Paper Report

gwasurvivr: an R package for genome-wide survival analysis

Table of contents

01

Introduction

Motivation and Background

02

Methods

Theory and Algorithms, Cox proportional hazard model

03

Datasets

Applicability and type of data

04

Results

Test of the R-package, positive and negative aspects

05

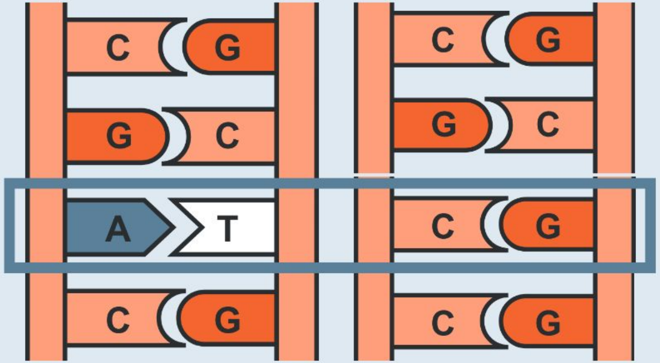
Conclusions

Final comments

01

Introduction

Background & Motivation



SINGLE NUCLEOTIDE POLYMORPHISM

Substitution of a single nucleotide at a specific position in the genome

SURVIVAL TIME ANALYSIS

time until an event occurs

GWASURVIVR

02

Methods

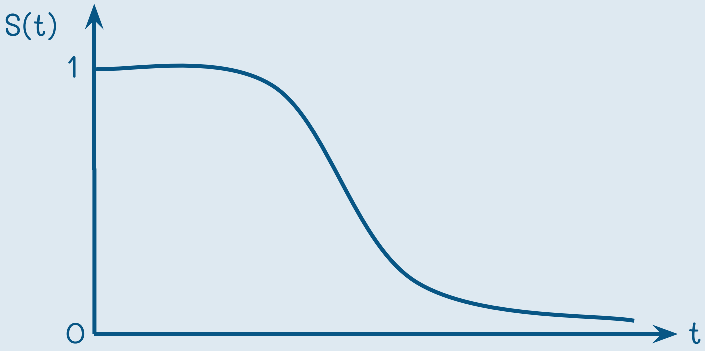
Theory and Algorithms : Survival Analysis and Cox proportional hazard model

Survival Analysis

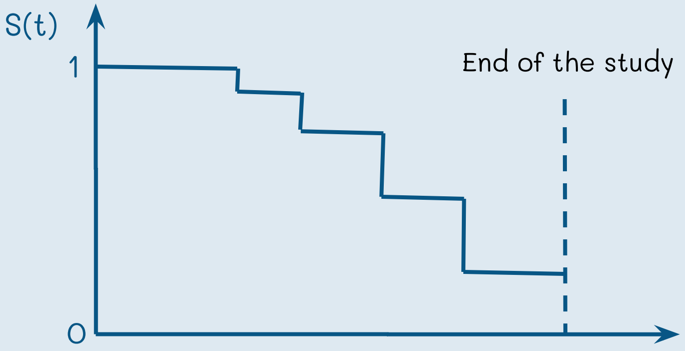
"analyse time-to-event data, i.e. estimate the time until an event occurs"

Hazard Function h(t) :
instantaneous potential at time t for getting the event, given survival up to time t

Survival Curve S(t) :
probability of the survival time to be greater or equal to t



Theoretical S(t)



Empirical S(t), Kaplan-Meier Curve

Cox Proportional Hazard Model

“Hazard function depending on time <u>and</u> others factors (covariates)”

$$h(t) = h_0(t) * e^{\sum x_i * \beta_i}$$

- $h_0(t)$: baseline hazard function depending only on the time
- X_i : factor i (do not depend on time)
- β_i : coefficient associated with X_i

How to estimate βi’s ? - - **The partial log-likelihood**

Maximizing the log-likelihood

=

Maximizing the probability of observing what we observed

GWASURVIVR : the trick

“When conducting survival analyses with million of SNPs the optimization of the partial log-likelihood takes a lot of time.”

1. Fit the Cox proportional hazard model with all the non-genetic covariates

2. Use those estimate parameters as initial points for fitting the model with the SNP covariate

→ **Great gain of time**

03

Dataset

Data formats, needed information

2 datasets are needed as input

SNP file

The SNP file contains the observed SNPs in the sample; it can be in 4 different formats (gds, bed, vcf, impute2), while vcf corresponds to files from the Michigan or the Sanger Imputations Server

Covariate data

A data that contains the express phenotypes (like sex, age, height) and covariates of the individuals in the SNP file

04

Results

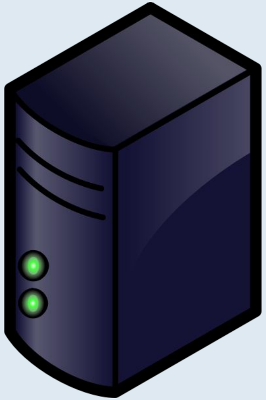
Use cases, performance, scalability, parallelization potential

Computational runtime simulation

Gwasurivr's performance was compared with the existing tools gnipe, GWASTools and SurvivalGWAS_SV

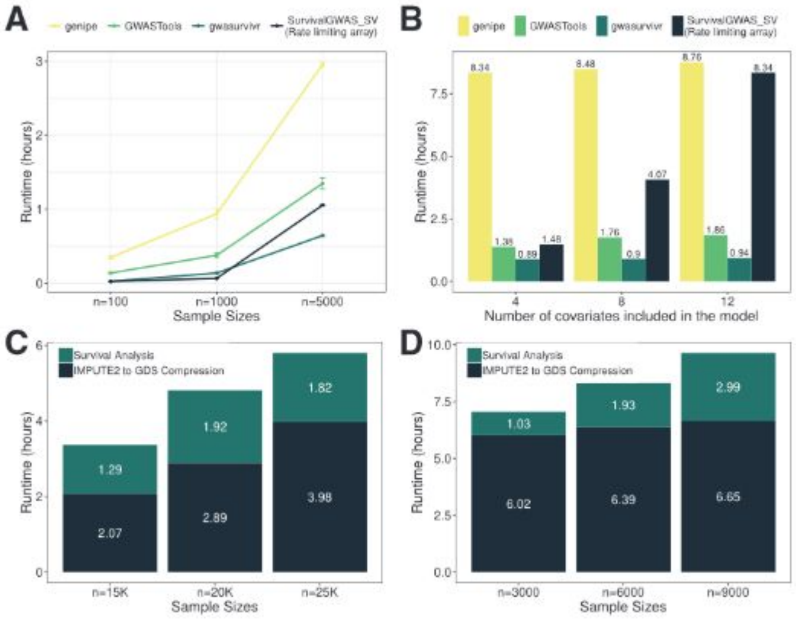
The **parameters** varied in the execution are:

- Number of covariates (4, 8 or 12)
- Number of samples (3000, 6000, 9000)



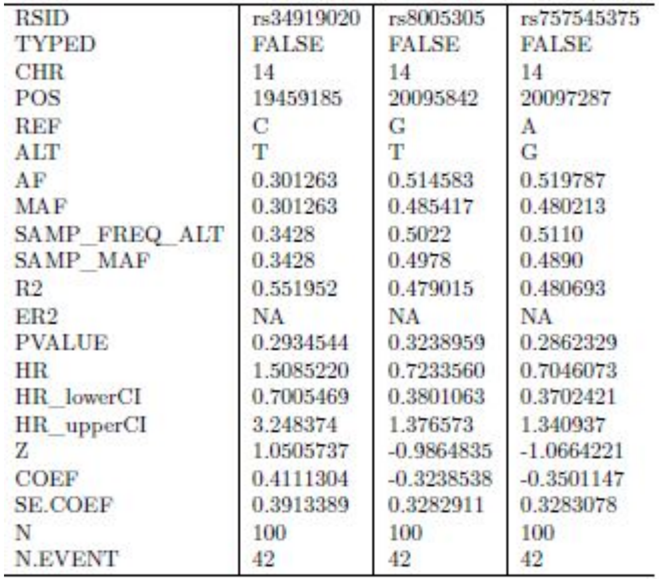
The **benchmarking** was executed with IMPUTE2 file format

Benchmarking

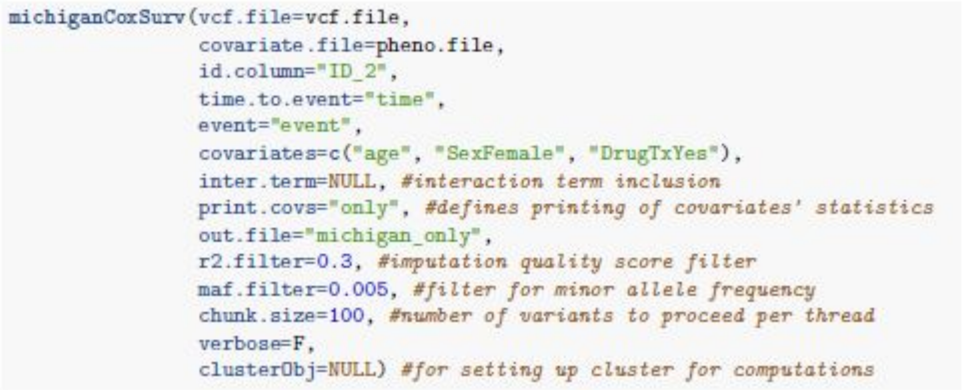


Gwasurivr uses data subsetting, CPU parallelization and cluster environment to get ahead over its competition, greatly reducing runtime of survival analysis.

Use cases and testing



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

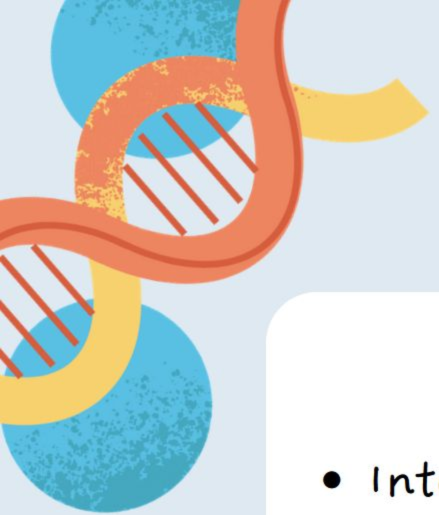


```
michiganCoxSurv(vcf.file=vcf.file,
                covariate.file=pheno.file,
                id.column="ID_2",
                time.to.event="time",
                event="event",
                covariates=c("age", "SexFemale", "DrugTxYes"),
                inter.term=NULL, #interaction term inclusion
                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
```

"Straightforward R syntax and uses cases described in the vignette make the package user-friendly."

05

Conclusions



GWASURVIVR

+

- Integrates GWAS results with survival analysis
- Fast
- Flexible
- Accurate
- Scalable

−

- Hard to integrate with other software
- No visualization tools