SnpRecode: A Versatile and Fast Genotype Recoding and Correlation Function

Andrew Marete (andrew.marete@agr.gc.ca)  
Agriculture and Agri-Food Canada

Nathalie Bissonnette  
Agriculture and Agri-Food Canada

Short report

Keywords: Imputation, Genomics, Pipeline

DOI: https://doi.org/10.21203/rs.3.rs-95704/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

SnpRecode is a tool meant to be implemented with the Fimpute imputation software. SnpRecode allows for fast and seamless conversion of genotypes to and from a format acceptable by Fimpute software. It also implements a fast genotype correlation function to estimate and plot the imputation accuracy. SnpRecode is implemented in Python and its source code and Linux executable is freely available for download at https://github.com/amarete/fimpute-utils

Background

The gradual decline in genotyping costs has led to an increase in the number of samples being genotyped irrespective of species. In the livestock industry, large scale genotyping is a major tool used during genomic selection. Studies have shown that it is cheaper to genotype samples with a lower density chip and impute them to higher densities using large reference panels. A popular, fast and versatile imputation software used in animal production is Fimpute (Sargolzaei et al. 2014). Fimpute has been shown to be robust when imputing many individuals genotyped with different panels. Apart from the obvious fast algorithm powering Fimpute, the speeds gained using the software can be attributed to its lack internal genotype recoding and allelic correlation and concordance estimation. This is in contrast to software such as Beagle (Browning et al. 2018), Minimac4 (Das et al. 2016), and Impute2 (Howie et al, 2009) which can do internal genotype recoding, estimate allelic R-square values and concordance. We have developed SnpRecode to bridge this gap by recoding genotypes from Variant Call Format (VCF) and/or Plink (Purcell et al. 2007) format to Fimpute format and recoding back to VCF or Plink format after the imputation procedure. Though this may seem trivial, format conversion bottleneck is a major concern in most next generation sequence analysis (Wang et al. 2014) and SnpRecode allows for addressing this bottleneck. Written in Python programming language v3.6, the software is highly customizable and the source code is freely available for modification as needed, however, it is recommended to use the compiled Linux executable.

SnpRecode can process data from up to 10 chips depending on the power of the machine and though there are various checks within the software to detect data preparation oversights, due diligence on quality control and general data preparation should be performed prior to using SnpRecode. As Python process data one row at a time, the function uses a linear (O(n)) and quasilinear (O(nLogₙ)) time complexity. This implies that using Plink formatted files will use less time as opposed to VCF files that require inversion. The sample size therefore remains to be the main limiting time factor. Furthermore, to allow the estimation of imputation accuracy, we have implemented a function that allows the user to correlate two VCF files. Most of the available software do not offer a direct allele-to-allele or genotype-to-genotype correlation, rather they can estimate variant overlap by comparing genomic coordinates (e.g. bedtools, Quinlan et al. 2010), duplicate sites or intersections (e.g. vcf-compare, bcftools isec), or estimate concordance (e.g. snpsift, Cingolani et. al (2012)). These tools were mainly created for instance to check genomic differences arising from variants being called by different tools. In contrast, SnpRecode allows for a direct genotype correlation estimate and provide the R-Square value between two genotypes.
Furthermore, it bins this result dependent on the minor allele frequency and produces an informative plot (Figure 1) that can allow one to discern the accuracy of the imputation run. A sample script of SnpRecode function is provided as supplementary file 1 and script of how to implement in a pipeline provided in supplementary file 2.

**Conclusions**

We have implemented a highly scalable function to allow seamless implementation in an Fimpute pipeline. The function can easily detect if multiple SNP or individuals are present within or between two genotype files thus streamlining such a pipeline. The function also allows one to quickly establish the imputation goodness of fit by estimating genotype correlations between a real and masked genotype.

**Abbreviations**

VCF – variant call format

SNP – single nucleotide polymorphism

**Declarations**

**Ethics approval and consent to participate**

Not applicable

**Consent for publication**

Not applicable

**Availability of data and materials**

All data and source code are freely available at [https://github.com/amarete/fimpute-utils](https://github.com/amarete/fimpute-utils)

**Competing interests**

The authors declare no competing interests

**Funding**

This study was funded by Agri-Food and Agriculture Canada (Project AAFC J0000–75).

**Authors' contributions**
AM wrote the source code and NB contributed to manuscript writing

Acknowledgements

We are thankful to Dr. Mehdi Sargolzaei for providing an academic version of Fimpute software.

References

1. Sargolzaei, M., J. P. Chesnais and F. S. Schenkel. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 2014;15:478.
2. Browning B. L., Zhou Y., and Browning S. R. A one-penny imputed genome from next generation reference panels. Am J Hum Genet 2018;103(3):338-348.
3. Das S., Forer L., Schonherr S., Sidore C., Locke A. E., Kwong A., Vrieze S. I., Chew E. Y., Levy S., McGue M., Schlessinger D., Stambolian D., Loh P. R., Iacono W. G., Swaroop A., Scott L. J., Cucca F., Kronenberg F., Boehnke M., Abecasis G. R., Fuchsberger C. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284-1287.
4. Howie B. N., Donnelly P., and Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 2009;5(6):e1000529.
5. The Variant Call Format (VCF) Version 4.2 Specification. (2020). https://samtools.github.io/hts-specs/VCFv4.2.pdf
6. Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M., Bender D., Maller J., Sklar P., de Bakker P., Daly M. and Sham P. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 2007;81.
7. Wang Y., Agrawal G., Ozer G., Huang K. (2014). Removing Sequential Bottlenecks in Analysis of Next-Generation Sequencing Data. IEEE International Parallel & Distributed Processing Symposium Workshops, Phoenix, AZ, 2014, pp. 508-517, doi.org/10.1109/IPDPSW.2014.64.
8. Quinlan A. and Ira M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841–842.
9. Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011;27(21):2987-2993
10. Cingolani P., Patel V., Coon M., Nguyen T., Land S.J., Ruden D.M., et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program - SnpSift. Front Genet 2012;3:35.

Figures
These tools were mainly created for instance to check genomic differences arising from variants being called by different tools. In contrast, SnpRecode allows for a direct genotype correlation estimate and provide the R-Square value between two genotypes.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfile2.pdf
- Supplementaryfile1.pdf