PopSize, a snpArcher module for population size change inference

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

PopSize, a snpArcher module for population size change inference

Thomas Forest [1,2,3], **Swan Portalier** [6], **Camille Steux** [7], **Timothy B. Sackton** [4], **and Guillaume Achaz** [3,5]

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

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

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

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

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

[6]Institute for Plant Sciences, Plant Ecological Genetics, University of Cologne, Zülpicher Straße 47b, Cologne, North Rhine-Westphalia, Germany

[7]Centre 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: Popsize is a Snakemake module of the snpArcher pipeline and is available on GitHub (https://github.com/tforest/popsize). 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 inteference methods in the context of a resequencing study.

To facilitiate 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.

Software description

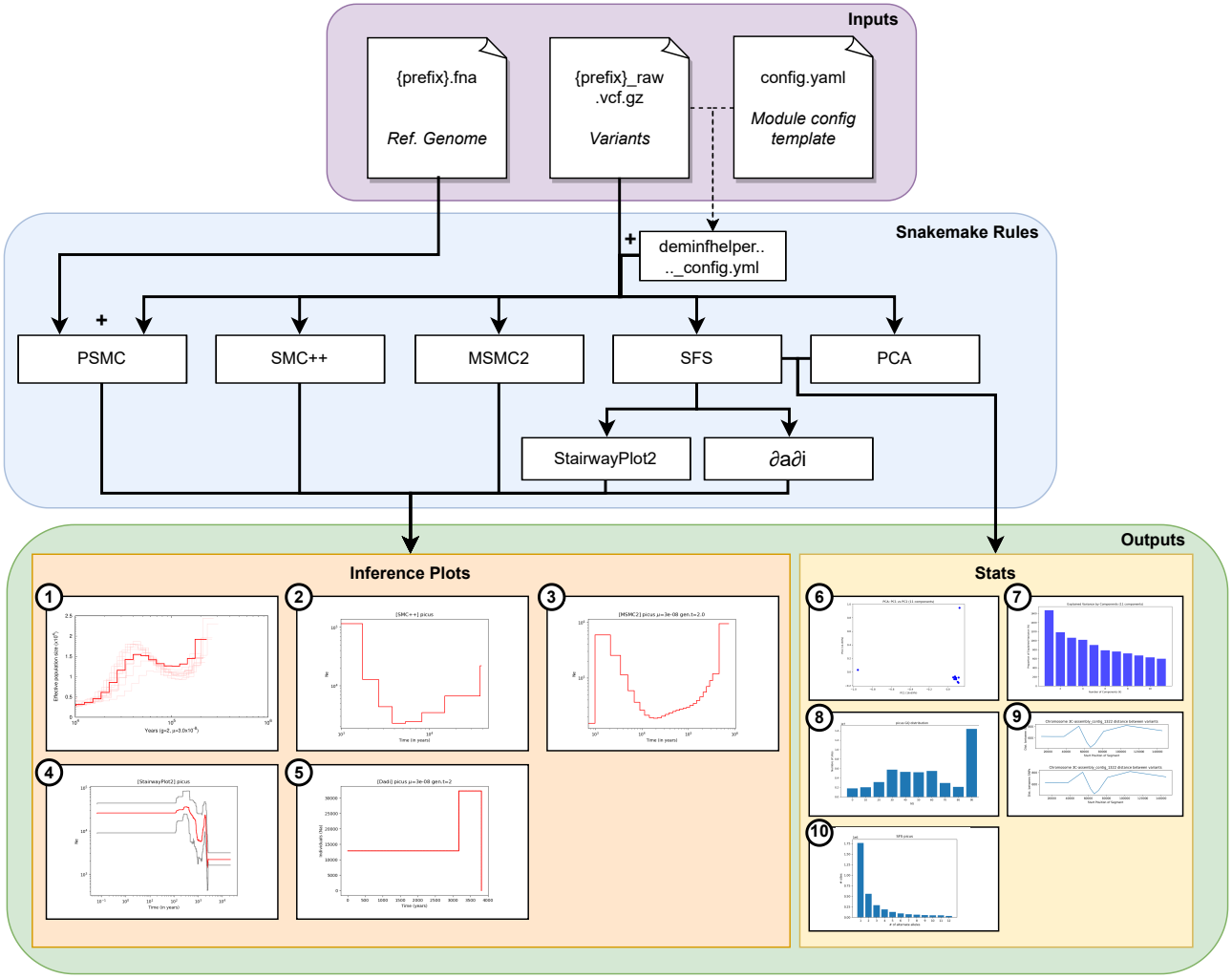

Figure 1. Workflow of the PopSize module. The module general process is divided in three main parts corresponding to the inputs, the different Snakemake rules that can be executed and the outputs. Arrows represent a dependency relation between parts of the workflow. Tasks requiring the availability of 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 $\partial a \partial i$(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 $\partial a \partial i$(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).

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

curves may not be straightforward. Indeed, it is known that these methods can be sensitive to sequencing quality, for instance, or to the presence of structure in the sampled population. Our module thus provides among others an estimation of the Genotyping Quality (GQ) distribution of the sites selected by the pipeline, as well as a genomic Principal Component Analysis (PCA), which is often used as a simple representation for visualizing population structure (van Waaij et al., 2023). The module's parametrisation also involves a minimalist configuration file for defining the parameters of the various tools used. This pipeline is based on Snakemake, which allows to handle parallelisation, dependencies installation using Conda, and resources allocation, based on the limits set by the user.

Input files

The main input is a Variant Call Format (VCF) file directly generated by snpArcher or provided by the user from another tool. Our pipeline utilizes information from this VCF file in all methods. By default, the user does not have to provide external data if the snpArcher pipeline was used, as it handles steps for genotyping, filtering and producing high quality variants in VCF format. Some tools like PSMC require a reference genome to produce their outputs, but this file is also present by default in the results directory of snpArcher. In most cases, the configuration file for the module is generated automatically, based on the template provided. However, it may be updated by the user depending on the results obtained after a first run.

Process

The process (Figure 1) begins with the parsing of the VCF file to extract the genetic data, and of the reference genome fasta file when PSMC is used (*Step 1*). Subsequently, we construct the Site Frequency Spectrum (SFS) for tools such as $\partial a \partial i$ and StairwayPlot2. This stage involves additional filtering of variants present in regions showing low density of SNPs. In parallel, the pipeline executes dedicated Snakemake rules for PSMC, SMC++ and MSMC2, which does not use the generated SFS (*Step 2*). Each tool ends by generating an inference plot illustrating temporal effective population size variations (Output Plots [1-5]). In parallel to these operations, we perform statistical analyses on the VCF data, including Principal Component Analysis (PCA) (Output Plots [6-10]) (*Step 3*).

Outputs

The outputs include detailed logs and command traces for each tool used, ensuring transparency in the parameters and methods applied. These logs are critical for reproducibility and for understanding nuances of the analysis. Additionally, we provide 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 $\partial a \partial i$ 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 IBEES (Initiative Biodiversity, Evolution, Ecology & Society) grant.

Bibliography

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

Forest, T., Achaz, G., Marbouty, M., Bignaud, A., Thierry, A., Koszul, R., Milhes, M., Lledo, J., Pons, J.-M., and Fuchs, J. (2024). Chromosome-level genome assembly of the european green woodpecker *Picus viridis*. *G3: Genes, Genomes, Genetics*, 14(5). doi: 10.1093/g3journal/jkae042.

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

Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. 475 (7357):493–496. doi: 10.1038/nature10231. Number: 7357 Publisher: Nature Publishing Group.

Liu, X. and Fu, Y.-X. (2020). Stairway plot 2: demographic 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.

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.