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[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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

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 entire proteomes (all protein coding sequences in a genome) from a large number of species.

Methods. 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 Perl and is designed to run within a UNIX environment.

Results. We have benchmarked VESPA and our results show that the method is consistent, performs well on both large scale and smaller scale datasets, and produces results in line with previously published datasets.

Discussion. 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

Supplementary information: The complete manual, tutorial, and user videos are all available at www.mol-evol.org/VESPA and software from https://github.com/aewebb80/VESPA

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

Introduction

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 (Kim, Korbel et al. 2007, Kosiol, Vinar et al. 2008, Alvarez-Ponce, Aguade et al. 2009). These calculations of selective pressure variation are performed on alignments of protein coding sequences (and not on other data types such as raw reads from NGS experiments). Codeml is part of the PAML package for the analyses of selective pressure variation in nucleotide sequence data in a maximum likelihood framework (Yang 2007). The models available in PAML for assessing selective pressure variation can simultaneously compare variation across sites and across lineages in the homologous protein coding gene dataset. In this way the "foreground lineage" is compared to all other lineages in the dataset in an attempt to determine lineage specific selective pressure variation. Some well-known examples of selective pressure variation on foreground lineages include the identification of positive selection in reproductive proteins that contribute to species divergence in mammals (Swanson, Yang et al. 2001), and the identification of molecular signatures of positive selection that govern protein functional divergence in a group of mammal enzymes (Loughran, Hinde 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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

65 sites in an alignment and (ii) across sites in a predefined, or "foreground" lineage on a
66 phylogenetic tree (Yang and dos Reis 2011).

67 Operating codeML requires a complex file structure to compute the parameters under multiple
68 nested models. Associated likelihood ratio tests (LRTs) must also be performed in the
69 identification of the model of best fit. These complexities are often compounded by the size of
70 study, which increasingly are genomic in scale (Liu, Lorenzen et al. 2014, Keane, Semeiks et al.
71 2015, Webb, Gerek et al. 2015). Other approaches to streamline the process of applying codon-
72 based models of evolution to homologous sequences sets focus on site-specific models such as
73 POTION (Hongo, de Castro et al. 2015).

74 To address these issues we have designed VESPA (Very large-scale Evolutionary and Selective
75 Pressure Analyses). VESPA automates selective pressure analyses and associated prerequisite
76 analyses and post-analysis summary statistics. VESPA can perform both lineage-site specific and
77 site-specific analyses whereas POTION presently performs the site-specific analyses. Therefore,
78 VESPA is unique in its capacity to perform the complex set of tasks involved in assessing
79 lineage specific selective pressure variation across homologous gene families and across
80 lineages. VESPA minimizes the majority of data manipulation requirements for standard
81 molecular evolutionary analyses and also automatically implements and analyzes selective
82 pressure variation analyses using codeML (Yang 2007). In addition, VESPA supplies an
83 assessment of potential false positives and produces summary files of the results that are easy to
84 interpret. VESPA allows the user to take advantage of the wealth of publically available genomic
85 data from model and non-model organisms to perform large-scale analyses of homology
86 searching, alignment, phylogeny reconstruction and selective pressure variation. All that VESPA
87 requires is the protein coding DNA sequences, which it will translate with the standard genetic

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

code and use to search and construct gene family alignments. This flexible toolkit can permit large-scale analyses to be performed in an efficient manner and with fewer errors.

Methods

VESPA helps automation by preparing input data files and processing results but program executions are initiated by the user (e.g. via submission to an HPC queuing system. VESPA has 5 major Phases (Table 1 and Figure 1) each of which is composed of a number of functions. VESPA was developed as a toolkit of various independent functions within each Phase and the primary goal is to simplify the procedures involved in large-scale selective pressure variation analyses. Each function either completes a specific Phase 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. Functions are categorized as either basic or advanced, e.g. the function to identify single gene orthologs is a basic function whereas confirming both SGOs and multi-gene families (MGFs) is an advanced function (Figure 2). This structure also provides users with a flexible and adaptable framework for more specialist tasks. For in-depth description and format requirements, please see the program manual, tutorials and documentation published on the program website (www.mol-evol-org/VESPA)

Depending on the Phase of VESPA input is accepted from any program capable of producing the supported file format(s) or a selected collection of third-party programs (Table 1). For example, Phase 2 (the homology searching phase) currently parses the output of BLAST (Altschul, Madden et al. 1997) or HMMER (Eddy 1998), whereas the alignment assessment and phylogeny

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

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). The homology searches are not fully automated by VESPA for two reasons: i) they are best run as individual serial tasks on large high-end computing clusters, or ii) the submission processes differ across compute clusters. However, VESPA can generate the BLAST formatted database for subsequent homology searches. All VESPA commands are issued on the command line and are readily executable via the scheduling system of the users server facility.

Results

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

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 et al. 1997) or HMMER3 (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 (Loytynoja and 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 and a single MSA of best fit is retained for each gene family (i.e. Gene Family A may have an MSA from MUSCLE while Gene Family B may have an MSA from PRANK). MSAs for each SGO family can then used in combination with the user-defined species phylogeny to create gene phylogenies using the function *infer_genetrees* or gene trees can be generated directly from the MSAs The MSAs and gene phylogenies are then used for the selective pressure analysis preparation phase. The function *create_branch* can be used to specifically label internal nodes as

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

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. VESPA creates a number of output files for each homologous gene family detailing the results of the codeML analysis. There are two primary output files for each gene family tested: 1) a single csv formatted summary text file which is readily formatted as seen in Table 2 for one gene family, and 2), a multiple sequence alignment for each model tested detailing the sites (protein/codon) proposed to be under positive selection is also provided in html format so it can be viewed with colour coding for ease of interpretation. For details of the summary file see Table 2. The summary text file (structured as a CSV file) contains this set of information for all genes tested, each within a clearly defined section. s

To compare the results of VESPA to other pipelines and predictions of positive selection, we used a dataset of 18 gene families from the Selectome database (Moretti et al. 2014). A tarball of the data and results of the VESPA analysis of 18 gene families (i.e. input sequences, alignments, trees, codeml output, VESPA summary at each phase) have been provided in supplementary file 1. Selectome is a publicly available database of genes under positive selection. Selectome was chosen to provide this benchmark dataset as it permits direct access to all relevant files used in its calculations, and it uses the codon based models of evolution implemented in codeML and integrated in VESPA facilitating a direct comparison of results.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

We carried out two tests with the dataset from Selectome (Release 6) (Moretti et al. 2014). Firstly we wished to assess if the way in which VESPA automatically sets up all codon models, labeling of foreground lineages and LRTs produced comparable results with those from the Selectome pipeline. To this end we ran VESPA using the masked DNA alignments and gene trees for each of the 18 homologous gene families from the Selectome database. VESPA produced identical results for these input alignments. This demonstrates that the VESPA automation of the process is reliable and robust.

To illustrate that different alignments for the same gene families can produce different results and therefore to highlight the importance of the alignment comparison feature in VESPA , *i.e.* the "metal_compare" function, we performed an additional test on this dataset of 18 homologous gene families from Selectome. We used the unmasked sequences and generated a set of alternative alignments through MUSCLE (Edgar 2004) and PRANK (Loytynoja and Goldman 2005) and using the metal_compare function of VESPA we identified the best (most statistically significant) alignment for each gene family. The VESPA functions 'setup_codeml' and 'codeml_reader' were then used to automate the codeML set up and analysis. The VESPA pipeline was able to replicate the lineages identified as under positive selection in the Selectome database. However, the results presented in Table 3 include small differences in the number and position of sites identified as positively selected, Table 3. The ATP5SL alternative alignment was found to have evidence of positive selection on two additional branches as compared to the original result, however, closer inspection of the alternative alignments found the additional branches to be false positives due to a poorly aligned segment of the MSA. Other slight variations in results between the original and alternative alignments included three instances where gene families (C12orf43, CACNA1I, and PASK) had a difference of one positively

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

selected site and one instance (TRIM40) of differences in two positively selected sites reported in the original alignment that was not reported in the alternative alignment for the same family. Finally, a single gene (SLC8A1) contained additional sites under positive selection following VESPA analysis using the alternative alignments, it should be noted that these sites are within a poorly conserved span of the protein. The remaining 13 genes were found to replicate the positively selected sites reported in Selectome.

To demonstrate the application of VESPA to very large datasets we have assembled a novel dataset of 7,918 homologous gene families (each containing >= 8 sequences) from the ENSEMBL 2016 genome database (release 82) (Yates et al. 2016). Using the VESPA "clean_ensembl" command we: (1) restricted the protein coding sequences per genome to the longest transcript for each gene, and (2) sequences containing internal stop codons or incomplete codons are logged and discarded. Multiple sequence alignments were made for all gene families identified and phylogenies for each homologous gene families were inferred by VESPA from the topology of the Ensembl Compara species tree demonstrating the flexibility of the VESPA package (Phase 3). Finally, the selective pressure analyses were carried out in VESPA Phase 5 with human and then mouse labeled as the foreground lineages of interest. Of the 7,918 genes analyzed by VESPA 1998 showed evidence of site-specific positive selection (model "m8"), 223 genes showed evidence of mouse-specific positive selection ("model A"), and finally 80 showed evidence of human-specific positive selection ("model A"), supplementary file 1 contains the full set of results.

For the dataset of 7,918 homologous gene families the majority of VESPA functions completed in under four minutes (on the following system: Intel Xeon CPU E5420 (2.50GHz), 16GB RAM, using Ubuntu OS 16.04). The exceptions were the 'codeml_setup' function which took

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

approximately 45 minutes for this dataset, this function is slightly slower as it is creating the complex directory/file structure for codeML. The second exception was the 'codeml reader' function which took approximately 6 hours (this function is to analyze the large number of files created by codeML and produce the output and summary files). The codeML component of the analyses took 36,180 CPU hours in total to complete. [Note: these time estimates are not inclusive of Phase 2].

Discussion

One of the primary goals of VESPA is to simplify and streamline basic comparative genomic procedures such as filtering poor quality protein coding sequences and generating the most appropriate alignment for each gene family. VESPA also simplifies and streamlines codeML-based selective pressure analyses. From our analysis of 7,918 homologous gene families we found that the majority of tasks could be completed within minutes using VESPA. However, the codeML-related functions for creating the input file structure and examining the output files takes considerably longer to complete. As these functions are an essential aspect of the pipeline, decreasing their execution time will be a primary goal in future updates to VESPA. Two possible approaches that will be explored are: i) increasing the overall efficiency of the functions and ii) developing a version of VESPA capable running these scripts (and possibly others) in parallel on multi-core processors. Future updates to VESPA will also explore additional functions and methods not currently implemented.

The modular nature of VESPA enables the pipeline (or specific functions) to be used in conjunction with existing workflow systems. The only requirements would be having the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

necessary data (MSAs, protein-coding transcripts, etc.) in a supported format. For example, VESPA could be used to perform a selective pressure analysis on protein-coding transcripts obtained from RNA-Seq. Also, it is possible to skip specific stages of the VESPA pipeline for a preferred approach/software package, e.g. it is possible to use an alternative approach for phylogenetic reconstruction and employ the resulting tree in your VESPA analysis. It should also be possible to integrate the majority of VESPA's functions (with the exception of the built-in tree pruning function) into workflow systems such as GALAXY (Afgan et al. 2016). This would allow VESPA to operate in conjunction with the scripts and tools already available on GALAXY, enabling greater freedom for the user. This integration will be implemented in future releases of VESPA.

We also employed VESPA in the analysis of 18 gene families from the Selectome database (Moretti et al. 2014). Our initial comparison used both the alignment (masked) and tree provided by Selectome. VESPA was able to confirm the findings reported within the database. Secondly, we re-aligned the 18 gene families using two methods (MUSCLE and PRANK) and then again performed the selective pressure analysis using VESPA. The analysis of these alternative alignments revealed minor differences in the reported positively selected sites of five genes (C12orf43, CACNA1I, PASK, SLC8A1, and TRIM40). These differences illustrate that the input alignment may have an impact on the genes and sites identified as positively selected as in (Blackburne and Whelan 2013). We therefore highly recommend that users are not biased in their choice of alignment method and we recommend the use of the "metal_compare" function or programs such as AQUA (Muller et al. 2010) to select which method is best for each gene family input alignment.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

265 It is important to note that the processes implemented in the 5 Phases of VESPA facilitates those
266 working on *de novo* sequence data or non-model organisms to perform large-scale comparative
267 genomic analyses without having pre-processed gene families, all that is required by VESPA is
268 that the protein coding DNA sequences are available.

269

270 Conclusion

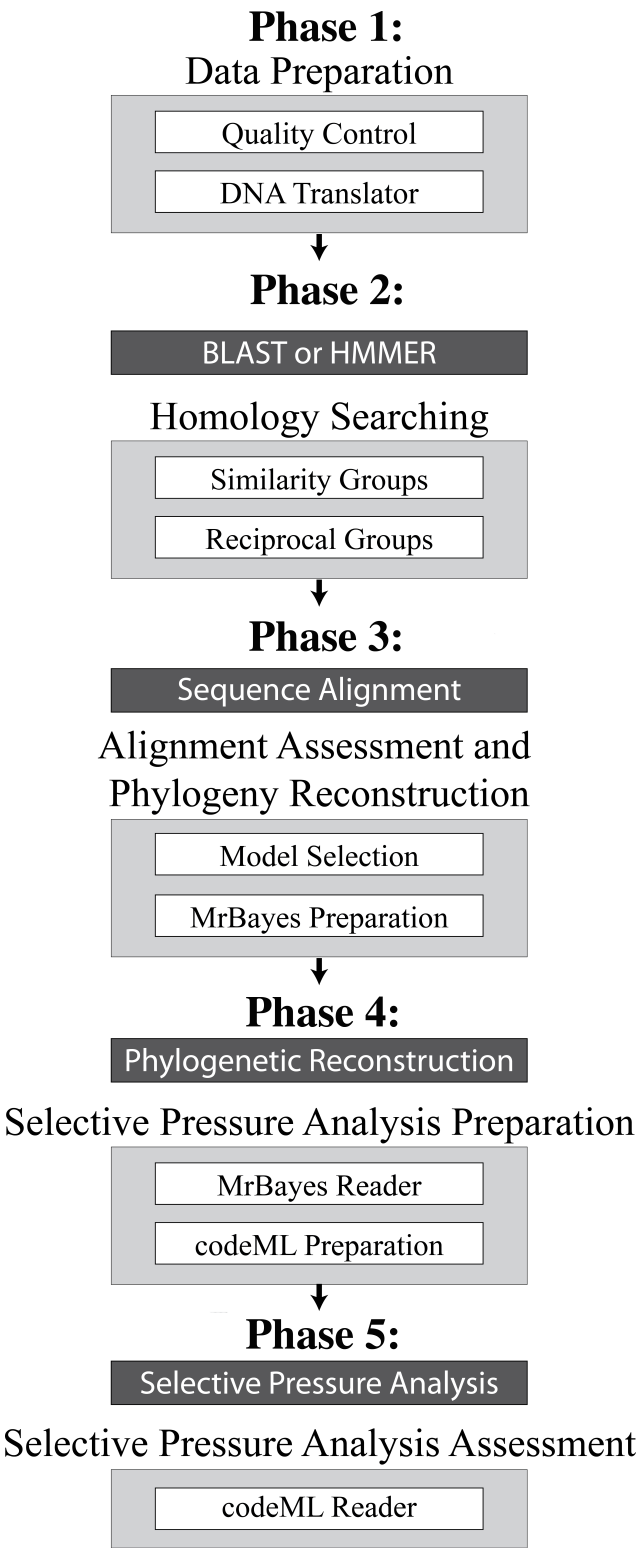
271 VESPA provides a flexible software package designed to simplify large-scale comparative
272 genomic analyses and specifically selective pressure variation analysis implemented in codeML
273 (Yang 2007). VESPA automates the entire comparative genomic process from data quality
274 checks and homology searching to phylogeny reconstruction and selective pressure analyses, and
275 it produces simple summary files for the user. VESPA offers users various functions that
276 automate many of the required prerequisite analyses and removes error-prone data manipulation
277 steps.

278

279

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

280 **Figure 1: Overview of the 5 Phases implemented in VESPA.**



281

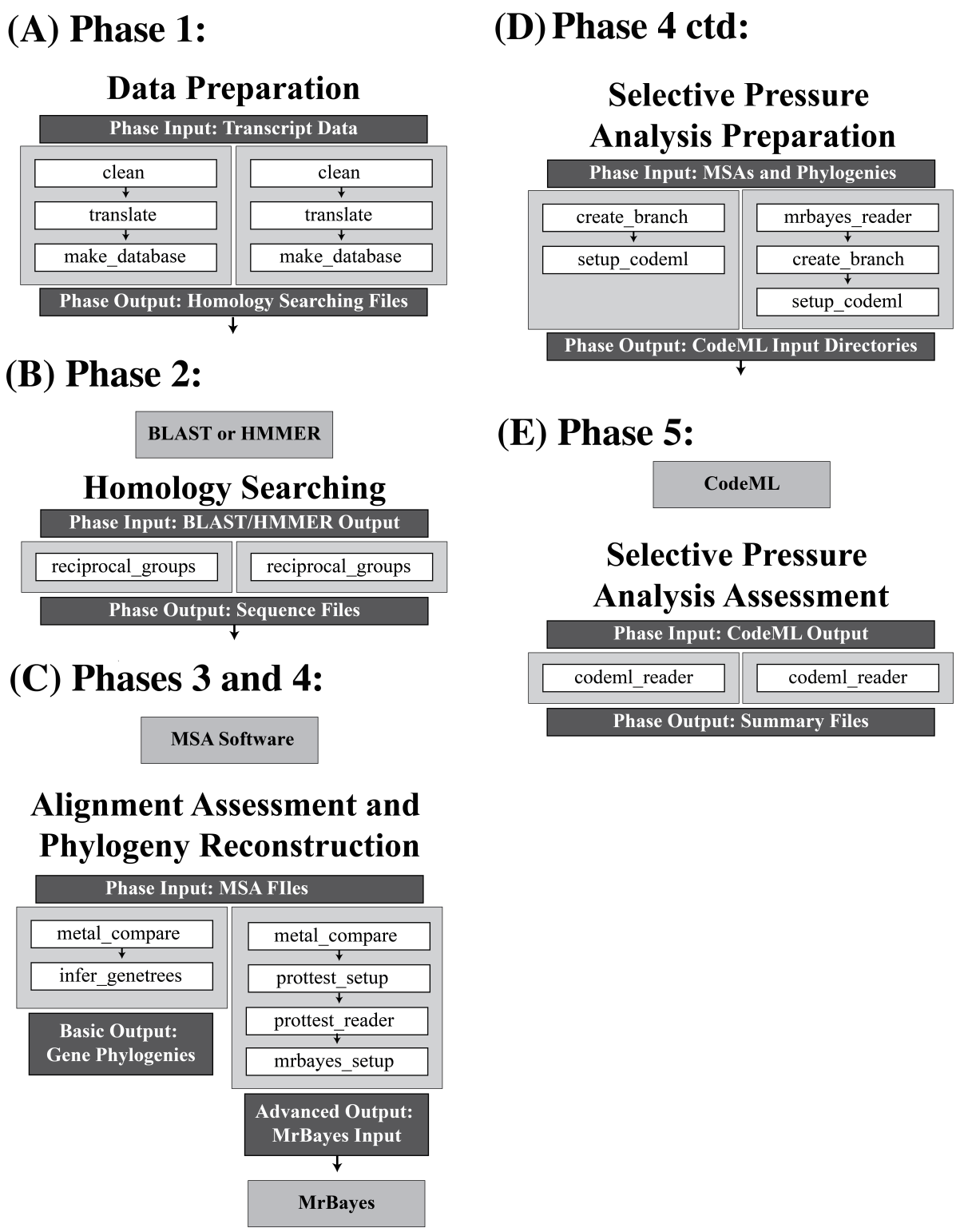
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

282 **Figure 1 legend:** The 5 Phases of VESPA are listed from "Data Preparation" to "Selective
283 Pressure Analysis Assessment". Underneath each is a grey box enclosing some representative
284 commands from that phase. Each phase concludes with a black box indicating the use of a third-
285 party program to perform the necessary task (e.g. sequence alignment or phylogenetic
286 reconstruction). The output of the first 4 phases is then used as the input of the next phase. The
287 final phase is written in Perl and concludes with the creation of summary files that contain all the
288 relevant information from the selective pressure analyses.

289

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

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



291

292

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

Figure 2 legend: An overview of the basic and advanced analysis options at each of the 5 Phases of VESPA. The functions of each phase are shown as white boxes, and are invoked in the order shown (Note: that not all functions are shown here, a complete set of VESPA functions are available in the manual). Within each phase the available alternatives for processing the data are given on the left and right hand columns. These only vary for Phase 3 and 4. The left most column may represent the processing of single gene orthologs that you wish to impose the species tree onto. If this is the case then VESPA will allow you to generate the phylogenies from the species tree (as shown on the left most side of (C)). However, you may wish to generate gene trees directly from the data for either multigene families or for single gene families of uncertain orthology (this option is shown on the right most column in (C) and involves selecting the substitution model of best fit and reconstructing the phylogeny). 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 is 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, Taboada et al. 2011) 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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

Tables

Table 1: Overview of the 5 Phases in the VESPA software package

Phase	Purpose	Supported Input Type	Supported File Formats
1	Data Preparation	Protein Coding DNA Sequences[1]	FASTA
2	Homology Searching	BLAST/HMMER output files	BLAST tabular, HMMER standard
3	Alignment Assessment and Phylogeny Reconstruction	Multiple Sequence Alignments[1]	FASTA, NEXUS, PHYLIP
4	Selective Pressure Analysis Preparation	Gene phylogenies (or species phylogenies) with corresponding Multiple sequence alignments[1]	Trees: Newick, NEXUS; Alignments: See above
5	Selective Pressure Analysis Assessment	Standard codeML output files generated directly by the software	VESPA formatted codeML output

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

The file formats required as input for each phase of VESPA are detailed. The numbering scheme is consistent with the numbering scheme for the phases as displayed in Figure 2.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

Table 2: Sample of a summary output file created by 'codeml_reader' in Phase 5 of the VESPA package

Model	Tree	Model Type	p	W (t=0)	lnL	LRT Result	Parameter Estimates	Positive Selection	Positively Selected Sites in 8x45 Alignment (P(w>1)>0.5)
m0	Sample_MSA	Homogeneous	1	2	-572.969394	N/A	w=0.61924	No	
m1Neutral	Sample_MSA	Site-specific	1	2	-568.572319	N/A	p0=0.35065 p1=0.64935 w0=0.08403 w1=1.00000	Not Allowed	
m2Selection	Sample_MSA	Site-specific	2	2	-568.521978	m1Neutral	p0=0.37699 p1=0.00000 p2=0.62301 w0=0.10174 w1=1.00000 w2=1.11245	No	
m3Discrtk2	Sample_MSA	Site-specific	3	2	-568.521978	m3Discrtk2	p0=0.37699 p1=0.62301 w0=0.10174 w1=1.11245	Yes	Alignment (28 NEB sites): 2 3 4 5 6 8 12 13 15 17 19 20 22 23 24 25 26 27 30 31 32 34 35 39 40 41 43 44
m3Discrtk3	Sample_MSA	Site-specific	5	2	-568.521978	m3Discrtk2	p0=0.37699 p1=0.53897 p2=0.08404 w0=0.10174 w1=1.11245 w2=1.11246	No	
m7	Sample_MSA	Site-specific	2	2	-568.764172	N/A	p=0.22135 q=0.11441	Not Allowed	
m8	Sample_MSA	Site-specific	4	2	-568.526486	m7, m8	p=11.60971 p0=0.37916 p1=0.62084 q=99.00000 w=1.11431	No	
m8a	Sample_MSA	Site-specific	4	1	-568.578069	N/A	p=9.38528 p0=0.35203 p1=0.64797 q=99.00000 w=1.00000	Not Allowed	
modelA	Sample_MSA_ Primates	Lineage-site	3	2	-568.572319	m1Neutral, modelAnull	p0=0.35065 p1=0.64935 p2=0.00000 p3=0.00000 w0=0.08403 w1=1.00000	No	

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

							w2=1.00000			
modelAnull	Sample_MSA_Primates	Lineage-site	3	1	-568.572319	N/A	p0=0.35065 p2=0.00000 w0=0.08403 w2=1.00000	p1=0.64935 p3=0.00000 w1=1.00000		
modelA	Sample_MSA_Chimp	Lineage-site	3	2	-557.052657	modelA	p0=0.33452 p2=0.02659 w0=0.1099 w2=999.000	p1=0.59186 p3=0.04704 w1=1.00000	Yes	Alignment (3 BEB sites): 24 25 32
modelAnull	Sample_MSA_Chimp	Lineage-site	3	1	-568.572319	N/A	p0=0.35065 p2=0.00000 w0=0.08403 w2=1.00000	p1=0.64935 p3=0.00000 w1=1.0000	Not Allowed	
modelA	Sample_MSA_Human	Lineage-site	3	2	-568.572319	m1Neutral, modelAnull	p0=0.35065 p2=0.00000 w0=0.08403 w2=1.00000	p1=0.64935 p3=0.00000 w1=1.0000	No	
modelAnull	Sample_MSA_Human	Lineage-site	3	1	-568.572319	N/A	p0=0.35065 p2=0.00000 w0=0.08403 w2=1.00000	p1=0.64935 p3=0.00000 w1=1.0000	No	

Table 2: The output file for every gene family details each site-specific and lineage-site-specific model from codeML tested within VESPA. The following information is provided for each model tested: the lineage (internal or terminal branches) tested as foreground; the type of model (i.e. site-specific or branch-specific) being tested; number of free parameters in the ω distribution that are estimated by codeML, the initial ω value used by codeML (each run within VESPA has multiple starting values to minimise the risk of reporting from a local minimum on the likelihood plane); the resulting log likelihood (lnL) of the analysis; the resulting model of the likelihood ratio test (LRT); the parameter

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

estimates of codeML; if positive selection was detected (yes/no); and the alignment coordinates for any positively selected sites. NEB = Naïve Empirical Bayes estimate and BEB = Bayes Empirical Bayes estimate.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

Table 3: Comparison of the results from 18 gene families from the Selectome database analysed in VESPA

	Original Masked Alignment		Alternative Alignment	
	Positive Selection (ModelA)	Positively Selected Sites	Positive Selection (ModelA)	Positively Selected Sites
ATG16L1	Yes (HPGPoNM)	Match	Yes (HPGPoNM , HP*, and HPG*)	Match
ATP5SL	Yes (HP)	Match	Yes (HP)	Match
AZGP1	Yes (HPGPoNM)	Match	Yes (HPGPoNM)	Match 9/10 (107)
C12orf43	Yes (HPG)	Match	Yes (HPG)	Match
CACNA1I	Yes (HP)	Match	Yes (HP)	Match 2/3 (1095)
CASP1	Yes (HPG)	Match	Yes (HPG)	Match
CD151	Yes (HPG)	Match	Yes (HPG)	Match
COBLL1	Yes (HPGPoN)	Match	Yes (HPGPoN)	Match
HUS1	Yes (HPGPoNM)	Match	Yes (HPGPoNM)	Match
IDH3B	Yes (HP)	Match	Yes (HP)	Match
IFIT2	Yes (HPGPoNM)	Match	Yes (HPGPoNM)	Match
INTS7	Yes (HPGPoNM)	Match	Yes (HPGPoNM)	Match
ODC1	Yes (HPGPoN)	Match	Yes (HPGPoN)	Match
PASK	Yes (HP)	Match	Yes (HP)	Match 2/3 (434)
RRP8	Yes (HPGPoN)	Match	Yes (HPGPoN)	Match
RTP2	Yes (HP)	Match	Yes (HP)	Match
SLC8A1	Yes (PG)	Match	Yes (PG)	Match

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

TRIM40	Yes (HPGPo)	Match	Yes (HPGPo)	Match 2/4 (140, 256)

*False positives

Table 3: The 18 homologous families chosen from the Selectome database are given their HUGO identifier on the left. The results of analysis in VESPA as compared to Selectome results using the same masked alignments are shown in column 2 and 3. The results from VESPA using the alternative alignment method as compared to the original alignments is shown in column 4 and 5. For both alignments (original and alternative) it is indicated if ModelA positive selection is identified in the same lineages, and if the sites identified as positively selected match. Using the alternative alignments, four cases where the sites identified as positively selected did not completely match the position in the original alignment are indicated in parenthesis. The extant lineages with evidence of positive selection throughout are shown in parentheses and are abbreviated as follows: Human (H), Chimp (P), Gorilla (G), Orangutan (Po), Gibbon (N) and Macaque (M). Ancestral nodes are denoted as a combination of the abbreviations for the extant lineages that the node includes, e.g. the ancestral node joining Human (H) and Chimp (P) is denoted as HP.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

344

345 **Supplementary file 1:** CSV file containing the results of the analysis of selective pressure
346 variation across 8,105 gene families labeling as foreground either human or mouse.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

Acknowledgements

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.

352 **Bibliography**

353

354 Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman
355 (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search
356 programs." Nucleic Acids Res **25**(17): 3389-3402.
357
358 Alvarez-Ponce, D., M. Aguade and J. Rozas (2009). "Network-level molecular evolutionary
359 analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes."
360 Genome Res **19**(2): 234-242.
361
362 Blackburne, B. P. and S. Whelan (2013). "Class of multiple sequence alignment algorithm
363 affects genomic analysis." Mol Biol Evol **30**(3): 642-653.
364
365 Darriba, D., G. L. Taboada, R. Doallo and D. Posada (2011). "ProtTest 3: fast selection of best-
366 fit models of protein evolution." Bioinformatics **27**(8): 1164-1165.
367
368 Delport, W., A. F. Poon, S. D. Frost and S. L. Kosakovsky Pond (2010). "Datamonkey 2010: a
369 suite of phylogenetic analysis tools for evolutionary biology." Bioinformatics **26**(19): 2455-
370 2457.
371
372 Eddy, S. R. (1998). "Profile hidden Markov models." Bioinformatics **14**(9): 755-763.
373
374 Edgar, R. C. (2004). "MUSCLE: a multiple sequence alignment method with reduced time and
375 space complexity." BMC Bioinformatics **5**: 113.
376
377 Hongo, J. A., G. M. de Castro, L. C. Cintra, A. Zerlotini and F. P. Lobo (2015). "POTION: an
378 end-to-end pipeline for positive Darwinian selection detection in genome-scale data through
379 phylogenetic comparison of protein-coding genes." BMC Genomics **16**: 567.
380
381 Keane, M., J. Semeiks, A. E. Webb, Y. I. Li, V. Quesada, T. Craig, L. B. Madsen, S. van Dam,
382 D. Brawand, P. I. Marques, P. Michalak, L. Kang, J. Bhak, H. S. Yim, N. V. Grishin, N. H.
383 Nielsen, M. P. Heide-Jorgensen, E. M. Oziolor, C. W. Matson, G. M. Church, G. W. Stuart, J. C.
384 Patton, J. C. George, R. Suydam, K. Larsen, C. Lopez-Otin, M. J. O'Connell, J. W. Bickham, B.
385 Thomsen and J. P. de Magalhaes (2015). "Insights into the evolution of longevity from the
386 bowhead whale genome." Cell Rep **10**(1): 112-122.
387
388 Kim, P. M., J. O. Korbel and M. B. Gerstein (2007). "Positive selection at the protein network
389 periphery: evaluation in terms of structural constraints and cellular context." Proc Natl Acad Sci
390 U S A **104**(51): 20274-20279.
391
392 Kosiol, C., T. Vinar, R. R. da Fonseca, M. J. Hubisz, C. D. Bustamante, R. Nielsen and A. Siepel
393 (2008). "Patterns of positive selection in six Mammalian genomes." PLoS Genet **4**(8): e1000144.
394 Liu, S., E. D. Lorenzen, M. Fumagalli, B. Li, K. Harris, Z. Xiong, L. Zhou, T. S. Korneliussen,
395 M. Somel, C. Babbitt, G. Wray, J. Li, W. He, Z. Wang, W. Fu, X. Xiang, C. C. Morgan, A.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

396 Doherty, M. J. O'Connell, J. O. McInerney, E. W. Born, L. Dalen, R. Dietz, L. Orlando, C.
397 Sonne, G. Zhang, R. Nielsen, E. Willerslev and J. Wang (2014). "Population genomics reveal
398 recent speciation and rapid evolutionary adaptation in polar bears." Cell **157**(4): 785-794.
399
400 Loughran, N. B., S. Hinde, S. McCormick-Hill, K. G. Leidal, S. Bloomberg, S. T. Loughran, B.
401 O'Connor, C. O'Fagain, W. M. Nauseef and M. J. O'Connell (2012). "Functional consequence of
402 positive selection revealed through rational mutagenesis of human myeloperoxidase." Mol Biol
403 Evol **29**(8): 2039-2046.
404
405 Loytynoja, A. and N. Goldman (2005). "An algorithm for progressive multiple alignment of
406 sequences with insertions." Proc Natl Acad Sci U S A **102**(30): 10557-10562.
407
408 Moretti, S., B. Laurenczy, W. H. Gharib, B. Castella, A. Kuzniar, H. Schabauer, R. A. Studer, M.
409 Valle, N. Salamin, H. Stockinger and M. Robinson-Rechavi (2014). "Selectome update: quality
410 control and computational improvements to a database of positive selection." Nucleic Acids Res
411 **42**(Database issue): D917-921.
412
413 Muller, J., C. J. Creevey, J. D. Thompson, D. Arendt and P. Bork (2010). "AQUA: automated
414 quality improvement for multiple sequence alignments." Bioinformatics **26**(2): 263-265.
415
416 Pond, S. L., S. D. Frost and S. V. Muse (2005). "HyPhy: hypothesis testing using phylogenies."
417 Bioinformatics **21**(5): 676-679.
418
419 Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under
420 mixed models." Bioinformatics **19**(12): 1572-1574.
421
422 Swanson, W. J., Z. Yang, M. F. Wolfner and C. F. Aquadro (2001). "Positive Darwinian
423 selection drives the evolution of several female reproductive proteins in mammals." Proc Natl
424 Acad Sci U S A **98**(5): 2509-2514.
425
426 Webb, A. E., Z. N. Gerek, C. C. Morgan, T. A. Walsh, C. E. Loscher, S. V. Edwards and M. J.
427 O'Connell (2015). "Adaptive Evolution as a Predictor of Species-Specific Innate Immune
428 Response." Mol Biol Evol **32**(7): 1717-1729.
429
430 Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Mol Biol Evol
431 **24**(8): 1586-1591.
432
433 Yang, Z. and M. dos Reis (2011). "Statistical properties of the branch-site test of positive
434 selection." Mol Biol Evol **28**(3): 1217-1228.
435
436 Yates, A., W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P.
437 Clapham, S. Fitzgerald, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek,
438 N. Johnson, T. Juettemann, S. Keenan, I. Lavidas, F. J. Martin, T. Maurel, W. McLaren, D. N.
439 Murphy, R. Nag, M. Nuhn, A. Parker, M. Patricio, M. Pignatelli, M. Rahtz, H. S. Riat, D.
440 Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, E. Birney, J. Harrow, M.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

441 Muffato, E. Perry, M. Ruffier, G. Spudich, S. J. Trevanion, F. Cunningham, B. L. Aken, D. R.
442 Zerbino and P. Flicek (2016). "Ensembl 2016." Nucleic Acids Res **44**(D1): D710-716.

443

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1895v2 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017