

The Popgen Pipeline Platform: A Software Platform for Facilitating Population Genomic Analyses

Andrew Webb^{*1}, Jared Knoblauch^{†1}, Nitesh Sabankar^{‡2}, Apeksha
Sukesh Kallur^{§2}, Jody Hey^{¶1} and Arun Sethuraman^{||1,2}

¹Center for Computational Genetics and Genomics, Temple
University

²Department of Biological Sciences, California State University
San Marcos

September 27, 2019

*tug41380@temple.edu
†jaredknoblauch@gmail.com
‡nsabankar@csusm.edu
§asukeshkall@csusm.edu
¶hey@temple.edu
||asethuraman@csusm.edu

Abstract

Here we present the Pop-Gen Pipeline Platform (PPP), a software platform with the goal of reducing the computational expertise required for conducting population genomic analyses. The PPP was designed as a collection of scripts that facilitate common population genomic workflows in a consistent and standardized **Python** environment. Functions were developed to encompass entire workflows, including: input preparation, file format conversion, various population genomic analyses, output generation, and visualization. By facilitating entire workflows, the PPP offers several benefits to prospective end users - it reduces the need of redundant in-house software and scripts that would require development time and may be error-prone, or incorrect. The platform has also been developed with reproducibility and extensibility of analyses in mind. The PPP is an open-source package that is available for download and use at https://ppp.readthedocs.io/en/latest/PPP_pages/install.html

25 Introduction

26 Since the advent of genomics, population genetics has quickly become domi-
 27 nated by complex statistical and computational methodologies [1, 2]. An un-
 28 fortunate consequence of this fact is that many investigators lack the necessary
 29 resources - computational, and time - to independently implement many of these
 30 methodologies. This inevitably requires investigators to select from a plethora
 31 of software (i.e. analytical tools) that have been developed by other researchers.
 32 While this is not inherently a problem, and a common practice among many
 33 professions, it is not without its own difficulties. Investigators frequently face
 34 bespoke input and output formats that may not be accompanied by an intuitive
 35 and easy-to-use file-format conversion software, implementations that may be
 36 complex and open to misinterpretation, and lastly implementations incapable
 37 of large-scale analyses. These challenges are further amplified as few analyses re-
 38 quire a single tool, but rather require an analytical pipeline. Analytical pipelines
 39 typically incorporate a number of methodologies and software designed specifi-
 40 cally to connect those methodologies in a specific order.

41 The challenges posed by analytical pipelines have been partially mitigated by
 42 the development of software packages or "tool-kits" that provide tools for a
 43 variety of methodologies. However, while popular packages such as `vcftools`
 44 [3], `bcftools` [4], and `plink` [5] have proven invaluable to many investigators,
 45 they cannot be all-encompassing. The absence of such tool-kits often requires
 46 investigators, if able, to create pipelines that are frequently recreated, infre-
 47 quently published, time consuming to develop, and susceptible to error. For
 48 these reasons, analyses based on such pipelines are often difficult or impossible
 49 to completely replicate [6, 7], which is an issue of growing concern in research
 50 [8].

51 In an attempt to greatly alleviate these obstacles we have developed the Pop-
52 Gen Pipeline Platform (PPP). The PPP was designed to be a comprehensive
53 platform wherein investigators can conduct many of the analytical pipelines in-
54 volved in population genomics in a simple and standardized environment. We
55 achieved this goal by incorporating and connecting various tool-kits, standard
56 tools/methods, and common analytical practices. To demonstrate both the sim-
57 plicity and the comprehensive nature of the PPP, we designed and implemented
58 population genomic analyses of publicly available data from chimpanzees [9]
59 using only the PPP.

60 **New Approach**

61 **Design**

62 The PPP was written in the `Python` programming language and designed to
63 operate using either `Python` versions 2 or 3. `Python` was selected primarily to
64 reduce the complexity of future development, take advantage of various relevant
65 and powerful `Python` libraries, and to minimize compatibility issues for prospec-
66 tive users. The PPP was designed as a collection of modular functions that may
67 be combined to offer a wide variety of analyses and pipelines required by pop-
68 ulation geneticists. The core functions of the PPP - i.e. functions commonly
69 used among analyses - were designed to operate using VCF-based file formats
70 [3]. This decision was due to the predominance of the VCF file format within
71 the population genomics community, specifically the frequent support for this
72 format among tools, and the likelihood of most publicly available datasets being
73 made available as VCF formatted files. Most hypothetical runs in the PPP will
74 begin with these core functions, and then branch off into the desired combina-
75 tion of analysis-specific functions. It should be stated that most analysis-specific

76 functions do not support VCF-based file formats, but rather incorporate a pre-
77 ceding file conversion core function to operate. This design was chosen to avoid
78 superfluous conversions, many of which are computationally intensive.

79 A fundamental aspect of the PPP's design is that if a specific technique (e.g.
80 tool, software package, statistic) is synonymous with an analysis, that technique
81 will be integrated into the function associated with the analysis. In some in-
82 stances we have integrated multiple techniques into a single function - e.g. we
83 have included both BEAGLE [10] and SHAPEIT [11] in our phasing function. As
84 prospective users may not be familiar with a technique, relevant information
85 and links to the original material may be found within the documentation and
86 appropriate references will be provided upon use of a technique.

87 The PPP was also designed to include other features to further simplify and
88 expedite analyses. For instance, the PPP integrates a versatile configuration
89 system that allows prospective users to configure functions in two ways: with
90 optional command-line arguments; or with optional arguments specified within
91 a configuration file. By using a configuration file it is possible for prospective
92 users to configure an entire analysis or pipeline. This is possible due to the stan-
93 dardized argument scheme designed for the PPP which allows the assignment
94 of global arguments - i.e. consistent among the entire platform - and function-
95 specific arguments - such as the explicit input and output for each function.

96 Another feature of the PPP is the use of the Model file format that we devel-
97 oped for use in the platform. The Model file is a JSON-based format that is able
98 to store multiple population models, including the relevant details of each model
99 (i.e. populations, individuals, population tree, and other relevant meta-data).

100 A primary benefit of the Model file is the ability to automatically assign infor-
101 mation from the specified model to functions, such as the populations and their
102 associated individuals. The file also simplifies record keeping as it becomes the

103 repository for model-related information.

104 Overview

105 A consequence of the design of the PPP is that a hypothetical analysis could
 106 use a combination of functions that do not demonstrate the comprehensive
 107 nature of the platform - see Figure 1. for an illustration of the initial release
 108 of the PPP. Therefore, to give a sufficient overview of the PPP, we have chosen
 109 to describe the functions required in the Isolation with Migration (IM) [12]
 110 pipeline we used for analyzing population genomic data from chimpanzees [9].
 111 As the demographic history of the chimpanzees have been extensively studied
 112 [13, 14, 15, 16], we selected two closely related populations - Central chimpanzees
 113 (*Pan troglodytes troglodytes*) and Western chimpanzees (*Pan troglodytes verus*)
 114 - to demonstrate the effectiveness of the PPP in comparison to similar analyses.
 115 In particular, we wished to explore the divergence of the two populations by
 116 estimating their population sizes, migration rates, and divergence time using
 117 multi-locus genomic data under an IM model.

118 The first procedure in our analysis pipeline was applying filters to remove sites
 119 with missing data and non-biallelic sites. The removal of non-biallelic sites (i.e.
 120 multiallelic sites) is of particular importance as they violate the Infinite Sites
 121 (IS) model [17] assumption of a single polymorphism per site assumed by the IM
 122 model [12] implemented in our analysis pipeline. It also bears mentioning that
 123 additional downstream procedures are also required to avoid other violations of
 124 model assumptions, and will be reported where relevant. The filter procedure
 125 of our analysis was completed using the PPP's **VCF-filter** function. **VCF-**
 126 **filter** was designed to perform filtering operations on VCF-based files and is
 127 expected to be the first function in most analyses. Prospective users are able
 128 to select from a comprehensive collection of filters that are assigned alongside a

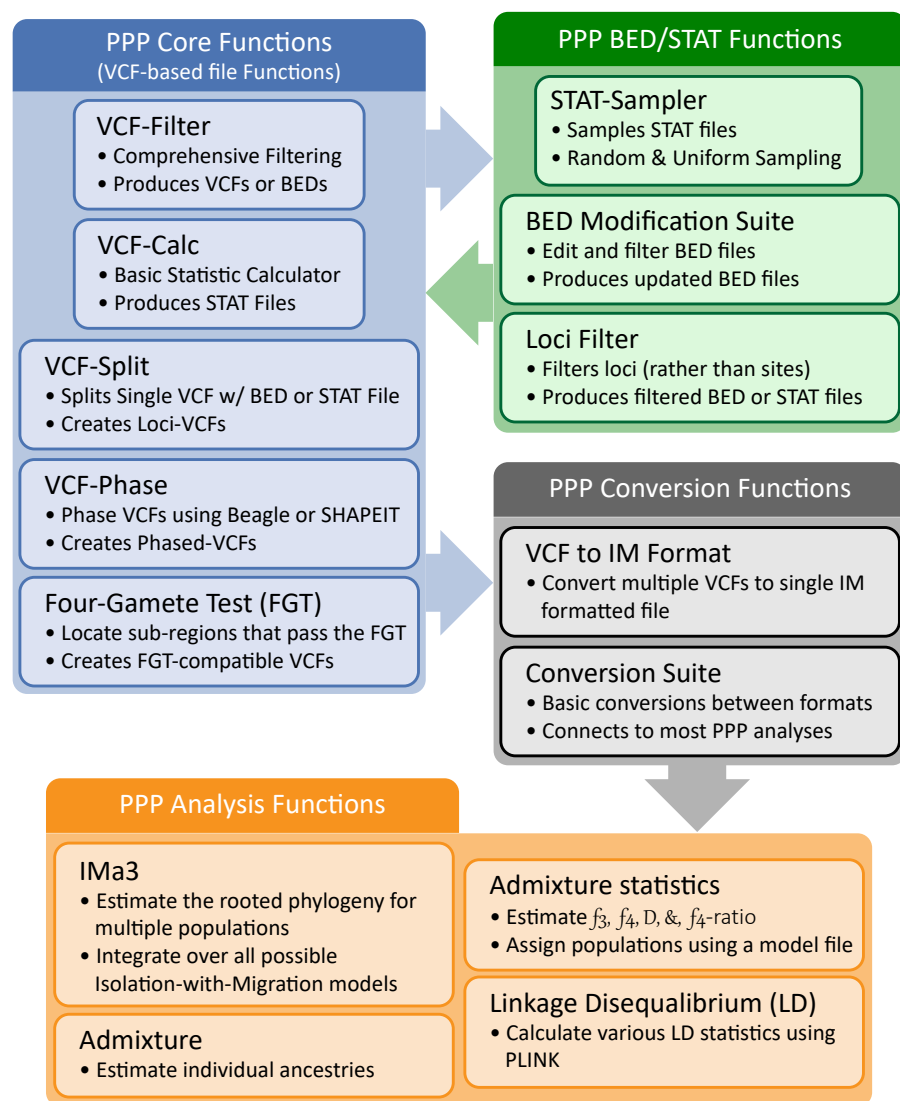


Figure 1: Structure of the PPP. PPP functions are grouped into four categories: i) the Core PPP functions that operate on VCF files; ii) the optional BED and STAT functions which may be used to sample, filter, and/or edit BED or STAT files; iii) the conversion functions which are required to convert from VCF to analysis-specific file formats; and iv) the analysis functions which are used to automate their respective analyses.

129 specified value (i.e. a threshold) or a specified file (e.g. a BED file containing
130 genomic coordinates of sites to remove, or include) (Table 1). Depending on the
131 needs of the prospective user, the function is capable of returning either a BED
132 file of the filtered sites, or an updated VCF-based file. If a model is specified
133 from a Model file, the function is designed to automatically remove non-relevant
134 samples before filters are applied.

135 Our analysis pipeline then proceeded to randomly sample 200, 1 kbps, non-
136 overlapping genomic loci by their respective F_{st} values between the two pop-
137 ulations. Due to subsequent analytical requirements, and the assumption of
138 putatively neutral sites [12], each sampled locus was required to be both in-
139 formative - i.e. in possession of at least four variant positions - and inter-
140 genic. This procedure, **VCF-calc** to calculate the F_{st} values within 1 kbps loci,
141 **informative-filter** to remove loci that were not informative, and **stat-sampler**
142 to pseudo-randomly sample 200 loci. **VCF-calc** was designed to calculate many
143 of the basic statistics used in population genetic analyses on VCF-based files
144 (Table 1). For most statistics, little to no configuration is necessary, however,
145 some statistics do require additional parameters (e.g. window length, window
146 step length) to operate. This function is designed to return a tabular statis-
147 tic output file that is usable by other functions within the PPP. If a model is
148 specified from a Model file, the function is designed to automatically assign the
149 relevant populations and/or individuals to compute the appropriate statistics.
150 The **informative-filter** function was designed to apply various locus-based fil-
151 ters often required by population genomic analyses on VCF-based files (Table
152 1). In comparison to **VCF-filter**, these filters evaluate and filter each locus
153 as a single entity. To operate, the majority of filters only require a BED or
154 statistic file to define the loci of interest. Filters were also designed to be eas-
155 ily configurable by altering default values or by enabling optional parameters.

informative-filter is designed to return a filtered copy of the original BED or statistic file. **stat-sampler** was designed to pseudo-randomly sample loci from statistic files produced by the **VCF-calc** function. Prospective users may select from one of two pseudo-random sampling schemes: a random scheme that samples loci from the entire file, and a uniform sampling scheme that samples loci from equally sized bins derived from the statistic of choice. **stat-sampler** may also be configured to alter both sampling schemes - e.g. samples to select and number of bins - and to reproduce previous results, if desired. The function is designed to return a sampled version of the statistic file as output.

The next procedure in our analysis pipeline was the creation of phased VCF-based files for each of the sampled loci. Phased chromosomes are required for our pipeline to identify potential recombination events by the Four-gamete Test [18]. It should be noted that phasing was possible prior to the creation of individual VCF-based files for each sampled locus, but is computationally demanding. Our procedure required the use of the **VCF-split** function to generate locus-specific VCF-based files and **VCF-phaser** function to phase the files. The **VCF-split** function was designed to split a single VCF-based file using either a BED or statistic file to define the coordinates for the loci of interest. If a model is specified from a Model file, the function is designed to only return the relevant individuals in the loci VCF-based files. **VCF-phaser** was designed to phase VCF-based files using either **SHAPEIT** [11] or **BEAGLE** [10]. Phasing with **VCF-phaser** only requires prospective users to specify a VCF-based file - which by default uses **SHAPEIT** [10]. However, **VCF-phaser** may be configured to instead phase VCF-based files with **BEAGLE** [10] or configure the settings of either algorithm. If a model is specified from a Model file, the function is designed to only phase and return the relevant individuals.

Our pipeline next required the identification of sub-regions of each locus without

183 recombination within our phased VCF-based files. This procedure was neces-
 184 sary to avoid violating the assumption of no recombination within loci of the
 185 IM model [19]. This was accomplished using the **Four-gamete Test** function
 186 of the PPP, which was designed to check for the presence of recombination
 187 events between pairs of segregating sites [18]. The PPP's implementation of
 188 the **Four-gamete Test** takes a VCF-based file of a kilobase-scale region in
 189 a chromosome, then finds sub-regions of the loci that have less than four ga-
 190 metes among them. Prospective users may configure the **Four-gamete Test**
 191 to: require a specific number of informative sites; return either a single or all
 192 compatible sub-regions; ignore multiallelic sites; and include sites with missing
 193 data. By default, the function is designed output a VCF file of a sub-region
 194 with at least two informative sites that passed the test.

195 The last procedure in our pipeline was performing an IM analysis using IMA3
 196 [16]. However, before we were able to proceed to the IM analysis of our pipeline
 197 we were required to convert the sub-region VCF-based files into a single IM
 198 formatted file that is compatible with our implementation of IMA3 [16]. This
 199 procedure was accomplished using the **vcf-to-ima** conversion function of the
 200 PPP. **vcf-to-ima** was designed to automatically generate an IM formatted file
 201 from a collection of sub-region VCF-based files, a model specified from a Model
 202 file, and additional parameters provided by the prospective user. This design
 203 allows for IM formatted files to be easily configured by specifying a different
 204 Model or altering parameters. Once the conversion process was finished we
 205 used the PPP function **ima3-wrapper** to perform all IM analyses. **ima3-**
 206 **wrapper** handles the passing of input parameters to IMA3, while also handling
 207 multi-threading in the subprocess calls if the user specifies. Most required input
 208 is specified in the IM input file, with additional options required to specify
 209 upper limits, priors for parameters to be estimated, and determine how long to

210 burn-in, and genealogy sampling run-time of the MCMC should be. The final
211 output is a file with estimates of population model parameters (migration rates,
212 population sizes, and divergence times), with confidence intervals around these
213 estimates.

214 Finally, while our pipeline focused on performing an IM analysis, the PPP was
215 designed to easily allow the implementation of additional analyses, if desired.
216 For example, we could use many of the files produced in our IM analysis to
217 estimate population structure using ADMIXTURE [20], test for introgression
218 using AdmixTools [21], or linkage disequilibrium using PLINK [5].

219 Results

220 To demonstrate the capabilities of the PPP we compared an Isolation with Mi-
221 gration analysis of two chimpanzees populations to previous reports [13, 14]. We
222 found our estimates of the divergence time, the ancestral chimpanzee population
223 size, migration rates, and the populations sizes of the extant chimpanzee pop-
224 ulations - central chimpanzees (*Pan troglodytes troglodytes*) and western chim-
225 panzees (*Pan troglodytes verus*) to be consistent with previous findings (Table
226 2).

227 Discussion

228 The primary goal behind the development of the PPP was to create a simple,
229 standardized, and robust platform for population genetic analyses. Ideally, an
230 end user would only require a specific combination of PPP functions to im-
231 plement their desired pipeline. To demonstrate this capability, we examined
232 the demographic history of two closely related chimpanzee population and com-
233 pared the results to previous findings [14, 15, 16]. We found that the PPP

234 greatly reduced the overall complexity of our analysis and was able to suc-
 235 cessfully reproduce previous findings. With the exception of downloading the
 236 necessary files (e.g. chimpanzee VCF input, BED files containing gene coordi-
 237 nates) all operations were completed using PPP functions alone. Assembling
 238 the pipeline was a straightforward process as the majority of functions could be
 239 invoked in tandem without requiring intermediate processing steps. We were
 240 also able to quickly process the VCF input for our IM analysis as the majority
 241 of PPP functions required less than 5 minutes to operate, with the exception
 242 being the initial filtering procedure which took roughly 50 minutes and the IM
 243 analysis which required approximately 400 hours of CPU time. We also found
 244 that repeating our analysis - either to explore the results of different parameters,
 245 reproduce our findings, or remedy errors - was a simple process and could be
 246 done rapidly if the initial filtering was not repeated. Taken together, the PPP
 247 has achieved its primary design goal, but that does not signify the platform is
 248 complete. Additionally, this sample pipeline, along with other examples have
 249 been published as Jupyter Notebooks on the PPP's development website.

250 Future development of the PPP will primarily be focused on improvements to
 251 the platform. First and foremost is the creation of a Galaxy Project [22] wrap-
 252 per to expand the user base of the platform, primarily to assist users more
 253 familiar with a graphical user interface and/or web applications. As the PPP
 254 was developed in consideration of an eventual Galaxy wrapper, implementing
 255 this improvement will be straightforward. We also intend to have ongoing re-
 256 leases of additional population genetic analyses for the platform. The modular
 257 structure of the platform should allow for the majority of these updates to only
 258 require creation of the function to automate the analysis and potentially updat-
 259 ing the file conversion suite. Future releases will also focus on improvements to
 260 the overall speed (and efficiency) of the platform. One potential improvement

currently being explored is the incorporation of Cython, which aims to achieve C-like performance among python scripts [23]. We also plan on exploring the possibility of using Jupyter notebooks [24] to store and share analysis pipelines. Jupyter notebooks are a simple and ideal format for analysis pipelines as they allow computer code - e.g. a PPP function - to be accompanied by textual elements, such as descriptions of each function and their overall purpose.

Acknowledgments

This work was supported by an NSF ABI Grant 1564659 to AS and JH. This research includes calculations carried out on Temple University's HPC resources and thus was supported in part by the National Science Foundation through major research instrumentation grant number 1625061 and by the US Army Research Laboratory under contract number W911NF-16-2-0189.

References

- [1] Sònia Casillas and Antonio Barbadilla. Molecular population genetics, 2017.
- [2] B Charlesworth and D Charlesworth. Population genetics from 1966 to 2016. *Heredity*, 118:2, jul 2016.
- [3] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011.

- 282 [4] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils
283 Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Se-
284 quence Alignment/Map format and SAMtools. *Bioinformatics*, 2009.
- 285 [5] Christopher C. Chang, Carson C. Chow, Laurent C.A.M. Tellier,
286 Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. Second-
287 generation PLINK: Rising to the challenge of larger and richer datasets.
288 *GigaScience*, 2015.
- 289 [6] Jill P Mesirov. Accessible Reproducible Research. *Science*, 327(5964):415–
290 416, 2010.
- 291 [7] Gordon J. Lithgow, Monica Driscoll, and Patrick Phillips. A long journey
292 to reproducible results. *Nature*, 2017.
- 293 [8] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 2016.
- 294 [9] Javier Prado-Martinez, Peter H Sudmant, Jeffrey M Kidd, Heng Li,
295 Joanna L Kelley, Belen Lorente-Galdos, Krishna R Veeramah, August E
296 Woerner, Timothy D O’Connor, Gabriel Santpere, Alexander Cagan,
297 Christoph Theunert, Ferran Casals, Hafid Laayouni, Kasper Munch,
298 Asger Hobolth, Anders E Halager, Maika Malig, Jessica Hernandez-
299 Rodriguez, Irene Hernando-Herraez, Kay Prüfer, Marc Pybus, Laurel John-
300 stone, Michael Lachmann, Can Alkan, Dorina Twigg, Natalia Petit, Carl
301 Baker, Fereydoun Hormozdiari, Marcos Fernandez-Callejo, Marc Dabad,
302 Michael L Wilson, Laurie Stevison, Cristina Camprubí, Tiago Carvalho,
303 Aurora Ruiz-Herrera, Laura Vives, Marta Mele, Teresa Abello, Ivanela
304 Kondova, Ronald E Bontrop, Anne Pusey, Felix Lankester, John A Kiyang,
305 Richard A Bergl, Elizabeth Lonsdorf, Simon Myers, Mario Ventura, Pas-
306 cal Gagneux, David Comas, Hans Siegmund, Julie Blanc, Lidia Agueda-
307 Calpena, Marta Gut, Lucinda Fulton, Sarah A Tishkoff, James C Mullikin,

- 308 Richard K Wilson, Ivo G Gut, Mary Katherine Gonder, Oliver A Ryder,
309 Beatrice H Hahn, Arcadi Navarro, Joshua M Akey, Jaume Bertranpetit,
310 David Reich, Thomas Mailund, Mikkel H Schierup, Christina Hvilsom,
311 Aida M Andrés, Jeffrey D Wall, Carlos D Bustamante, Michael F Hammer,
312 Evan E Eichler, and Tomas Marques-Bonet. Great ape genetic diversity
313 and population history. *Nature*, 499:471, jul 2013.
- 314 [10] Sharon R Browning and Brian L Browning. Rapid and Accurate Haplo-
315 type Phasing and Missing-Data Inference for Whole-Genome Association
316 Studies By Use of Localized Haplotype Clustering. *The American Journal*
317 *of Human Genetics*, 81(5):1084–1097, 2007.
- 318 [11] Jared O’Connell, Deepti Gurdasani, Olivier Delaneau, Nicola Pirastu,
319 Sheila Ulivi, Massimiliano Cocca, Michela Traglia, Jie Huang, Jennifer E
320 Huffman, Igor Rudan, Ruth McQuillan, Ross M Fraser, Harry Campbell,
321 Ozren Polasek, Gershim Asiki, Kenneth Ekoru, Caroline Hayward, Alan F
322 Wright, Veronique Vitart, Pau Navarro, Jean-Francois Zagury, James F
323 Wilson, Daniela Toniolo, Paolo Gasparini, Nicole Soranzo, Manjinder S
324 Sandhu, and Jonathan Marchini. A General Approach for Haplotype Phas-
325 ing across the Full Spectrum of Relatedness. *PLOS Genetics*, 10(4):1–21,
326 2014.
- 327 [12] Jody Hey and Rasmus Nielsen. Integration within the Felsenstein equation
328 for improved Markov chain Monte Carlo methods in population genetics.
329 *Proceedings of the National Academy of Sciences*, 104(8):2785–2790, 2007.
- 330 [13] Yong-Jin Won and Jody Hey. Divergence population genetics of chim-
331 panzees. *Molecular biology and evolution*, 22(2):297–307, feb 2005.
- 332 [14] Arun Sethuraman and Jody Hey. IMA2p—parallel MCMC and inference
333 of ancient demography under the Isolation with migration (IM) model.

- 334 *Molecular ecology resources*, 16(1):206–215, jan 2016.
- 335 [15] Yujin Chung and Jody Hey. Bayesian analysis of evolutionary divergence
336 with genomic data under diverse demographic models. *Molecular Biology*
337 *and Evolution*, 34(6):1517–1528, 2017.
- 338 [16] Jody Hey, Yujin Chung, Arun Sethuraman, Sarah Tishkoff, Joseph
339 Lachance, Vitor C Sousa, and Yong Wang. Phylogeny Estimation by In-
340 tegration over Isolation with Migration Models. *Molecular Biology and*
341 *Evolution*, 35(11):2805–2818, 2018.
- 342 [17] M Kimura. The number of heterozygous nucleotide sites maintained in a
343 finite population due to steady flux of mutations. *Genetics*, 61(4):893–903,
344 apr 1969.
- 345 [18] R R Hudson and N L Kaplan. Statistical properties of the number of re-
346 combination events in the history of a sample of DNA sequences. *Genetics*,
347 111(1):147–164, sep 1985.
- 348 [19] Jody Hey and Rasmus Nielsen. Multilocus Methods for Estimating Popu-
349 lation Sizes, Migration Rates and Divergence Time, With Applications to
350 the Divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*,
351 167(2):747–760, 2004.
- 352 [20] David H Alexander, John Novembre, and Kenneth Lange. Fast model-
353 based estimation of ancestry in unrelated individuals. *Genome research*,
354 19(9):1655–1664, sep 2009.
- 355 [21] Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Roh-
356 land, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich.
357 Ancient Admixture in Human History. *Genetics*, 192(3):1065–1093, 2012.

- bioRxiv preprint doi: <https://doi.org/10.1101/785774>; this version posted September 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Function	Purpose	Capabilities
VCF-filter	Filtering Variants	<i>Include/exclude variants sites by:</i> genomic position, missing data count and percentage, allele count, MAF, MAC, presence of indels, SNP IDs, associated with a specific flag (i.e. PASS)
VCF-calc	Statistic Calculator	<i>Calculate the following statistics:</i> F_{st} (site- and window-based), Tajima's D , Nucleotide Diversity (site- and window-based), allele frequency, inbreeding coefficients (F_{IT} and F_{IS}), tests of Hardy-Weinberg Equilibrium
informative-loci-filter	Filtering Loci	<i>Include/exclude loci by:</i> informative site count, variant site count, missing data count, ignoring indels, ignoring multiallelic variants, ignoring CpG sites

Table 1: Capabilities of the PPP Filters and Statistic Calculator.

Parameter	Mean	Highest Posterior
q0	1.219	1.204
q1	0.3469	0.3400
q2	0.7531	0.7640
$m0 \rightarrow m1$	0.5860	0.5675
$m1 \rightarrow m0$	0.8330	0.7925
t	0.4155	0.4494

Table 2: Evolutionary history of Central and Western Chimpanzees, estimated using PPP and IMA3. The mean and highest posterior parameter estimated population sizes (q), migration rates (m), and divergence time (t) between *P. t. troglodytes* (population 0) and *P. t. verus* (population 1).