ANALYSIS OF GENOME-WIDE ASSOCIATION STUDIES

PRACTICAL FOR FOS COURSE **MOLECULAR DATA SCIENCE: FROM DISEASE MECHANISMS TO PERSONALIZED MEDICINE**, WEDNESDAY OCTOBER 24, 2018

FACULTY

Marian Beekman (Molecular Epidemiology)

Please send your answers to M.Beekman@lumc.nl with *FOS2018GWAS* as subject.

INTRODUCTION

The first genome-wide association study studies (GWAS) within the LUMC focused on human longevity and was performed by the researchers from Molecular Epidemiology, Gerontology and Geriatrics, and Medical Statistics. In the primary analysis, individuals older than 90 years from long-lived families from the Leiden Longevity Study were compared with controls, namely the partners of their children who represent 'normal' families. Particularly the genotyped partners aged around 60 years can also be used to perform GWAS tot detect associations with intermediate phenotypes. (Deelen *et al.* Aging Cell. 2011 Aug;10(4):686-98; Deelen *et al.* Hum Mol Genet. 2014 Aug 15;23(16):4420-32)

TODAY

This is what you will do in the current practical. From the 500k SNPs genotyped in the controls, you will analyze a small proportion for association with the intermediate phenotype plasma insulin level. Members of long-lived families have lower insulin levels, suggesting a role in ageing and high plasma insulin is indicative of insulin resistance, a pre-stage of type II diabetes.

BEFORE YOU START

1. Create a working folder, for example H:\FOS2018GWAS
2. Download from files from https://github.com/molepi/Molecular-Data-Science/tree/master/GWASday to your working folder.
3. Copy the following 4 software files from C:\plink to your working folder: *gPLINK-2.050.zip*, *plink-1.07-dos.zip*
4. Extract gPLINK-2.050.zip and plink-1.07-dos.zip to your working folder, so that all files are in your working folder.
5. To properly work with Plink, you need to adjust the Regional and Language options accessible though the Control panel (Region and Language, formats, additional settings, decimal symbol) so that the decimal character is . (dot) and the digit grouping character is a , (comma).

6. Also make sure that you do NOT 'Hide extensions for known file types' (in Windows Control Panel, Appearance and Personalization, Folder options \view).

7. You will do the statistical analysis using gPLINK. Everything you wish to know about this software including a description of all options can be found at http://zzz.bwh.harvard.edu/plink/ .

IT IS CRUCIAL THAT YOU PERFORM STEPS 4 AND 5. DOUBLE CHECK WHETHER YOU CORRECTLY PERFORMED STEP 4.

1. EXAMINING DATA

1. The .ped file contains the genotypes for all individuals. Open the file in Excel (which cannot show all columns of this file but that doesn't matter now). The genotypes are coded with the numbers 1, 2, 3 and 4, where 1=A, 2=C, 3=G and 4=T. Go to the Plink website and click on the left of the page on '5. Basic usage/data formats' to get information on the format of .ped files.

A) WHAT INFORMATION IS IN THE DIFFERENT COLUMNS? IN WHICH COLUMN DO THE GENOTYPES START? HOW MANY CASES AND HOW MANY CONTROLS ARE THERE IN THE FILE?

2. The .map file contains data on the SNPs that were measured. Open it in Excel and use the Plink website to find out what is in which column.

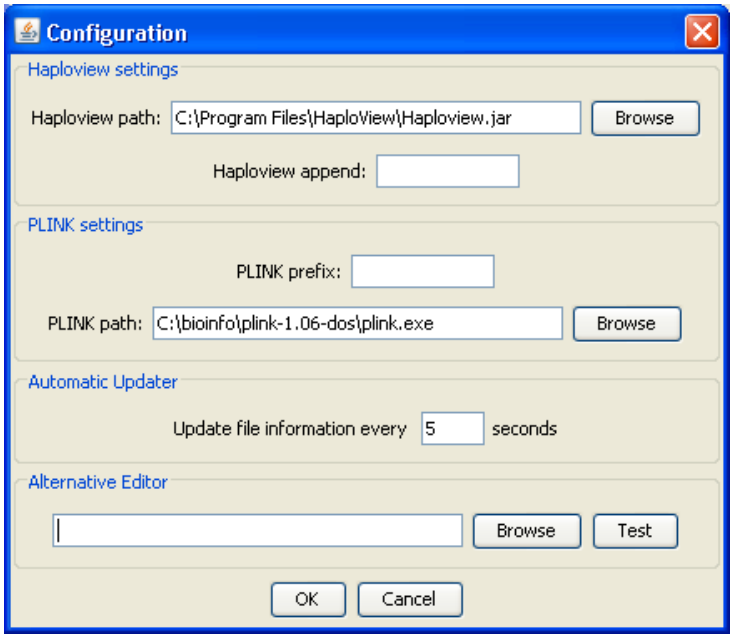
B) WHAT INFORMATION IS IN THE DIFFERENT COLUMNS?

C) THERE IS NO INFORMATION ON THE GENETIC DISTANCE OF THE SNPS. FOR WHAT KIND OF STUDY DESIGN WOULD THIS HAVE BEEN A PROBLEM?

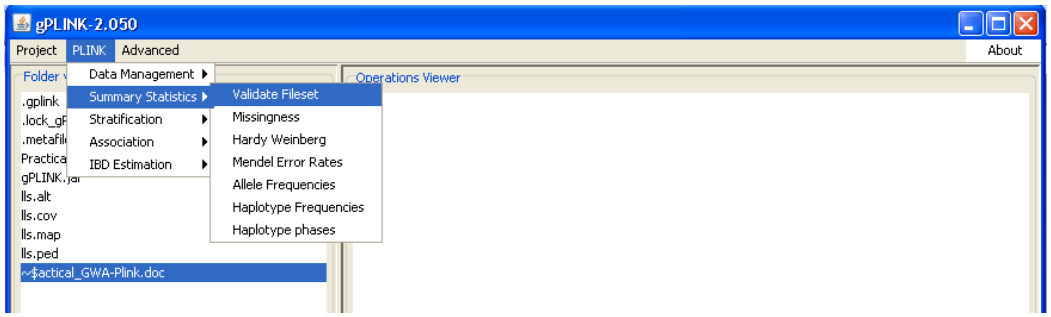
2. USING PLINK AND SOME CHECKS

1. Open Gplink by double clicking Gplink2.jar in your C:\Users\YOURNAME folder or a similar name. If that doesn't work, click-right on Gplink2.jar, 'open with' and select Java. You may want to create a shortcut to Gplink on your desktop.

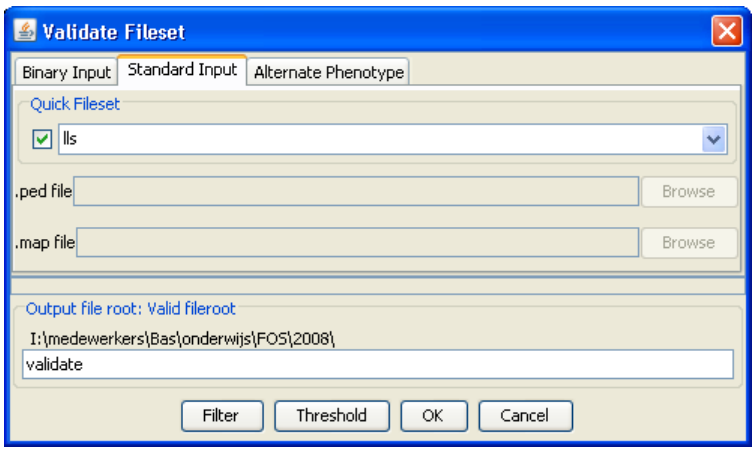
2. After starting the program, click Project\Open and tell Gplink where it can find the data files (namely in H:\FOS2018GWAS). Then click project\configure and tell Gplink where it can find haploview.jar (in H:\FOS2018GWAS \Haploview.jar) and Plink.exe (in H:\FOS2018GWAS\plink.exe). See below for a screen shot. Gplink is now ready to use!



3. First we will check whether everything works fine (Plink and data-files). This can be done using PLINK\Summary Statistics\Validate Fileset. See the next page for a screen shot.



4. After selecting this option, click the tab Standard Input. You will see that it recognized your lls-data (Quick Fileset). Enter a name for your output in the box Output file root, e.g. validate.

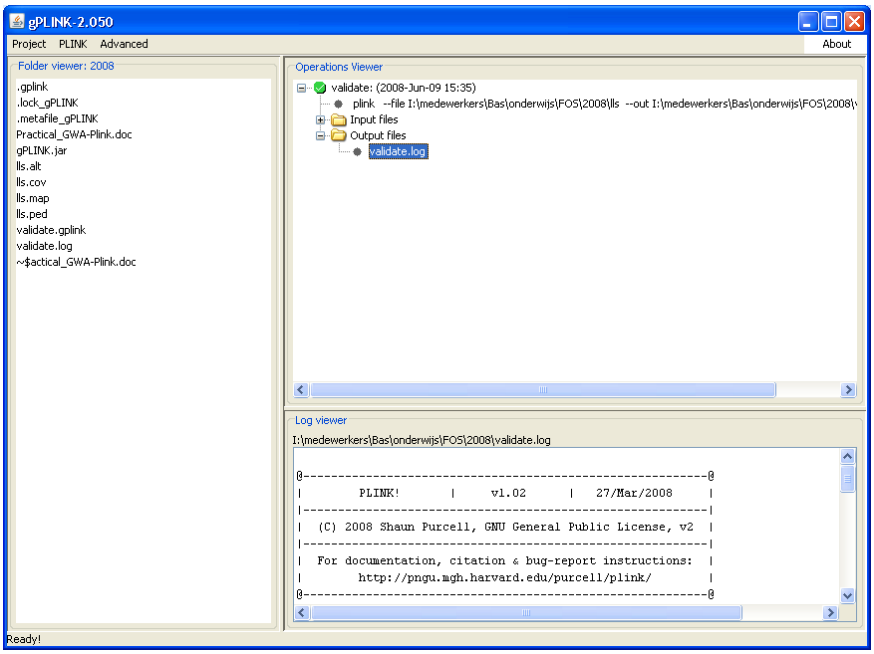


5. Click OK and you will see how the Plink command looks like (more experienced researchers do not use the graphical user interface Gplink but directly type in Plink commands).

6. Again click OK. Now you will see that Plink performed the requested analysis called 'validate'. Click until you see the output files and then on validate.log. Read this file.

A) HOW MANY MARKERS (SNPS) ARE IN THE FILE?

B) HOW MANY INDIVIDUALS ARE IN THE FILE AND WHAT PERCENTAGE IS MALE?



7. Now use PLINK\Summary Statistics\HWE to test for Hardy-Weinberg Equilibrium. Enter hwe in the 'Output file root' in the Standard Input tab. Instead of opening the output-file with the result of the HWE test in Gplink, it is better to open hwe.hwe using Excel (space delimited). The file can of course be found in your folder containing today's analyses (H:\FOS2018GWAS).

8. Note: the principle of HWE will be explained again centrally.

C) WHY IS THERE NO HWE TEST DONE FOR AFF (= AFFECTED INDIVIDUALS)?

D) WHAT IS THE LOWEST P-VALUE FOR THE HWE TEST? (SORT IN EXCEL FIRST ON TEST THEN ON P-HWD; THE D IN HWD STANDS FOR DISEQUILIBRIUM (YES, IT OFTEN SEEMS AS IF THEY TRY TO CONFUSE YOU ON PURPOSE).

E) IS THE LOWEST P-VALUE STATISTICALLY SIGNIFICANT AFTER ADJUSTMENT FOR MULTIPLE TESTING USING A SIMPLE (AND INADEQUATE) BONFERRONI ADJUSTMENT

(MULTIPLYING P WITH THE NUMBER OF TESTS)? AND THE SAME BUT NOW FOR A COMPLETE GWA OF 500,000 SNPS?

F) WHAT PERCENTAGE OF P-VALUES IS LOWER THAN 0.05? IS THIS WHAT YOU WOULD EXPECT ACCORDING TO CHANCE AND DOES IT CONCERN YOU?

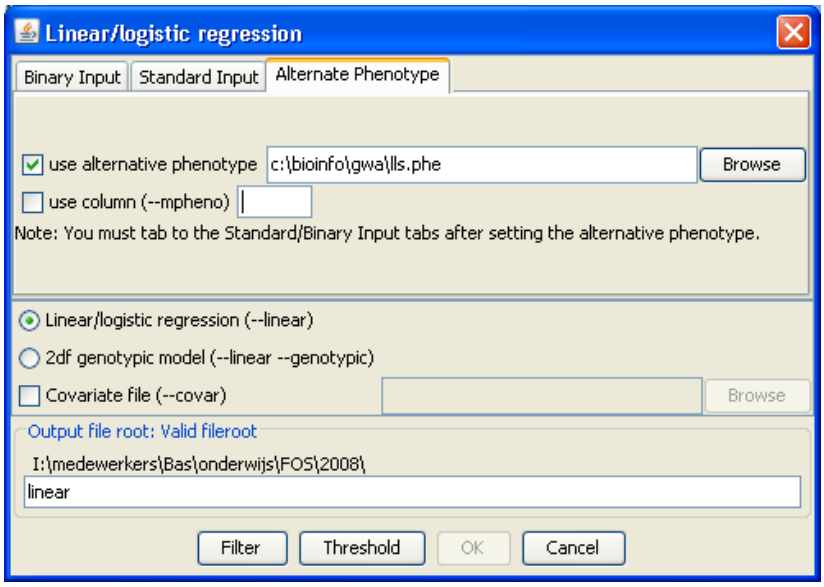
3. ASSOCIATION ANALYSIS

For the analysis we do not only need the ped-and map-files but also the phe-file with the quantitative phenotype and the cov-file with the covariates.

1. The quantitative trait under study is plasma insulin. Open the file lls.phe in excel. Note that values were transformed using the natural logarithm to obtain a normal distribution (insulin levels are positively skewed).

A) HOW IS MISSING DATA CODED? (LOOK FOR STRANGE NUMBER)

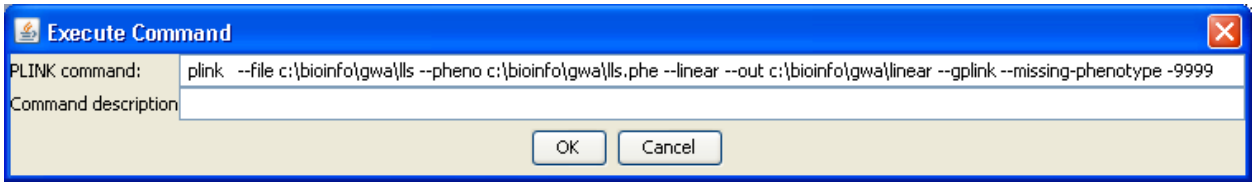
2. Now perform a genetic association analysis for quantitative traits. Select PLINK\Association\Linear/logistic regression. Tell Plink to give the output files the name linear and, in the tab Alternate Phenotype, where to find the data on insulin levels (lls.phe) as shown below.



3. Go back to the Standard input tab and click OK. Now you have to manually type in how missing data is defined. This you can do by adding the following in the Execute command window at the end of the Plink command line:

--missing-phenotype -9999

4. make sure that there is a <space> between your new command and the previous command (--gplink). Click OK to start the analysis.

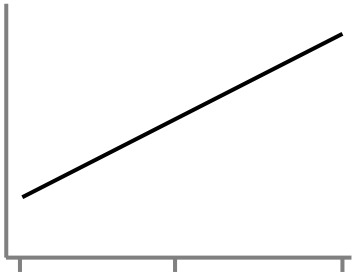


5. Open the output file linear.assoc.linear in Excel.

The default test in linear analysis is additive (ADD).

A. WRITE DOWN THE REGRESSION MODEL (GENERAL LINEAR REGRESSION MODEL: Y = B0 (CONSTANT) + B1*X).

Remember that such a regression model can be drawn as a graph. Asking yourself what numbers you need to describe the graph will give you the regression model (so describe in words what is y, b, x and the constant).



B. WHICH SNP SHOWS THE LOWEST P-VALUE FOR THE ASSOCIATION WITH INSULIN LEVEL AND WHAT IS IT. IS THE LATTER SIGNIFICANT AFTER A CRUDE MANUAL BONFERRONI ADJUSTMENT FOR MULTIPLE TESTING?

C. WHAT PERCENTAGE OF SNPS HAVE A P-VALUE <0.05 AND HOW DO YOU INTERPRET THIS?

6. Open the log file and note at what time the analysis started and finished.

D. HOW LONG DID THIS ANALYSIS TAKE?

E. HOW LONG WILL THE ANALYSIS OF A COMPLETE GWA TAKE CONSISTING OF 500,000 SNPS?

Apart from an additive model, you can test a dominant and recessive model.

F. DESCRIBE WHAT YOU DO WHEN TESTING RECESSIVE AND DOMINANT (I.E. WHAT GENOTYPES ARE GROUPED)?

7. Now perform a recessive and dominant test by doing the same as in the previous question but now not only add --missing-phenotype -9999, but also --recessive or --dominant. Note: this cannot be done in one command so run them separately, one after the other. Give the output a logical name (e.g. linear_rec and linear_dom).

G. LOOK UP THE P-VALUES FOR THE RECESSIVE AND DOMINANT TESTS FOR THE SNP SHOWING THE STRONGEST ASSOCIATION IN THE ADDITIVE MODEL.

The plasma insulin levels for the common homozygote, heterozygote and rare homozygote genotype are 3.1, 2.7 and 2.6, respectively (to be precise: the natural logarithm of the levels).

H. WHICH GENETIC MODEL DESCRIBES THESE DATA BEST AND DID YOU EXPECT THIS LOOKING AT THE P-VALUES?

4. VISUALIZATION WITH HAPLOVIEW

Without visualization it is very difficult to get an overview of large amounts of analysis results and, even more important, is it required for the first steps of biological interpretation. First you will use Haploview to visualize the data.

1. Before doing so, redo the analysis using an additive genetic model but now also adjust for covariates age and sex. This can be done by doing the same analysis as in 3b but now tell Gplink to use the file lls.cov by ticking Covariate file in the Standard input tab (and do not forget to add --missing-phenotype -9999!) . If you would open the lls.cov file you will see data on the covariates Sex and Age.

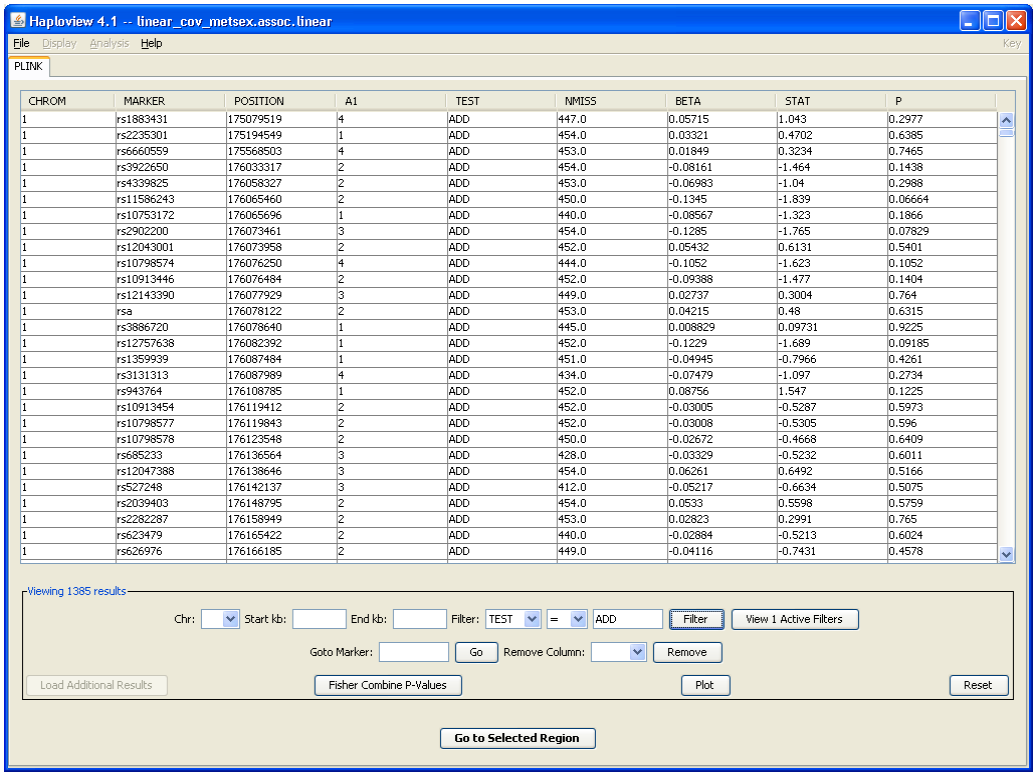
A. CHECK THE OUTPUT. ARE THE COVARIATES SEX AND AGE ASSOCIATED WITH INSULIN LEVEL?

B. IN GENETIC STUDIES SUCH COVARIATES CAN NEVER BE CONFOUNDERS IN THE CLASSICAL SENSE. EXPLAIN WHY KEEPING IN MIND THE DIFFERENCE BETWEEN A GENETIC STUDY AND STUDIES FOCUSING ON ENVIRONMENTAL FACTORS. TO HELP YOUR THINKING: COMPARE THE CURRENT STUDY WITH A STUDY ON INSULIN LEVEL AS A RISK FACTOR FOR HEART ATTACKS BEFORE THE AGE OF 50 IN A POPULATION INCLUDING BOTH SEXES.

C. WHY WOULD YOU STILL WANT TO ADJUST FOR CONFOUNDERS? ACTUALLY, THIS IS THE SAME REASON WHY THIS IS DONE IN RANDOMIZED TRIALS.

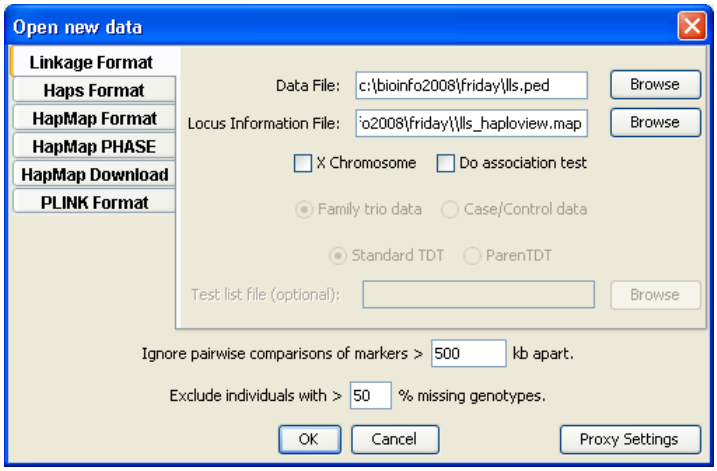
2. Now right click on the output of the previous analysis in Gplink (e.g. lls_cov.assoc.linear). The second option is open in Haploview. Do so.

3. Now do the following to create a so called Manhattan Plot. First, apply a filter in the lower part of the Haploview screen to include the results of the additive test only (TEST = ADD and then click the Filter button). Next, click the Plot button. In screen that opens (see screenshot), choose to plot P on the y-axis as ($-\log_{10}$) transformed so that 10^{-5} will be plotted as 5. Finally click OK.



D. INSPECT THE PLOT AND IDENTIFY THE LOWEST P-VALUES. ARE THEY SCATTERED OR CLUSTERED? MENTION TWO POSSIBLE EXPLANATIONS FOR THIS.

4. Now use Haploview to study linkage disequilibrium (LD) patterns in the region associated with insulin levels. To do this click File\Open new data set in Haploview. The files are in linkage format. The data file simply is lls.ped you used in Plink. The locus information file is called lls_question4c.map (this is lls.map but without the chromosome number and genetic distance columns). Click OK and then wait a while (particularly after all markers were loaded).



You will see all 1,492 markers in the files in the order of their physical location on the chromosome in basepairs.

5. Now Deselect All in the Check Markers tab and look up the most strongly associated SNP. Tick this SNP in the column Rating as well as all SNPs located up to about 50kb before and after this SNP. Next, go to the LD Plot tab.

You will see a lot of red indicating high LD. The colors are based on the LD measure D'. If D'<1 the value for D' is given.

E. WHAT IS THE POPULATION GENETIC EXPLANATION FOR THIS REGION WITH HIGH D' VALUES?

6. Now use Display\LD color scheme and Display\Show LD values to get r^2 instead of D'. You will see less dark colors and lower values as compared to the D' picture.

F. FOR R2 TO BE 1, AN ADDITIONAL POPULATION GENETIC CONDITION SHOULD BE SATISFIED AS COMPARED TO D'. WRITE DOWN THIS CONDITION.

7. Look up the r^2 values between the SNPs that showed a low p-value close to strongest associated SNP.

G. WHICH OF THE TWO POSSIBLE EXPLANATIONS YOU MENTIONED IN QUESTION 4B IS CORRECT?

5. VISUALISATION IN LOCUSZOOM

Now you will use LocusZoom to plot the regional association results.

1. First, you will need to change a result file from space to tab-delimited and you need to make sure that the file only contains p-values for the SNPs and not for covariates (for example by choosing the results file lls.qassoc or modify another file). You can do all this using Excel. Then go to the LocusZoom website http://csg.sph.umich.edu/locuszoom/ (or google on 'locuszoom').

2. Click *Plot Your Data: Single Plot*. Take some time to go over the page you see now and to inspect the options.

3. Click Set for Plink data. Select the file with your results. In *Specify region to Display*, fill in the name of your top SNP at *SNP:SNP reference name*. The plot will then be centered around your top-SNP. Now you are ready to plot the data (*Plot Data*).

A. TO WHAT EXTEND MAY LD EXPLAIN THAT ADJACENT SNPS WERE ASSOCIATED WITH INSULIN LEVEL?

B. IN WHICH GENE IS THE TOP SNP LOCATED?

C. IN WHICH REGION WOULD YOU LIKE TO TYPE ADDITIONAL SNPS TO FINE MAP THE RESULTS, WHICH PLOTTED DATA PLOTTED ARE USEFUL TO DETERMINE THIS? ARE THERE OTHER SNPS YOU CAN GENOTYPE?

5. VISUALIZATION WITH UCSC GENOME BROWSER.

A first start in getting more biological insight is to look up information about the associated SNPs in genome browsers. We will use the UCSC genome browser (http://genome.ucsc.edu/ and click Genomes on the left).

A. LOCATE THE MOST STRONGLY ASSOCIATED SNP IN DECEMBER 2013 ASSEMBLY OF THE HUMAN GENOME.

B. IN WHICH GENE IS THE SNP LOCATED?

C. IS THE SNP LOCATED IN AN EXON OR INTRON (MAKE SURE THAT IN GENES AND PREDICTION TRACKS THE TRACK UCSC GENES IS SET ON PACK)?

D. IS IT LOCATED IN A CONSERVED REGION (MAKE SURE THAT IN COMPARATIVE GENOMICS THE TRACK CONSERVATION IS ON SQUISH)?

E. IS IT LOCATED IN A REGION THAT MAY BE INVOLVED IN REGULATING TRANSCRIPTION (REGULATION, ENCODE REGULATION)? YOU MAY WANT TO ZOOM OUT (UPPER RIGHT) TO GET A BETTER VIEW.

1. Click on the gene name in the genome browser

2. Scroll down and find in which tissues the gene is expressed.

A. IS THIS WHAT YOU EXPECT FOR A GENE INFLUENCING INSULIN LEVEL?

3. Click the Entrez Gene link and scroll down to Pathways.

B. IS THIS WHAT YOU EXPECT FOR A GENE INFLUENCING INSULIN LEVEL?

4. Search pubmed using the gene name (not the abbreviation) and (in pubmed type AND (capitol letters) insulin.

C. IS THIS WHAT YOU EXPECT FOR A GENE INFLUENCING INSULIN LEVEL?

5. Go to the great database of GWASs on http://www.genome.gov/gwastudies/. Or even easier: the UCSC genome browser includes a track with this information called NHGRI catalog of published genome-wide association studies.

D. HAVE THERE BEEN ARE OTHERS FOUND ASSOCIATIONS WITH THE TOP-SNP AND GENE FOR RELATED TRAITS (E.G. TYPE 2 DIABETES)?

6. REPLICATION STUDIES

The ultimate proof is to perform an independent study and find the association again.

A. WOULD YOU USE A REPLICATION STUDY THAT IS SMALLER, BIGGER OR OF SIMILAR SIZE AS THE CURRENT ONE AND WHY?

B. WHAT IS YOUR PROVISIONAL VERDICT: DO YOU BELIEVE THAT THE TOP SNPS IS TRULY ASSOCIATED WITH INSULIN LEVELS OR DO YOU FIND IT MORE LIKELY THAT IT IS A FALSE POSITIVE?