



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



Arun Durvasula¹, Tyler Kent¹, Siddharth Bhadra-Lobo¹, Jeffrey Ross-Ibarra²

¹Dept. of Plant Sciences, University of California Davis

²Dept. 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 ?? and related programs ?? 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 ?. 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 ?? and Shiny ?. 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, and after ANGSD has finished calculations, ANGSD-wrapper provides interactive graphing of the results ?.

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 ?). After the 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
PEST=results/${TAXON}_DerivedSFS
SLIDING_WINDOW=true
WIN=100
STEP=50
```

Figure: Example configuration file

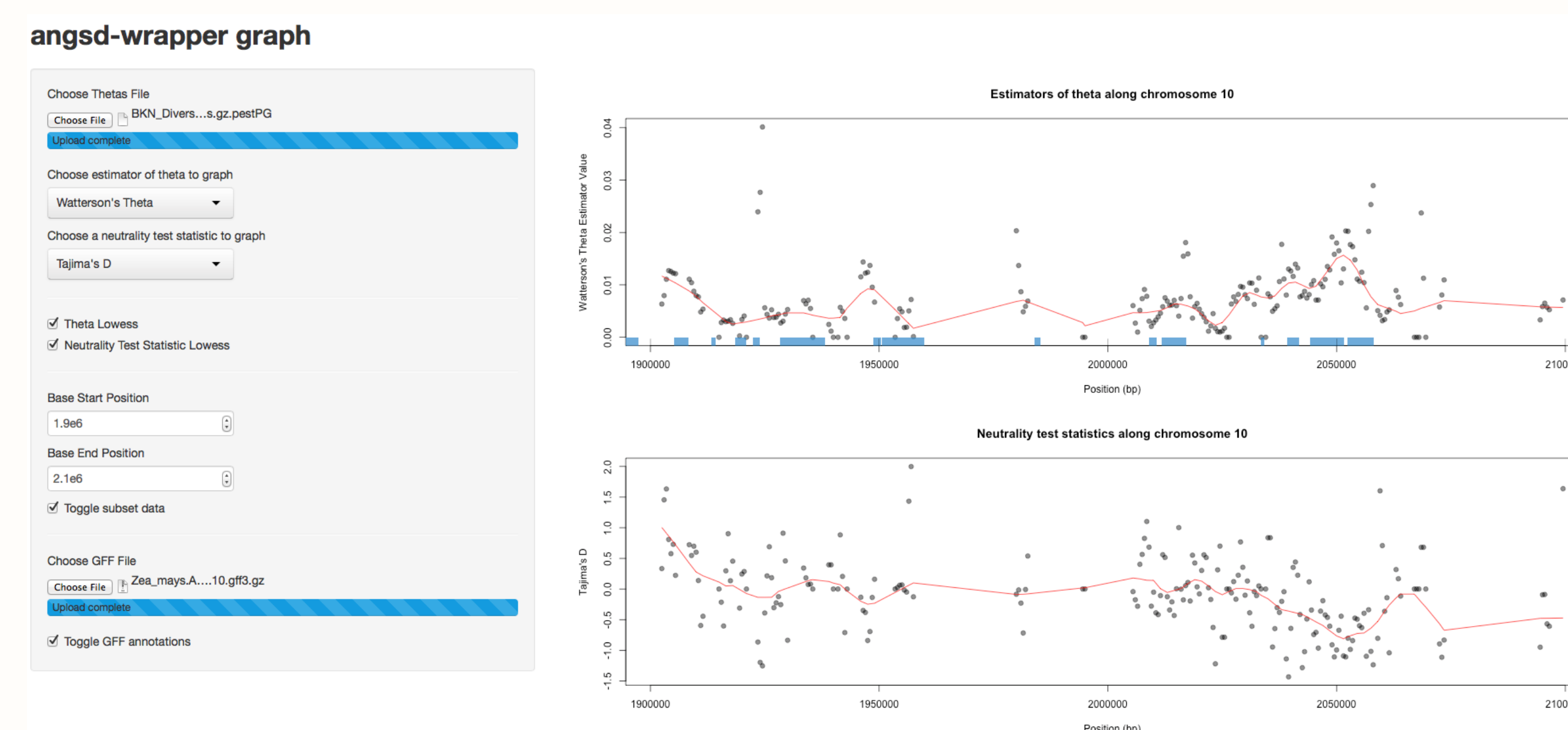


Figure: Interactive plotting with Shiny

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 [4] native to Central and South America (Vaughan et al. 2003). Both species share a AA genome (Vaughan et al. 2003) 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 (Rhymer Simberloff 1996). Preliminary analysis using ANGSD-wrapper is suggestive of introgression as evidenced by higher correlation of nucleotide diversity (π) between domesticated rice and the sympatric population than between domesticated rice and allopatric *O. glumaepatula* (Fig. 2).

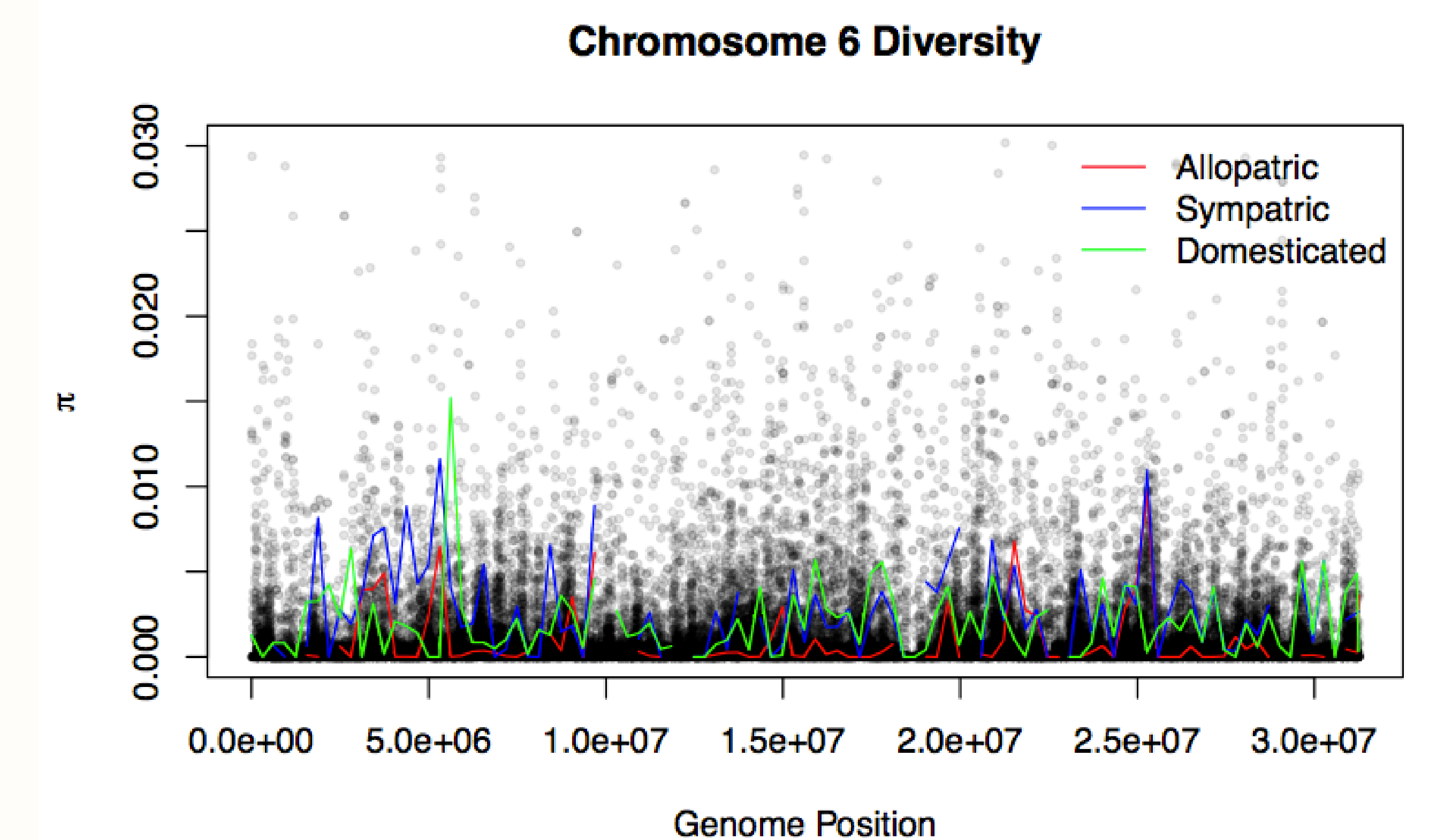


Figure: Figure 2. Nucleotide diversity (π) along chromosome 6. Correlation of π between domesticated rice and the sympatric population is 0.417, while between domesticated rice and the allopatric population it is 0.077.

References

- [1] Akimoto M, Shimamoto Y and Morishima H (1998). Mol Ecol 7:1371-1381.
 - [2] Korneliussen et al. (2014). BMC Bioinformatics. 15:356
 - [3] R Core Team (2014). R Foundation for Statistical Computing, Vienna, Austria <http://www.R-project.org/>
 - [4] Zhang Q-J, et al. (2014). PNAS 111:E4954-E4962.
- Rhymer JM and Simberloff D (1996) Extinction by Hybridization and Introgression. Annu Rev Ecol Syst 27:83-109.
- Vaughan DA, Morishima H, and Kadowaki K (2003) Diversity in the Oryza genus. Current Opinion in Plant Biology 6:139-146.