Snprcode: A versatile and fast genotype recoding and correlation function

Andrew Marete (✉ amarete@lactanet.ca)

Lactanet Canada

Short report

Keywords: Imputation, SNP, Converter, Parsing

Posted Date: July 8th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-95704/v3>

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

[Read Full License](#)

Abstract

Genotype imputation is an essential tool used in genomic selection in plants and animals. A popular imputation tool used in animal genomics is FImpute. FImpute, however, accepts a specific genotype format and produces dosages whose conversion to VCF or Plink format requires multiple software packages in a pipeline with a large amount of processing time. We have developed SnpRecode as a helper tool that bridges the gap between regular genotype files and the FImpute imputation software by allowing for fast and seamless conversion of genotypes to-and-from FImpute format. SnpRecode also implements a fast genotype correlation function to estimate and plot the imputation accuracy. We run tests on 6,000 samples with a step of 1,000 to determine the performance of SnpRecode on various sample sizes and runtime and memory usage used as performance measures. The performance of SnpRecode was modest at 10sec/1,000 samples. Written in Python programming language, SnpRecode provides users with great flexibility in implementation with other software packages in a pipeline.

Main Text

The gradual genotyping costs decline has led to an increase in genotyped samples irrespective of species. In the livestock industry, large scale genotyping is an essential 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). In comparison to imputation software that imputes between a reference and a validation file (Browning et al. 2018; Das et al. 2016; Howie et al. 2009), FImpute is robust because it can impute many individuals genotyped with different SNP chip panels (Figure 1A). Compared to most genotype imputation software, FImpute's algorithm is faster at a magnitude of greater than one. However, one can hypothesize that this fast imputation speed is due to FImpute's lack of internal genotype recoding, allelic correlation, and concordance estimation. The absence of an internal genotype conversion method is of significant concern (Yi Wang 2014) and a bottleneck for most imputation pipelines involving Fimpute. To address this, we have developed SnpRecode to bridge this gap by recoding genotypes from Variant Call Format (VCF) and Plink format to FImpute format and recoding back to VCF or Plink format after the imputation procedure (GDT 2020; Purcell et al. 2007). Written in Python programming language v3.6, the SnpRecode is

highly customizable because the source code is freely available for modification as needed; however, we recommend to use the compiled Linux executable.

Snprerecode can process data from up to 10 chips depending on the power of the machine. There are various checks within the software to detect data preparation oversights; however, it's better to perform quality control and general data preparation before using Snprerecode. As Python process data one row at a time, Snprerecode uses a linear ($O(n)$) and quasilinear ($O(n \log n)$) time complexity. This dynamic implies that using Plink formatted files will use less time than VCF files that require inversion and the sample size, therefore, remains the main time-limiting factor.

Furthermore, to allow the estimation of imputation accuracy, Snprerecode allows the user to correlate two VCF files for a direct genotype correlation estimate and provide the R-Square value between two genotypes. It then bins this result using the minor allele frequency and produces an informative plot that can allow one to discern the imputation run's accuracy (Figure 1B). In contrast, most genotype comparison software can estimate variant overlap by comparing genomic coordinates (e.g., bedtools), check duplicate sites or intersections (e.g., vcf-compare, bcftools isec), or estimate concordance (e.g., snpsift) (Cingolani et al. 2012; Li 2011; Quinlan and Hall 2010). Supplementary file 1 illustrates the functions of Snprerecode, and supplementary file 2 illustrates how to incorporate Snprerecode in a FImpute pipeline.

Conclusion

We have implemented a highly optimizable tool to allow seamless implementation in a FImpute pipeline. The tool allows one to convert genotypes to-and-fro FImpute formats and detect data consistency errors, thus streamlining such a pipeline. Additionally, the tool allows one to quickly establish the imputation goodness of fit by estimating genotype correlations between a real and masked genotype.

Declarations

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

Acknowledgements

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

Competing interests

The authors declare no competing interests

Funding

This work is supported by Agri-Food and Agriculture Canada (Project AAFC J0000-75).

Authors' contributions

AM wrote the source code and NB contributed to manuscript writing

References

- Browning BL, Zhou Y, Browning SR (2018) A One-Penny Imputed Genome from Next-Generation Reference Panels *Am J Hum Genet* 103:338-348 doi:10.1016/j.ajhg.2018.07.015
- Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X (2012) Using *Drosophila melanogaster* as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift *Front Genet* 3:35 doi:10.3389/fgene.2012.00035
- Das S et al. (2016) Next-generation genotype imputation service and methods *Nat Genet* 48:1284-1287 doi:10.1038/ng.3656
- GDT (2020) The Variant Call Format (VCF) Version 4.2 Specification Global Alliance for Genomics and Health doi:samtools.github.io/hts-specs
- Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies *PLoS Genet* 5 doi:10.1371/journal.pgen.1000529
- Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data *Bioinformatics* 27:2987-2993 doi:10.1093/bioinformatics/btr509

- Purcell S et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses *American Journal of Human Genetics* 81:559-575 doi:10.1086/519795
- Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features *Bioinformatics* 26:841-842 doi:10.1093/bioinformatics/btq033
- Sargolzaei M, Chesnais JP, Schenkel FS (2014) A new approach for efficient genotype imputation using information from relatives *BMC Genomics* 15:478-478 doi:10.1186/1471-2164-15-478
- Yi Wang GA, Gulcin Ozer, Kun Huang (2014) Removing Sequential Bottlenecks in Analysis of Next-Generation Sequencing Data *IEEE International Parallel & Distributed Processing Symposium Workshops*:508-517 doi:10.1109/IPDPSW.2014.64

Figures

Figure 1

Graphical representation of Genomic correlation between a true and masked vcf file

Supplementary Files

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

- [Supplementaryfile1.pdf](#)
- [Supplementaryfile2.pdf](#)