Bioinformatics, 32(23), 2016, 3661–3663
doi: 10.1093/bioinformatics/btw487
Advance Access Publication Date: 6 August 2016
Applications Note

OXFORD

Genome analysis

genipe: an automated genome-wide imputation pipeline with automatic reporting and statistical tools

Louis-Philippe Lemieux Perreault[1,*], Marc-André Legault[1,2], Géraldine Asselin[1] and Marie-Pierre Dubé[1,3,*]

[1]Beaulieu-Saucier Université de Montréal Pharmacogenomics Centre, Montreal Heart Institute Research Center, Montréal, Canada H1T 1C8, [2]Department of Biochemistry and molecular medicine, Université de Montréal, Montreal, Canada H3T 1J4, and [3]Department of Medicine, Université de Montréal, Montreal, Canada H3T 1J4

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on December 21, 2015; revised on June 30, 2016; accepted on July 18, 2016

Abstract

Summary: Genotype imputation is now commonly performed following genome-wide genotyping experiments. Imputation increases the density of analyzed genotypes in the dataset, enabling fine-mapping across the genome. However, the process of imputation using the most recent publicly available reference datasets can require considerable computation power and the management of hundreds of large intermediate files. We have developed *genipe*, a complete genome-wide imputation pipeline which includes automatic reporting, imputed data indexing and management, and a suite of statistical tests for imputed data commonly used in genetic epidemiology (Sequence Kernel Association Test, Cox proportional hazards for survival analysis, and linear mixed models for repeated measurements in longitudinal studies).

Availability and Implementation: The *genipe* package is an open source Python software and is freely available for non-commercial use (CC BY-NC 4.0) at https://github.com/pgxcentre/genipe. Documentation and tutorials are available at http://pgxcentre.github.io/genipe.

Contact: louis-philippe.lemieux.perreault@statgen.org or marie-pierre.dube@statgen.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) are usually performed on datasets containing over 1 million genetic markers. Those markers are typed using high-throughput genotyping arrays that target a small fraction of all possible genetic variants. Imputation is a low-cost and popular statistical method to infer genotypic information at up to 80 million known genetic variants (including single nucleotide variants and insertions/deletions). Imputation is often used to boost statistical power and it can be used to infer missing data and standardize variant sets for meta-analysis (Marchini *et al.*, 2010).

Large sequencing projects spanning multiple human populations greatly increase the availability and quality of public imputation panels, but the statistical methods needed for haplotype phasing and imputation at the genome-wide level are computationally intensive. In order to streamline this process, we have developed a genome-wide imputation pipeline that automates all necessary computational steps including quality control and reporting while providing support for high performance computing environments and statistical tests for imputed data.

2 Methods

2.1 Main pipeline

The main pipeline uses three commonly used bioinformatics tools: PLINK (Purcell *et al.*, 2007) for the initial genetics data management, SHAPEIT (Delaneau *et al.*, 2012) for the loci strand

verification and the phasing step and IMPUTE2 (Howie *et al.*, 2009) for the imputation process. The *genipe* package orchestrates the pipeline according to the best practices for genome-wide imputation analysis (as described by SHAPEIT and IMPUTE2), and manages the intermediate files created by the tools at the different stages. The following steps are performed (most of them in parallel, see Supplementary Table S1) in a typical imputation analysis with *genipe*.

Quality metrics and statistics on the initial dataset are obtained, including the missing call rate which is computed using PLINK for all loci and all samples of the study dataset. Genetic loci from the study panel are then filtered according to their genetic location and allele composition. Starting from a binary *ped* file (PLINK format), loci located on the Y and Mt chromosomes, or with ambiguous alleles (A/T or G/C) are filtered out of the dataset. At the same time, a single version of a genetic marker is kept for downstream analysis (i.e. duplicated variants). An optional strand check step can be performed to ensure consistency between the human genome reference and the microarray derived alleles. This means that the nucleotides that are inconsistent with the reference will be complemented with respect to the Watson Crick pairings. The reference strands of the study data and that of the imputation reference panels are then compared using SHAPEIT and inconsistent loci are flipped. A second and last strand verification is performed and the remaining loci that are discordant with the imputation reference are excluded from the dataset.

The remaining markers are phased in parallel for each chromosome using SHAPEIT. The phasing tool allows for multiple threads to run concurrently for each chromosome, increasing the speed of the analysis.

Finally, IMPUTE2 is used to perform genome-wide imputation. Each chromosome is split into segments of 5 Mb (by default) for the imputation process in accordance with the segment size limitation of the IMPUTE2 program. For each segment, IMPUTE2 produces cross validation metrics, where a subset of genotyped loci are masked, imputed and compared to their expected value. The *genipe* package aggregates those statistics at a chromosomal and genomic level using a weighted arithmetic mean.

Results for all segments are merged to produce one result file per chromosome. Multiple companion files, including important information such as minor allele frequency and information value (INFO) are also included. The merged IMPUTE2 results file can optionally be compressed using BGZIP, a variant of GZIP that allows indexing.

The pipeline can be run on a desktop computer or on a DRMAA compliant high-performance computing server. The phasing and imputation steps are customizable to fit any study design. A database keeps track of all executed steps and enables post-failure relaunch of the pipeline where it last stopped, saving processing time and resources. At completion, a LATEX report containing important imputation quality metrics and processing times is automatically created.

2.2 Output file management

Given the high number of variants imputed using a genome-wide approach, file management becomes a computational challenge that needs to be addressed. To facilitate this task, we have developed the *impute2-extractor* utility that uses file indexing to accelerate extraction of variants from IMPUTE2 files by name or by genomic region. Allele frequency filtering as well as quality control thresholds based on the completion rate, the INFO field and the imputation probability can also be specified. This script can be used as a standalone utility and is automatically installed and added to the system path after setting up *genipe*.

2.3 Statistical analysis

Bindings for linear and logistic regression using the *statsmodels* package (Seabold *et al.*, 2010) are included for common variant association testing with continuous and discrete outcomes. Other more sophisticated models, which were not previously available for dosage data in an efficient tool, now have bindings with *genipe*. These include the Cox proportional hazards model (*lifelines* package) which had no suitable implementations for pharmacogenomics GWAS (Syed *et al.*, 2016), an optimization (Sikorska *et al.*, 2015) of the linear mixed model for repeated measurements (*statsmodels* package, see Supplementary Figs S2 and S3), and the Sequence Kernel Association Test [*SKAT* R package (Ionita-Laza *et al.*, 2013)]. All these tests are now available for dosage data through the companion script *imputed-stats*. See Supplementary Figure S1 for execution times for the different models.

3 Application

The pipeline's documentation provides a typical imputation analysis tutorial along with required files. Those files includes a dataset of 2 278 357 markers genotyped on 90 HapMap samples. Using the 1000 Genomes Phase 3 reference panels, the dataset was imputed on two different systems: a computing server (10 nodes of 8 Intel® Xeon® E5620 CPUs 2.40 GHz, 48G of RAM per node) using the DRMAA API for automatic task submission and a desktop computer (Intel® Core™ i7-3770 CPU 3.40 GHz, 16G of RAM). Using a maximum of 50 simultaneous tasks, the pipeline took a total of 4.25 h on the computing server (including a waiting period of 0.08 h in queue). Using a maximum of four simultaneous tasks, the pipeline took 10.62 h to complete on the desktop computer.

4 Conclusion

Although online imputation pipeline exists (e.g. the Michigan Imputation Server that uses Minimac3; https://imputationserver.sph.umich.edu/ and the Sanger Imputation Service that uses PBWT; https://imputation.sanger.ac.uk/), *genipe* is advantageous for the users who cannot upload genotypic data on an off-site server for ethical or legal restrictions. Also, as public servers gain in popularity, the high workload can add significant time to the imputation analysis (queue time). The *genipe* pipeline can be efficiently executed on a local high-performance computing server or on a single desktop computer.

Finally, *genipe* provides a unified interface to statistical analysis packages that did not have existing tools to automate the use of dosage data (e.g. linear mixed models from *statsmodels*, Cox proportional hazards from *lifelines* and SKAT).

Funding

This work was supported by the Montreal Heart Institute Foundation; Genome Canada and Genome Quebec.

Conflict of Interest: none declared

References

Delaneau,O. *et al*. (2012) A linear complexity phasing method for thousands of genomes. *Nat. Methods*, **9**, 179–181.

Howie,B.N. *et al*. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*., **5**, e1000529.

Ionita-Laza,I. *et al*. (2013) Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet*., **92**, 841–853.

Marchini,J. *et al*. (2010) Genotype imputation for genome-wide association studies. *Nat. Rev. Genet*., **11**, 499–511.

Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*., **81**, 559–575.

Seabold,S. *et al*. (2010). Statsmodels: econometric and statistical modeling with python. In: *Proceedings of the Ninth Python in Science Conference*, pp. 57–61. http://conference.scipy.org/proceedings/scipy2010/seabold.html.

Sikorska,K. *et al*. (2015) GWAS with longitudinal phenotypes: performance of approximate procedures. *Eur. J. Hum. Genet*., **23**, 1384–1391.

Syed,H. *et al*. (2016) Evaluation of methodology for the analysis of 'time-to-event' data in pharmacogenomic genome-wide association studies. *Pharmacogenomics*, **17**, 907–915.