Human dataset analysis details

Irineo Cabreros and John D. Storey

4/18/2019

Summary

In this document we detail the steps used to analyze the human datasets studied in Cabreros and Storey 2018 (https://www.biorxiv.org/content/early/2019/03/27/240812.full.pdf). All referenced files can be found in the public `GitHub` repository: https://github.com/StoreyLab/alstructure_paper.

Data acquisition

There are four human datasets analyzed in this work:

1. TGP (1000 Genomes Project) (The 1000 Genomes Project Consortium 2015)
2. HGDP (Human Genome Diversity Project) (Cavalli-Sforza 2005)
3. HO (Human Origins) (Lazaridis and others 2014)
4. IND (India) (Basu, Sarkar-Roy, and Majumder 2016)

The TGP, HGDP, and HO datasets are publically available, while the the IND dataset is not. For the publically available datasets, we have included the exact data analyzed throughout Cabreros and Storey 2018 in the `GitHub` repository: https://github.com/StoreyLab/alstructure_paper/tree/master/data. The filtering done to obtain these datasets is detailed in Cabreros and Storey 2018. The IND dataset must be obtained from the authors of (Basu, Sarkar-Roy, and Majumder 2016). In the same `GitHub` directory containing the publically available datasets, we have included the script used to filter the IND dataset: `clean_IND.R`. Someone trying to reproduce our results will need to first obtain the IND dataset and then run `clean_IND.R`.

Method acquisition

There are four methods compared in this work:

1. `Admixture` (Alexander, Novembre, and Lange 2009)
2. `fastSTRUCTURE` (Raj, Stephens, and Pritchard 2014)
3. `terastructure` (Gopalan et al. 2016)
4. `ALStructure` (Cabreros and Storey 2018)

The `Admixture` software can be obtained from http://software.genetics.ucla.edu/admixture/publications.html. The `fastSTRUCTURE` software can be obtained from https://rajanil.github.io/fastStructure/. The `terastructure` software can be obtained from https://github.com/StoreyLab/terastructure. The `ALStructure` software is an `R` package, which can be obtained by executing the following commands into an `R` console:

```
library("devtools")
install_github("storeylab/alstructure", build_vignettes=TRUE)
```

Original dataset fits

Each of the four datasets (TGP, HGDP, HO, and IND) are first fit by each of the four methods (`Admixture`, `fastSTRUCTURE`, `terastructure`, and `ALStructure`). Each of these 16 total fits were obtained by running the script: `/scripts/original_data_fits.R`.

This script assumes 16 parallel jobs are being submitted to a server through a `.slurm` script with array indices 1-16. The array index `AI` is read, which specifies the particular dataset and method used. Running these jobs in parallel is highly recommended, as some of the individual fits required several days or weeks.

Simulating datasets from original dataset fits

In order to assess the performace of each method on human data, we compare their performance on datasets simulated from the fitted model parameters obtained in the previous section by evaluating `/scripts/original_data_fits.R` (this is described in greater detail in Cabreros and Storey 2018). We produce and store the datasets by running the script: `/scripts/simulate_datasets.R`.

This script assumes k parallel jobs are being submitted to a server through a `.slurm` script with array indices 1-k. The array index `AI` is read, which specifies replication ID of the datasets produced. In Cabreros and Storey 2018, there were four different replicate datasets produced for each method-dataset pair ($k = 4$), and we excluded the IND dataset. The output `/scripts/original_data_fits.R` is therefore 3(number of methods) $\times$ 4(number of datasets) $\times$ 4(number of replications) = 48 total simulated datasets.

Fitting simulated datasets

Each of the 48 simulated datasets produced by `/scripts/simulate_datasets.R` are then fitted by each of the four methods, using the script `/scripts/fit_simulated_datasets.R`. This produces 48(number of simulated datasets) $\times$ 4(number of methods) = 192 total model fits.

This script assumes 192 parallel jobs are being submitted to a server through a `.slurm` script with array indices 1-192. The array index `AI` is read, which specifies the particular dataset and method used. Running these jobs in parallel is highly recommended, as some of the individual fits required several days or weeks.

References

Alexander, D. H., J. Novembre, and K. Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research* 19: 1655–64.

Basu, Analabha, Neeta Sarkar-Roy, and Partha P. Majumder. 2016. "Genomic Reconstruction of the History of Extant Populations of India Reveals Five Distinct Ancestral Components and a Complex Structure." *Proceedings of the National Academy of Sciences* 113 (6). National Academy of Sciences: 1594–9. doi:10.1073/pnas.1513197113.

Cavalli-Sforza, L. L. 2005. "The Human Genome Diversity Project: Past, Present and Future." *Nature Reviews Genetics* 6: 333–40.

Gopalan, P., W. Hao, D. M. Blei, and J. D. Storey. 2016. "Scaling Probabilistic Models of Genetic Variation to Millions of Humans." *Nature Genetics* 48 (12): 1587–92.

Lazaridis, I., and others. 2014. "Ancient Human Genomes Suggest Three Ancestral Populations for Present-Day Europeans." *Nature* 513: 409–13.

Raj, Anil, Matthew Stephens, and J. K. Pritchard. 2014. "fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets." *Genetics* 197 (2). Genetics: 573–89. doi:10.1534/genetics.114.164350.

The 1000 Genomes Project Consortium. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526: 68–74.