

Weekly reports during Research Assistant

Zhang Leyi

Supervisor: Doug Speed

2023-11

Contents

1	Week 1	6
1.1	Comparison among MegaPRS, QuickPRS, ClassicalPRS	6
1.2	Plan	6
2	Week 2	6
2.1	-make-pheno Implementation	6
2.2	MegaPRS Implementation	6
2.3	Comments	6
3	Week 3	7
3.1	New MegaPRS test	7
3.2	Liftover	7
3.3	Using FinnGen SS	8
3.4	Comments	8
4	Week 4	9
4.1	Tasks done	9
4.2	Liftover and make prediction model	9
4.2.1	Error Occuring	9
4.2.2	Solution	9
4.3	Run 100 SS on Sumher and Prediction Model of New Mega	10

4.4	To find UKBB phenotypes corresponding to FinnGen	10
4.5	About non-iid data analysis (for pig data)	10
4.6	Comments	11
5	Week 5	12
5.1	finngen in icd10	12
5.2	non-iid	12
5.3	Python Read Plink	13
5.4	Summary Statistics changing format to LDAK	13
6	Week 6: Progress of Proj 1 and Proj 3	13
6.1	Proj1 Testing each PRS on UKBB phenotypes with FinnGen Summary Statistics . .	13
6.1.1	Data	13
6.1.2	Progress	14
6.2	Proj3 Converting the format of any Summary Statistics into LDAK format	14
6.2.1	Data	14
6.2.2	Progress	14
7	Week 7	15
7.1	About 100 phenos	15
7.2	Steps for running megaPRS on 2400 FinnGen SS and 1000 UKBB Pheno	15
7.3	Week 7: Summary and work after Presentation	15

7.4	Comments	15
7.5	Finnish Population reference panel	16
7.5.1	Process	16
7.6	FinnGen map to UKBB	16
8	Week 8	18
8.1	MAF Comparison of 33KG and 1000G	18
8.2	Description of MegaPRS	20
8.3	FinnGen and UKBB PRS, 33KG_FIN as reference panel	20
8.4	Discussion with Ole	20
9	Week 9	20
9.1	MAF Comparison of 1000G and 33KG	20
9.2	formatting	21
9.3	PRS calculation	21
9.4	megaprs code	21
9.5	Two tasks	22
10	Week 10 Summary	22
10.1	FinnGen -> UKBB	22
10.1.1	Data	22
10.1.2	Methods	23

10.1.3 Results	23
11 Week 11	24
11.1 Formatting	24
11.2 PRS tools comparison results	24
11.3 FinnGen to hg19	25
12 Week 12	25
12.1 Run LDpred2 and Lassosum in parallel	25
12.2 Compare PRS results of 4 tools	25
12.3 LDpred2 computational usage	26
12.4 About liability scale R^2 correlation	26
12.5 Plan	27
13 Week 13	27
13.1 SS with decent R2	27
13.2 About heterogeneity paper	27
13.3 Difference between prevalence and ascertainment	29
13.4 Plan for next week	29

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 h^2 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 *LDAC_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

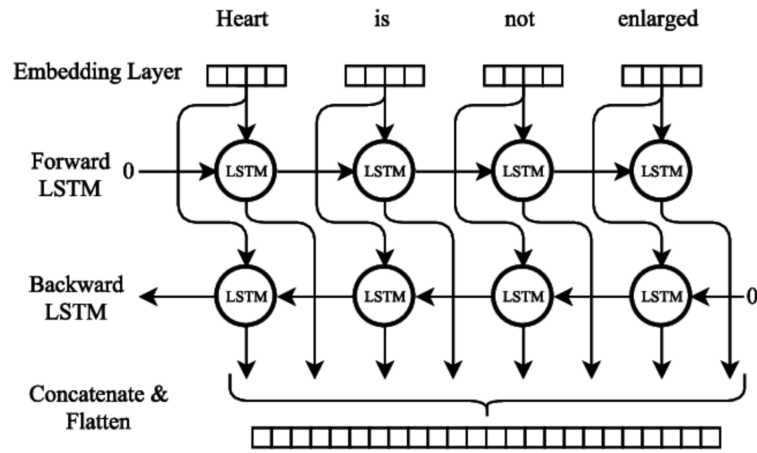
Names in FinnGen hardly coordinate with UKBB.

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, etc). 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

4.6 Comments

```
Here where you can find the PGC sumstats

/faststorage/project/dsmwpred/takiy/analysis8_gbat_pgc_traits/1_data/pgc_raw/

###

List of icd10 traits with prevalence > 0.005

/home/doug/snpher/faststorage/biobank/icd10/icdtraits.details

the column two index refers to files in

/home/doug/snpher/faststorage/biobank/newphens/icdphens/
```

```
###

there is some mhc data in

/home/doug/dsmwpred/ml

about 200k individuals and 160k snps
plus 100s of phenotypes

aim is to make use machine learning methods on to make prediction models for the phenotypes
```

5 Week 5

5.1 finngen in icd10

The subsequence of finngen in icd10 list.

finngen_subset_in_icd10.txt

The last column contains the row numbers in finngen.

Progress:

1. Got index in finngen and icd10
2. Copied icd10 phenotypes
3. Summary Statistics script standardization

Progress:

5.2 non-iid

1. By Transformer, as Time series.

See noniid(NLP) in Latex.

[Link to Transformer](#)

2. By Bayesian Neural Network, individual data.

5.3 Python Read Plink

See readPlink.py

I used the Python package "hail" to read Plink files and convert them into numpy matrix.

5.4 Summary Statistics changing format to LDK

see from [Link to Format Changing](#).

For the similarity between numerics, I used Euclidean Distance:

$$Dist(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

For the similarity where there is character, I used the Jaccard Similarity Coefficient:

$$J(p, q) = \frac{|p \cap q|}{|p \cup q|}$$

6 Week 6: Progress of Proj 1 and Proj 3

Date: 2024/1/4

6.1 Proj1 Testing each PRS on UKBB phenotypes with FinnGen Summary Statistics

6.1.1 Data

100 Summary Statistics from FinnGen in hg38;
UKBB phenotypes

Problem: Cannot find the correct UKBB phenotype matching with FinnGen GWAS.

6.1.2 Progress

1. By Sequence matching, I've found 40 out of 100 UKBB with similar names to FinnGen.
2. But by running `-score`, the correlations were very low.
3. Question: Can we use UKBB phenotypes directly to make GWAS, and predict on UKBB as well (train and test)? There are 7,221 phenotypes in UKBB.

6.2 Proj3 Converting the format of any Summary Statistics into LDAK format

6.2.1 Data

Base GWAS: LDAK format

Target GWAS:

1. SNP info included but in a random order.
2. Some columns missed.
3. SNP info + score columns.

The score columns refer to p-value, logP, beta, sebeta, etc.

6.2.2 Progress

1. Combine LDAK format + some score columns (for Z-score).
2. Order.
3. Similarity + Jaccard Similarity.
4. `Cor()`.
4. Maybe use Machine Learning methods? like k-means cluster. But here comes a problem, since it's based on prediction, it can cause problems.

7 Week 7

7.1 About 100 phenos

Added file: `finngen_ss_add_n` to `proj1`.

Function: in FinnGen SS there is no column N, while the file `R8_manifest` contains `n_cas` and `n_con`. So I added them together.

7.2 Steps for running megaPRS on 2400 FinnGen SS and 1000 UKBB Pheno

1. Download (`wget`) 2400 FinnGen SS and unzip
2. Get a list including the directory of 1000 UKBB
3. Convert 2400 FinnGen SS into LDAK format, by 2 codes
(`Proj1: finngen_ss_add_n.R` to get `N_cas + N_con` and `Proj3: ss_to_ldak_format.R` to convert into LDAK format)
4. Run 2 steps in MegaPRS to get every prediction model.
5. Calc-score, without using `-pheno`
6. jackknife, and include `-pheno`, for faster calculation.

7.3 Week 7: Summary and work after Presentation

2024/1/19

7.4 Comments

1. About Z-score:

Try not to use p-value, since there would be very small numbers, $\log(P)$ could be an alternative value.

Use `qt()` instead `qnorm()`

Note that this is technically correct, but will have no impact (because a t distribution with > 100 parameters is almost identical to a normal) and will make coding harder - so please ignore

2. About FG -> UKBB Use the ICD-10 code to get the mapping directly.

7.5 Finnish Population reference panel

See the link to 33000 Genomes: <https://drive.google.com/drive/folders/1jXFwhTWoPpFMHkMOUHbvVju7IYckW>

This data contains 33000 individuals from many countries, and FIN is in the 12th column.

Details are in [Link to Week7.md](#)

7.5.1 Process

In this directory: `/faststorage/project/dsmwpred/zly/RA/data/33KG`

1. Download (wget) the full genome data.
2. Extract the 12th column out, which are 0/1/2 for 3529 individuals, and the format is called sp format.

See this link: <https://doudspeed.com/file-formats/>

3. Make a sudo fam file.
4. Get .bim file.
5. `-make-bed`

And wait for the map in finngen -> ukbb, then make the prs model and prediction.

7.6 FinnGen map to UKBB

By the ICD-10 code, and the result is in the following file:

```
/faststorage/project/dsmwpred/zly/RA/proj1_testprs_finngen_ukbb/data/finngen_icd10/  
finngen_ukbb_mapping.txt
```


In the FinnGen Excel sheet, each summary statistics has a certain ICD 10 code; and in the UKBB table, each phenotype has a range of ICD 10 codes (StartCode to EndCode), in Figure 2. So after mapping, some FinnGen SS may map to the same UKBB phenotype.

Microsoft Excel - [Book1] - [Sheet1]														Comments		Sh			
File	Home	Insert	Draw	Page Layout	Formulas	Data	Review	View	Automate	Help									
Ribbon tabs: File, Home, Insert, Draw, Page Layout, Formulas, Data, Review, View, Automate, Help														Comments				Sh	
Font settings: Arial, 11, Bold, Italic, Underline, Color, Background Color, Text Color, Paragraph, Styles, Conditional Formatting, Table, Styles, Font, Editing, Add-ins, AutoFill, AutoFill Options, AutoFill Options																			
Clipboard: Copy, Paste, Format Painter, Fill, Fill Color, Fill Effects, Font Color, Font Effects, Paragraph, Styles, Conditional Formatting, Table, Styles, Font, Editing, Add-ins, AutoFill, AutoFill Options, AutoFill Options																			
Alignment: Left, Center, Right, Justify, Indent, Decrease Indent, Increase Indent, Wrap Text, Merge Cells, Unmerge Cells, Sort, Filter, Show/Hide Rows, Show/Hide Columns, Show/Hide Tables, Show/Hide Charts, Show/Hide Comments, Show/Hide Hyperlinks, Show/Hide Images, Show/Hide Objects, Show/Hide Shapes, Show/Hide Text Boxes, Show/Hide Tables, Show/Hide Charts, Show/Hide Comments, Show/Hide Hyperlinks, Show/Hide Images, Show/Hide Objects, Show/Hide Shapes, Show/Hide Text Boxes																			
Number: General, Number, Currency, Accounting, Date, Time, Percentage, Fraction, Scientific, Text, Custom, More Numbering, More Styles, More Font, More Paragraph, More Table, More Chart, More Comment, More Hyperlink, More Image, More Object, More Shape, More Text Box, More Table, More Chart, More Comment, More Hyperlink, More Image, More Object, More Shape, More Text Box																			
Formulas: AutoSum, Insert Function, Paste Formulas, Paste Values, Paste All, Paste Formulas and Values, Paste Formulas and Values and Styles, Paste Formulas and Values and Styles and Hyperlinks, Paste Formulas and Values and Styles and Hyperlinks and Images, Paste Formulas and Values and Styles and Hyperlinks and Images and Objects, Paste Formulas and Values and Styles and Hyperlinks and Images and Objects and Shapes, Paste Formulas and Values and Styles and Hyperlinks and Images and Objects and Shapes and Text Boxes																			
Data: Sort, Filter, Advanced Filter, Data Validation, Conditional Formatting, Table, PivotTable, PivotChart, PowerPivot, Data Model, Data Connections, Data Sources, Data Refresh, Data Load, Data Sync, Data Sync Settings, Data Sync Status, Data Sync History, Data Sync Log, Data Sync Settings, Data Sync Status, Data Sync History, Data Sync Log																			
Review: Undo, Redo, Font, Paragraph, Styles, Conditional Formatting, Table, Styles, Font, Editing, Add-ins, AutoFill, AutoFill Options, AutoFill Options																			
View: Show/Hide Rows, Show/Hide Columns, Show/Hide Tables, Show/Hide Charts, Show/Hide Comments, Show/Hide Hyperlinks, Show/Hide Images, Show/Hide Objects, Show/Hide Shapes, Show/Hide Text Boxes																			
Automate: Macros, Record Macro, Stop Recording, Run Macro, Stop Running Macro																			

(a)

ICD 10 code	Description	Other details
22.2	Intestinal infectious diseases	
23.1	Cholera	
24.1	Typhoid and paratyphoid fevers	
25.1	Other salmonella infections	
26.1	Shigellosis	
27.1	In other bacterial intestinal infections	
28.1	Amoebiasis	
29.1	Other protozoal intestinal diseases	
30.1	Cviral and other specified intestinal infections	
31.1	Diarrhoea and gastroenteritis of presumed infectious origin	
32.1	Tuberculosis	
33.1	Respiratory tuberculosis	
34.1	Respiratory tuberculosis, bacteriologically and histologically confirmed	
35.1	Respiratory tuberculosis, not confirmed bacteriologically or histologically	
36.1	Tuberculosis of nervous system	
37.1	Tuberculosis of other organs	
38.1	Miliary tuberculosis	
39.1	Certain zoonotic bacterial diseases, not elsewhere classified	
40.1	Plague	
41.1	Tularaemia	
42.1	Anthrax	
43.1	Brucellosis	
44.1	Glanders and melioidosis	
45.1	Battle-borne fevers	
46.1	Erysipelas	
47.1	Leptospirosis	
48.1	Other zoonotic bacterial diseases, not elsewhere classified	
49.1	Other bacterial diseases	
50.1	Leprosy (Hansen's disease)	
51.1	Infection due to other mycobacteria	
52.1	Listeriosis	
53.1	Tetanus neonatorum	

(b)

Figure 2: ICD 10 information presented in FinnGen and UKBB. (a) FinnGen sheet. (b) UKBB table.

And the result of the mapping is shown in Figure 3.

```

finngen_ukbb_mapping.txt
1  NAME Letter Number Symbol ukbb_Phen
2  ENTEROPATH_E_COLI A 040 26
3  EHEC A 043 26
4  CAMPYLOENTERITIS A 045 26
5  YERSINIAENTERITIS A 046 26
6  C DIFFICILE ENTEROCOLITIS A 047 26
7  BACT_INTEST_INFECTION_NOS A 04 23
8  IBD_CMV B 25 36
9  CMV_NOS B 25 36
10 IBD_TUBERCULOSIS A 1 23
11 AB1_INTESTINAL_INFECTIONS A 0 23
12 AB1_CHOLERA A 00 23
13 AB1_TYPHOIDPARATHYPHOID A 01 23
14 AB1_SALMONELLA_OTH A 02 23
15 AB1_SHIGELLOSIS A 03 23
16 AB1_BACT_INTEST_OTH A 0 23
17 AB1_AMOEBIASIS A 06 23
18 AB1_OTHER_PROTOZOAL A 07 23
19 AB1_VIRAL_OTHER_INTEST_INFECTIONS A 08 23
20 AB1_GASTROENTERITIS_NOS A 09 23
21 AB1_TUBERCULOSIS A 1 23
22 AB1_ZOONOTIC_BACTERIAL A 2 23
23 AB1_LEPROSY A 30 26
24 AB1_OTHER_MYCOBAC A 31 26
25 AB1_LISTERIOSIS A 32 26
26 AB1_TETANUS_NEONAT A 33 26
27 AB1_TETANUS_OBSTR A 34 26
28 AB1_TETAN_OTH A 35 26
29 AB1_DIBTHERIA A 36 26
30 AB1_WHOOPCOUGH A 37 26
31 AB1_SCARLET_FEVER A 38 26
32 AB1_MENINGOCOCCAL A 39 26
33 AB1_STREPTO_SEPSIS A 40 26
34 AB1_OTHER_SEPSIS A 41 26
35 AB1_STREPTOMYCOSE A 42 26

```

(a)

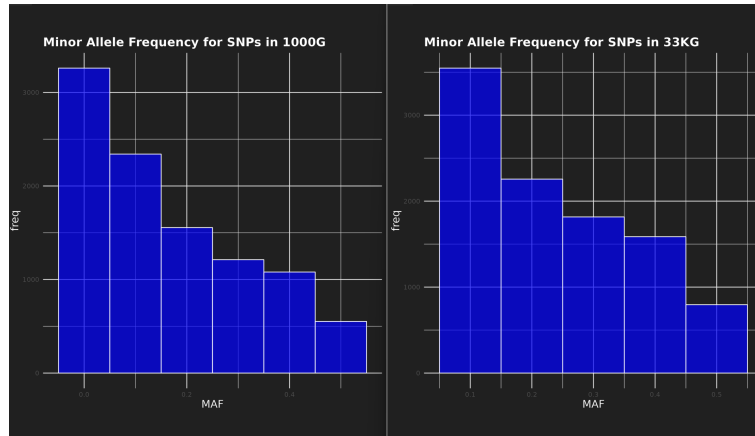
Figure 3: The result of FinnGen mapping to UKBB. ukbb_Phen presents the suffix number of UKBB phenotype in the phenotype folder.

3 - improve the description of megaprs See Week 8.

8 Week 8

8.1 MAF Comparison of 33KG and 1000G

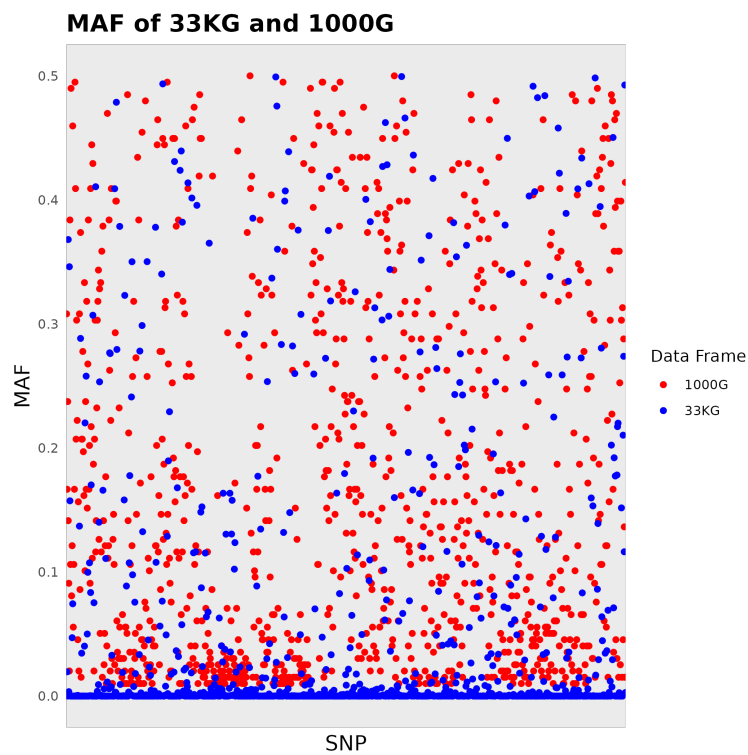
Since I've gotten the plink format of 33KG Finnish individuals, and there is data of 98 Finnish individuals from 1000G. I need to compare the MAF in between. In Figure 4, I deleted $MAF < 0.05$ and made the histogram.



(a)

Figure 4: This is the histogram of MAF between 33KG and 1000G.

And I compare the MAF of different consortiums at the same SNP position, Figure 5.



(a)

Figure 5: This is the comparison of MAF between 33KG and 1000G at the same position.

8.2 Description of MegaPRS

See the pdf report.

8.3 FinnGen and UKBB PRS, 33KG_FIN as reference panel

Finished.

Results are in the following directory:

```
/faststorage/project/dsmwpred/zly/RA/proj1_testprs_finngen_ukbb/  
fg_ukbb_33kg/megaprs_new/jackknife
```

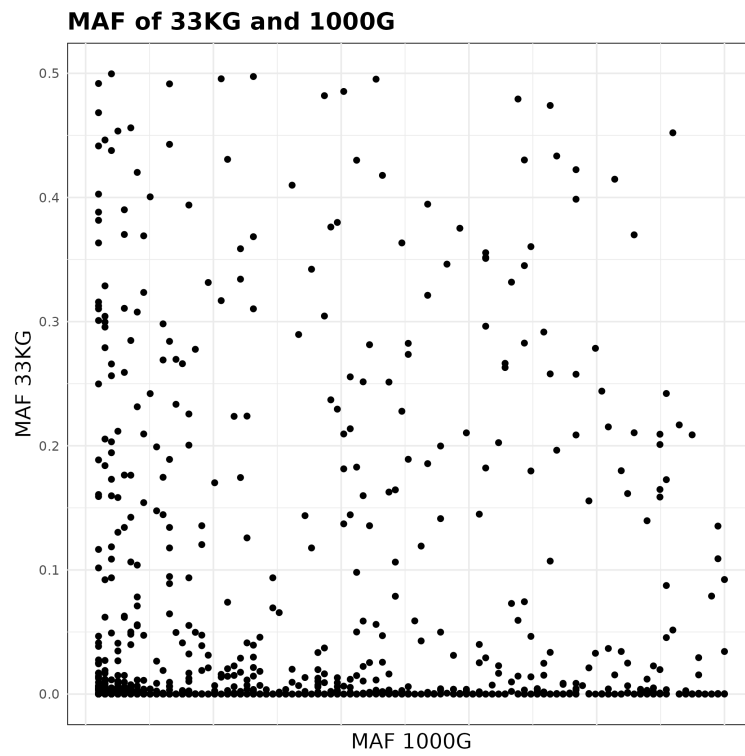
8.4 Discussion with Ole

1. the meaning of A1, A2 are different in different software. A1 can be either the effect allele or the alternative allele.
2. Is it necessary to make an automatic pipeline, to run everything all at once? My answer: No, it would decline the efficiency as we can reuse some files if we run MegaPRS on multiple traits.
3. Lassosum fits better in cross-ancestry populations than LDpred2.

9 Week 9

9.1 MAF Comparison of 1000G and 33KG

See in Figure [6](#).



(a)

Figure 6: This is the histogram of MAF between 33KG and 1000G.

9.2 formatting

Have done: FinnGen, Bolt-lmm-inf, LDAK format.

9.3 PRS calculation

Running

9.4 megaprs code

123

9.5 Two tasks

1 - a script / R function / python program to format summary statistics This should take as input the summary statistic file (eg downloaded from pgc, finnngen) and also the reference bim file

Then it should output summary statistics in ldak format, keeping only those snps consistent with the bim file

Ideally, it should be easy to run (eg so that me, or Caroline can use it)

2 - a list of X (eg 5, 10, 15) phenotypes where we have a good match between finnngen and ukbb, and decent prediction (eg correlation >0.1)

It might be that we cant find this many, but I currently think it should be possible

10 Week 10 Summary

10.1 FinnGen -> UKBB

10.1.1 Data

The UK Biobank genotype data comprises genetic information from approximately 500,000 participants ([Bycroft et al. \(2018\)](#)). This large cohort size provides significant statistical power for genetic analyses. After cleaning, there remains 392214 unrelated white population in UKBB, $MAF > 0.1$, with 628694 genotyped SNPs.

33000 Genomes Reference Panel (33KG) is a reference dataset containing genome sequence data from approximately 33,000 individuals from different populations. We extracted the Finnish population out for usage, with 32953 individuals and 24 million SNPs remaining.

FinnGen Summary Statistics refers to aggregated data and statistical summaries derived from the FinnGen project. FinnGen is a large-scale research effort in Finland aimed at utilizing genetic data to better understand diseases, their genetic underpinnings, and potential treatment avenues. The

summary statistics from FinnGen typically include information about genetic variants, their frequencies in the population, associations with various diseases or traits, and other relevant statistics. In our project, we are using version DF10, which consists of >412,000 individuals, > 21 M variants, and 2,408 disease endpoints.

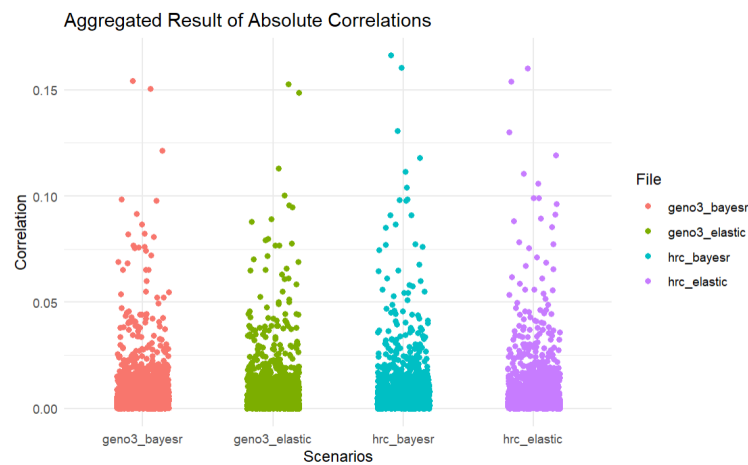
10.1.2 Methods

BayesR

Elastic

10.1.3 Results

The figure shows the absolute correlations of each trait, Fig. 7. In general, using Finnish as reference panel and using bayes R would be a good choice. However, most correlation values are below 2%, and only 10 traits got a correlation > 10%.



(a)

Figure 7: This is the Aggregated Result of Absolute Correlations.

11 Week 11

11.1 Formatting

Previously, I found that only 100K SNPs got matched between `hrc_fin` and `ss_ldak_format`. I've updated the code, and this is the result of the number of matched SNPs, Fig. 8.

```
[lezh@fe-open-02 ldak_format]$ awk 'NR==FNR {a[$2]; next} $1 in a {count++} END {print count}' /home/lezh/dsmwpred/data/ukbb/geno3.bim /faststorage/project/dsmwpred/zly/RA/proj1_testprs_finngen_ukbb/data/finngen_icd10/ldak_format/finngen_R10_I9_HYPTENS.ldak
622333
[lezh@fe-open-02 ldak_format]$ awk 'NR==FNR {a[$2]; next} $1 in a {count++} END {print count}' /faststorage/project/dsmwpred/zly/RA/data/33KG/fin/hrc_geno_fin.bim /faststorage/project/dsmwpred/zly/RA/proj1_testprs_finngen_ukbb/data/finngen_icd10/ldak_format/finngen_R10_I9_HYPTENS.ldak
7859570
```

(a)

Figure 8: This shows the number of matched SNPs.

It becomes correct now, with 600K SNPs found in `geno3`, and 7M SNPs found in `hrc_fin`.

11.2 PRS tools comparison results

I did 2 simple tests by using `height_ukbb` and `height_FinnGen` as summary statistics, then predicted and compared with *height.test*. Here is the R^2 result, Table 1.

Summary	Bayesr	Elastic	Lassosum	LDpred2
height ukbb	0.343	0.341	0.311	NULL
height fin	0.144	0.149	0.008	0.135

Table 1: This shows the R^2 result of prs comparison.

I need to see if someone gets distinct results from me in Lassosum or LDpred2.

11.3 FinnGen to hg19

I've updated all 2409 FinnGen summary statistic datasets into version hg19, see from the directory:

```
/faststorage/project/dsmwpred/zly/RA/proj1_testprs_finngen_ukbb/data/finngen_icd10/ss_liftover/hg19
```

12 Week 12

12.1 Run LDpred2 and Lassosum in parallel

See week 12 command.

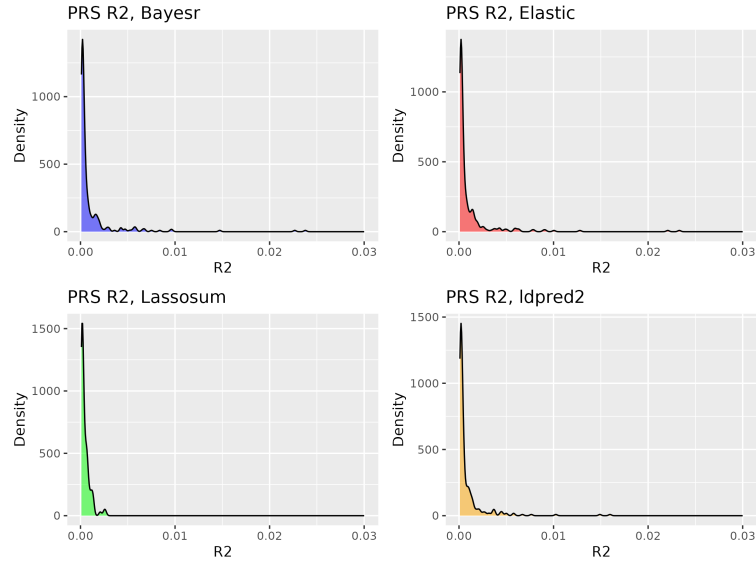
12.2 Compare PRS results of 4 tools

The four tools are Megaprs with Bayesr model, Megaprs with Elastic model, Ldpred2, and Lassosum.

I tested it based on 2409 FinnGen summary statistics and tested on UKBB phenotypes, and I used UKBB genotype data as the reference panel only. I will change the code to use both genotype data.

The R^2 results, in Figure 9, show that Megaprs with the Bayesr model performs the best, and Lassosum works the worst. I'll check if it's due to my Lassosum code being not robust.

It should be noted that none of the 2000 traits performs well, with the highest R^2 being lower than 3%. The correlation scores were calculated on the observation scale.



(a)

Figure 9: **This is the observation scale R^2 test of all matched phenotypes on 4 tools.**

12.3 LDpred2 computational usage

The calculating time varies from 1 hour to 3 hours: jobinfo 33762398 and 33762374.

The time was mostly spent calculating LD, and the time for calculating PRS was around 20 mins.

And the max memory usage was below 1.2 GB.

12.4 About liability scale R^2 correlation

Higher than observation scale R^2 , while bad performance. Consistent with the results by calc-score and square.

NEED to find out why. The profiles of LDpred2 are running now.

12.5 Plan

1. Get profiles of LDpred2 and Lassosum
2. liability scale R^2
3. use 14 ukbb continuous traits for PRS.

13 Week 13

13.1 SS with decent R2

I've made a table containing the liability scale R^2 above 3.5%, Fig. 10.

For some PGC summary statistics, I cannot find the corresponding phenotype. ADHD, ALZ2, Panic Anxiety.

SS	R2	Prevalence		R2_Obs	Software
finngen_R10_K11_MALABSORB	0.003148	0.003148	0.007409	0.059108	LDpred2
finngen_R10_T2D_WIDE	0.010307	0.010307	0.076828	0.03771	LDpred2
finngen_R10_M13_DUPUTRYEN	0.006715	0.006715	0.013118	0.071336	Bayesr
finngen_R10_M13_FIBROBLASTIC	0.00569	0.00569	0.014388	0.060448	Bayesr
finngen_R10_T2D_WIDE	0.014722	0.014722	0.071211	0.052606	Bayesr
finngen_R10_I9_HYPTENS	0.023779	0.023779	0.285773	0.042	Bayesr
finngen_R10_I9_HYPTENSESS	0.022663	0.022663	0.285773	0.040028	Bayesr
finngen_R10_I9_VARICVE	0.006796	0.006796	0.033275	0.039786	Bayesr
finngen_R10_E4_THYROID	0.009685	0.009685	0.065352	0.036489	Bayesr
finngen_R10_I9_AF	0.008948	0.008948	0.069493	0.034645	Elastic
PGC3_SCZ_wave3.european.autosome	0.001131	0.001131	0.001795	0.06137	Bayesr
daner_bip_pgc3_nm_noukbiobank	0.001284	0.001284	0.003363	0.041845	Elastic

(a)

Figure 10: The R^2 scores of FinnGen and PGC sumstats in good performance.

13.2 About heterogeneity paper

The paper talks about using PRS and diagnosis of finding comorbid, then using GWAS to find significant SNPs.

I tried to use Major depressive disorder (MDD) and schizophrenia (SCZ), but I didn't get a good prediction, Fig. 11. I

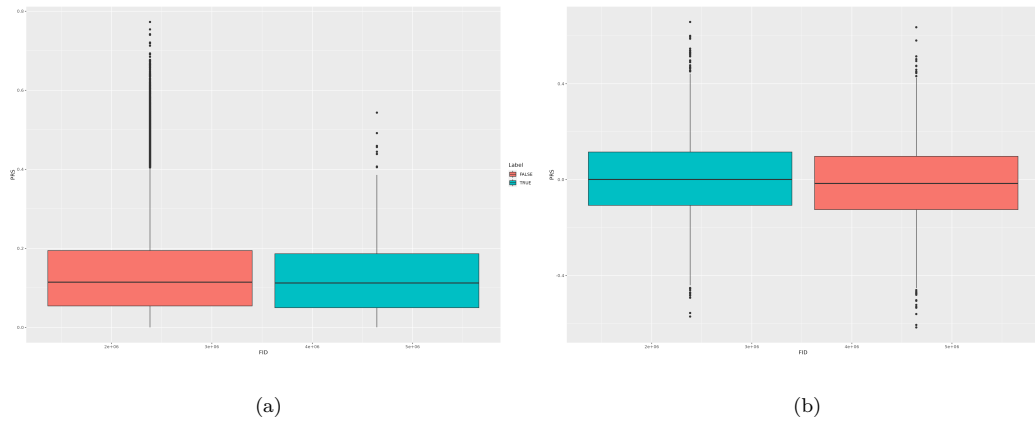


Figure 11: The PRS score of MDD, trying to find comorbid. (a) SCZ PRS scores of MDD diagnosis, (b) BMI PRS scores of MDD diagnosis

I then used Bipolar Disorder (BIP) to replace the use of MDD, and I got a better correlation result, Fig. 12.

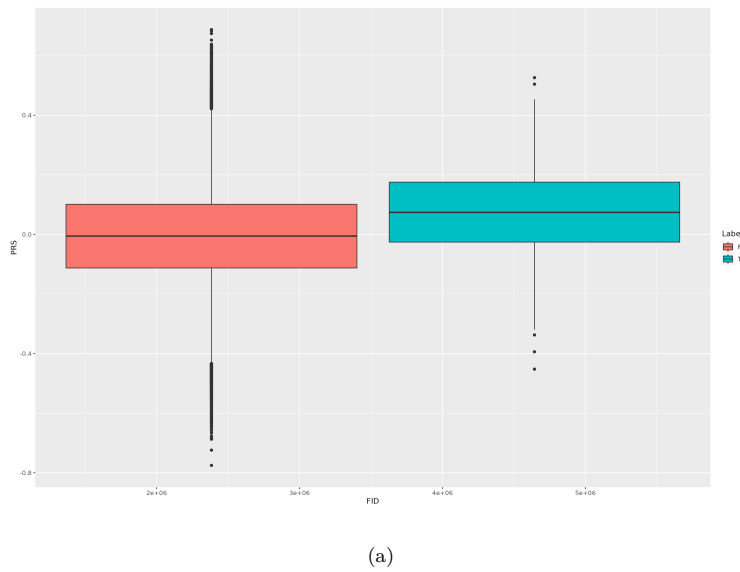


Figure 12: The PRS score of BIP, trying to find comorbid, based on SCZ PRS scores of BIP diagnosis

One possible hypothesis: cross-population comorbid study?

13.3 Difference between prevalence and ascertainment

13.4 Plan for next week

References

- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ... Marchini, J. (2018, 10). The uk biobank resource with deep phenotyping and genomic data. *Nature*, *562*, 203-209.
DOI: 10.1038/s41586-018-0579-z