



HAL
open science

PopSize, a snpArcher module for population size change inference

Thomas Forest, Swan Portalier, Camille Steux, Timothy B Sackton,
Guillaume Achaz

► To cite this version:

Thomas Forest, Swan Portalier, Camille Steux, Timothy B Sackton, Guillaume Achaz. PopSize, a snpArcher module for population size change inference. 2024. hal-04599797

HAL Id: hal-04599797

<https://hal.science/hal-04599797v1>

Preprint submitted on 4 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

PopSize, a snpArcher module for population size change inference

Thomas Forest^{1,2,3}, Swan Portalier⁶, Camille Steux⁷, Timothy B. Sackton⁴, and Guillaume Achaz^{3,5}

¹Éco-anthropologie CNRS (UMR 7206), Muséum National d'Histoire Naturelle, Musée de l'Homme, 75006, Paris, France

²Institut de Systématique Evolution Biodiversité (UMR 7205), Muséum national d'Histoire naturelle CNRS SU EPHE UA, CP 51, 55 rue Buffon, 75005, Paris, France

³Center for Interdisciplinary Research in Biology (CIRB) UMR 7241, Collège de France, 11 place Marcelin Berthelot, 75005, Paris, France

⁴Informatics Group, Harvard University, Cambridge, 02138, MA, USA

⁵Université Paris-Cité, 85 boulevard Saint-Germain, 75006, Paris, France

⁶Institute for Plant Sciences, Plant Ecological Genetics, University of Cologne, Zùlpicher StraÙe 47b, Cologne, North Rhine-Westphalia, Germany

⁷Centre de Recherche sur la Biodiversité et l'Environnement (CRBE), UMR 5300, Université de Toulouse, CNRS, IRD, Toulouse INP, Université Toulouse 3 – Paul Sabatier (UT3), 118, route de Narbonne, 31062, France

Motivation: Simplify the use of common demographic inference tools for a cluster infrastructure making it more scalable and robust for large datasets, taking advantage of a pre-built pipeline for non-model organisms called snpArcher.

Results: Runs up to 5 of the most cited demographic inference tools in parallel, using input directly from the snpArcher pipeline, without the need to add additional material.

Availability and implementation: Popsiz is a Snakemake module of the snpArcher pipeline and is available on GitHub (<https://github.com/tforest/popsiz>). This software is freely available under the same conditions as the main pipeline snpArcher.

Population genetics, demography inference, pipeline

Correspondence: thomas.forest1@edu.mnhn.fr

Introduction

In the context of the sixth mass extinction crisis, it is crucial to focus on species lacking conservation status, especially those absent from resources like the IUCN Red List of threatened species. Indeed, these kind of resources are not exhaustive and display a taxonomic bias (Cowie et al., 2022). As genomic data becomes increasingly available, including for many species that lack official conservation status, there is an opportunity to make use of this already available information. Multiple population genetic approaches have been developed to make use of these data. Among them are efficient methods attempting to monitor effective population size changes using genomic data, which is especially relevant in this context of mass extinction in poorly known taxa.

Demographic inference methods are commonly employed for simulating and understanding demographic scenarios, such as migration and population size fluctuations (Gutenkunst et al., 2009; Li and Durbin, 2011). In order to make these studies robust and reliable, a lot of work has to be done in precursor steps like sampling, sequencing, and

extracting information from the sequences. Hence, these analysis will benefit from high quality reference genomes and from adapted computational methods, to extract genomic information in accordance with observed levels of genomic diversity. A large number of these inference methods exist, and it is common to wish to run multiple inference methods in the context of a resequencing study.

To facilitate optimal use of a variety of demographic inference methods, we present here a Snakemake pipeline, which will call PopSize, optimized to multiple tools in parallel, including methods that focus on descriptive statistics such as Stairwayplot2 (Liu and Fu, 2020) and $\partial a \partial i$ (Gutenkunst et al., 2009) which use the Site Frequency Spectrum (SFS), and methods relying on the rate of heterozygous sites, like PSMC (Li and Durbin, 2011) or MSMC2 (Schiffels and Wang, 2020).

This tool is optimized to complement snpArcher (Mirchandani et al., 2023), a recently published Snakemake pipeline for variant calling in non-model organisms. This approach aims to enhance user experience in using conservation genomics approaches by providing a comprehensive framework for demographic analysis.

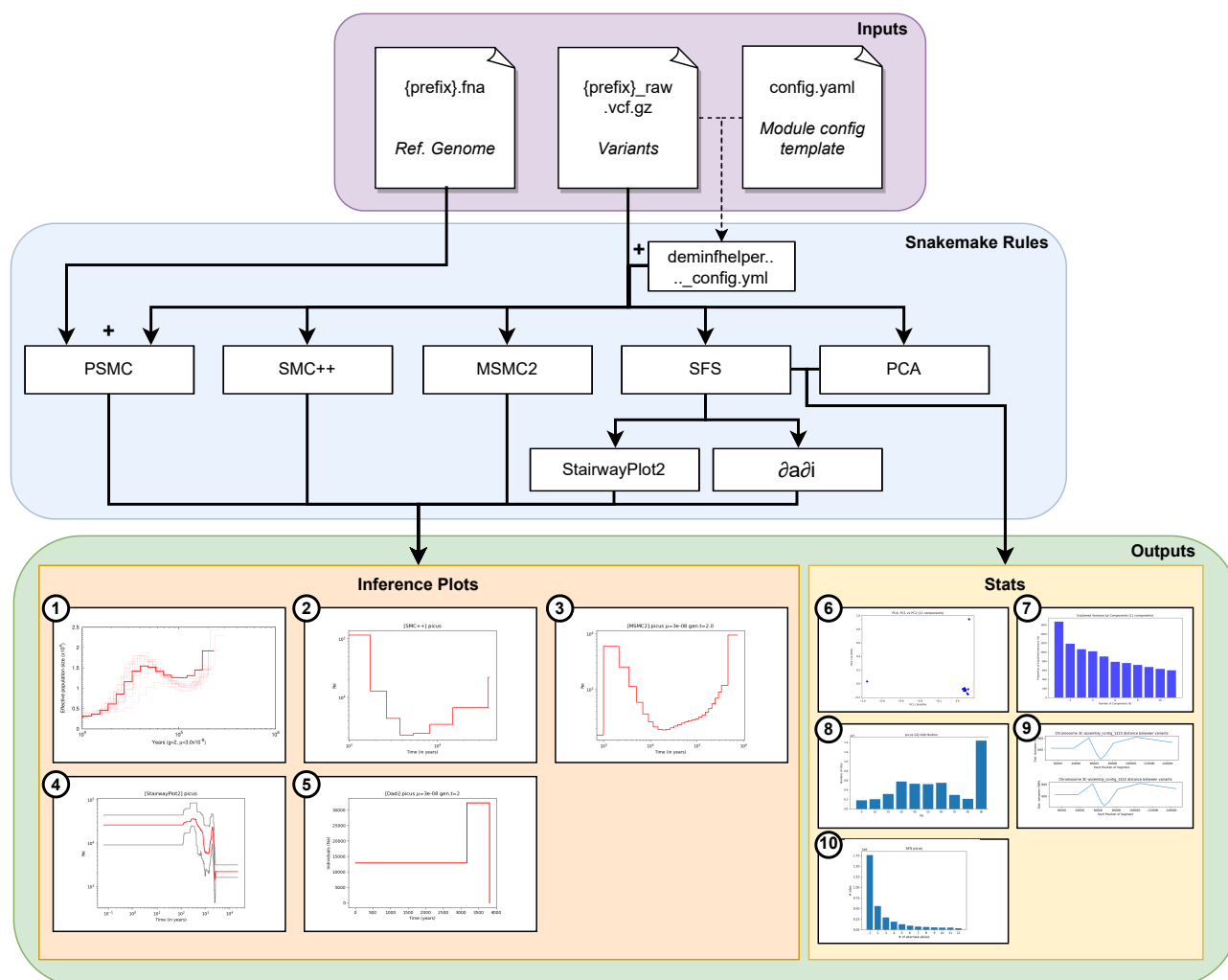


Figure 1. Workflow of the PopSize module. The module general process is divided in three main parts corresponding to the inputs, the different Snakemake rules that can be executed and the outputs. Arrows represent a dependency relation between parts of the workflow. Tasks requiring the availability of multiple dependencies at the same time are represented with a "+". Outputs 1 to 5 are the different inference plots of PSMC(1), SMC++(2), MSMC2(3), StairwayPlot2(4) and *daði*(5); whereas plots 6 to 10 correspond to the genomic PCA(6), the proportion of explained variance per component of the PCA(7), the distribution of Genotyping Quality (8), the distance between two variants along each chromosome representing genotyping coverage(9) and the Site Frequency Spectrum(10).

The module general process is divided in three main parts corresponding to the inputs, the different Snakemake rules that can be executed and the outputs. Arrows represent a dependency relation between parts of the workflow. Tasks requiring the availability of multiple dependencies at the same time are represented with a "+". The displayed outputs were obtained for the European Green Woodpecker *Picus viridis*. Outputs 1 to 5 are the different inference plots of PSMC(1), SMC++(2), MSMC2(3), StairwayPlot2(4) and *daði*(5). Red curves (plot 1-5) correspond to the effective population size; grey curves on plot 4 correspond to a [2.5 ; 97.5]% confidence interval. Outputs 6 to 10 correspond to the genomic PCA(6), the proportion of explained variance per component of the PCA(7), the distribution of Genotyping Quality (8), the distance between two variants along each chromosome representing genotyping coverage(9) and the Site Frequency Spectrum(10).

73 The objective of the PopSize module is to run these 78 the parameters. Moreover, this module returns statistics
 74 tools in a straightforward manner, without altering their 79 to assist the user in determining the relevance of the
 75 internal methods, and their outputs. It allows for some 80 obtained results. We stress that, though the programs
 76 flexibility in configuration, requiring to set the initial 81 implemented in our pipeline can infer changes in
 77 parameters accordingly, while providing transparency in 82 effective population size, interpreting the output of such

83 curves may not be straightforward. Indeed, it is known 137
84 that these methods can be sensitive to sequencing 138
85 quality, for instance, or to the presence of structure 139
86 in the sampled population. Our module thus provides 140
87 among others an estimation of the Genotyping Quality 141
88 (GQ) distribution of the sites selected by the pipeline, 142
89 as well as a genomic Principal Component Analysis 143
90 (PCA), which is often used as a simple representation 144
91 for visualizing population structure (van Waaij et al., 145
92 2023). The module's parametrisation also involves a 146
93 minimalist configuration file for defining the parameters 147
94 of the various tools used. This pipeline is based 148
95 on Snakemake, which allows to handle parallelisation, 149
96 dependencies installation using Conda, and resources 150
97 allocation, based on the limits set by the user. 151

98 Input files

99 The main input is a Variant Call Format (VCF) file 154
100 directly generated by snpArcher or provided by the user 155
101 from another tool. Our pipeline utilizes information 156
102 from this VCF file in all methods. By default, the 157
103 user does not have to provide external data if the 158
104 snpArcher pipeline was used, as it handles steps for 159
105 genotyping, filtering and producing high quality variants 160
106 in VCF format. Some tools like PSMC require a 161
107 reference genome to produce their outputs, but this 162
108 file is also present by default in the results directory 163
109 of snpArcher. In most cases, the configuration file for 164
110 the module is generated automatically, based on the 165
111 template provided. However, it may be updated by the 166
112 user depending on the results obtained after a first run. 167

113 Process

114 The process (Figure 1) begins with the parsing of 168
115 the VCF file to extract the genetic data, and of the 169
116 reference genome fasta file when PSMC is used 170
117 (*Step 1*). Subsequently, we construct the Site 171
118 Frequency Spectrum (SFS) for tools such as *daði* and 172
119 StairwayPlot2. This stage involves additional filtering 173
120 of variants present in regions showing low density of 174
121 SNPs. In parallel, the pipeline executes dedicated 175
122 Snakemake rules for PSMC, SMC++ and MSMC2, 176
123 which does not use the generated SFS (*Step 2*). Each 177
124 tool ends by generating an inference plot illustrating 178
125 temporal effective population size variations (Output 179
126 Plots [1-5]). In parallel to these operations, we perform 180
127 statistical analyses on the VCF data, including Principal 181
128 Component Analysis (PCA) (Output Plots [6-10]) (*Step*
129 *3*).

130 Outputs

131 The outputs include detailed logs and command 182
132 traces for each tool used, ensuring transparency in 183
133 the parameters and methods applied. These logs 184
134 are critical for reproducibility and for understanding 185
135 nuances of the analysis. Additionally, we provide 186
136 a range of inference visualisations (Fig. 1, Output

Plots [1-6]). Original plots from tools with built-in
plotting capabilities are included directly. For tools
lacking such features, we generate plots using Python's
Matplotlib library, offering consistency in visual analysis
across different tools. Furthermore, the pipeline
features an interactive PCA plot, created using the
Plotly Python library, which displays k-means clustering
results. This interactive visualisation allows for an
intuitive exploration of the data. The corresponding
cluster assignments are also made available in a
`clusters.csv` file, providing a detailed breakdown
of the k-means clustering results. Moreover, the
module plots the proportion of explained variance per
component of the PCA, the distribution of Genotyping
Quality, the distance between two variants along each
chromosome representing genotyping coverage, and the
Site Frequency Spectrum (Fig. 1, Output Plots
[7-10]). These diverse outputs, combining detailed logs
with interactive and static visualisations, enhance the
interpretability and utility of the pipeline. All the outputs,
log files and temporary files are kept in the `/popsize`
folder of snpArcher's `/results` top-level folder.

Results

The pipeline was executed on a SLURM scheduler on
the genome of the European Green Woodpecker *Picus
viridis* (Forest et al., 2024). From this genome of 1.1Gb,
snpArcher called 8815631 variants from the 12 samples
of *Picus viridis* that were used. The resources available
for each job were set to a default value of 15 CPU
threads and 8GB of memory. In the configuration file,
the mutation rate and the generation time were set to
 $\mu = 5 \times 10^{-9}$ per site per generation and 5.6 years,
respectively.

The parsing of the VCF and the generation of all the
statistics took only five minutes. Followed by *daði* and
StairWayplot2 which finished in 30 minutes. The longest
task was performed by PSMC, which finished in 3 hours.
The process generated 3227 files, corresponding to
4.8Gb of output.

Acknowledgements

Most of the bioinformatic analyses were carried out
through the PCIA cluster (Plateforme de Calcul Intensif
et Algorithmique PCIA, Muséum national d'histoire
naturelle, Centre national de la recherche scientifique,
UAR 2700 2AD).

Funding

This work is part of a doctoral thesis funded by
Sorbonne University, through the IBES (Initiative
Biodiversity, Evolution, Ecology & Society) grant.

186 **Bibliography** 204

187 Cowie, R. H., Bouchet, P., and Fontaine, B. (2022). The sixth 206
188 mass extinction: fact, fiction or speculation? *Biological* 207
189 *Reviews*, 97(2):640–663. doi: 10.1111/brv.12816. 208

190 Forest, T., Achaz, G., Marbouty, M., Bignaud, A., Thierry, A., 209
191 Koszul, R., Milhes, M., Lledo, J., Pons, J.-M., and Fuchs, 210
192 J. (2024). Chromosome-level genome assembly of the 211
193 european green woodpecker *Picus viridis*. *G3: Genes*, 212
194 *Genomes, Genetics*, 14(5). doi: 10.1093/g3journal/ 213
195 jkae042. 214

196 Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and 215
197 Bustamante, C. D. (2009). Inferring the joint demographic 216
198 history of multiple populations from multidimensional SNP 217
199 frequency data. 5(10):e1000695. doi: 10.1371/journal. 218
200 pgen.1000695. Publisher: Public Library of Science. 219

201 Li, H. and Durbin, R. (2011). Inference of human population 220
202 history from individual whole-genome sequences. 475 221
203 (7357):493–496. doi: 10.1038/nature10231. Number: 222

7357 Publisher: Nature Publishing Group.

Liu, X. and Fu, Y.-X. (2020). Stairway plot 2: demographic 205
history inference with folded SNP frequency spectra. 21(1):
280. doi: 10.1186/s13059-020-02196-9.

Mirchandani, C. D., Shultz, A. J., Thomas, G. W. C., Smith,
S. J., Baylis, M., Arnold, B., Corbett-Detig, R., Enbody,
E., and Sackton, T. B. (2023). A fast, reproducible,
high-throughput variant calling workflow for population
genomics. page msad270. doi: 10.1093/molbev/
msad270.

Schiffels, S. and Wang, K. *MSMC and MSMC2:
The Multiple Sequentially Markovian Coalescent*, page
147–166. Springer US, (2020). ISBN 9781071601990.
doi: 10.1007/978-1-0716-0199-0_7. URL [http://dx.
doi.org/10.1007/978-1-0716-0199-0_7](http://dx.doi.org/10.1007/978-1-0716-0199-0_7).

van Waaij, J., Li, S., Garcia-Erill, G., Albrechtsen, A., and
Wiuf, C. (2023). Evaluation of population structure inferred
by principal component analysis or the admixture model.
GENETICS, 225(2). doi: 10.1093/genetics/iyad157.