gwasurvivr: an R package for genome wide survival analysis

Abbas A Rizvi[1*], Ezgi Karaesmen[1*], Martin Morgan[2], Leah Preus[1], Junke Wang[1], Michael Sovic[1] , Lara E Sucheston-Campbell[1,3]
[1]College of Pharmacy, The Ohio State University, Columbus, OH, [2]Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, [3]College of Veterinary Medicine, The Ohio State University, Columbus, OH

*These authors contributed equally

ABSTRACT

Summary: Researchers are increasingly interested in evaluating time-to-event outcomes such as survival in the context of genetic variation. However, there are limited software options for performing survival analyses with millions of SNPs. To address this, we developed gwasurvivr, an R/Bioconductor package to conduct fast and efficient genome wide survival analyses. gwasurvivr accepts data in VCF (as outputted from Michigan or Sanger imputation servers) and IMPUTE2 format, and provides a simple interface to run large-scale analyses. We benchmarked gwasurvivr with other GWAS software capable of conducting genome wide survival analysis (genipe, SurvivalGWAS_SV, and GWASTools) and have demonstrated improved scalability that includes shorter run times for large sample sizes and larger number of SNPs.

Availability and implementation: gwasurvivr, including source code, documentation, and vignette are available at: https://github.com/suchestoncampbelllab/gwasurvivr

1 INTRODUCTION

Genome-wide association studies (GWAS) are population-level experiments that investigate genetic variation in individuals to observe single nucleotide polymorphism (SNP) associations with a phenotype. Genetic variants tested for association are genotyped on an array and imputed from a reference panel of sequenced genomes using 1000 Genomes Project or Haplotype Reference Consortium (HRC) (Das, et al., 2016; Genomes Project, et al., 2015). Imputation increases genome coverage from hundreds of thousands or a few million to upwards of 30 million SNPs, improves power to detect genetic associations, and/or homogenizes variant sets for meta-analyses (Das, et al., 2016). Imputed SNPs can be tested for association with binary outcomes (case/control) and quantitative outcomes (i.e. height) using a range of available software packages including SNPTEST or PLINK (Marchini, et al., 2007; Purcell, et al., 2007). However, existing software options for performing survival analyses across millions of imputed SNPs (genipe, SurvivalGWAS_SV, and GWASTools) require user interaction with raw output, were not initially designed for survival and/or have other limitations that could deter more introductory users (Gogarten, et al., 2012; Lemieux Perreault, et al., 2016; Syed, et al., 2017). We developed an R/Bioconductor package, gwasurvivr, for genome wide survival analyses of imputed data in multiple formats with flexible analysis and output options.

2 DATA STRUCTURES

Gwasurvivr can be used to analyze IMPUTE2 data or VCF derived from Michigan or Sanger imputation servers. Data from each are prepared in gwasurvivr by leveraging existing Bioconductor packages GWASTools or VariantAnnotation depending on the imputation output.

IMPUTE2 IMPUTATION

IMPUTE2 format is a standard genotype (.gen) file that is (typically) partitioned into 5 Mb chunks which store genotype probabilities (GP). We utilized GWASTools in R to compress files into genomic data structure (GDS) format. This allows for efficient, iterative accession to subsets of the data while simultaneously converting GP into dosages (DS) (Gogarten, et al., 2012) for use in subsequent survival analyses.

MICHIGAN or SANGER SERVER IMPUTATION

VCF files generated from these servers include a DS field and server specific meta-fields (INFO score, reference panel allele frequencies) which are read in by VariantAnnotation (Obenchain, et al., 2014) specifying user-defined chunk sizes (number of SNPs). The default chunk size is 5000 SNPs but can easily be scaled up.

3 FUNCTIONALITY

gwasurvivr implements a Cox proportional hazards regression model to test each SNP with an outcome, with or without covariates and/or SNP-covariate interaction (Cox, 1992). To potentially decrease the number of iterations needed for convergence when optimizing the parameter estimates in the Cox model, we modified the survival package in R such that the covariates in the model are first fit without the SNP and those parameter estimates are used as initial points (Therneau and Grambsch, 2000). If no additional covariates are added to the model, the parameter estimation optimization begins with null initial value.

Survival analyses are run using genetic data in either VCF or IMPUTE2 formats. For both formats, survival time, survival status and additional covariates are stored in a phenotype file indexed by sample ID. The VCF files contain genomic data and IDs, while IMPUTE2 requires .gen and sample files. gwasurvivr functions for IMPUTE2 (impute2CoxSurv) or VCF (michiganCoxSurv or sangerCoxSurv) include arguments for the survival model (event of interest, time to event, and covariates) and quality control based on minor allele frequency (MAF) and INFO score filtering. Users can also provide a list of sample IDs for gwasurvivr to internally subset the data, an option which is not available for SurvivalGWAS_SV. Gwasurvivr outputs two files: (1) a .snps_removed file listing all SNPs that failed QC parameters including MAF and INFO of each dropped SNP and (2) a .coxph file with the results from the analyses, which includes parameter estimates, p-values, MAF, INFO score for sample set analyzed, the number of events and total sample N for each SNP. Verbose arguments in the function allow the user to obtain more detailed information about any covariates and SNP*covariate interactions included in the model. Gwasurvivr is well suited for multi-core processors and users can specify the number of cores used during computation on Windows and Unix based systems and if run on a high performance computing cluster, chromosomes can be batched. Lastly, gwasurvivr

overcomes potential memory limitations that are often attributed to R by iteratively reading in data on subsets of the entire data making it possible to conduct genome-wide analyses on a typical laptop computer.

3 BENCHMARKING PERFORMANCE

We benchmarked gwasurvivr with genipe, SurvivalGWAS_SV v.1.3.2, and GWASTools using IMPUTE2 data (the comparison packages do not take VCF) for varying sample sizes (n=100, n=1000, n=5000) and number of SNPs (p=1,000, p=10,000, p=100,000), and included three non-genetic covariates. Survival time outcomes and patient status (alive/dead) were simulated using a normal and a binomial distribution, respectively. Genetic data were simulated using HAPGENv2 with 1000G as a backbone and SNP sets selected from chromosome 18. To achieve the sample sizes for simulations, we combined the simulated cases and controls to form each of the 9 datasets. All simulations were performed using identical CPU constraints, 1 node (2.27 GHz Clock Rate) and 8 cores with 24 GB of RAM, on the University at Buffalo Center for Computational Research supercomputer. Genipe, SurvivalGWAS_SV, and GWASTools analyses were performed as specified by the authors in the online manual or vignette. All 4 software packages showed excellent agreement between MAF estimates, coefficient estimates, and p-values. For all 9 simulated datasets, gwasurvivr performed faster than genipe, SurvivalGWAS_SV, and GWASTools (Figure 1). Genipe timed out at 72 hours for the final simulation set of 5000 patients and 100,000 SNPs.

ACKNOWLEDGEMENTS

The authors would like to thank Amy Webb and Guy Brock at The Ohio State University Department of Biomedical Informatics for listening to discussions and offering suggestions for the development of the package. This work was performed in part at the University at Buffalo's Center for Computational Research.

FUNDING

This work was supported by the National Institute of Health R03CA188733-01 and R01HL102278-01A1.

REFERENCES

Cox, D.R. Regression Models and Life-Tables. Springer; 1992.

Das, S.*, et al.* Next-generation genotype imputation service and methods. *Nat Genet* 2016;48(10):1284-1287.

Genomes Project, C.*, et al.* A global reference for human genetic variation. *Nature* 2015;526(7571):68-74.

Gogarten, S.M.*, et al.* GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* 2012;28(24):3329-3331.

Lemieux Perreault, L.P.*, et al.* genipe: an automated genome-wide imputation pipeline with automatic reporting and statistical tools. *Bioinformatics* 2016;32(23):3661-3663.

Marchini, J.*, et al.* A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;39(7):906-913.

Obenchain, V.*, et al.* VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* 2014;30(14):2076-2078.

Purcell, S.*, et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81(3):559-575.

Syed, H., Jorgensen, A.L. and Morris, A.P. SurvivalGWAS_SV: software for the analysis of genome-wide association studies of imputed genotypes with "time-to-event" outcomes. *BMC Bioinformatics* 2017;18(1):265.

Therneau, T.M. and Grambsch, P.M. Modeling Survival Data: Extending the Cox Model. Springer; 2000.
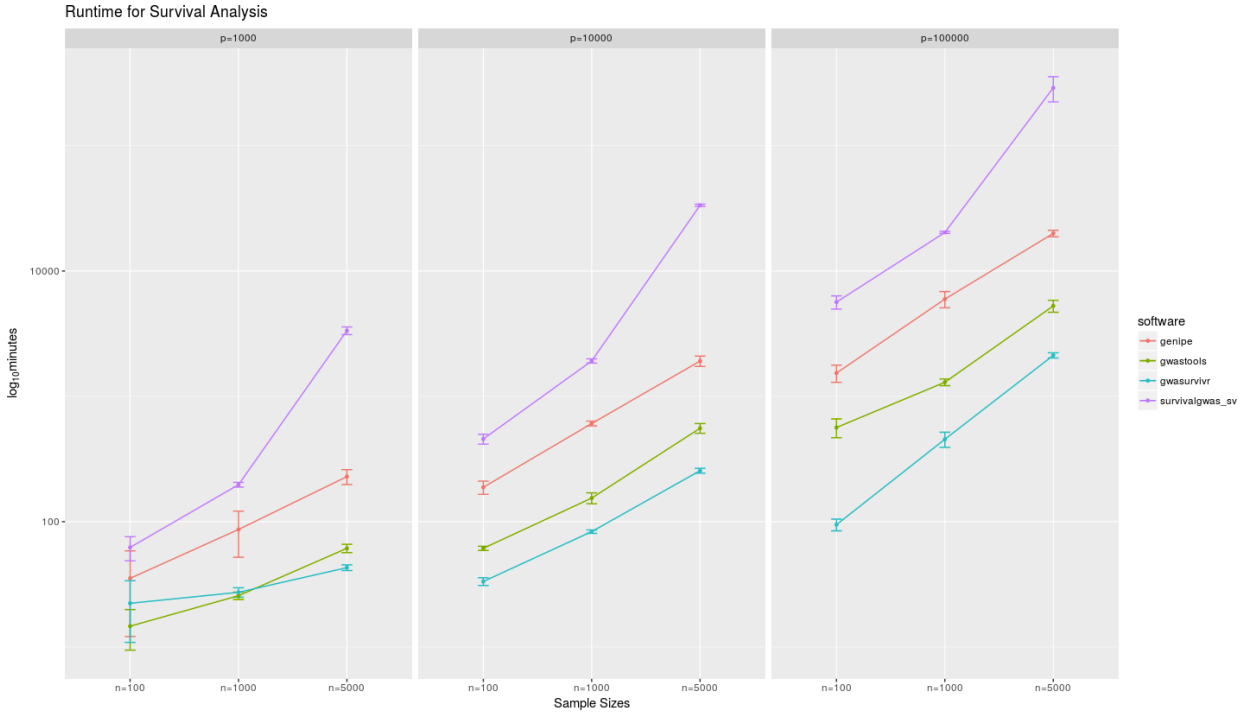
Figure 1. Run time for survival analyses. The x axis shows the three sample sizes (n=100, 1000, 5000) tested with SNP set sizes listed across the top (p=1000, 10000 and 100000). The y axis is the total run time in $\log_{10}$minutes. The mean run time and 95% confidence intervals for each n and p combination are show for genipe (red), GWAStools (green), survivalgwas_sv (purple) and gwasurvivr (cyan). Each of the nine simulations were run in triplicate with identical CPU constraints (1 node, 8 cores).