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:
$E\left[h^2\right] = R^2 = Cor^2(exp, obs)$

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

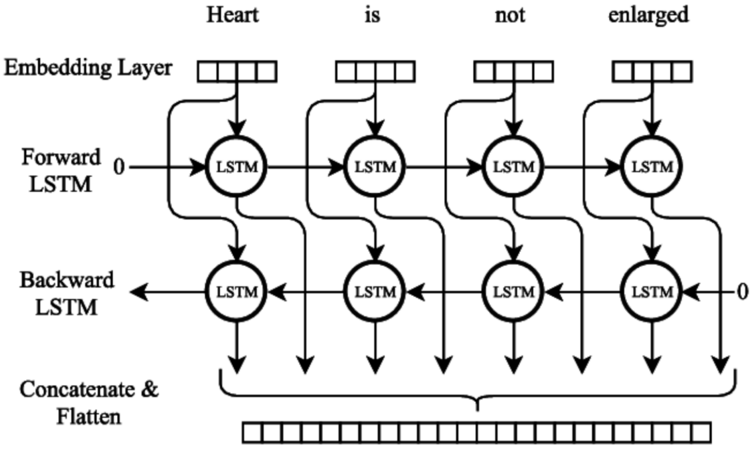
Names in FinnGen hardly coordinate with UKBB.

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

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 LDAK

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://dougspeed.com/file-formats/
3. Make a sudo fam file.
4. Get .bim file.
5. –make-bed

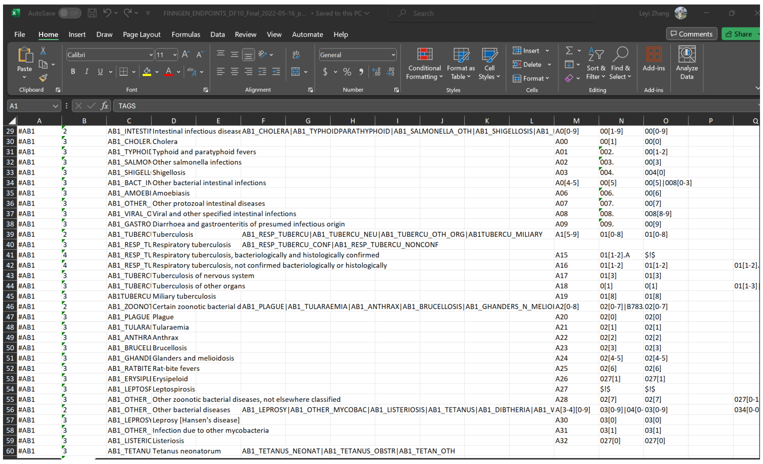
And wait for the map in finngen -> ukbb, then make the prs model and prediction.

7.6 FinnGen map to UKBB

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

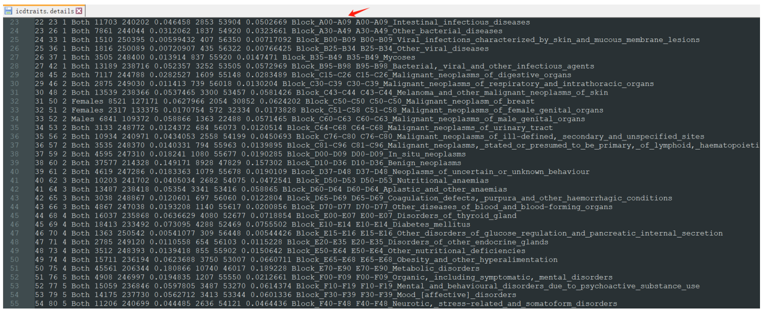
```
/faststorage/project/dsmwpred/zly/RA/proj1_testprs_finngen_ukbb/data/finngen_icd10/
                                                    finngen_ukbb_mapping.txt
```

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.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
29	#AB1	2	AB1_INTESTIN	Intestinal infectious disease		AB1_CHOLERA|AB1_TYPHOIDPARATHYPHOID|AB1_SALMONELLA_OTH|AB1_SHIGELLOSIS|AB1_I							A0[0-9]	00[1-9]	00[0-9]		
30	#AB1	3	AB1_CHOLER.	Cholera									A00	00[1]	00[0]		
31	#AB1	3	AB1_TYPHOIE	Typhoid and paratyphoid fevers									A01	002.	00[1-2]		
32	#AB1	3	AB1_SALMON	Other salmonella infections									A02	003.	00[3]		
33	#AB1	3	AB1_SHIGELL	Shigellosis									A03	004.	004[0]		
34	#AB1	3	AB1_BACT_IN	Other bacterial intestinal infections									A0[4-5]	00[5]	00[5]|008[0-3]		
35	#AB1	3	AB1_AMOEBI	Amoebiasis									A06	006.	00[6]		
36	#AB1	3	AB1_OTHER_	Other protozoal intestinal diseases									A07	007.	00[7]		
37	#AB1	3	AB1_VIRAL_C	Viral and other specified intestinal infections									A08	008.	008[8-9]		
38	#AB1	3	AB1_GASTRO	Diarrhoea and gastroenteritis of presumed infectious origin									A09	009.	00[9]		
39	#AB1	2	AB1_TUBERC	Tuberculosis		AB1_RESP_TUBERCU|AB1_TUBERCU_NEU|AB1_TUBERCU_OTH_ORG|AB1TUBERCU_MILIARY							A1[5-9]	01[0-8]	01[0-8]		
40	#AB1	3	AB1_RESP_TU	Respiratory tuberculosis		AB1_RESP_TUBERCU_CONF|AB1_RESP_TUBERCU_NONCONF											
41	#AB1	4	AB1_RESP_TU	Respiratory tuberculosis, bacteriologically and histologically confirmed									A15	01[1-2].A	$!$		
42	#AB1	4	AB1_RESP_TU	Respiratory tuberculosis, not confirmed bacteriologically or histologically									A16	01[1-2]	01[1-2]		01[1-2].
43	#AB1	3	AB1_TUBERC	Tuberculosis of nervous system									A17	01[3]	01[3]		
44	#AB1	3	AB1_TUBERC	Tuberculosis of other organs									A18	0[1]	0[1]		01[1-3]
45	#AB1	3	AB1TUBERCU	Miliary tuberculosis									A19	01[8]	01[8]		
46	#AB1	2	AB1_ZOONOT	Certain zoonotic bacterial d		AB1_PLAGUE|AB1_TULARAEMIA|AB1_ANTHRAX|AB1_BRUCELLOSIS|AB1_GHANDERS_N_MELIOI							A2[0-8]	02[0-7]|B783.	02[0-7]		
47	#AB1	3	AB1_PLAGUE	Plague									A20	02[0]	02[0]		
48	#AB1	3	AB1_TULARAI	Tularaemia									A21	02[1]	02[1]		
49	#AB1	3	AB1_ANTHRA	Anthrax									A22	02[2]	02[2]		
50	#AB1	3	AB1_BRUCELI	Brucellosis									A23	02[3]	02[3]		
51	#AB1	3	AB1_GHANDE	Glanders and melioidosis									A24	02[4-5]	02[4-5]		
52	#AB1	3	AB1_RATBITE	Rat-bite fevers									A25	02[6]	02[6]		
53	#AB1	3	AB1_ERYSIPLE	Erysipeloid									A26	027[1]	027[1]		
54	#AB1	3	AB1_LEPTOSF	Leptospirosis									A27	$!$	$!$		
55	#AB1	3	AB1_OTHER_	Other zoonotic bacterial diseases, not elsewhere classified									A28	02[7]	02[7]		027[0-1
56	#AB1	2	AB1_OTHER_	Other bacterial diseases		AB1_LEPROSY|AB1_OTHER_MYCOBAC|AB1_LISTERIOSIS|AB1_TETANUS|AB1_DIBTHERIA|AB1_V							A[3-4][0-9]	03[0-9]|04[0-	03[0-9]		034[0-0
57	#AB1	3	AB1_LEPROSY	Leprosy [Hansen's disease]									A30	03[0]	03[0]		
58	#AB1	3	AB1_OTHER_	Infection due to other mycobacteria									A31	03[1]	03[1]		
59	#AB1	3	AB1_LISTERIC	Listeriosis									A32	027[0]	027[0]		
60	#AB1	3	AB1_TETANU	Tetanus neonatorum		AB1_TETANUS_NEONAT|AB1_TETANUS_OBSTR|AB1_TETAN_OTH											

(a)

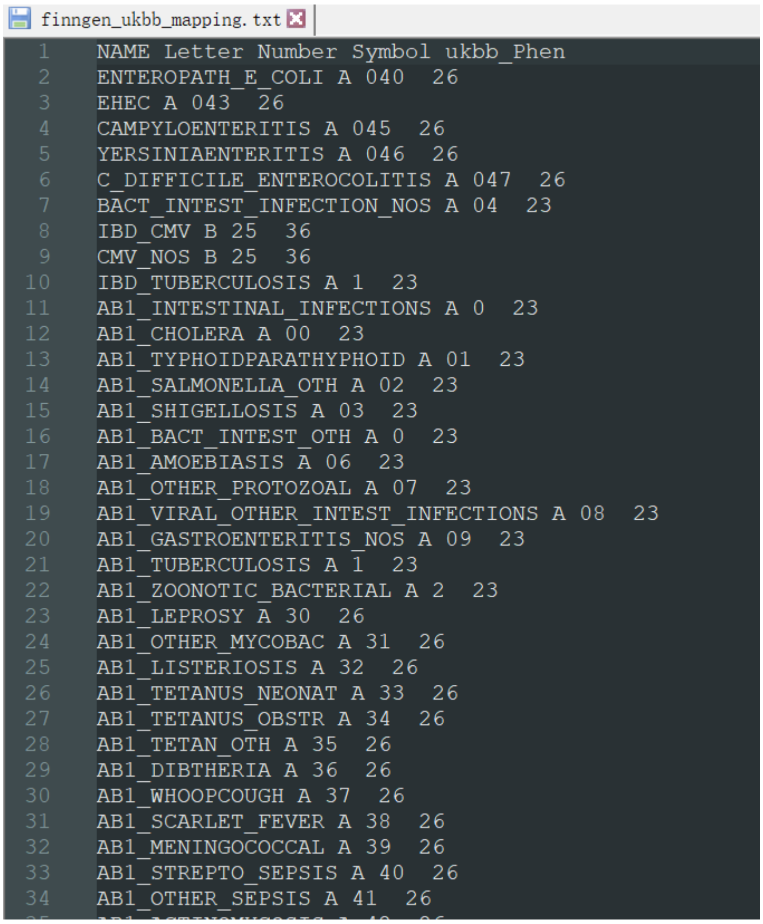


```
22 23 1 Both 11703 240202 0.046458 2853 53904 0.0502669 Block_A00-A09 A00-A09_Intestinal_infectious_diseases
23 26 1 Both 7861 244044 0.0312062 1837 54920 0.0323661 Block_A30-A49 A30-A49_Other_bacterial_diseases
24 33 1 Both 1510 250395 0.00599432 407 56350 0.00717092 Block_B00-B09 B00-B09_Viral_infections_characterized_by_skin_and_mucous_membrane_lesions
25 36 1 Both 1816 250089 0.00720907 435 56322 0.00766425 Block_B25-B34 B25-B34_Other_viral_diseases
26 37 1 Both 3505 248400 0.013914 837 55920 0.0147471 Block_B35-B49 B35-B49_Mycoses
27 42 1 Both 13189 238716 0.052357 3252 53505 0.0572969 Block_B95-B98 B95-B98_Bacterial,_viral_and_other_infectious_agents
28 45 2 Both 7117 244788 0.0282527 1609 55148 0.0283489 Block_C15-C26 C15-C26_Malignant_neoplasms_of_digestive_organs
29 46 2 Both 2875 249030 0.011413 739 56018 0.0130204 Block_C30-C39 C30-C39_Malignant_neoplasms_of_respiratory_and_intrathoracic_organs
30 48 2 Both 13539 238366 0.0537465 3300 53457 0.0581426 Block_C43-C44 C43-C44_Melanoma_and_other_malignant_neoplasms_of_skin
31 50 2 Females 8521 127171 0.0627966 2054 30852 0.0624202 Block_C50-C50 C50-C50_Malignant_neoplasm_of_breast
32 51 2 Females 2317 133375 0.0170754 572 32334 0.0173828 Block_C51-C58 C51-C58_Malignant_neoplasms_of_female_genital_organs
33 52 2 Males 6841 109372 0.058866 1363 22488 0.0571465 Block_C60-C63 C60-C63_Malignant_neoplasms_of_male_genital_organs
34 53 2 Both 3133 248772 0.0124372 684 56073 0.0120514 Block_C64-C68 C64-C68_Malignant_neoplasms_of_urinary_tract
35 56 2 Both 10934 240971 0.0434053 2558 54199 0.0450693 Block_C76-C80 C76-C80_Malignant_neoplasms_of_ill-defined,_secondary_and_unspecified_sites
36 57 2 Both 3535 248370 0.0140331 794 55963 0.0139895 Block_C81-C96 C81-C96_Malignant_neoplasms,_stated_or_presumed_to_be_primary,_of_lymphoid,_haematopoieti
37 59 2 Both 4595 247310 0.018241 1080 55677 0.0190285 Block_D00-D09 D00-D09_In_situ_neoplasms
38 60 2 Both 37577 214328 0.149171 8928 47829 0.157302 Block_D10-D36 D10-D36_Benign_neoplasms
39 61 2 Both 4619 247286 0.0183363 1079 55678 0.0190109 Block_D37-D48 D37-D48_Neoplasms_of_uncertain_or_unknown_behaviour
40 62 3 Both 10203 241702 0.0405034 2682 54075 0.0472541 Block_D50-D53 D50-D53_Nutritional_anaemias
41 64 3 Both 13487 238418 0.05354 3341 53416 0.058865 Block_D60-D64 D60-D64_Aplastic_and_other_anaemias
42 65 3 Both 3038 248867 0.0120601 697 56060 0.0122804 Block_D65-D69 D65-D69_Coagulation_defects,_purpura_and_other_haemorrhagic_conditions
43 66 3 Both 4867 247038 0.0193208 1140 55617 0.0200856 Block_D70-D77 D70-D77_Other_diseases_of_blood_and_blood-forming_organs
44 68 4 Both 16037 235868 0.0636629 4080 52677 0.0718854 Block_E00-E07 E00-E07_Disorders_of_thyroid_gland
45 69 4 Both 18413 233492 0.073095 4288 52469 0.0755502 Block_E10-E14 E10-E14_Diabetes_mellitus
46 70 4 Both 1363 250542 0.00541077 309 56448 0.00544426 Block_E15-E16 E15-E16_Other_disorders_of_glucose_regulation_and_pancreatic_internal_secretion
47 71 4 Both 2785 249120 0.0110558 654 56103 0.0115228 Block_E20-E35 E20-E35_Disorders_of_other_endocrine_glands
48 73 4 Both 3512 248393 0.0139418 855 55902 0.0150642 Block_E50-E64 E50-E64_Other_nutritional_deficiencies
49 74 4 Both 15711 236194 0.0623688 3750 53007 0.0660711 Block_E65-E68 E65-E68_Obesity_and_other_hyperalimentation
50 75 4 Both 45561 206344 0.180866 10740 46017 0.189228 Block_E70-E90 E70-E90_Metabolic_disorders
51 76 5 Both 4908 246997 0.0194835 1207 55550 0.0212661 Block_F00-F09 F00-F09_Organic,_including_symptomatic,_mental_disorders
52 77 5 Both 15059 236846 0.0597805 3487 53270 0.0614374 Block_F10-F19 F10-F19_Mental_and_behavioural_disorders_due_to_psychoactive_substance_use
53 79 5 Both 14175 237730 0.0562712 3413 53344 0.0601336 Block_F30-F39 F30-F39_Mood_[affective]_disorders
54 80 5 Both 11206 240699 0.044485 2636 54121 0.0464436 Block_F40-F48 F40-F48_Neurotic,_stress-related_and_somatoform_disorders
```

(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.



```
NAME Letter Number Symbol ukbb_Phen
ENTEROPATH_E_COLI A 040  26
EHEC A 043  26
CAMPYLOENTERITIS A 045  26
YERSINIAENTERITIS A 046  26
C_DIFFICILE_ENTEROCOLITIS A 047  26
BACT_INTEST_INFECTION_NOS A 04  23
IBD_CMV B 25  36
CMV_NOS B 25  36
IBD_TUBERCULOSIS A 1  23
AB1_INTESTINAL_INFECTIONS A 0  23
AB1_CHOLERA A 00  23
AB1_TYPHOIDPARATHYPHOID A 01  23
AB1_SALMONELLA_OTH A 02  23
AB1_SHIGELLOSIS A 03  23
AB1_BACT_INTEST_OTH A 0  23
AB1_AMOEBIASIS A 06  23
AB1_OTHER_PROTOZOAL A 07  23
AB1_VIRAL_OTHER_INTEST_INFECTIONS A 08  23
AB1_GASTROENTERITIS_NOS A 09  23
AB1_TUBERCULOSIS A 1  23
AB1_ZOONOTIC_BACTERIAL A 2  23
AB1_LEPROSY A 30  26
AB1_OTHER_MYCOBAC A 31  26
AB1_LISTERIOSIS A 32  26
AB1_TETANUS_NEONAT A 33  26
AB1_TETANUS_OBSTR A 34  26
AB1_TETAN_OTH A 35  26
AB1_DIBTHERIA A 36  26
AB1_WHOOPCOUGH A 37  26
AB1_SCARLET_FEVER A 38  26
AB1_MENINGOCOCCAL A 39  26
AB1_STREPTO_SEPSIS A 40  26
AB1_OTHER_SEPSIS A 41  26
```

(a)

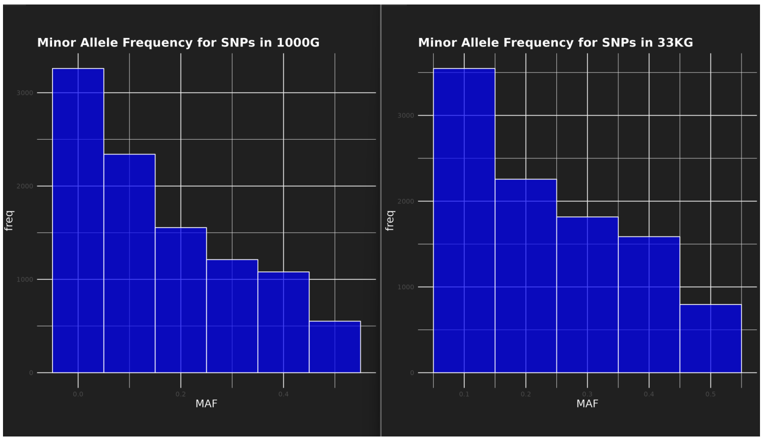
Figure 3: **The result of FinnGen mapping to UKBB.** ukbb_Phen presents the suffix number of UKBB phenotype in the phenotype folder.

3 - improve the description of megaprs See Week 8.

8 Week 8

8.1 MAF Comparison of 33KG and 1000G

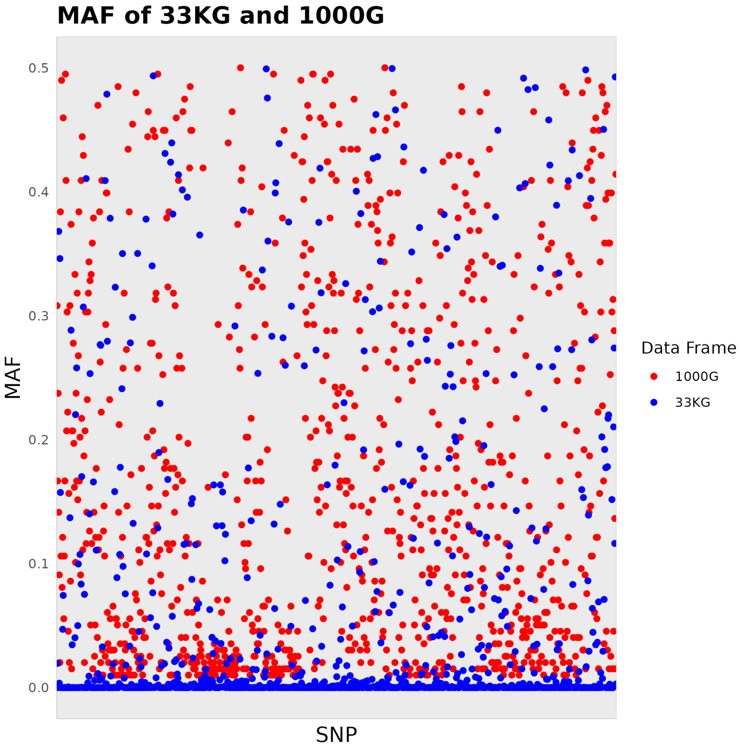
Since I've gotten the plink format of 33KG Finnish individuals, and there is data of 98 Finnish individuals from 1000G. I need to compare the MAF in between. In Figure 4, I deleted $MAF < 0.05$ and made the histogram.



(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.

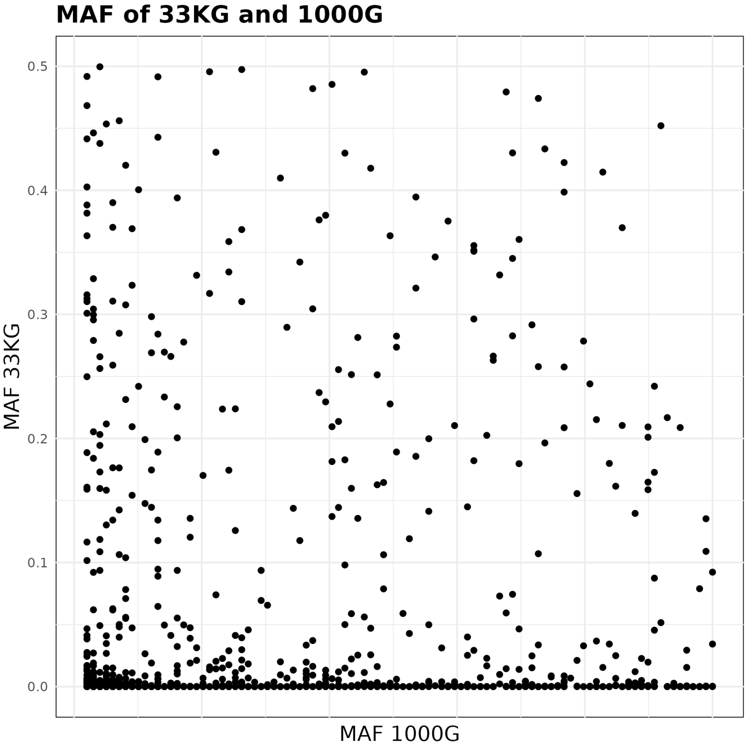
2. Is it necessary to make an automatic pipeline, to run everything all at once? My answer: No, it would decline the efficiency as we can reuse some files if we run MegaPRS on multiple traits.

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, finngenn) 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 finngen 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.

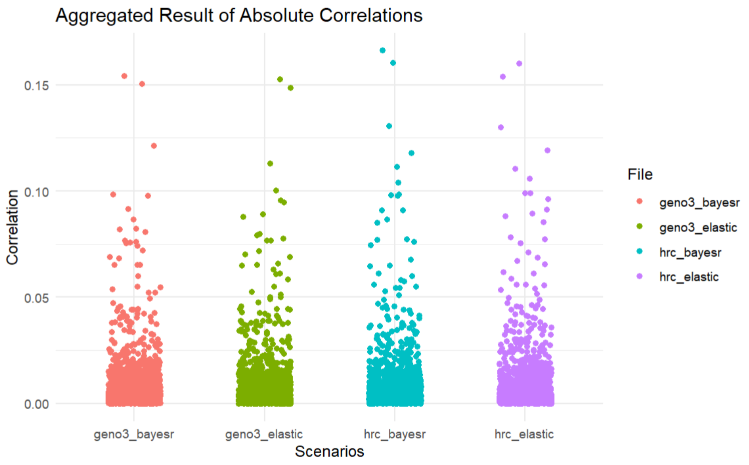
10.1.2 Methods

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%.



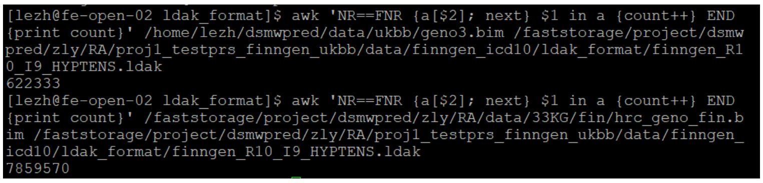
(a)

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} $1 in a {count++} END
{print count}' /home/lezh/dsmwpred/data/ukbb/geno3.bim /faststorage/project/dsmw
pred/zly/RA/proj1_testprs_finngen_ukbb/data/finngen_icd10/ldak_format/finngen_R1
0_I9_HYPTENS.ldak
622333
[lezh@fe-open-02 ldak_format]$ awk 'NR==FNR {a[$2]; next} $1 in a {count++} END
{print count}' /faststorage/project/dsmwpred/zly/RA/data/33KG/fin/hrc_geno_fin.b
im /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 result, Table 1.

Summary	Bayesr	Elastic	Lassosum	LDpred2
height ukbb	0.586	0.5847	0.311	NULL
height fin	0.38	0.387	0.008	0.166

Table 1: **This shows the 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
```

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