Weekly reports during Research Assistant

Zhang Leyi

Supervisor: Doug Speed

2023-11

Contents

1 Week 1

1.1 Comparison among MegaPRS, QuickPRS, ClassicalPRS

When the phenotypes are standardized, heritability and correlation follows:

$E[h^2] = R^2 = Cor^2(exp, obs)$

1.2 Plan

1. how to implement –make-pheno, by R or python
2. MegaPRS document
3. Summary Statistics script standardization

2 Week 2

2.1 –make-pheno Implementation

Shown in the following link: Link to –make-pheno Report

And the Python code is shown in the following link: Link to MakePheno.py

2.2 MegaPRS Implementation

Shown in the following link: Link to MegaPRS Report

To see the code via: Link to MegaPRS main.py

2.3 Comments

Main idea is to

1 - show we have a simpler version of Megaprs

2 - promote megaprs!

Pick 100 traits!

UKBB ICD10, PGC, MVP, FinnGen, Japan Biobank etc (e.g., Takiy used 109 phenotypes in GBAT paper)

Compare Old MegaPRS and Simplified MegaPRS and QuickPRS for each trait

New MegaPRS skips (*d*) in MegaPRS latex (SumHer step) - includes estimation of h2 in –mega-prs step

Also, megaprs can use elastic net (instead of bolt / bayesr / lasso)

New MegaPRS has a function to format summary statistics

Can we remove highld regions?

New MegaPRS can combine PRS?

Compare to other methods? LDpred / PRS-CS

Cross ancestry

Here is a relevant paper

https://www.medrxiv.org/content/10.1101/2023.11.20.23298215v1

Look at the high ld regions - do they tend to have much higher ld scores?

(the column "tagging" of the .tagging file gives the ld scores)

3 Week 3

3.1 New MegaPRS test

Tested on phenotype height.

$Score.cor = 0.59$

3.2 Liftover

1. get FinnGen summary statistics downloaded, 100 files.
2. QC
3. title -> notitle

4. make .bed

5. convert hg38 to hg19, by using liftover

About QC, thanks to Takiy's script, 20M snps been reduced to 8M.
Question about converting, is it okay to keep only 4 columns (chr, pos_start, pos_end, rsid))? Or should I keep more columns?
Guess the next step would be: chr:pos_end to coordinate with the FinnGen SS file?

3.3 Using FinnGen SS

1. Genotyping, to reduce $\#SNPs$ by using $LDAK\_thin$ tagging file.

3.4 Comments

Aim 1 - find some traits that are in both finn gen and ukbb

https://www.medrxiv.org/content/10.1101/2023.11.20.23298215v1.full.pdf

Aim 2 - how can we compute the SD of the accuracy of a PRS?
Assuming A as the accuracy of PRS ($\hat{Y} - Y$).

$$\bar{A} = \frac{1}{n}\sum A_i$$
$$SD = \sqrt{\frac{1}{n}\sum (A_i - \bar{A})^2}$$

3 ECTS from journal club

https://qgg.au.dk/en/education/qgg-phd-programme

4. How to solve the non-iid problems?

4 Week 4

4.1 Tasks done

1. 100 traits converting to hg19.

4.2 Liftover and make prediction model

4.2.1 Error Occuring

After using liftover, I got the FinnGen SS under hg19 assembly.
I compared the SNPs between FinnGen SS and the geno2.bim file. G6_HEADACHE as an example, in extract file: 517049 snps, in exclude file: 111647 snps, in the tagging file: 577460 snps, and in geno2.bim file: 628695 snps. While by running Megaprs_new for prediction model, I got the error: In total, 530 predictors have inconsistent alleles.

I checked the geno2.bim file for inconsistency: duplicated chr:pos, and alleles that in the situation : A1 == 'A' and A2 != 'T'. I deleted these, and got a new bim file geno2_v2. By running plink –make-bed, got new plink files. I ran highld and cor_matrix again, and used all these to make a new prediction model.

Error: 536 predictors have inconsistent alleles.

4.2.2 Solution

SummaryStatistics Column A1 == Reference Panel Column A1 and
SummaryStatistics Column A2 == Reference Panel Column A2 or

SummaryStatistics Column A1 == Reference Panel Column A2 and
SummaryStatistics Column A1 == Reference Panel Column A1,

4.3 Run 100 SS on Sumher and Prediction Model of New Mega

done.

4.4 To find UKBB phenotypes corresponding to FinnGen

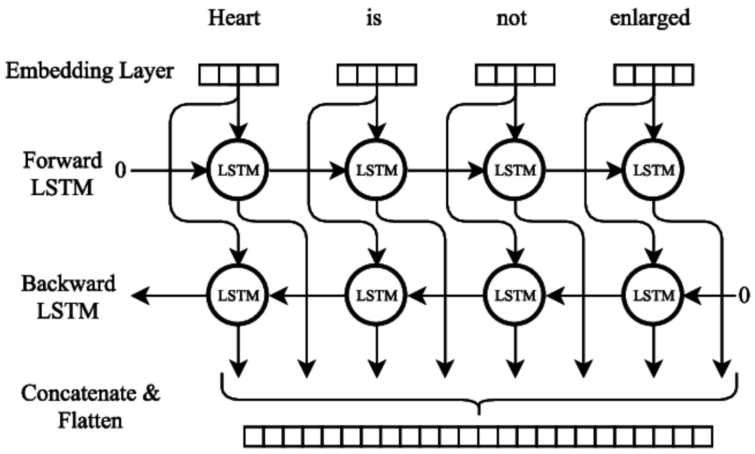
Names in FinnGen hardly coordinate with UKBB.

4.5 About non-iid data analysis (for pig data)

If the SNPs have high ld in every region, the basic Gaussian distribution with the Law of large numbers may not be harmonious with it. When I was doing my Bachelor's graduation research, I did something about Sentiment Analysis.

In NLP, the words do not occur independently, since there is a rule to follow (Part of Word, Syntax, Named Entity, Context, etcs). About a word in a sentence, we assume the words showing previously and forwardly would impact the occurrence of this word. That's why the sequence of words is non-iid data.

I used Bi-LSTM model to predict the sentiment :), :(, :| Bi-LSTM is similar to RNN, while the improvement is that it can process on long sequences, RNN can only perform on short sequences.



(a)

Figure 1: **The workflow of Bi-LSTM model.**

Besides Bi-LSTM, the Transformer is a mainstream model to perform such a Time Series as well.

In general, RNN, BiLSTM, and Transformer are used for time series, while genotyping data is normally seen as static data. However, if genotype data is non-iid, we need to find some other models best.

Demo: Transformer_GWAS.py, LR_GWAS.py, RNN_PRS.py

References