MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 1449–1454 doi: 10.1111/1755-0998.12578

ANGSD-wrapper: utilities for analysing next-generation sequencing data

ARUN DURVASULA,*[1] PAUL J. HOFFMAN,†[1] TYLER V. KENT,* CHAOCHIH LIU,† THOMAS J. Y. KONO,† PETER L. MORRELL† and JEFFREY ROSS-IBARRA*‡

Department of Plant Sciences, University of California, Davis, Davis, CA 95616, USA, †Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108, USA, ‡Center for Population Biology and Genome Center, University of California, Davis, Davis, CA 95616, USA

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. ANGSD uses probabilistic approaches which can directly make use of genotype likelihoods; thus, SNP calling is not required for comparative analyses. This takes advantage of all the sequencing data and produces more accurate results for samples with low sequencing depth. Here, we present ANGSD-wrapper, a set of wrapper scripts that provides a user-friendly interface for running ANGSD and visualizing results. ANGSD-wrapper supports multiple types of analyses including estimates of nucleotide sequence diversity neutrality tests, principal component analysis, estimation of admixture proportions for individual samples and calculation of statistics that quantify recent introgression. ANGSD-wrapper also provides interactive graphing of ANGSD results to enhance data exploration. We demonstrate the usefulness of ANGSD-wrapper by analysing resequencing data from populations of wild and domesticated *Zea*. ANGSD-wrapper is freely available from https://github.com/mojaveazure/angsd-wrapper.

Keywords: domestication, population genetics, software, visualization, *Zea*

Received 1 February 2016; revision received 30 May 2016; accepted 23 June 2016

Introduction

High-throughput sequencing has revolutionized evolutionary genetics, allowing researchers to quickly assay large numbers of individuals and examine genomewide variation. Application of these approaches has led to changes in both experimental design and data analysis (Ekblom & Galindo 2011). Many popular software packages used by researchers for analysis of comparative resequencing data (see Excoffier & Heckel 2006) were not designed to handle these novel data types or efficiently analyse 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 (Pluzhnikov & Donnelly 1996; Felsenstein 2006). While these experimental design challenges in molecular population genetic studies have existed for at least two decades, high-throughput sequencing brings the added 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 (Hutter *et al.* 2006; Purcell *et al.* 2007; Danecek *et al.* 2011; Garrigan 2013). The software package ANGSD (Korneliussen *et al.* 2014) is noteworthy because it enables users to 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 of neutrality test statistics (Korneliussen *et al.* 2013). ANGSD works directly with alignment formats produced from standard high-throughput sequence analysis pipelines, which removes the need for the user to transform the data into a software-specific format. One of the most important features of ANGSD is that analyses can be integrated over per-site genotype likelihoods, rather than on predetermined polymorphic sites a priori. This permits ANGSD to

Correspondence: Jeffrey Ross-Ibarra; E-mail: rossibarra@ucdavis.edu; Peter L. Morrell, Fax: (612) 625-1268, E-mail: pmorrell@umn.edu

[1]These authors contributed equally to this study.

calculate common population genetic descriptive statistics on low-coverage sequencing data and handle missingness due to variation in coverage. Further, this probabilistic framework is appealing to researchers studying nonmodel organisms where the lack of existing genomic resources or cost considerations precludes the ability to obtain high-quality data. Also, many existing tools do not implement statistical models that account for biological issues such as nonrandom mating. Recent use of ANGSD highlights these benefits, allowing population genetic analyses in non-model organisms such as the blue-eyed black lemur (Meyer *et al.* 2015) and *Ficedula* flycatchers (Burri *et al.* 2015) and inbred lines of wild and cultivated maize (Beissinger *et al.* 2016).

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 (Fig. 1) that streamline the execution of multi-step pipelines required for data analysis in ANGSD. ANGSD-wrapper also assists with configuration of ANGSD-related programmes such as ngsTools (Fumagalli *et al.* 2014), ngsF (Vieira *et al.* 2013), ngsAdmix (Skotte *et al.* 2013) and PCA (Fumagalli *et al.* 2013). As in ANGSD, ANGSD-wrapper allows users to perform whole genome analysis or analyse a set of user-defined windows across the genome. The wrapper scripts are written against frozen versions of ANGSD (currently v0.902-48-g8b89ba4) 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 resequencing using 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 2016) is required for some light, dynamic preprocessing of the data, and the statistical software R (R Core Team 2014) is required to make use of the visualization tools incorporated in ANGSD-wrapper. The visualization of ANGSD-wrapper requires the isntallation of 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) GENOMEINTERVALS (Gagneur *et al.* 2015) from Bioconductor (Huber *et al.* 2015) as well as other packages (for a full list see https://github.com/mojaveazure/angsd-wrapper/wiki/Dependencies); these are installed automatically upon first run of the visualization interface.

ANGSD-wrapper is divided into scripts associated with analytical approaches implemented in ANGSD and associated software. It 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-

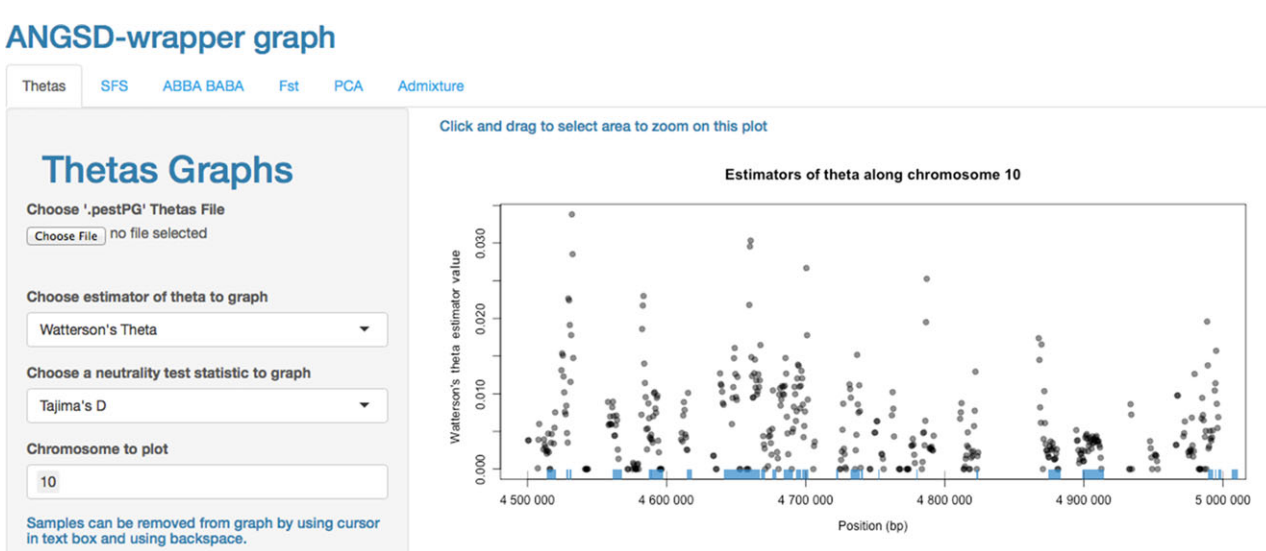

Fig. 1 Visualization of Watterson's θ estimated by ANGSD across a 1.5 Mb region of chromosome 10 in *Zea mays* ssp. *mays* using ANGSD-wrapper. Darker colours indicate a higher density of points. Blue boxes indicate gene annotations provided by a GFF file.

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 flow chart of each of these workflows is shown in Fig S1, and additional documentation, a tutorial and a wiki can be found on the project GitHub page: https://github.com/mojaveazure/angsd-wrapper/wiki.

The visualization software included with ANGSD-wrapper is contained in a 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. To use the plotting software, the user navigates to the desired tab and uploads the appropriate results file. The SHINY server automatically parses standard ANGSD output files and creates the resulting plot(s) (Fig. 1), which can be saved using the browser's built-in image export options.

Results and discussion

As a demonstration of analyses in ANGSD-wrapper, we explore patterns of pairwise nucleotide 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 (Chia *et al.* 2012) and calculated summary statistics using a 10 Mb region on chromosome 10 (Table 1). The data are available at https://figshare.com/articles/Example_Data_tar_bz2/2063442 and can be downloaded from within ANGSD-wrapper using the command

```
$ angsd-wrapper setup data
```

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 by the 'Ancestral Sequence' method from a single resequenced genome of *Tripsacum dactyloides*. We show that the maize SFS is skewed towards intermediate-frequency variants (Fig. 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 (Eyre-Walker *et al.* 1998; Beissinger *et al.* 2016). Using the 'Thetas' method, we find further evidence of the domestication bottleneck, with mean levels of pairwise nucleotide diversity in this region in maize ≈25% lower than in teosinte (0.0061 and 0.0082, respectively; Fig. 2C). Using the 'F_{ST} Estimations', which includes an F_{ST} calculation, we find a mean F_{ST} in this region of 0.116, nearly identical to the genomewide value of 0.11 reported in Hufford *et al.* (2012). None of the genes in this region have been identified as potential domestication candidates (Hufford *et al.* 2012), consistent with the lack of extended regions of high F_{ST} in our analysis (Fig. S3).

Finally, we include two samples of the related wild teosinte *Z. mays* ssp. *mexicana* to assess evidence for admixture. We ran the 'Admixture' method, which implements an estimate of admixture proportions from genotype likelihoods (Skotte *et al.* 2013). We identify structure within domesticated maize separating three high-latitude temperate landraces from the other tropical accessions (Fig. 3). *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). Consistent with an independent analysis from SNP genotyping (Hufford *et al.* 2013), the lowland maize samples included here show no evidence of admixture with ssp. *mexicana*. Most ssp. *parviglumis* accessions fall primarily into a single (light blue) cluster, but two accessions show assignment to multiple clusters, perhaps due to the limited resolution resulting

Table 1 Methods implemented in ANGSD-wrapper

Method	Calculations	Interactive graphing
SFS	Site frequency spectrum	Yes
F_{ST} Estimations	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

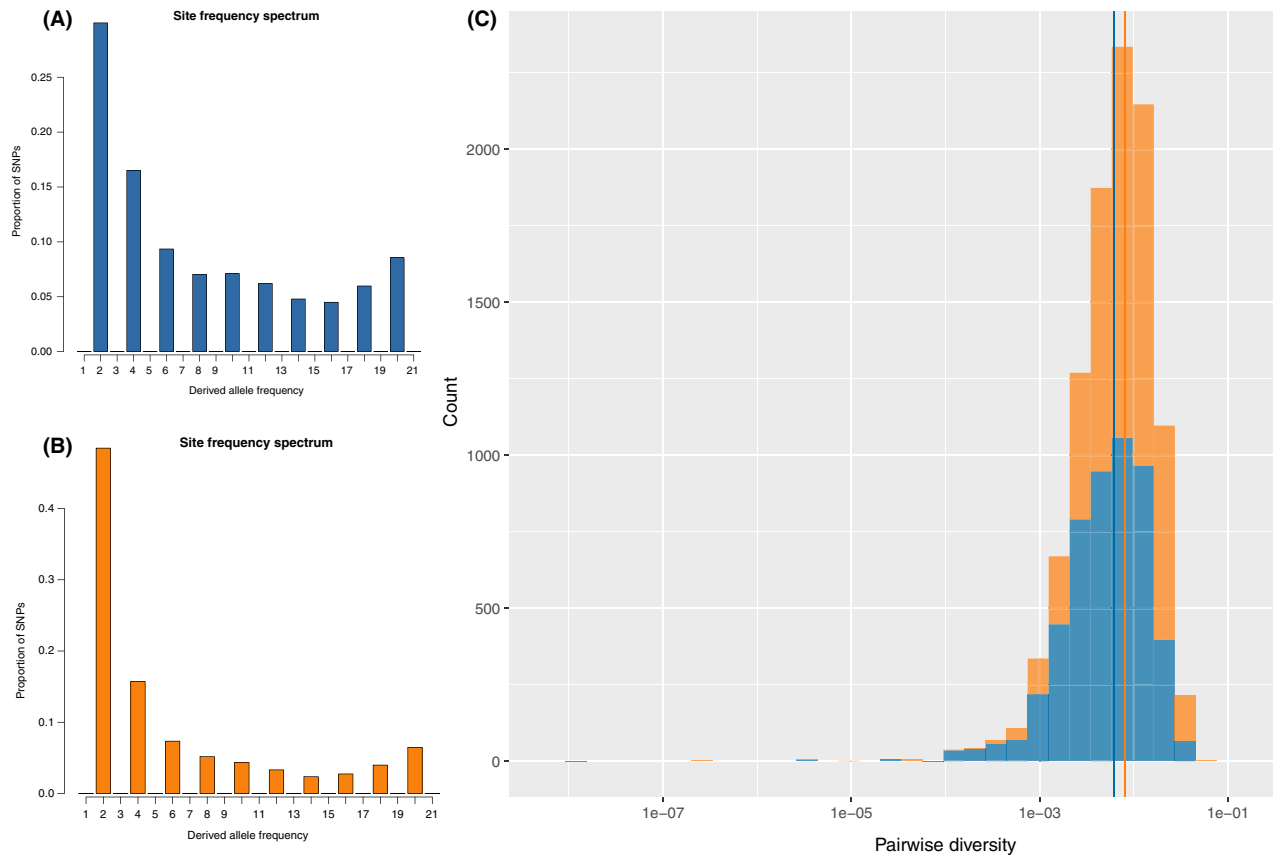

Fig. 2 Summary statistics for *Zea mays*. Derived site frequency spectra for (A) maize and (B) teosinte. (C) Distribution of pairwise nucleotide diversity for maize (blue) and teosinte (orange). Mean values for each taxon are represented by corresponding vertical lines. Pairwise nucleotide diversity results are visualized separately from the interactive graphics, and colours were added to A and B using a custom script.

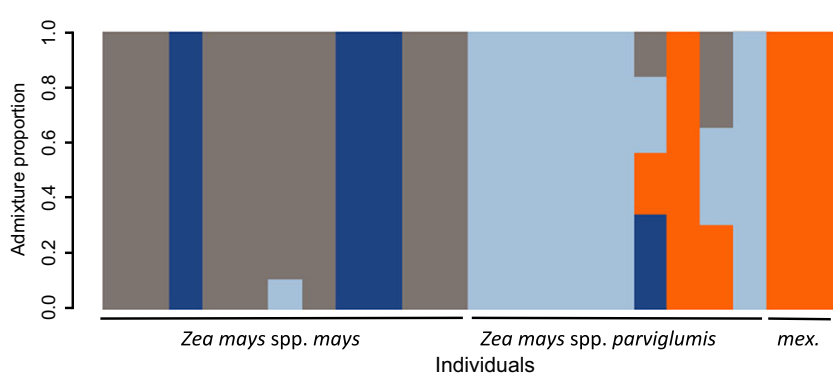

Fig. 3 Admixture analysis for *Zea mays* ssp. *mays*, *Z. mays* ssp. *parviglumis*, and *Z. mays* ssp. *mexicana* (*mex*). Colours represent the proportion of each individual's genome assigned to one of the $K = 4$ source populations. Individuals are shown in the same order as Table S1. For results using other values of K see Fig. S2.

from analysis of a single genomic region and restricted geographic 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 new analyses and 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, from USDA Hatch project CA-D-PLS-2066-H and from the University of California Davis Plant Sciences Department. Support was also provided by the Minnesota Agricultural Experiment Station Variety Development fund. We thank members of the Ross-Ibarra and Morrell Labs for discussion and software testing. We thank the authors of ANGSD and related programmes for answering questions, particularly Matteo Fumagalli and Filipe Vieira. Matteo Fumagalli and an anonymous reviewer provided comments that improved both the manuscript and the implementation of ANGSD-wrapper. Finally, we would like to thank Felix Andrews for statistical advice, although we did not follow it.

References

Beissinger TM, Wang L, Crosby K, Durvasula A, Hufford MB, Ross-Ibarra J (2016) Recent demography drives changes in linked selection across the maize genome. *Nature Plants*, **2**, 16084.

Burri R, Nater A, Kawakami T *et al*. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of ficedula flycatchers. *Genome Research*, **25**, 1656–1665.

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

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

Chia J-M, Song C, Bradbury PJ *et al*. (2012) Maize hapmap2 identifies extant variation from a genome in flux. *Nature Genetics*, **44**, 803–807.

Danecek P, Auton A, Abecasis G *et al*. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

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

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

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

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

Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS (1998) Investigation of the bottleneck leading to the domestication of maize. *Proceedings of the National Academy of Sciences of the USA*, **95**, 4441–4446.

Fang Z, Pyhäjärvi T, Weber AL *et al*. (2012) Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics*, **191**, 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**, 691–700.

Fumagalli M, Vieira FG, Korneliussen TS *et al*. (2013) Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, **195**, 979–992.

Fumagalli M, Vieira FG, Linderoth T, Nielsen R (2014) ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, **30**, 1486–1487.

Gagneur J, Toedling J, Bourgon R, 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*, pp. 35–36. IEEE Computer Society, Washington, District of Columbia.

Huber W, Carey JV, Gentleman R *et al*. (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, **12**, 115–121.

Hufford MB, Xu X, Van Heerwaarden J *et al*. (2012) Comparative population genomics of maize domestication and improvement. *Nature Genetics*, **44**, 808–811.

Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J (2013) The genomic signature of crop-wild introgression in maize. *PLoS Genetics*, **9**, 1–13.

Hutter S, Vilella AJ, Rozas J (2006) Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics*, **7**, 409.

Jette MA, Yoo AB, 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* (eds Feitelson D, Rudolph L, Schwiegelshohn U), pp. 44–60. Springer-Verlag, Berlin.

Jr FEH, with contributions from Charles Dupont et al. (2015) *Hmisc: Harrell Miscel-laneous*. R package version 3.17-1.

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

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*, **15**, 356.

Meyer WK, Venkat A, Kermany AR, de Geijn B, Zhang S, Przeworski M (2015) Evolutionary history inferred from the de novo assembly of a nonmodel organism, the blue-eyed black lemur. *Molecular Ecology*, **24**, 4392–4405.

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

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

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

Purcell S, Neale B, Todd-Brown K *et al*. (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

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

van Rossum G (2016) *Python*. Available from https://www.python.org (accessed 27 January 2016).

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

Skotte L, Korneliussen TS, Albrechtsen A (2013) Estimating individual admixture proportions from next generation sequencing data. *Genetics*, **195**, 693–702.

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

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

Xie Y (2015) *DT: a wrapper of the JavaScript library 'DataTables'*. R package version 0.1.

J.R.I., and A.D. conceived the study. A.D., P.J.H., C.L., T.V.K., and T.J.Y.K. wrote code. A.D. analysed the data. A.D., J.R.I., and P.L.M. wrote the initial draft of the manuscript. All authors contributed to final manuscript preparation.

Data accessibility

Data is available at: https://figshare.com/articles/Example_Data_tar_bz2/2063442 and within ANGSD-wrapper with the command `angsd-wrapper setup data`.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 Workow diagram for all methods available in ANGSD-wrapper.

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

Fig. S3 F_{ST} values plotted against base pair position on chromosome 10 of maize.

Table S1 Table of samples used in analysis with mean depth over the region 15 000 000–25 000 000 on chromosome 10.