A peer-reviewed version of this preprint was published in PeerJ on 5 June 2017.

View the peer-reviewed version (peerj.com/articles/cs-118), which is the preferred citable publication unless you specifically need to cite this preprint.

Webb AE, Walsh TA, O'Connell MJ. 2017. VESPA: Very large-scale Evolutionary and Selective Pressure Analyses. PeerJ Computer Science 3:e118 https://doi.org/10.7717/peerj-cs.118

VESPA: Very large-scale Evolutionary and Selective Pressure Analyses

Andrew E. Webb, Thomas A. Walsh, Mary J O'Connell

Large-scale molecular evolutionary analyses of protein coding sequences requires a number of preparatory inter-related steps from finding gene families, to generating alignments and phylogenetic trees and assessing selective pressure variation. Each phase of these analyses can represent significant challenges particularly when working with the entire genome of large sets of species. We present VESPA, software capable of automating a selective pressure analysis using codeML in addition to the preparatory analyses and summary statistics. VESPA is written in python and is designed to run within a UNIX environment. Large-scale gene family identification, sequence alignment, and phylogeny reconstruction are all important aspects of large-scale molecular evolutionary analyses. VESPA provides flexible software for simplifying these processes along with downstream selective pressure variation analyses. The software automatically interprets results from codeML and produces simplified summary files to assist the user in better understanding the results. VESPA may be found at the following website: www.mol-evol.org/VESPA

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v1 | CC-BY 4.0 Open Access | rec: 25 Mar 2016, publ: 25 Mar 2016

VESPA: Very large-scale Evolutionary and Selective Pressure Analyses.

Andrew E. Webb[1], Thomas A. Walsh[1] and Mary J. O'Connell[1,2]*

[1]Bioinformatics and Molecular Evolution Group, School of Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland

[2]Computational and Molecular Evolutionary Biology Group, School of Biology, Faculty of Biological Sciences, The University of Leeds, Leeds LS2 9JT, UK

*Corresponding author: m.oconnell@leeds.ac.uk

DR MARY J. O'CONNELL, PHD,

COMPUTATIONAL & MOLECULAR EVOLUTIONARY BIOLOGY GROUP,

SCHOOL OF BIOLOGY,

FACULTY OF BIOLOGICAL SCIENCES,

THE UNIVERSITY OF LEEDS,

LEEDS, LS2 9JT.

UNITED KINGDOM

EMAIL: m.oconnell@leeds.ac.uk

PHONE: +44 (0) 113 34 34890

Abstract:

Background: Large-scale molecular evolutionary analyses of protein coding sequences requires a number of preparatory inter-related steps from finding gene families, to generating alignments and phylogenetic trees and assessing selective pressure variation. Each phase of these analyses can represent significant challenges particularly when working with the entire genome of large sets of species.

Results: We present VESPA, software capable of automating a selective pressure analysis using codeML in addition to the preparatory analyses and summary statistics. VESPA is written in python and is designed to run within a UNIX environment.

Conclusion: Large-scale gene family identification, sequence alignment, and phylogeny reconstruction are all important aspects of large-scale molecular evolutionary analyses. VESPA provides flexible software for simplifying these processes along with downstream selective pressure variation analyses. The software automatically interprets results from codeML and produces simplified summary files to assist the user in better understanding the results. VESPA may be found at the following website: www.mol-evol.org/VESPA

Contact: m.oconnell@leeds.ac.uk

Keywords: Selective pressure analysis, protein molecular evolution, larges-scale comparative genomics.

Supplementary information: The complete manual, tutorial, and user videos are all available at www.mol-evol.org/VESPA

1 Background

Estimating selective pressure variation across homologous protein-coding genes from different species is typically done by assessing the ratio of Dn/Ds, i.e. the number of non-synonymous substitutions per non-synonymous site (Dn) as a function of the number of synonymous substitutions per synonymous site (Ds). The ratio of Dn/Ds is commonly referred to as omega (ω), and is routinely used to assess selective pressure variation or constraints across protein families or protein-interaction networks (Hurst 2002, Kim et al. 2007, Kosiol et al. 2008, Alvarez-Ponce et al. 2009). Some well-known examples of selective pressure variation include the identification of positive selection in reproductive proteins that contribute to species divergence in mammals (Swanson et al. 2001), and the identification of molecular signatures of positive selection that govern protein functional divergence in a group of mammal enzymes (Loughran et al. 2012). A number of software packages estimate selective pressure variation (Pond, Frost et al. 2005, Yang 2007, Delport, Poon et al. 2010). One of the most popular methods is codeML from the PAML software package (Yang 2007). The strength of this approach is the application of flexible codon-based models capable of assessing variation in selective pressures at two levels: (i) across sites in an alignment and (ii) across sites in a predefined lineage on a phylogenetic tree (Yang and dos Reis 2011).

Operating codeML requires a complex file structure to compute the parameters under multiple nested models. Associated likelihood ratio tests (LRTs) must also be performed in the identification of the model of best fit. These complexities are often compounded by the size of study, which increasingly are genomic in scale [Keane et al. 2015, Webb et al. 2015, Liu et al. 2014]. To take advantage of the wealth of publically available genomic data, VESPA (Very large-scale Evolutionary and Selective Pressure Analyses) is capable of performing large-scale

analyses of homology searching, alignment, phylogeny reconstruction and selective pressure variation. This flexible toolkit can permit larger-scale analyses to be performed in an efficient manner and with fewer errors.

Here we present VESPA, which is designed to automate selective pressure analyses and associated prerequisite analyses and post-analysis summary statistics. VESPA is designed primarily to minimize the majority of data manipulation requirements for standard molecular evolutionary analyses and also to automatically implement and analyze selective pressure variation analyses using codeML (Yang 2007). In addition, VESPA supplies an assessment of potential false positives and produces summary files of the results that are easy to interpret.

2 Implementation

VESPA was developed as a toolkit of various independent functions with the primary goal of simplifying the various procedures involved in large-scale selective pressure variation analyses. Each function of the toolkit either completes a specific stage of the analysis (e.g. homologous gene identification) or facilitates/automates the use of third-party software packages to complete more specialized procedures. The majority of functions are written in Python 2.7 and are designed to operate on a UNIX command-line. VESPA categorizes functions into two analyses, a basic analysis for confirmed single gene orthologs (SGOs) and an advanced analysis for both confirmed SGOs and multi-gene families (MGFs) (Figure 1). Functions are further separated into five phases (Table 1 and Figure 1). This structure also provides users with a flexible and adaptable framework for more specialist tasks. An in-depth description of these functions can be found in the program manual on the VESPA website along with tutorials for each command to

demonstrate usage, input format requirements, and command options. Here we provide a summary of the operation and current functionality of VESPA. More information on functionality can be found online (www.mol-evol.org/VESPA).

VESPA operates using a standardized data input and command-line organization. Each analysis phase dictates the supported input data (e.g. sequences, alignments, phylogenies, etc.) and the supported file formats of its functions (e.g. FASTA, NEXUS, Newick, etc.) (Figure 1 and Table 1). Depending on the phase of analysis, VESPA processes input from any program capable of producing the supported file format(s) or a selected collection of third-party programs (Table 1). For example, the homology searching phase currently parses the output of BLAST (Altschul, Gish et al. 1990) or HMMER (Eddy 1998), whereas the alignment assessment and phylogeny reconstruction phase is limited only by file format requirements (e.g. FASTA, NEXUS, PHYLIP). Functions in VESPA are invoked following the program call (i.e. vespa.py) along with arguments to indicate the phase-relevant input data and function-specific optional arguments. Depending on the function, optional arguments enable users to modify parameter values (e.g. BLAST search thresholds, phylogenetic reconstruction settings) or alter command-specific settings.

Functions in VESPA complete by producing the relevant output files without modifying the original input files. While this design results in the generation of a number of intermediate files (especially in the later stages of selection analysis), it enables users to easily keep track of all data modifications. Each phase of VESPA's analysis produces the necessary data files for

conducting a specialized analysis using third-party software (Figure 1). Some of these packages are not fully automated by VESPA for two reasons: i) they are best suited for individual serial tasks on large high-end computing clusters, or ii) the submission processes differ across compute clusters.

3 Example implementation from a mammal dataset

As detailed above, VESPA incorporates two analyses, a basic analysis for analyzing SGOs and an advanced analysis for analyzing both SGOs and MGFs (Figure 2). Here we provide an example of an application of the basic analysis using ten genes from eleven species as a small test dataset.

As seen in Figure 2, the process begins with the user supplying transcript data for the data preparation phase. The first phase begins with the *clean* function, a basic quality control (QC) filtering step, followed by *translate*, to translate the filtered transcripts. VESPA then proceeds to the *make_database* function to create a sequence database for homology searching with either BLAST (Altschul, Gish et al. 1990) or HMMER (Eddy 1998). Upon completion of homology searching, the function *reciprocal_groups* is used to identify proteins that share reciprocal similarity. Then files containing these families of sequences are produced. This function is highly configurable by optional arguments so that users can evaluate various different similarity scenarios (i.e. different e-value cutoffs) with only a single output file. The produced sequences files are then aligned using any multiple sequence alignment (MSA) method that can produce a supported file format (e.g. programs such as MUSCLE [Edgar 2004] and PRANK [Löytynoja and Nick Goldman 2005] are supported). It is advisable to explore a variety of MSA methods for

every gene family (Muller et al. 2010), and VESPA facilitates this the user to compare these different approaches. The *metal_compare* function (within the Alignment Assessment and Phylogeny Reconstruction phase) in VESPA allows alignment approaches to be compared. MSAs are then used in combination with the user-defined species phylogeny to create gene phylogenies using the function *infer_genetrees*. The MSAs and gene phylogenies are then used for the selective pressure analysis preparation phase. The function *create_branch* can be used to specify label internal nodes as ancestral lineages that the user may wish to explore. The MSAs and gene phylogenies are then used by the function *setup_codeml* to automatically create the complex codeML file structure and a task file for automating codeML (Yang 2007). Upon completion of codeML the *codeml_reader* function is used to automate the interpretation of the results and producing summary files of the results.

4 Discussion

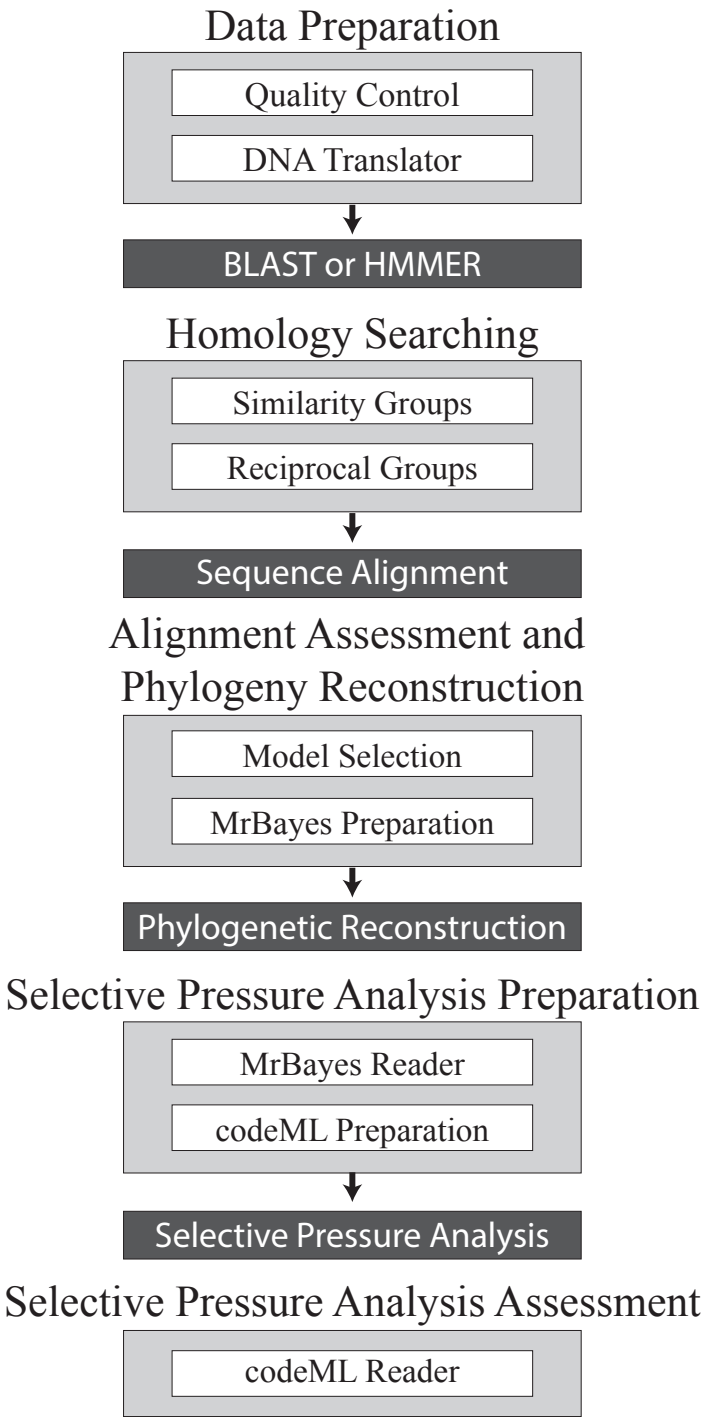
The VESPA toolkit was designed to both simplify and streamline large-scale comparative genomic analyses including codeML-based selective pressure analyses. The goal of the toolkit was to provide the community with functions capable of performing the associated prerequisite analyses and to minimize the error-prone or technically challenging procedures associated with selective pressure analyses. VESPA also provides a flexible frameowrk for analyzing both single gene orthologous families and multigene families. Users can apply the entire suite of available options within VESPA or only specific functions that are of interest (e.g. homology searching). VESPA is also capable of directly interpreting and presenting all relevant information from a selective pressure analysis within simplified summary files.

5 Conclusion:

VESPA provides a flexible software package designed to simplify large-scale selective pressure variation analysis, including those using the program codeML (Yang 2007), by automating the entire comparative genomic process from data quality checks and homology searching to phylogeny reconstruction and selective pressure analyses, and it produces simple summary files for the user. VESPA offers users various functions that automate many of the required prerequisite analyses and removes error-prone data manipulation steps.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v1 | CC-BY 4.0 Open Access | rec: 25 Mar 2016, publ: 25 Mar 2016

162 **Figure 1:** Overview of the phases implemented in VESPA.



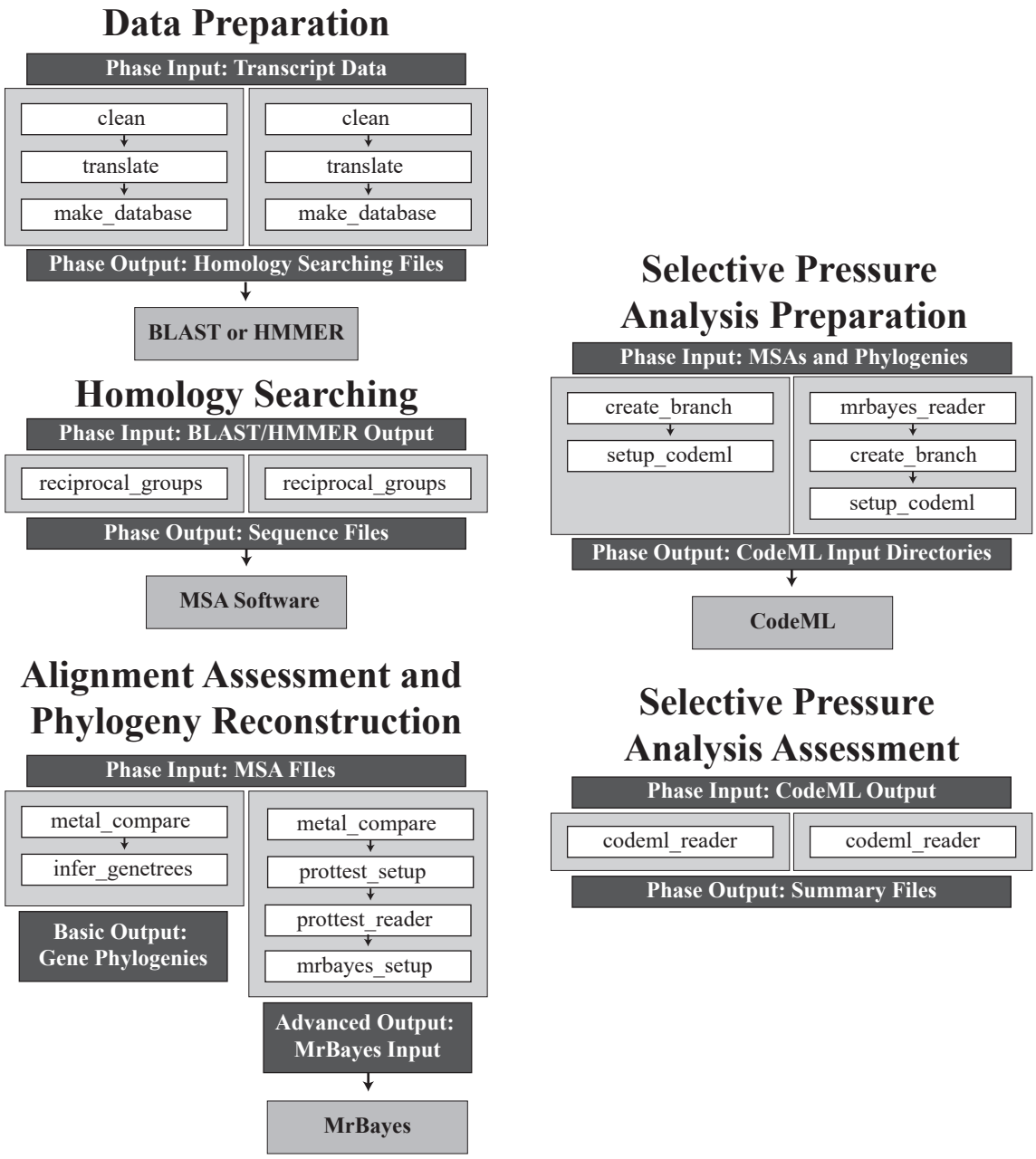
163

164 **Figure 1 legend:** The 5 phases of VESPA are listed from "Data Preparation" to "Selective

165 Pressure Analysis Assessment". Underneath each is a grey box enclosing some representative

commands from that phase. Each phase concludes with a black box indicating the use of a third-party program to perform the necessary task (e.g. sequence alignment or phylogenetic reconstruction). The output of the first 4 phases is then used as the input of the next phase. The final phase concludes with the creation of summary files that contain all the relevant information from the selective pressure analyses.

172 **Figure 2:** Overview of the options available in the VESPA package.



173

174 **Figure 2 legend:** An overview of both the basic (on left) and advanced (on right) analysis
175 options at each phase of VESPA highlighting key differences. The functions of each phase are
176 shown as white boxes, and are invoked in the order shown (Note: that not all functions are

shown). In addition to the functions, the input and output of each phase are shown in dark grey boxes and if a third-party program is required to analyze the output of the phase, the program will be specified below the phase in a light grey box. For three of the five phases (data preparation, homology searching, and selective pressure analysis assessment) the functions invoked in both the basic and advanced options are identical. The primary difference between the analyses (basic/advanced) is found in the alignment assessment and phylogeny reconstruction phase. The advanced option uses ProtTest (Darriba et al. 2001) for substitution model selection and MrBayes (Ronquist and Huelsenbeck 2003) for phylogenetic reconstruction. Beyond this major difference, the selective pressure analysis preparation simply requires a function to import the output of MrBayes.

Tables

Table 1: Overview of the Phases in the VESPA software package

Phase	Purpose	Supported Input Type	Supported File Formats
1	Data Preparation	Sequences[1]	FASTA
2	Homology Searching	BLAST/HMMER output	BLAST tabular, HMMER standard
3	Alignment Assessment and Phylogeny Reconstruction	Alignments[1]	FASTA, NEXUS, PHYLIP
4	Selective Pressure Analysis Preparation	Phylogenies with alignments[1]	Trees: Newick, NEXUS; Alignments: See above
5	Selective Pressure Analysis Assessment	codeML output	LaMP formmated codeML output

[1]Indicates phases of VESPA that incorporate third-party programs.

Acknowledgements

The authors would like to thank The Irish Centre for High-End Computing (ICHEC). Science Foundation Ireland Research Frontiers Programme grant (SFI RFP EOB2673) to MJO'C, and DCU Pierse Trust Award and SoBT awards (to TAW and MJO'C). We would also like to thank Louisse Mirabueno (funded by the Wellcome Trust Vacation Scholarship programme) and other members of the community for their help in trouble-shooting, testing and providing feedback on the VESPA software package and associated manual and tutorials.

Bibliography

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Alvarez-Ponce, D., Aguadé, M., & Rozas, J. (2009). Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes. *Genome Research*, *19*(2), 234–242. http://doi.org/10.1101/gr.084038.108

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)*, *27*(8), 1164–1165. http://doi.org/10.1093/bioinformatics/btr088

Delport, W., A. F. Poon, S. D. Frost and S. L. Kosakovsky Pond (2010). "Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology." Bioinformatics **26**(19): 2455-2457.

Eddy, S. R. (1998). "Profile hidden Markov models." Bioinformatics **14**(9): 755-763.

Hurst, L. D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in Genetics : TIG*, *18*(9), 486–487.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*, 113. http://doi.org/10.1186/1471-2105-5-113

Keane, M., Semeiks, J., Webb, A. E., Li, Y. I., Quesada, V., Craig, T., et al. (2015). Insights into the Evolution of Longevity from the Bowhead Whale Genome. *Cell Reports*, *10*(1), 112–122. http://doi.org/10.1016/j.celrep.2014.12.008

Kim, P. M., Korbel, J. O., & Gerstein, M. B. (2007). Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(51), 20274–20279. http://doi.org/10.1073/pnas.0710183104

Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., & Siepel, A. (2008). Patterns of positive selection in six Mammalian genomes. *PLoS Genet*, *4*(8), e1000144. http://doi.org/10.1371/journal.pgen.1000144

Liu, S., Lorenzen, E. D., Fumagalli, M., Li, B., Harris, K., Xiong, Z., et al. (2014). Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, *157*(4), 785–794. http://doi.org/10.1016/j.cell.2014.03.054

Loughran, N. B., Hinde, S., McCormick-Hill, S., Leidal, K. G., Bloomberg, S., Loughran, S. T., et al. (2012). Functional Consequence of Positive Selection Revealed Through Rational

234 Mutagenesis of Human Myeloperoxidase. *Molecular Biology and Evolution*.
235 http://doi.org/10.1093/molbev/mss073

236 Löytynoja, A., & Goldman, N. (2005). An algorithm for progressive multiple alignment of
237 sequences with insertions. *Proceedings of the National Academy of Sciences of the United*
238 *States of America*, *102*(30), 10557–10562. http://doi.org/10.1073/pnas.0409137102

239 Muller, J., Creevey, C. J., Thompson, J. D., Arendt, D., & Bork, P. (2010). AQUA: automated
240 quality improvement for multiple sequence alignments. *Bioinformatics (Oxford, England)*,
241 *26*(2), 263–265. http://doi.org/10.1093/bioinformatics/btp651

242 Pond, S. L., S. D. Frost and S. V. Muse (2005). "HyPhy: hypothesis testing using phylogenies."
243 Bioinformatics **21**(5): 676-679.
244
245 Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under
246 mixed models. *Bioinformatics (Oxford, England)*, *19*(12), 1572–1574.

247 Sukumaran, J. and M. T. Holder (2010). "DendroPy: a Python library for phylogenetic
248 computing." Bioinformatics **26**(12): 1569-1571.
249
250 Swanson, W. J., Yang, Z., Wolfner, M. F., & Aquadro, C. F. (2001). Positive Darwinian
251 selection drives the evolution of several female reproductive proteins in mammals.
252 *Proceedings of the National Academy of Sciences of the United States of America*, *98*(5),
253 2509–2514. http://doi.org/10.1073/pnas.051605998

254 Webb, A. E., Gerek, Z. N., Morgan, C. C., Walsh, T. A., Loscher, C. E., Edwards, S. V., &
255 O'Connell, M. J. (2015). Adaptive Evolution as a Predictor of Species-Specific Innate
256 Immune Response. *Molecular Biology and Evolution*.
257 http://doi.org/10.1093/molbev/msv051

258 Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Mol Biol Evol
259 **24**(8): 1586-1591.
260
261 Yang, Z. and M. dos Reis (2011). "Statistical properties of the branch-site test of positive
262 selection." Mol Biol Evol **28**(3): 1217-1228.

263