Review and testing of gwasurvivr: an R package for genome-wide survival analysis

Nadia Ivy Gabriel Piotr

2023-10-05

Introduction

Survival analysis has an important place in biomedical research, facilitating the exploration of time-to-event outcomes such as mortality or relapse.

An essential component in exploring the genetic basis of diseases involves investigating Single Nucleotide Polymorphisms (SNPs). These occasional variations in a single letter of DNA can have a significant impact on susceptibility to certain diseases or on response to medical treatments.

Integrating SNPs into survival analysis enables the discovery of genetic factors linked to time-to-event outcomes, shedding light on the genetic factors that influence disease progression and other critical events.

However, the major challenge is that our genome is made up of millions of SNPs, making large-scale survival analysis (GWAS) extremely complex. The existing software options for conducting such analyses are limited in several aspects (the need to interact with raw data, software not suited for survival analysis, and long execution times that hinder scalability). Consequently, researchers often face practical difficulties when conducting large-scale survival analyses.

gwasurvivr, an R/Bioconductor package was designed to surmount these challenges. This library offers a significant advancement in allowing researchers to perform survival analysis on large SNP datasets with remarkable efficiency and accuracy and with multiple file input formats such as VCF, IMPUTE2 or PLINK.

In this paper, we thoroughly examine the functionalities of gwasurvivr and offer an extensive evaluation of its operational mechanisms and effectiveness in unraveling the genetic factors that influence disease survival.

Methods

Datasets

Results

Michigan Imputation Server pre-phases typed genotypes using HAPI-UR, SHAPEIT, or EAGLE (default is EAGLE2), imputes using Minimac3 imputation engine and outputs Blocked GNU Zip Format VCF files (`.vcf.gz`). These `vcf.gz` files are used as input for `gwasurvivr`. Process of importing `vcf.gz` and phenotype files from `gwasurvivr` is shown below.

```r
vcf.file <- system.file(package="gwasurvivr",
                        "extdata",
                        "michigan.chr14.dose.vcf.gz")
pheno.fl <- system.file(package="gwasurvivr",
                        "extdata",
                        "simulated_pheno.txt")
```

The simulated phenotype file can be represented in the table.

ID_1	ID_2	event	time	age	DrugTxYes	sex	group
1	SAMP1	0	12.00	33.93	0	male	control
2	SAMP2	1	7.61	58.71	1	male	experimental
3	SAMP3	0	12.00	39.38	0	female	control
4	SAMP4	0	4.30	38.85	0	male	control
5	SAMP5	0	12.00	43.58	0	male	experimental
6	SAMP6	1	2.60	57.74	0	male	control

Now using phenotype file as covariate file and given VCF file we can run `michiganCoxSurv` wrapper for Cox regression model.

```r
#decoding sex into binary format
pheno.file$SexFemale <- ifelse(pheno.file$sex=="female", 1L, 0L)

#running Cox regression
michiganCoxSurv(vcf.file=vcf.file,
                covariate.file=pheno.file,
                id.column="ID_2",
                time.to.event="time",
                event="event",
                covariates=c("age", "SexFemale", "DrugTxYes"),
                inter.term=NULL, #interaction term inclusion
                print.covs="only", #defines printing of covariates' statistics
                out.file="michigan_only",
                r2.filter=0.3, #imputation quality score filter
                maf.filter=0.005, #filter for minor allele frequency
                chunk.size=100, #number of variants to proceed per thread
                verbose=F,
                clusterObj=NULL) #for setting up cluster for computations
```

Functions saves the outputed model with the .coxph extension as a seperate file as well as SNPs removed (pvalues below 0.05) in the .snps_removed file. Accessed results of the performed regression are showcased below (`print covs = "only"` was chosen as a printing option).

RSID	rs34919020	rs8005305	rs757545375
TYPED	FALSE	FALSE	FALSE
CHR	14	14	14
POS	19459185	20095842	20097287
REF	C	G	A
ALT	T	T	G
AF	0.301263	0.514583	0.519787
MAF	0.301263	0.485417	0.480213
SAMP_FREQ_ALT	0.3428	0.5022	0.5110
SAMP_MAF	0.3428	0.4978	0.4890
R2	0.551952	0.479015	0.480693
ER2	NA	NA	NA
PVALUE	0.2934544	0.3238959	0.2862329
HR	1.5085220	0.7233560	0.7046073
HR_lowerCI	0.7005469	0.3801063	0.3702421
HR_upperCI	3.248374	1.376573	1.340937
Z	1.0505737	-0.9864835	-1.0664221
COEF	0.4111304	-0.3238538	-0.3501147
SE.COEF	0.3913389	0.3282911	0.3283078
N	100	100	100
N.EVENT	42	42	42

To decipher most column names and extract knowledge from the output table in the appendix section explains the meaning of certain variables.

Conclusions

Appendix

RSID	SNP ID
TYPED	Imputation status: TRUE (SNP IS TYPED)/FALSE (SNP IS IMPUTED)
CHR	Chromosome number
POS	Genomic Position (BP)
REF	Reference Allele
ALT	Alternate Allele
AF	Minimac3 output Alternate Allele Frequency
MAF	Minimac3 output of Minor Allele Frequency
SAMP_FREQ_ALT	Alternate Allele frequency in sample being tested
SAMP_MAF	Minor allele frequency in sample being tested
R2	Imputation R2 score (minimac3 R^2)
ER2	Minimac3 ouput empirical R^2
PVALUE	P-value of single SNP or interaction term
HR	Hazard Ratio (HR)
HR_lowerCI	Lower bound 95% CI of HR
HR_upperCI	Upper bound 95% CI of HR
COEF	Estimated coefficient of SNP
SE.COEF	Standard error of coefficient estimate
Z	Z-statistic
N	Number of individuals in sample being tested
NEVENT	Number of events that occurred in sample being tested