Sgkit is awesome

This manuscript (permalink) was automatically generated from tomwhite/manubot-test@aa933ae on January 23, 2023.

Authors

- **Tom White**
  XXXX-XXXX-XXXX-XXXX · Tom White · tom_e_white
  TWC · Funded by Grant XXXXXXXX

- **Jane Roe** ✉
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

✉ — Correspondence possible via GitHub Issues or email to Jane Roe <jane.roe@whatever.edu>.

Abstract

Popgen

Audience

[TODO]

Overview of sgkit’s API methods

Sgkit provides a number of methods for computing statistics in population genetics. Before running the methods, the dataset is usually divided into windows along the genome, using the `window_by_*` functions, which tell sgkit to produce per-window statistics. For example, `window_by_position` creates windows that are a fixed number of base pairs, while `window_by_interval` creates windows corresponding to arbitrary user-defined intervals.

It’s common in population genetics to group samples into populations, which in sgkit are referred to as *cohorts*. There are two types of statistics: one-way statistics where there is a single statistic for each cohort, and multi-way statistics where there is a statistic between each pair, triple, etc of cohorts. [TODO: do we need to say how cohorts are defined?]

The methods for one-way statistics include `diversity` for computing mean genetic diversity, `Tajimas_D` for computing Tajima’s D, and `Garud_H` for computing the H1, H12, H123 and H2/H1 statistics defined in [1].

The methods for multi-way statistics include `divergence` and `Fst` for computing mean genetic divergence and F[ST] (respectively) between pairs of cohorts, and `pbs` for computing the population branching statistic between cohort triples.

Example

We converted phased Ag1000G hypotype data in Zarr format [2] to sgkit’s Zarr format using the `read_scikit_allel_vcfzarr` function. The data contained 1,164 samples at 39,604,636 sites, and was [TODO] MB on disk before conversion, and Y MB after conversion to sgkit’s Zarr format. Data for the X chromosome was discarded since it was not available for all samples. The conversion took [TODO] minutes Y seconds, including a postprocessing `rechunk` step to ensure that the data was suitably chunked for the subsequent analysis.

References

1. **Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps**
Nandita R Garud, Philipp W Messer, Erkan O Buzbas, Dmitri A Petrov
PLOS Genetics (2015-02-23) https://doi.org/f67qcv
DOI: 10.1371/journal.pgen.1005004 · PMID: 25706129 · PMCID: PMC4338236

2. **Ag1000G phase 2 AR1 data release.**
The Anopheles gambiae 1000 Genomes Consortium
MalariaGEN (2017) https://www.malariagen.net/data/ag1000g-phase-2-ar1