ANGSD-wrapper: scripts to streamline and visualize NGS population genetics analysis

Arun Durvasula[1], Tyler Kent[1], Siddharth Bhadra-Lobo[1], Jeffrey Ross-Ibarra[2]

[1]Dept. of Plant Sciences, University of California Davis

[2]Dept. of Plant Sciences, Center for Population Biology, and Genome Center, University of California Davis

Introduction

The advent of highly multiplexed sequencing has opened a number of exciting avenues for evolutionary biologists. One of the powerful approaches enabled by inexpensive sequencing is the ability to sequence a large number of individuals, each to relatively low sequencing depth. However, this approach also presents statistical challenges in the analysis of low coverage data. The software ANGSD [3] and related programs [2] were developed to deal with low coverage sequence data. Rather than call genotypes at variable sites, ANGSD performs a number of population genetic analyses on genotype likelihoods, including estimation of the population mutation rate θ, the site frequency spectrum, neutrality tests, inbreeding coefficients, and population structure. ANGSD has already been used in several studies to analyze genome sequence data [1] [4]. However, ANGSD requires considerable familiarity with command line tools and remains inaccessible to many biologists that are not from a computational background. Here we present a software package that aids in the preparation of analyses for ANGSD and provides interactive graphing software implemented in R [5] and Shiny [7]. ANGSD-wrapper simplifies multistep analyses such as calculating Tajima's D into a single step. Users supply all the needed information in a single configuration file 1, and after ANGSD has finished calculations, ANGSD-wrapper provides interactive graphing of the results 2.

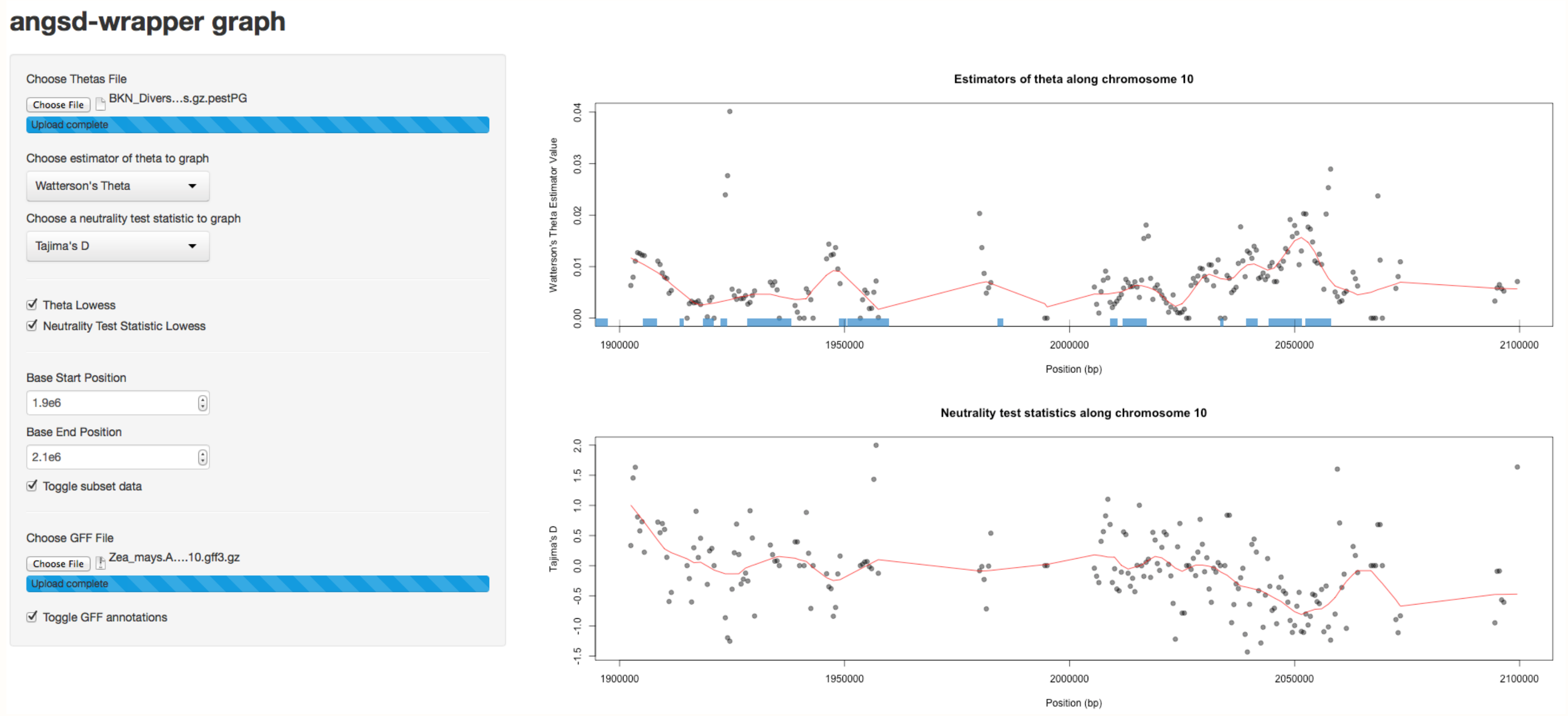
Implementation

ANGSD-wrapper is implemented using bash scripts that call ANGSD methods and handle saving intermediate files between the initial data preparation and the final data analysis. Each overall method in ANGSD, such as calculating estimates of θ, follows a specific order of program calls. Thus, we have abstracted away the running of each step and provided a set of default values for parameters and instead require the user to supply the data using a configuration file (Figure 1). The user can override the default values of the parameters in the configuration file as well.

Additionally, ANGSD-wrapper contains a powerful graphing application based on R and Shiny (Figure 2). After analysis is done by ANGSD, the user can load the resulting statistics into a web-based application hosted on the user's computer. This application allows the interactive plotting of values such as estimators of θ and Tajima's D as well as the ability to load gene annotations from Ensembl. These features make it easy and intuitive to analyze next generation sequencing data.

```
TAXON=BKN
TAXON_LIST=${DATA_DIR}/${TAXON}_samples.txt
TAXON_INBREEDING=${DATA_DIR}/${TAXON}_F.txt
PEST=${RESULTS_DIR}/${TAXON}_DerivedSFS
#   If you would like to change the default parameters of the analysis,
#   you may declare the appropriate variables below.
#   See ANGSD_Thetas.sh for all possible parameters and their defaults
SLIDING_WINDOW=true
WIN=100
STEP=50
```

Figure 1: Example configuration file



angsd-wrapper graph

Choose Thetas File
Choose File BKN_Divers...s.gz.pestPG
Upload complete

Choose estimator of theta to graph
Watterson's Theta

Choose a neutrality test statistic to graph
Tajima's D

☑ Theta Lowess
☑ Neutrality Test Statistic Lowess

Base Start Position
1.9e6

Base End Position
2.1e6

☑ Toggle subset data

Choose GFF File
Choose File Zea_mays.A....10.gff3.gz
Upload complete

☑ Toggle GFF annotations

Figure 2: Interactive plotting with Shiny showing a 200kb region along chromosome 10 of *Zea mays*. Blue rectangles are annotated genes.

Example Application

We sequenced 8 genomes of the wild rice *Oryza glumaepatula*, 4 each from populations allopatric and sympatric to cultivated fields of the domesticated *Oryza sativa* ssp. *indica*, in order to investigate evidence of crop-wild introgression. *O. glumaepatula* is a potentially endangered species 1.8 MY diverged from domesticated rice [9] and native to Central and South America [8]. Both species share the AA genome [8] and we have successfully crossed them experimentally, providing reason to believe the possibility of natural hybridization and thus the risk of wild population decline and extinction [6]. Preliminary analysis using ANGSD-wrapper is suggestive of introgression (Figure 3) as evidenced by the higher correlation of nucleotide diversity (π) between domesticated rice and the sympatric population (Spearman's $\rho = 0.537$) than between domesticated rice and allopatric O. glumaepatula (0.139).

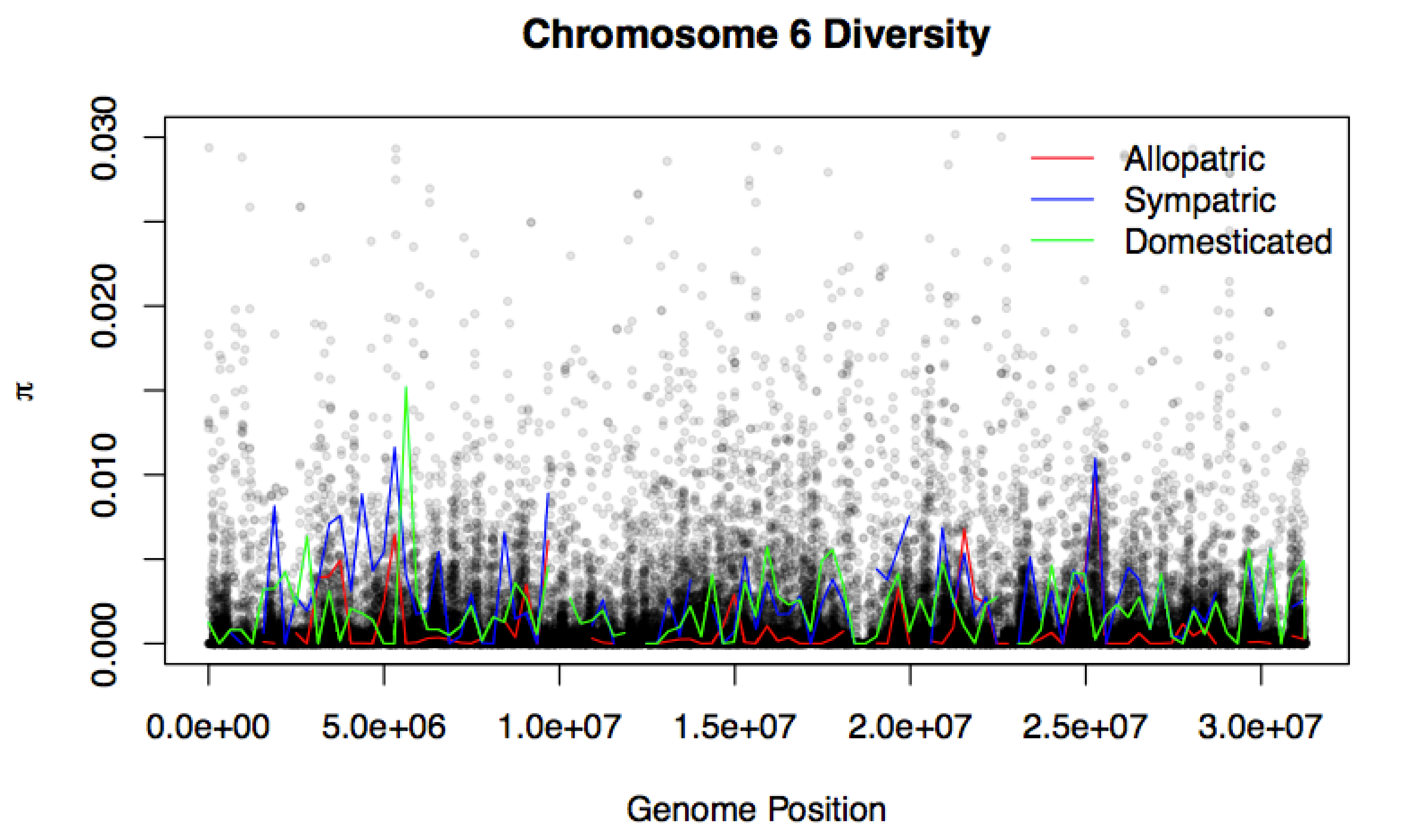


Figure 3: Nucleotide diversity (π) along chromosome 6 of *Oryza glumaepatula* and *Oryza sativa*

References

[1] Crawford, J., M. M. Riehle, W. M. Guelbeogo, A. Gneme, N. f. Sagnon, K. D. Vernick, R. Nielsen and B. P. Lazzaro (2014). Reticulate speciation and adaptive introgression in the Anopheles gambiae species complex. bioRxiv.

[2] Fumagalli, M., F. G. Vieira, T. Linderoth and R. Nielsen (2014). ngsTools: methods for population genetics analyses from next-generation sequencing data. Bioinformatics 30(10): 1486-1487.

[3] Korneliussen et al. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics. 15:356

[4] Lohmueller, et al. (2013). Whole-Exome Sequencing of 2,000 Danish Individuals and the Role of Rare Coding Variants in Type 2 Diabetes. American Journal of Human Genetics, 93(6), 1072–1086.

[5] R Core Team (2014). R Foundation for Statistical Computing, Vienna, Austria http://www.R-project.org/

[6] Rhymer JM and Simberloff D (1996) Extinction by Hybridization and Introgression. Annu Rev Ecol Syst 27:83-109.

[7] Rstudio, Inc. (2013). Easy web applications in R. http://www.rstudio.com/shiny/

[8] Vaughan DA, Morishima H, and Kadowaki K (2003) Diversity in the Oryza genus. Current Opinion in Plant Biology 6:139-146.

[9] Zhang, Q.-J., et al. (2014). Rapid diversification of five Oryza AA genomes associated with rice adaptation. PNAS 111(46): E4954-E4962.

BAPG XI, 6 December 2014, Davis, CA, USA