MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2013) doi: 10.1111/1755-0998.12160

HeFPipe: a complete analytical pipeline for heterozygosity-fitness correlation studies

MARK A. FISHER

Rm, 117 Riverbend Research North, University of Georgia, 110 Riverbend Road, Athens, GA 30602, USA

Abstract

As the body of heterozygosity-fitness correlation (HFC) research grows, more and increasingly complicated tests have become an integral part of a typical HFC analysis (Chapman *et al.* 2009). Currently, no software is available to undertake conversion between the file formats required to conduct all of these tests and to conduct the main regression analyses at the core of all HFCs. Heterozygosity-Fitness Pipeline (*HeFPipe*) is a script written in *Python* that accomplishes both of these tasks for studies based on microsatellite data. *HeFPipe* is designed to be used from the command line terminal and will run on any Mac OSX computer. The script takes input in the form of allele reports from either the genotype-calling software, *GeneMapper* or *GeneMarker*, and reconfigures the data into *GENEPOP* (Raymond & Rousset 1995), *Rhh* (Alho *et al.* 2010), *RMES* (David *et al.* 2007) and *GEPHAST* (Amos & Acevedo-Whitehouse 2009) formats. The script is also equipped to reformat the output from *GENEPOP* on the Web (option 5) and *Rhh* into csv spreadsheets that can be incorporated into downstream analyses. *HeFPipe* accommodates user-provided lists of samples and markers to be included in or excluded from analyses. *HeFPipe* is equipped to create generalized linear models (GLMs) from both the main data set and subsets of the data. Finally, *HeFPipe* allows users to explore single-marker effects and conduct correlation analyses. The script, a comprehensive manual, a link to a series of video tutorials, and an example data set are available from GitHub (http://github.com/Atticus29/*HeFPipe*_repos).

Keywords: heterozygosity-fitness correlation, pipeline, population genetics, *Python*

Received 29 May 2013; revision received 6 August 2013; accepted 8 August 2013

Identifying associations between heterozygosity and fitness is a lynchpin of studies intending to clarify the role that genetic diversity plays in the survival and reproductive success of individuals. HFCs have a long history in the population genetics literature (Mitton & Grant 1984; Pogson 1991; Britten 1996; Coltman & Slate 2003; Chapman *et al.* 2009), and they were originally conducted by genotyping samples at a modest panel of allozyme loci and regressing a trait(s) associated with fitness (e.g. survival, growth rate) on multilocus heterozygosity (MLH) as measured by the panel (Britten 1996). Modern HFC studies almost exclusively employ microsatellite markers rather than allozymes, and an emphasis has been placed on the use of large numbers of markers (Balloux *et al.* 2004), although few studies have yet to fulfil this recommendation (Chapman *et al.* 2009). The shift to larger marker panels containing potentially neutral loci has brought with it a wider availability of statistical tests that help researchers explore the nature of heterozygosity-fitness correlations in their study systems. These tests make it possible to determine (i) whether the MLH of the marker panel is reflective of genome-wide MLH (Balloux *et al.* 2004; Alho *et al.* 2010), (ii) whether there is identity disequilibrium (ID) among the markers and consequently inbreeding *sensu lato* in the study system (David *et al.* 2007; Szulkin *et al.* 2010) and (iii) whether there is evidence for single-marker effects on the trait(s) of interest (David 1997; Amos & Acevedo-Whitehouse 2009; Szulkin *et al.* 2010) (Fig. 1). The software now available to conduct these tests as well as run the regressions and correlations that are the core of HFC analyses require input files of different formats, and there is currently no software that provides ecumenicism across these formats.

Heterozygosity-Fitness Pipeline (*HeFPipe*) is a script written in *Python* that conducts analyses typically performed in HFC studies. It also tests for evidence of single-marker effects on a trait(s). More specifically, *HeFPipe* takes input in the form of allele reports in the 'Marker Table' style from the microsatellite genotype-calling software, *GeneMarker*, or from the microsatellite genotype-calling software, *GeneMapper*, and reconfigures the data into *GENEPOP* (Raymond & Rousset 1995), *Rhh* (Alho *et al.* 2010), *RMES* (David *et al.* 2007)

Correspondence: Mark A. Fisher, Fax: 706-542-2279;
E-mail: mark.aaron.fisher@gmail.com

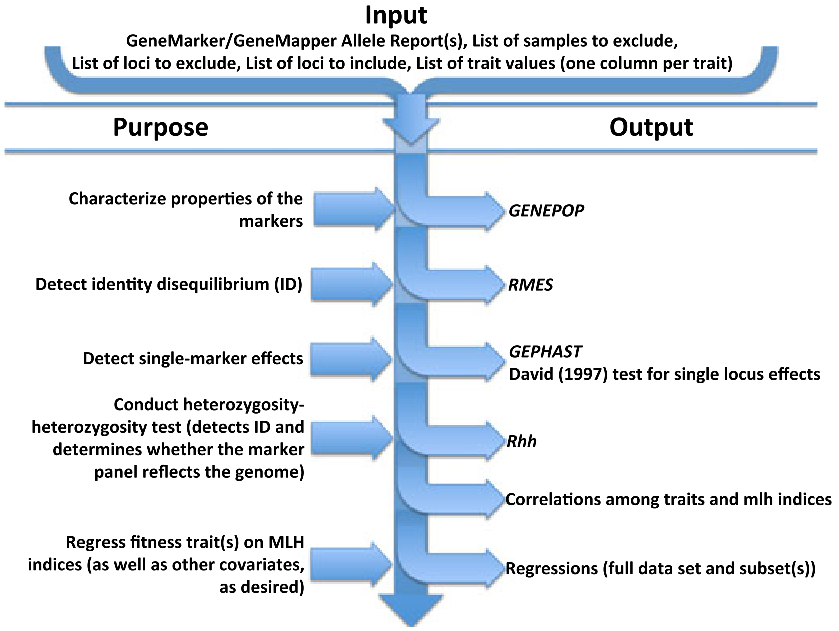
Input
GeneMarker/GeneMapper Allele Report(s), List of samples to exclude,
List of loci to exclude, List of loci to include, List of trait values (one column per trait)
Purpose
Output
Characterize properties of the markers
GENEPOP
Detect identity disequilibrium (ID)
RMES
Detect single-marker effects
GEPHAST
David (1997) test for single locus effects
Conduct heterozygosity-heterozygosity test (detects ID and determines whether the marker panel reflects the genome)
Rhh
Correlations among traits and mlh indices
Regress fitness trait(s) on MLH indices (as well as other covariates, as desired)
Regressions (full data set and subset(s))

Fig. 1 Simplified flowchart depicting the Heterozygosity-Fitness Pipeline (*HeFPipe*). The input files listed in the top section are used at various relevant points throughout the pipeline. The chronological flow of the pipeline is depicted by the direction of the arrow in the middle of the figure, and the output of the pipeline is depicted on the right, while a brief description of the relevance of each output item to an heterozygosity-fitness correlation (HFC) analysis is described on the left. The items listed under 'Output' are files generated by *HeFPipe* that are either useable by external programs (*GENEPOP*, *RMES*, *GEPHAST*, *Rhh*) or are themselves core results of HFC analyses (correlations, regressions). Other, less-essential output files that are also products of *HeFPipe* are not described in this figure but are discussed in the *HeFPipe* manual and tutorial videos along with instructions for how to use the output from the external programs listed in this figure as input in subsequent steps of the *HeFPipe* pipeline.

and *GEPHAST* (Amos & Acevedo-Whitehouse 2009) formats (Fig. 1). The script is also equipped to reformat the *output* from *GENEPOP* on the Web (option 5) and *Rhh* into comma-separated values (csv) formatted spreadsheets and incorporate them into downstream analyses. The *HeFPipe* script accommodates user-provided lists of markers to be included in or excluded from analyses, a list of samples to exclude from analyses, and a spreadsheet containing trait values on which to perform the HFCs and search for single-marker effects. These input files allow the user to refine and repeat each analysis with ease. With regard to the analyses that require regression—HFCs and one of the tests for single-marker effects—*HeFPipe* is equipped to run generalized linear models (GLMs) using the *Python* package *PypeR* (Xia *et al.* 2010), a package that enables the statistics software R (R Development Core Team 2011) to be used in the context of a *Python* script. Using GLMs, the user is able to assign a link function and error distribution appropriate for the response variable in a particular model, which can be used to relax some of the assumptions of general linear regression. The script is also equipped to conduct the regression analyses on subsets of the data set, which might be desirable in various scenarios, such as where HFCs are predicted to appear in stressed individuals (e.g. Pujolar *et al.* 2006; Schmeller *et al.* 2007). Single-marker effects are explored using methodologies described in David (1997) and using *GEPHAST* (Amos & Acevedo-Whitehouse 2009). Correlations (both Pearson and Spearman) among the traits provided are reported in several different formats (as text, spreadsheets, and images); significance tests are conducted on these correlations, and the *P*-values (both adjusted and unadjusted for multiple comparisons) are also reported in the various formats.

Heterozygosity-Fitness Pipeline is designed to be used from the command line terminal and will run on any Mac OSX computer that has *Python* v 2.7.3 and *R* v 2.15.1 (or compatible versions) installed. Several features of the pipeline depend on properties of UNIX-based operating systems, and these properties are not native to Windows-based operating systems. The script, a users' manual, a link to a video tutorial and an example data set are available at GitHub (https://github.com/Atticus29/*HeFPipe*_repos). Software dependencies, including packages in both *R* and *Python* required for the pipeline (e.g. *PypeR*), are listed in the manual, as are brief instructions for their installation.

Related manuscripts

An empirical manuscript for which the pipeline was developed is in preparation.

Acknowledgements

I thank Xiao-Qin Xia for help with *PypeR* and Emily Bewick, Kerin Bentley, and Brian Shamblin for providing additional data sets on which *HeFPipe* was tested. I also thank Kenneth Ross and DeWayne Shoemaker, who reviewed an early version of the manuscript, as well as three anonymous reviewers for constructive comments. This work was funded in part by the Georgia Agricultural Experiment Stations, University of Georgia.

References

Alho JS, Välimäki K, Merilä J (2010) Rhh: an R extension for estimating multilocus heterozygosity and heterozygosity-heterozygosity correlation. *Molecular Ecology Resources*, **10**, 720–722.

Amos W, Acevedo-Whitehouse K (2009) A new test for genotype-fitness associations reveals a single microsatellite allele that strongly predicts the nature of tuberculosis infections in wild boar. *Molecular Ecology Resources*, **9**, 1102–1111.

Balloux F, Amos W, Coulson T (2004) Does heterozygosity estimate inbreeding in real populations? *Molecular Ecology*, **13**, 3021–3031.

Britten HB (1996) Meta-analyses of the association between multilocus heterozygosity and fitness. *Evolution*, **50**, 2158–2164.

Chapman JR, Nakagawa S, Coltman DW, Slate J, Sheldon BC (2009) A quantitative review of heterozygosity-fitness correlations in animal populations. *Molecular Ecology*, **18**, 2746–2765.

Coltman DW, Slate J (2003) Microsatellite measures of inbreeding: a meta-analysis. *Evolution; International Journal of Organic Evolution*, **57**, 971–983.

David P (1997) Modeling the genetic basis of heterosis: tests of alternative hypotheses. *Evolution*, **51**, 1049–1057.

David P, Pujol B, Viard F, Castella V, Goudet J (2007) Reliable selfing rate estimates from imperfect population genetic data. *Molecular Ecology*, **16**, 2474–2487.

Mitton JB, Grant MC (1984) Associations among protein heterozygosity, growth rate, and developmental homeostasis. *Annual Review of Ecology and Systematics*, **15**, 479–499.

Pogson GH (1991) Expression of overdominance for specific activity at the Phosphoglucomutase-2 locus in the pacific oyster, Crassostrea gigas. *Genetics*, **128**, 133–141.

Pujolar JM, Maes GE, Vancoillie C, Volckaert FAM (2006) Environmental stress and life-stage dependence on the detection of heterozygosity-fitness correlations in the European eel, Anguilla anguilla. *Genome*, **49**, 1428–1437.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raymond M, Rousset F (1995) GENEPOP (Version 1.2): population genetics software for exact tests and ecumenicism. *Heredity*, **86**, 248–249.

Schmeller DS, Schregel J, Veith M (2007) The importance of heterozygosity in a frog's life. *Naturwissenschaften*, **94**, 360–366.

Szulkin M, Bierne N, David P (2010) Heterozygosity-fitness correlations: a time for reappraisal. *Evolution*, **64**, 1–16.

Xia X-Q, McClelland M, Wang Y (2010) PypeR, A Python package for using R in Python. *Journal of Statistical Software*, **35**, 1–8.

Fisher wrote the software and manuscript, curated and tested the *GeneMarker* and *GeneMapper* example data sets and recorded the tutorial video series.

Data Accessibility

HeFPipe scripts, a users' manual and example data are deposited in GitHub: http://github.com/Atticus29/HeFPipe_repos.

A series of tutorial videos are available on YouTube: http://www.youtube.com/watch?v=cKhKmeqjG6I&feature=share&list=PLv-e9CNPZr-o34dIwUKi-Eew-t6A643tX.