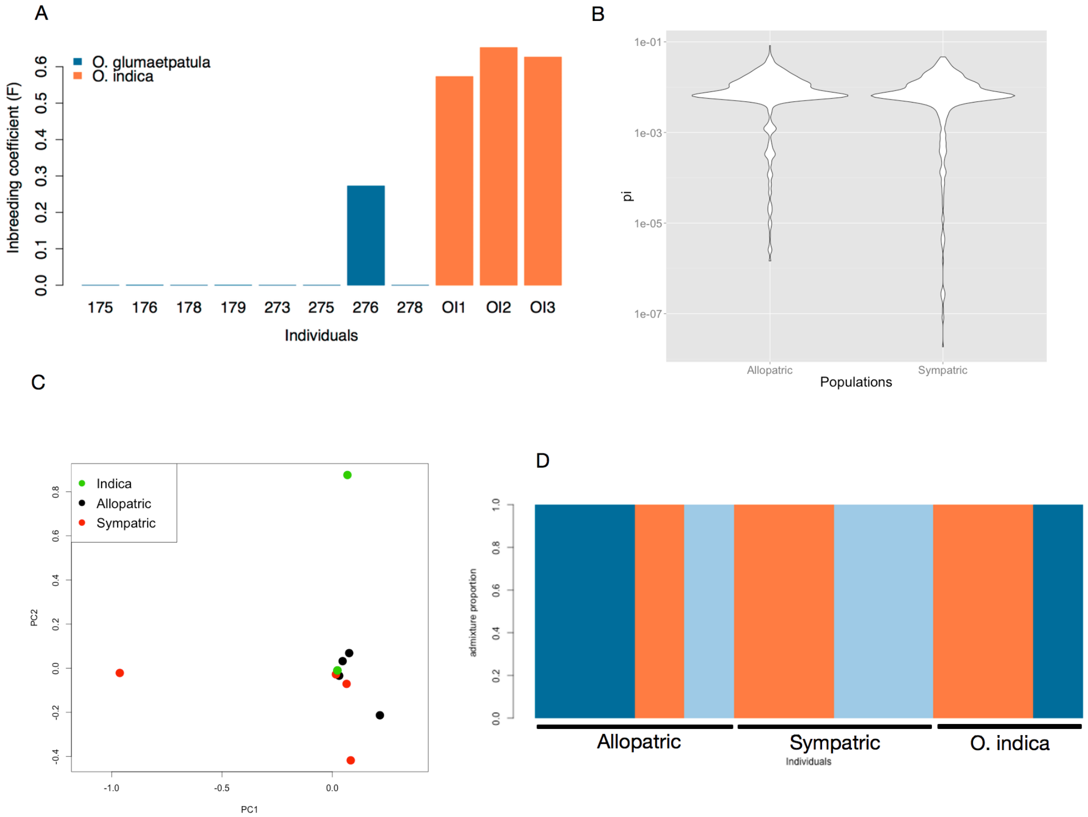


Figure 3. A. Inbreeding coefficients (F) for all samples. B. A comparison of θ_{π} for the *O. glumaepatula* allopatric and sympatric populations. Points represent mean θ_{π} across chromosome 12 and bars represent 1 standard deviation. C. Principle component analysis performed with ngsCovar. D. Admixture analysis performed with ngsAdmix for k=3.

We used paired end 100 base pair Illumina resequencing data constituting $\approx 1.25X$ coverage of 4 individuals in each of the populations. Further details of sample collection and sequencing are presented in Kent et al. (in preparation). We compare these samples to data from domesticated rice *O. sativa* taken from (Li et al., 2014) All samples were aligned to the *O. sativa* ssp. *japonica* reference genome using bwa-mem v0.7.5a-r405, (Li and Durbin, 2009) and the samples were sorted and indexed. Data is made available and the analyses depicted below are described in more detail in a tutorial at the project's Github page (`https://github.com/arundurvasula/angsd-wrapper/wiki/Tutorial`). A schematic of the analysis workflows shown here can be seen in Figure 2.

Because *O. sativa* is a predominantly self-pollinated species and *O. glumaepatula* is thought to have a mixed mating system, we began by estimating inbreeding coefficients for each sample (Figure 3A) Consistent with reports of a mixed-mating system, our data reveal little evidence for inbreeding in *O. glumaepatula* , but identify all the domesticated rice as highly inbred. We note that one *O. glumaepatula* individual, sample 276, shows some evidence of inbreeding, although less than domesticated rice. We incorporated our inbreeding estimates into the analysis of overall patterns of genetic diversity, and found notably higher diversity in *O. glumaepatula* than domesticated rice (mean $\theta_{\pi} = 0.00087$), and little difference in diversity between the two wild populations ($\theta_{\pi} = 1.6x10^{-}4$ for the allopatric population and $\theta_{\pi} = 1.4x10^{-}4$ for the sympatric population, see Figure 3B).

We then sought to explicitly test for gene flow, applying ANGSD's implementation of a principle component analysis of these samples as well as a STRUCTURE (Skotte et al., 2013)-like admixture analysis (Figure 3C and D, respectively). These analyses fail to find any strong evidence of gene flow between populations, suggesting that fears of genetic swamping by cultivated rice may be misplaced.

CONCLUSIONS

Our software ANGSD-wrapper provides an intuitive and easy-to-use interface to employ the powerful and flexible suite of population genetic analyses developed in ANGSD (Korneliussen et al., 2014) and permits the exploration of genome-scale results through interactive visualization. ANGSD-wrapper is under active development, and work is already underway on a new release that will 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) and funding from the UC Davis Plant Sciences Department. We thank members of the Ross-Ibarra and Morrell labs for discussion and software testing. We thank Eric Fuchs for access to the test dataset. 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

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

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.

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.

Fuchs, E. J., Martínez, A. M., Calvo, A., Muñoz, M., and Arrieta-Espinoza, G. (2015). Genetic structure of oryza glumaepatula wild rice populations and evidence of introgression from o. sativa in costa rica. *PeerJ PrePrints*.

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.

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.

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.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, J.-Y., Wang, J., and Zeigler, R. S. (2014). The 3,000 rice genomes project: new opportunities and challenges for future rice research. *GigaScience*, 3(1):1–3.

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.

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.

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.

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.

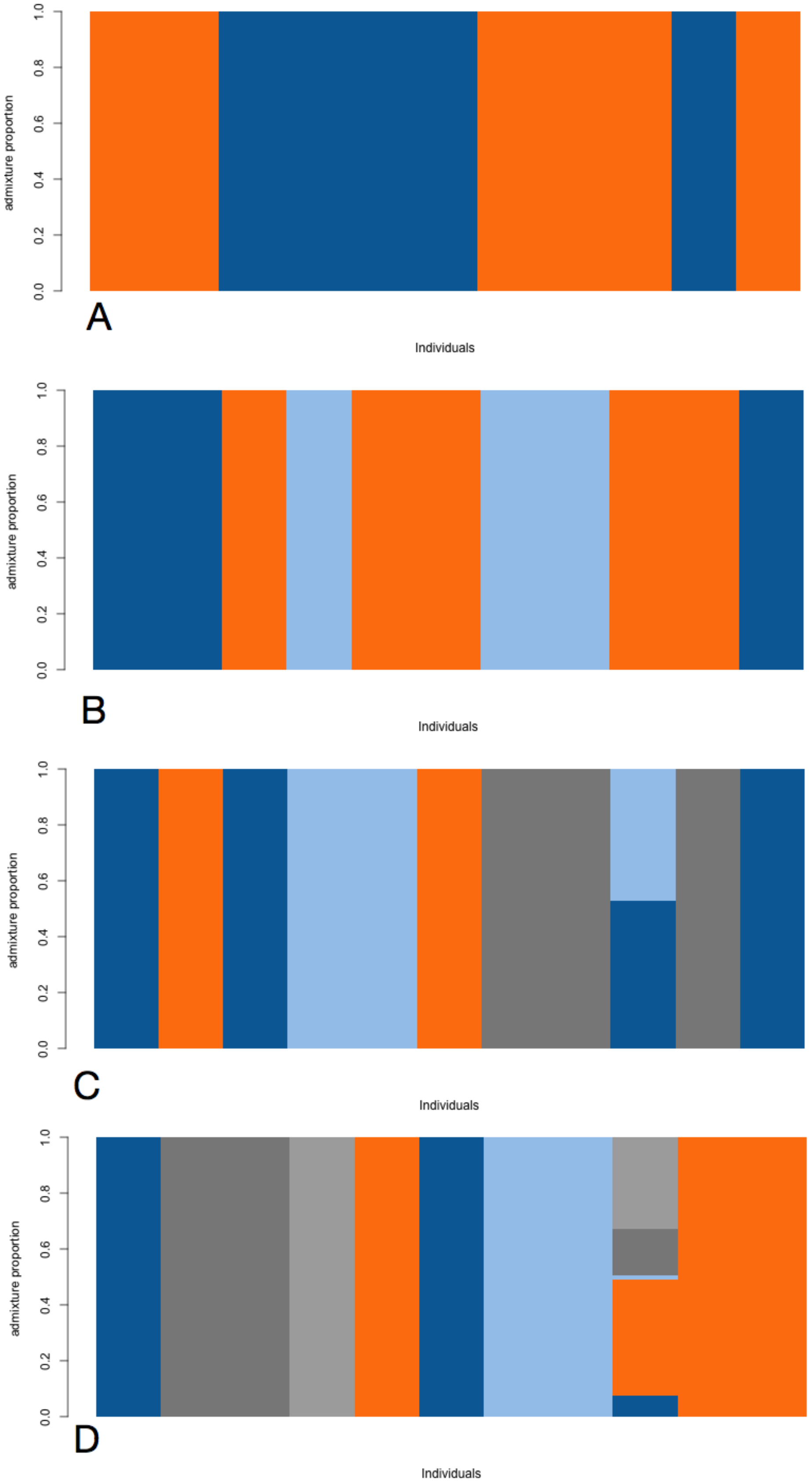


Figure S1. ngsAdmix results for different values of K. A. K=2 B. K=3 C. K=4 D. K=5

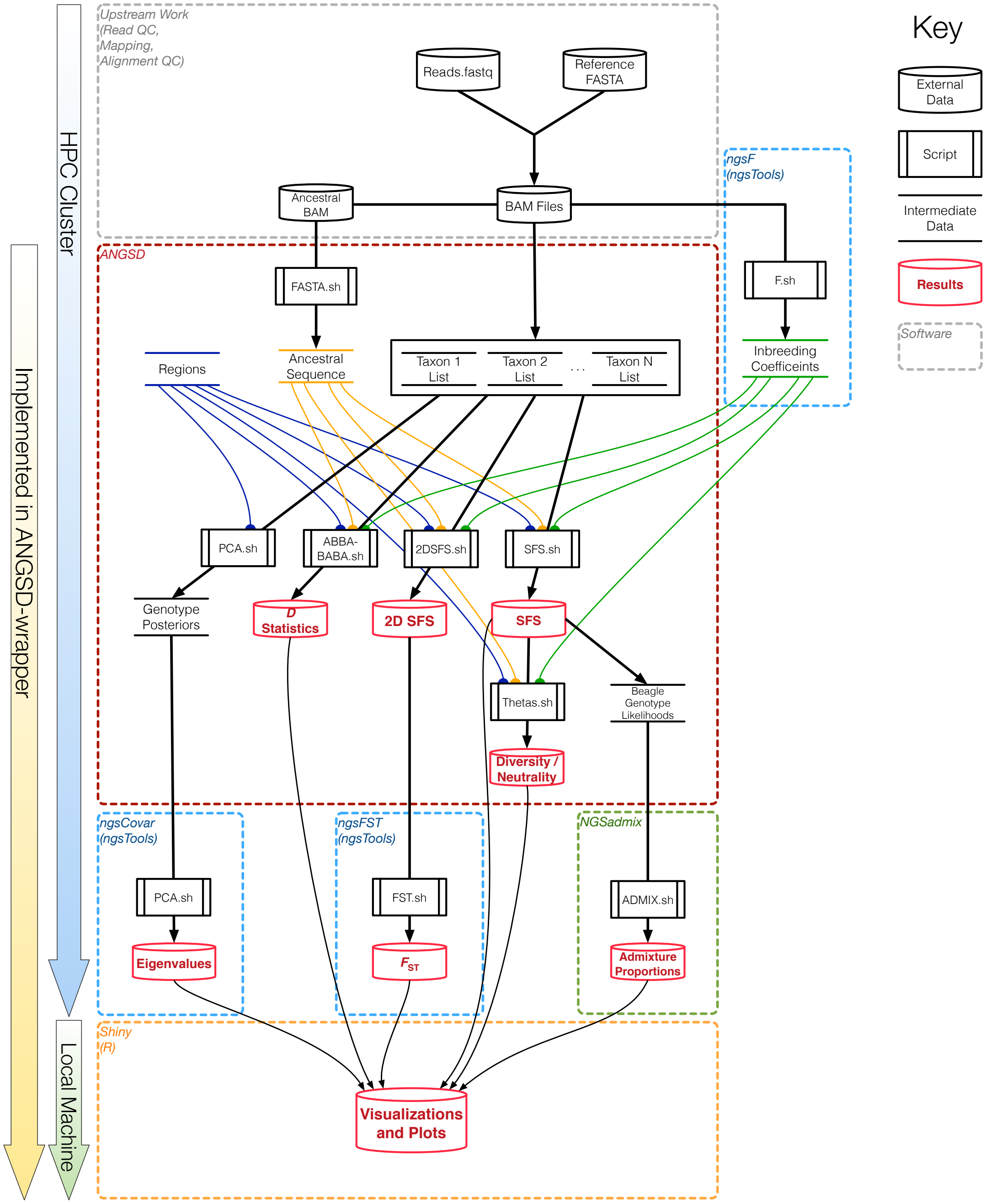


Figure S2. Workflow diagram for all methods available in ANGSD-wrapper.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1472v1 | CC-BY 4.0 Open Access | rec: 3 Nov 2015, publ: 3 Nov 2015