Assignment 1

Niklas Rindtorff

2/3/2019

Background questions:

1.) What are the diseases of interest ascertained in the WTCCC study? Draw the E, G, P, and D diagram and annotate what is investigated in the WTCCC GWASs.(2 points)

Next to two control groups, the study included seven groups with the following diseases:

- BD - bipolar disorder
- CAD - coronary artery disease
- CD - crohn's disease
- HT - primary hypertension
- RA - rheumatoid arthritis
- T1D - type 1 diabetes mellitus
- T2D - type 2 diabetes mellitus

The study investigated the relationships between variations in the genome and diseases. Diseases themselves are a subgroup of a person's phenome. The **X** (genome-wide SNPs) are highlighted in green, the **y** (cohort) is shown in blue.

2.) How are the each of the seven diseases assayed phenotypically (7 points)?

The sampples for this study were accrued from different, already existing, cohorts that represented the UK population with European ancestry. The patients were clinically diagnosed with a given condition but not further phenotypically profiled.

Further details can be found in the citation below from the WTCCC paper:

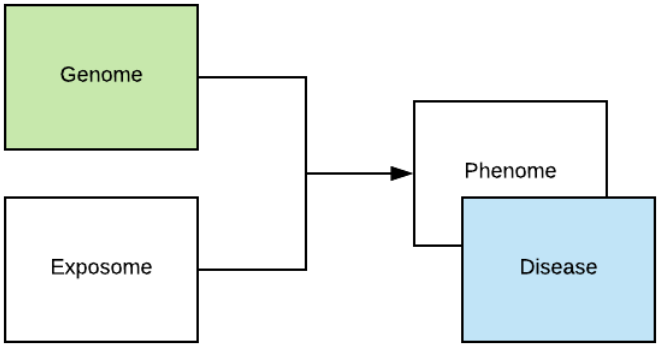


Figure 1: P=E+G theorem in this study

Case series came from previously established collections with nationally represent- ative recruitment: 2,000 samples were genotyped for each. The control samples came from two sources: half from the 1958 Birth Cohort and the remainder from a newUK Blood Service sample. The latter collection was established specifically for this study and is a UK national repository ofanonymized DNA samples from 3,622 consenting blood donors. The vast majority ofsubjects were self-reported as of European Caucasian ancestry.

3.) What is the assay platform for this study? Briefly describe the concept behind the assay technology (2points).

The authors use the Affymetrix GeneChip 500K Mapping Array Set to genotype patients. Briefly, this chip contains ~500,000 oligonucleotide DNA probes that hybridize with a patients flourescently labeled DNA according to Watson-Crick base pairing. If a given individual has a certain SNP that is represented on the chip, its DNA will bind a probe for this polymorphism while the wildtype probe will not bind labeled DNA (assuming homozygosity of the SNP). After hybridization, the individual's genotype can be measured and reconstructed using flourescent microscopy.

4.) How many SNPs are measured for each individual for all the chromosomes (you can do a line count of the data files themselves) (1 point)?

I log into orechstra, start a tmux session and execute the following code in the command line:

```
wc -l ~/ > ~/ds_mdm/NTR4_assignment1/data/row_count.txt
```

Now I load the data into R and create a plot.

```
# Initiating R session
library(tidyverse)
```

```
## -- Attaching packages ----------------------------- tidyverse 1.2.1 --
## v ggplot2 3.1.0       v purrr   0.3.0
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.3.0
## Warning: package 'tibble' was built under R version 3.5.2
## Warning: package 'purrr' was built under R version 3.5.2
## Warning: package 'dplyr' was built under R version 3.5.2
## Warning: package 'stringr' was built under R version 3.5.2
## -- Conflicts -------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)
```

```
## here() starts at /Users/nrindtor/GitHub/ds_mdm
```

```
#read.delim()
```

I don't run word count, instead -after completing the homework- I load a final result without any QC filtering.

```
tmp <- list.files(here("NTR4_assignment1/data/"), pattern = "T1D", full.names = TRUE) %>%
  .[!grepl(., pattern = "pgr")] %>%
  .[grepl(., pattern = ".csv")] %>%
```

```
  map(., ~ .x %>% read_csv() %>% mutate(chr_n = as.character(chr_n))) %>%
  bind_rows()

tmp %>%
  nrow()
```

```
## [1] 500568
```

The printed number is the total number of SNPs measured on each array.

5.) Let's start small. The year is 1999, and an investigator has painstakingly genotyped 1 SNP called (rsGOINGALLIN) in individuals with bipolar disorder and without.

He asks you how to execute an allele-based test of association "allele frequency test", whereby the reference allele is the major allele (the allele, C or T, that is most prevalent in the healthy population).

A.) What is the allele frequency (C and T) in bipolar population? And in Controls? (4 points)

We can calculate the allele frequency by solving the genotypes according to the Hardy-Weinberg theorem. I validate this approach by:

- summing the allele frequencies
- calculating the heterozygous genotype abundance

```
bipolar <- c(270, 957, 771)
ctrl <- c(436, 1398, 1170)
# scaling the genotypes by total count
bipolar_scaled <- bipolar/sum(bipolar)
ctrl_scaled <- ctrl/sum(ctrl)
# calculating the sqrt of the first and last genotype. The heterozygous genotype should have the abunda
allele_freq_bipolar <- sqrt(bipolar_scaled[c(1,3)])
allele_freq_ctrl <- sqrt(ctrl_scaled[c(1,3)])

# I perform a quick test for the bipolar subgroup:
# The sum of frequencies should be almost 1
sum(allele_freq_bipolar) %>% round(2)
```

```
## [1] 0.99
```

```
# The heterozygous allele frequency should close to our observed. The difference is very small
((2*allele_freq_bipolar[1]*allele_freq_bipolar[2])-bipolar_scaled[2]) %>% round(2)
```

```
## [1] -0.02
```

```
# I plot the final table
allele_freq <- tibble(allele = rep(c("C", "T"), times = 2),
       freq = c(allele_freq_bipolar %>% round(2), allele_freq_ctrl %>% round(2)),
       condition = c("bipolar", "bipolar", "ctrl", "ctrl"))

allele_freq
```

```
## # A tibble: 4 x 3
##   allele  freq condition
##   <chr>  <dbl> <chr>
```

```
## 1 C       0.37 bipolar
## 2 T       0.62 bipolar
## 3 C       0.38 ctrl
## 4 T       0.62 ctrl
```

B.) Execute a chi-squared test to test the association of the allele frequency in bipolar vs. controls in R and report the pvalue.

Is there evidence to conclude there is an association between rsGOINGALLIN and bipolar disorder? (4 points)

I create a matrix from the allele frequency counts. The C allele is our allele of interest. I create a 2x2 table of subjects that carry the allele and subjects that do not carry the allele.

```
matrix(c(270 + 957, 771, 436 + 1398, 1170), byrow = TRUE, ncol = 2) %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 0.051037, df = 1, p-value = 0.8213
```

We can not reject the null, that there is no link between the abundance of C in rsGOINGALLIN and bipolar disorder. The p-value for the Chi-squared test is *0.82*.

C.) Execute a chi-squared test amongst healthy controls to test for deviation from Hardy-Weinberg equilibrium (4 points).

Now we take the allele frequencies estimated from diseased individuals, and calculate the predicted number of genotypes for the control group. After that, we calculate the Chi-Square test to test for a deviation from the HWE.

```
# I calculate the expected distribution of genotypes
ctrl_expected <- c(allele_freq_bipolar[1]^2,
                   2*allele_freq_bipolar[1]*allele_freq_bipolar[2],
                   allele_freq_bipolar[2]^2)*sum(ctrl)
# I create a matrix of observed vs. expected and test it
matrix(c(ctrl_expected, ctrl), byrow = TRUE, ncol = 3) %>% chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 0.61461, df = 2, p-value = 0.7354
```

We can not reject the null hypothesis which assumes a Hardy-Weinberg equalibrium. The p-value is 0.7354.

Execution of GWAS:

Preparation

After logging in, I start a tmux session, import R and organize my packages.

```
srun --pty -p interactive -t 0-12:00 --mem 8G -c 1 bash
tmux new -s assignment1

module load gcc/6.2.0
module load spider R/3.5.1
```

I set up a local R package repository once

```
mkdir -p ~/R-3.5.1/library
echo 'R_LIBS_USER="~/R-3.5.1/library"' >  $HOME/.Renviron
export R_LIBS_USER="~/R-3.5.1/library"
```

I submit jobs by running the small loop in the */bash* directory

```
for FILE in call_pgrT2D_*; do echo ${FILE}; sbatch ${FILE}; sleep 1; done
```

1.) We are now going to execute a GWAS on T2D, a moderately heritable disease in which both the genome and exposome are said to play a role.

A.) After filtration, how many SNPs are available for analysis? Why filter out SNPs that deviate from Hardy-Weinberg? What is the Bonferroni-threshold of significance (round to 1 significant figure) (2points).

I import the data

```
library(here)

t2d <- list.files(here("NTR4_assignment1/data/"), pattern = "T2D_", full.names = TRUE) %>%
  # removing the X chromosome file
  .[!grepl(., pattern = "X")] %>%
  .[!grepl(., pattern = "pgr")] %>%
  .[!grepl(., pattern = "lambda")] %>%
  map(., ~ .x %>% read_csv() %>% mutate(chr_n = as.character(chr_n))) %>%
  bind_rows()
```

Now I filter the data according to my criteria and print the number of rows in the remaining dataset.

```
t2d_f <- t2d %>%
  filter(!hw_p.value < 0.05) %>%
  filter(!(minor_disease < 0.01 | minor_ctrl < 0.01))

t2d_f %>% nrow()
```

```
## [1] 1219779
```

B.) Compile the GWAS results into one .csv table and call it 'T2D.csv' (15 points=10 points for correct findings and format+5 points for correct filtration]).

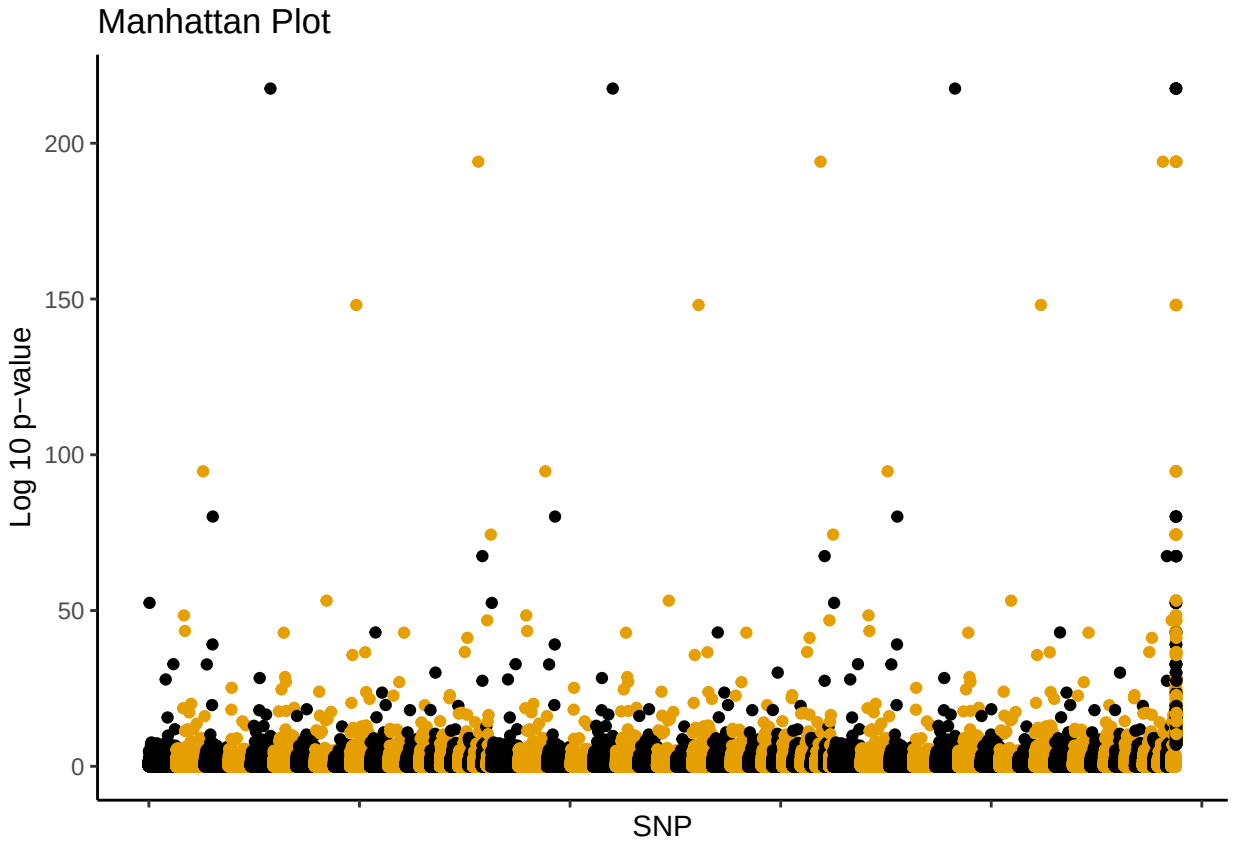
```
t2d_f %>% write.csv(here("NTR4_assignment1/data/T2D.csv"))
```

2.) Analysis of GWAS results:

A.) Produce a manhattan plot and qqplot of the pvalues (2 points).

Manhattan Plot

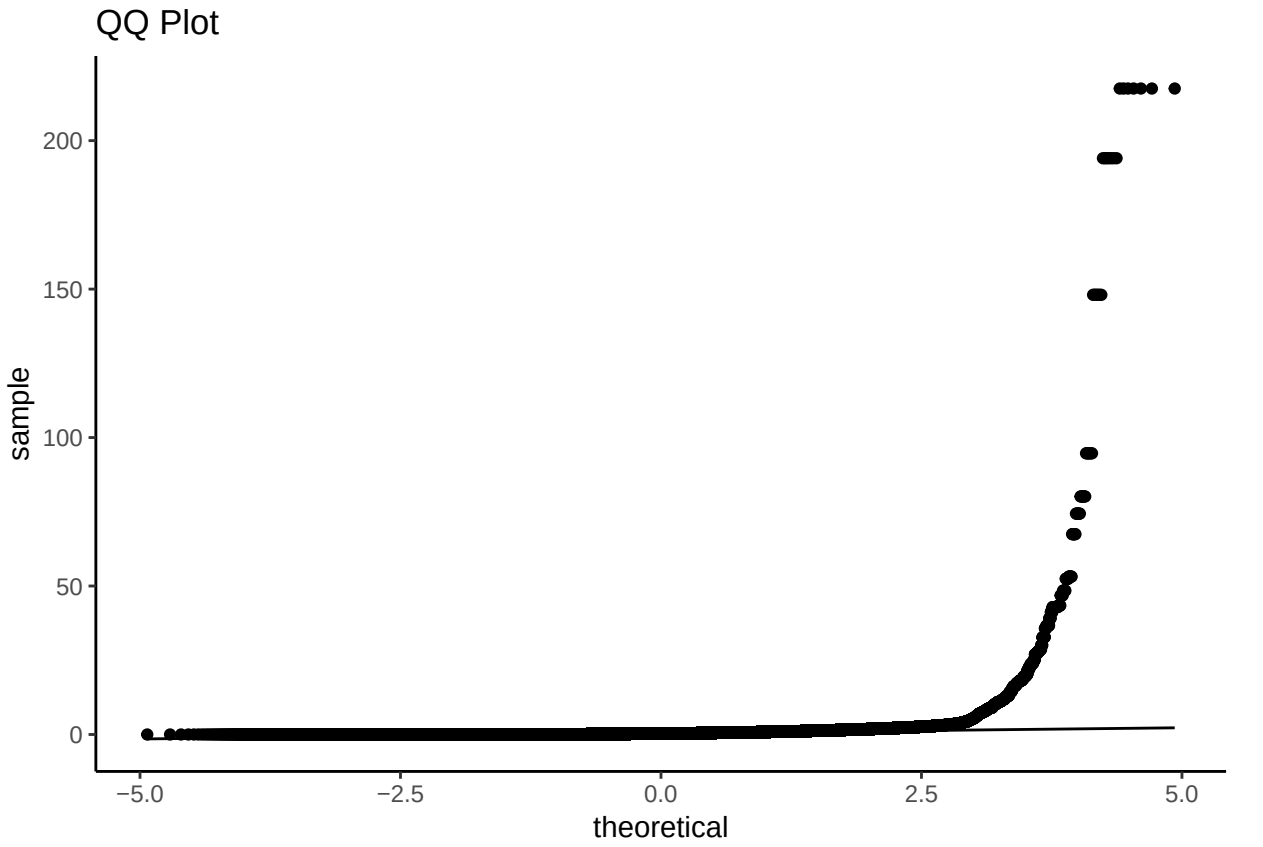
```
t2d_f %>%
  mutate(log_p = -log10(cd_p.value)) %>%
  mutate(n = c(1:nrow(.))) %>%
  ggplot(aes(x = n,
             color = chr_n,
             y = log_p)) +
  geom_point() +
  theme_classic() +
  theme(axis.text.x = element_blank(),
        legend.position = "none") +
  scale_colour_manual(values=rep(c("#000000", "#E69F00"), times = 11)) +
  labs(title = "Manhattan Plot",
       y = "Log 10 p-value",
       x = "SNP")
```



QQ Plot

```
t2d_f %>%
  mutate(log_p = -log10(cd_p.value)) %>%
  ggplot(aes(sample = log_p)) +
  stat_qq() +
  stat_qq_line() +
```

```
  theme_classic() +
  labs(title = "QQ Plot")
```



I am suprised to see these extreme deviations from the theoretical normal distribution. The strongest hits in the dataset have not been linked to specific traits according to dbsnp. Perhaps they are caused by technical artefacts, dysbalances in ethnicicty between groups or an implementation error.

B.) How many SNPs were found that exceeded the Bonferroni-level of significance for T2D? (1point)

```
t2d_f %>%
  mutate(fwer = p.adjust(cd_p.value, method = "bonferroni")) %>%
  filter(fwer < 0.05) %>%
  nrow()
```

```
## [1] 1260
```

C.) What could be biasing the associations and how might one control for this phenomenon? (1point)

The associations could be biased by differences in ethnicity between patient cohorts. Moreover, the analysis could be obscured by highly correlated SNPs, that are in a linkage disequalibrium.

D.) What is the p-value of association and odds ratio for rs4506565 for T2D? What is the interpretation of the odds ratio? What gene is this SNP associated with and its putative function in T2D pathogenesis? (10 points)

```
t2d_f %>%
  mutate(fwer = p.adjust(cd_p.value, method = "bonferroni")) %>%
  filter(rs_id == "rs4506565") %>%
  knitr::kable()
```

chr_n	wtccc_id	start	end	subject	genotype	a1	a2	group	hw_p.value	cd_p.value	cd_or
10	SNP_A-2005462	NA	NA	NA	NA	NA	NA	NA	0.8354306	0	1.369667
10	SNP_A-2005462	NA	NA	NA	NA	NA	NA	NA	0.8354306	0	1.369667
10	SNP_A-2005462	NA	NA	NA	NA	NA	NA	NA	0.8354306	0	1.369667
10	SNP_A-2005462	NA	NA	NA	NA	NA	NA	NA	0.8354306	0	1.369667
10	SNP_A-2005462	NA	NA	NA	NA	NA	NA	NA	0.8354306	0	1.369667
10	SNP_A-2005462	NA	NA	NA	NA	NA	NA	NA	0.8354306	0	1.369667
10	SNP_A-2005462	NA	NA	NA	NA	NA	NA	NA	0.8354306	0	1.369667

The SNP in question is located in TCF7L2, a transcription factor linked to the canonical Wnt signaling pathway, which is relevant for proper beta-cell function. It has been linked to an increased risk of T2D in prior studies.

The odds ratio for T2D in our data is 1.37 for the minor "T" allele. Thus, individuals with a "T" variant, instead of "A" have a 37% higher risk of developing T2D compared to unaffected controls.

3.) You just executed an allele frequency test of association. How would the genotypic test of association be different? Second, how would you test both of these types of allele configurations using a logistic regression? (2 points).

A genotype test for association, would be based on a 3x2 matrix of genotype abundance in the control and diseased subgroup. Similar to the allele-based test (with a 2x2 matrix), we could perform a Chi-Square test of association.

In case of a logistic regression, we would dumbify 2 of the 3 genotype configurations, while excluding the most frequent genotype configuration. The presence or absence of a given genotype would be a binary feature in the logistic regression model. In contrast, for an allele-frequency based test, we assume a linear "allele-dose" relationship in which a coefficient is multiplied by the number of alleles counted in a given individual.

4.) What is the main analytical differences between the test in Background Question 4 (candidate SNP) and GWAS? Why would one want to execute one or the other (1 point)?

The main analytical difference is the scale of evaluation. In a GWAS we perform thousands of tests, while a candidate evalutation only relies on one statistical test. Thus, we need to take great care to control our type-1 error in GWAS studies by multiple-test correction, which reduces our power. This does not apply for the evaluation of a single candidate SNP. Thus, if we have a strong hypothesis and only a limited number of samples, performing a candidate evaluation instead of a GWAS is the most successful strategy.

5.) In question 2B.), you identified sources of biases for GWASs. One source of bias can be controlled for using information from the SNP array. What is this factor? (2 points).

We can compute the principal components of the complete SNP matrix and correct for the first set of prinicpal components. Prior work has shown that prinicpal components are driven by differences in ethnicity. Thus, by accounting for differences in the first prinicipal components, we correct for ethnicity related bias in the analysis.

6.) The genomic inflation factor is arithmetically defined as the median of the observed chi-squared values (a function of the pvalues in GWAS) divided by the median of the expected chi-squared values. Estimate the genomic inflation factor for your GWAS in T2D. What does this quantity estimate? (10 points [5 points for estimate, 5 points for answer]).

```
# We expect the median chi-squared value to be:
exp_median = qchisq(0.5,1, lower.tail = FALSE)

# I calculate the chi-square statistic based on the test's p-value
t2d_f <- t2d_f %>%
  mutate(cd_statistic = qchisq(cd_p.value,1, lower.tail = FALSE))

# We calculate the inflation factor and correct by it.
lambda = median(t2d_f$cd_statistic)/exp_median

lambda
```

```
## [1] 1.041121
```

7.) BONUS. Modify your code to take into account this bias and re-run your associations for 'Executing GWAS', Question 1 (10 point bonus).

I add the following peace of code to my analysis to calculate lambda, correct for it and estimate more interpretable p-values after filtering for low-abundance and non-HW-equal. SNPs. I store the resulting table as *T2D_lambda*. I do not need to implement this code in the gwas.R file, as it does not take many ressources to run and the estimation of lambda is more robust when considering the whole genome, instead of a single chromosome at a time.

```
t2d_fc <- t2d_f %>%
  mutate(cd_statistic_correct = cd_statistic/lambda) %>%
  mutate(cd_p.value_correct = pchisq(cd_statistic_correct, df=1,lower.tail=FALSE))

t2d_fc %>%
  write.csv(here("NTR4_assignment1/data/T2D_lambda.csv"))

t2d_fc %>%
  dplyr::select(wtccc_id, cd_p.value, cd_p.value_correct) %>%
  head() %>%
  knitr::kable()
```

wtccc_id	cd_p.value	cd_p.value_correct

wtccc_id	cd_p.value	cd_p.value_correct
SNP_A-1938722	0.0718070	0.0776599
SNP_A-4217222	0.2424182	0.2519417
SNP_A-4196224	0.8819231	0.8842617
SNP_A-2286934	0.4901930	0.4988894
SNP_A-2056136	0.8620337	0.8647590
SNP_A-4241328	0.4895176	0.4982213

GWAS and Medical Decision Making: Introduction to the Polygenic Risk Score (25 points)

1.) What is a "polygenic risk score"? What is the underlying assumption? (1 point)

A polygenic risk score describes the odds of developing a given complex disease based on an individual's genotype. In practice, a polygenic risk score is often calculated by fitting a logistic regression model based on pre-selected SNPs and applying it to an individuals genotype data.

Polygenic risk scores assume a causal influence of genotype on disease occurance and independence between model parameters (SNP mutation status), if not corrected otherwise.

2.) Suppose you have a set of M SNPs associated with a disease/phenotype (say identified at p-value threshold of P), each with an effect size x_i (i indexes the ith SNP of M total SNPs). Specifically, x_i is the odds ratio of the risk allele relative to the non-risk allele for M independent SNPs that are not in LD with one another. You also have SNP data for one individual with the number of risk alleles n_i that correspond to the effect sizes x_i assuming the effects are additive. Write down the formula to compute the PRS for this individual. (2 points)

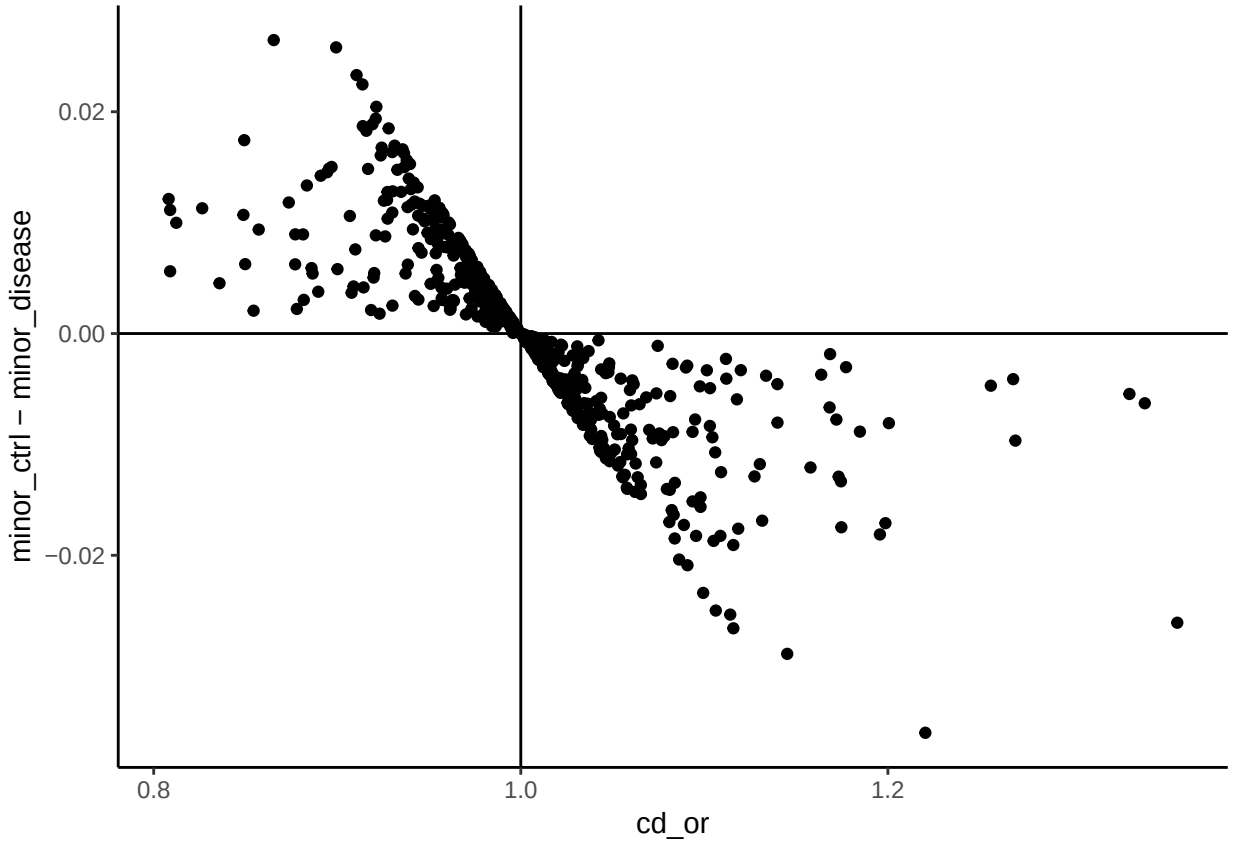
The effect size in a polygenic-risk-model is described in terms of a log odds-ratio for a given allele status. We can calculate an individuals risk score by multiplying the (modfified) odds-ratios for disease-specific SNPs.

3.) Using your GWAS summary statistics for T2D, compute the PRS for each individual in the cases and controls for T2D. Plot a histogram of the distribution of the PRS in cases and controls (in different colors) and indicate the quintiles (20, 40, 60, 80, 100 percentiles of the distribution). (5 points)

- BE careful! Make sure all of your ORs are for the risk allele (reference allele is the non-risk allele)!
- There is an emerging literature on what SNPs to choose for a PRS. For this exercise, use a pvalue threshold of 1e-7, HWE pvalue > 0.05 and minor allele frequency of greater than 1% (in both cases and controls)

I perform a quick sanity check: The odds-ratio for minor alleles that are more abundant in the control group than the diseased group should be less than 1.

```
t2d_fc %>%
  dplyr::select(cd_or, major, minor, minor_ctrl, minor_disease) %>%
  head(500) %>%
  mutate(delta = minor_ctrl - minor_disease) %>%
  ggplot(aes(cd_or, delta)) +
  geom_point() +
  theme_classic() +
  labs(y = "minor_ctrl - minor_disease") +
  geom_hline(yintercept = 0) +
  geom_vline(xintercept = 1)
```



Now I filter my data to identify the SNPs I am interested in.

```
t2d_fc_r <- t2d_fc %>%
  filter(hw_p.value > 0.05) %>%
  filter(!(minor_disease < 0.01 | minor_ctrl < 0.01)) %>%
  filter(cd_p.value < 1e-7)

# I store the file for use on o2
t2d_fc_r %>% write_csv(here("NTR4_assignment1/data/T2D_pgr.csv"))
```

With this list of SNPs, I load a set of chromosomes and left_join the locations that are relevant for my condition of interest. I run this operation on the cluster, as it consumes a lot of compute.

Below is an excerpt of the pgr.R script that I submit using sbatch.

```
library(parallel)
list.files(path = here("NTR4_assignment1/R"), full.names = TRUE) %>% lapply(., source)
```

```
args <- list(path = here("NTR4_assignment1/data"),
             ctrl = c("58C", "NBS"),
             disease = "T2D",
             chr_n = "X")

chr <- c("01", "02", "03", "04", "05", "06", "07", "08", "09", "10",
         "11", "12", "13", "14", "15", "16", "17", "18", "19", "20",
         "21", "22")

args_list <- lapply(chr[length(chr)], function(x){args$chr_n <- x; return(args)})

tmp <- args_list %>%
  mclapply(., function(x) import_data(x) %>%
           mutate(chr_n = as.character(chr_n)) %>%
           filter(wtccc_id %in% t2d_fc_r$wtccc_id)) %>%
  bind_rows()
```

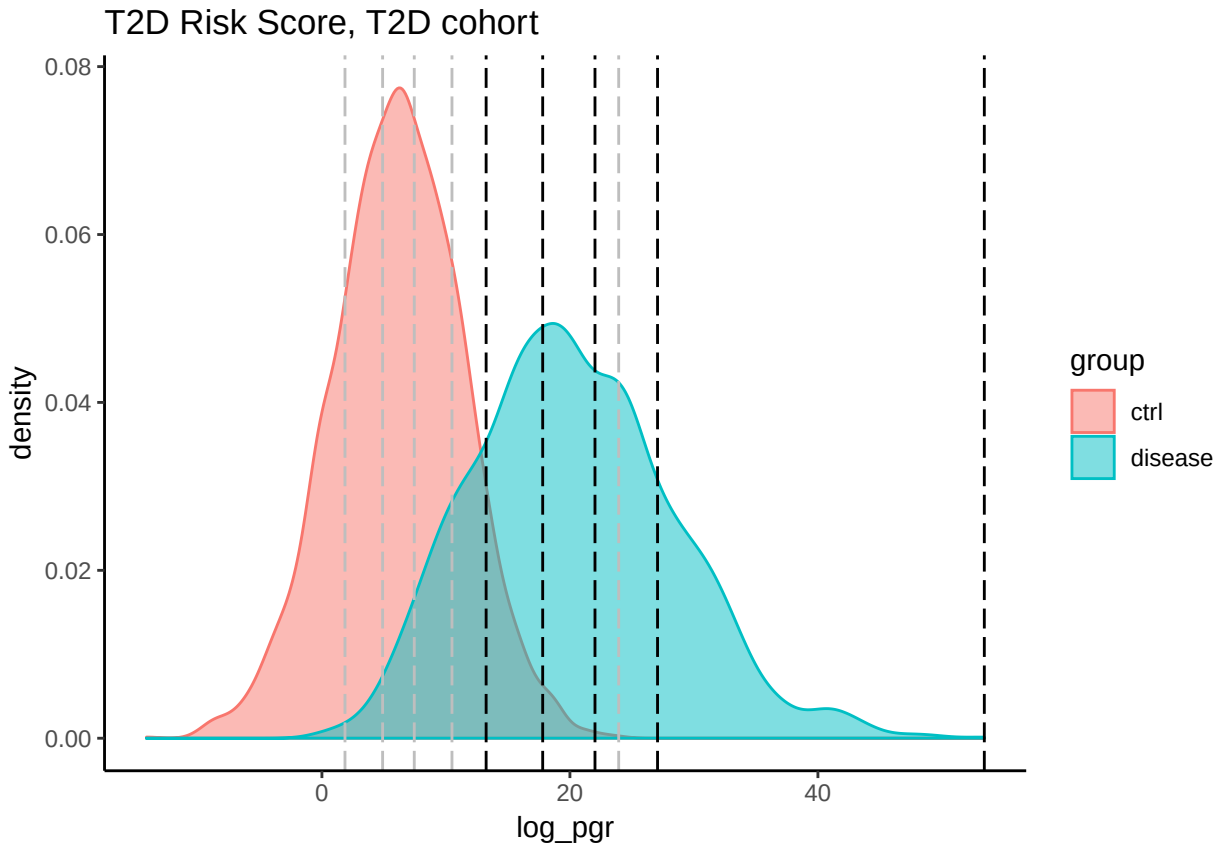
After running the jobs on o2, I collect my data. I group by patient and calculate the associated risk.

```
t2d_pgr <- list.files(here("NTR4_assignment1/data/"), pattern = "T2D_", full.names = TRUE) %>%
  # removing the X chromosome file
  .[!grepl(., pattern = "X")] %>%
  .[grepl(., pattern = "pgr")] %>%
  .[!grepl(., pattern = "T2D_pgr")] %>%
  map(., ~ .x %>% read_csv() %>% mutate(chr_n = as.character(chr_n))) %>%
  bind_rows()

library(stringr)
t2d_pgr_result <- t2d_pgr %>%
  dplyr::select(-start, -end, -chr_n) %>%
  left_join(t2d_fc_r, by = "wtccc_id") %>%
  dplyr::select(-contains("p.value")) %>%
  mutate(log_or = log(cd_or),
         n_minor = stringr::str_count(genotype, pattern = minor),
         or_minor = n_minor*log_or) %>%
  nest(-subject, -group) %>%
  mutate(log_pgr = map(data, ~ .x %>% .$or_minor %>% sum())) %>%
  unnest(log_pgr) %>%
  mutate(pgr = exp(log_pgr))

saveRDS(t2d_pgr_result, file = "t2d_pgr_result.Rds")
```

Now I plot the overlapping histograms

```
t2d_pgr_result <- readRDS("t2d_pgr_result.Rds")

# creating quantile data
q_ctrl <- quantile(t2d_pgr_result %>%
                     filter(group == "ctrl") %>%
                     .$log_pgr, probs = seq(0, 1, 0.2))[-1]
q_t2d <- quantile(t2d_pgr_result %>%
                    filter(group == "disease") %>%
                    .$log_pgr, probs = seq(0, 1, 0.2))[-1]
```

```r
t2d_pgr_result %>%
  ggplot(aes(log_pgr, fill = group, color = group)) +
  geom_density(alpha = 0.5) +
  theme_classic() +
  labs(title = "T2D Risk Score, T2D cohort") +
  geom_vline(xintercept =q_ctrl, linetype = "longdash", color = "grey") +
  geom_vline(xintercept =q_t2d, linetype = "longdash", color = "black")
```



4.) Suppose you have an individual in the 85th percentile of the PRS distribution for T2D and you are the individual's physician–what do you advise your patient? Suppose another individual is in the 43rd percentile–what do you advise this patient? (5 points)

As of today, the predicitive value of polygenic risk scores in the clinical setting just starts to become established. While there are cases in which patients with a high monogenetic risk for cardiovascular disease have been preventatively medicated, there is no evidence to plan interventions based on a high polygenetic risk for type 2 diabetes.

While no data suggests a benefit for preventative interventions based on a high T2D polygenic risk, there are two clinical actions that can be taken:

- Recommend high-risk patients to reduce other, non-genetic, risk factors for T2D and its complications, such as high BMI, high waist circumference, blood-pressure, LDL cholesterol etc.
- Recommend high-risk patients to undergo early-detection screenings at a higher and earlier frequency. As the pre-test probability for these patients is increased, the value of diagnostics might extend beyond the current general guidelines.

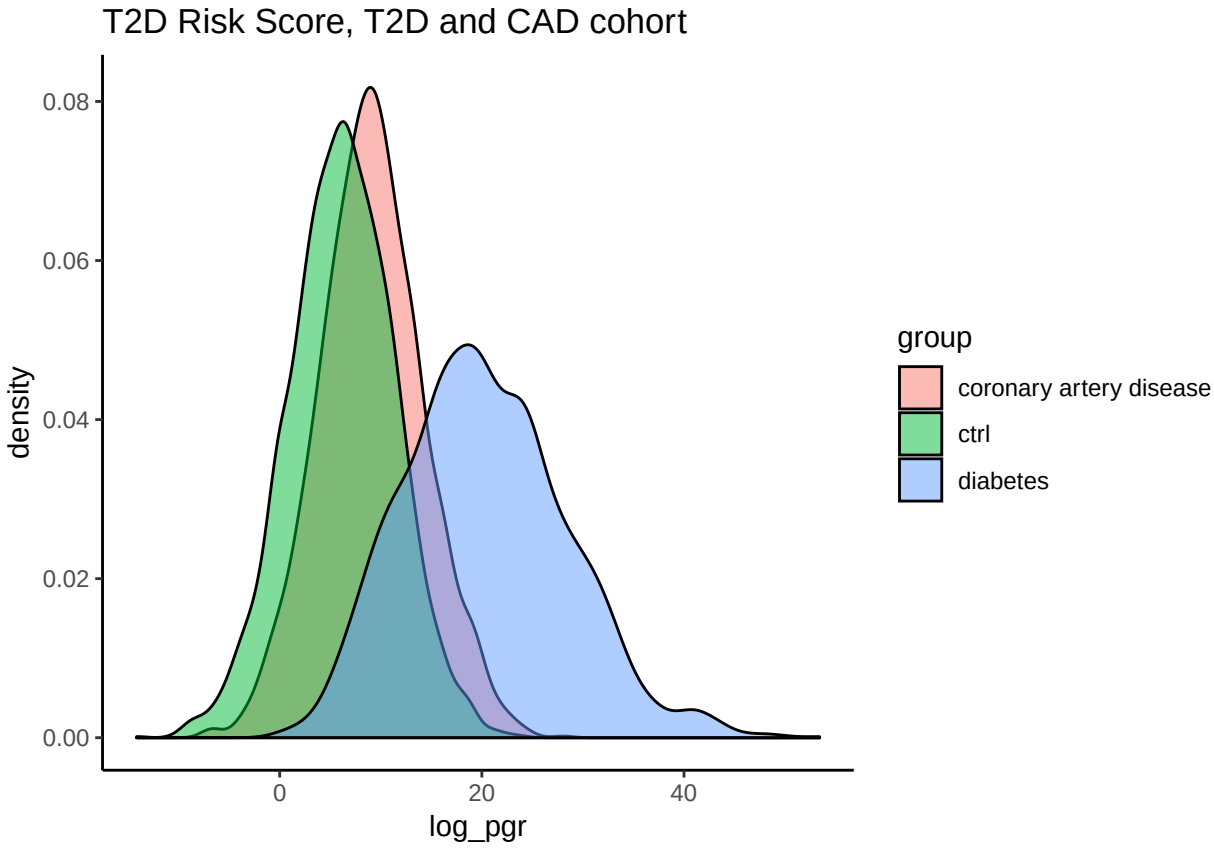
Based on the reasoning above, I would recommend the following: * the 85% patient should reduce all other risk-factors that can be modified by a healthy lifestyle. The patient should present him/herself with a higher frequency and take advantage of screening programs. * the 43% patient should reduce all other risk-factors that can be modified by a healthy lifestyle. Although the genetic risk for that patient is below average, lifestyle choices still account for the majority of variation for T2D and associated complications.

5.) T2D is a risk factor for cardiovascular disease (see: https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/heart-disease-s troke). Using your GWAS summary statistics for T2D, compute the PRS for each individual for the CVD cases . Add this to the histogram. Are individuals with the highest T2D PRS at genetic risk for CVD? (5 points)

```
cad_pgr_result <- list.files(here("NTR4_assignment1/data/"), pattern = "CAD_", full.names = TRUE) %>%
  # removing the X chromosome file
  .[!grepl(., pattern = "X")] %>%
  .[grepl(., pattern = "pgr")] %>%
  .[!grepl(., pattern = "T2D_pgr")] %>%
  map(., ~ .x %>% read_csv() %>% mutate(chr_n = as.character(chr_n))) %>%
  bind_rows() %>%
  dplyr::select(-start, -end, -chr_n) %>%
  left_join(t2d_fc_r, by = "wtccc_id") %>%
  dplyr::select(-contains("p.value")) %>%
  mutate(log_or = log(cd_or),
         n_minor = stringr::str_count(genotype, pattern = minor),
         or_minor = n_minor*log_or) %>%
  nest(-subject, -group) %>%
  mutate(log_pgr = map(data, ~ .x %>% .$or_minor %>% sum())) %>%
  unnest(log_pgr) %>%
  mutate(pgr = exp(log_pgr))

saveRDS(cad_pgr_result, file = "cad_pgr_result.Rds")
```

Now I plot the histograms

```
cad_pgr_result <- readRDS("cad_pgr_result.Rds")

cad_pgr_result %>%
  filter(group != "ctrl") %>%
  mutate(group = "coronary artery disease") %>%
  rbind(t2d_pgr_result) %>%
  mutate(group = if_else(group == "disease", "diabetes", group)) %>%
  ggplot(aes(log_pgr, fill = group)) +
  geom_density(alpha = 0.5) +
  theme_classic() +
  labs(title = "T2D Risk Score, T2D and CAD cohort")
```



6.) Suppose you are given SNP data for a set of new patients with and without T2D. How would you expect your PRS to perform on the new patients (ie, more discriminative, less discriminative, or the same)? Why or why not? Further, you did not account for LD when estimating the PRS. How does not accounting for LD influence the PRS? (4 points)

- The score is going to be less discriminative. The model has been optimized to separate healthy from diseased individuals in the original cohort. When applied to the new cohort, it is highly unlikely that it will demonstrate a better driscriminatory power. Instead, we will observe a "regression towards the mean" effect. The score will be more average.
- LD causes correlations between SNPs. As I did not account for these effects, there is a strong possibility that multiple SNPs that made it into the final model were correlated with eachother. However, simple polygenic risk models expect features to be independent. Thus, the model is currently biased and will perform less optimal on a new dataset.

7.) Encapsulate your work above into an R function that accepts a two column data frame of SNPs and genotypes for a single individual as input and returns a polygenic risk score for that individual. (3 points)

```
return_risk <- function(df, #input, contains genotype and wtccc_id
                        risk_score_df = t2d_fc_r){
  df %>%
  left_join(risk_score_df, by = "wtccc_id") %>%
```

```
  mutate(log_or = log(cd_or),
         n_minor = stringr::str_count(genotype, pattern = minor),
         or_minor = n_minor*log_or) %>%
  .$or_minor %>%
    sum() %>%
    log() %>%
    return()
  # the function can handle
}
```

Genetic architecture and disease similarity (25 point bonus):

Execute question 2 for all 7 diseases and answer the following questions.

1.) Produce a qqplot for each disease using ggplot. Rank each disease by the number of significant findings (pvalue threshold of 1x10-7) (10 point bonus).

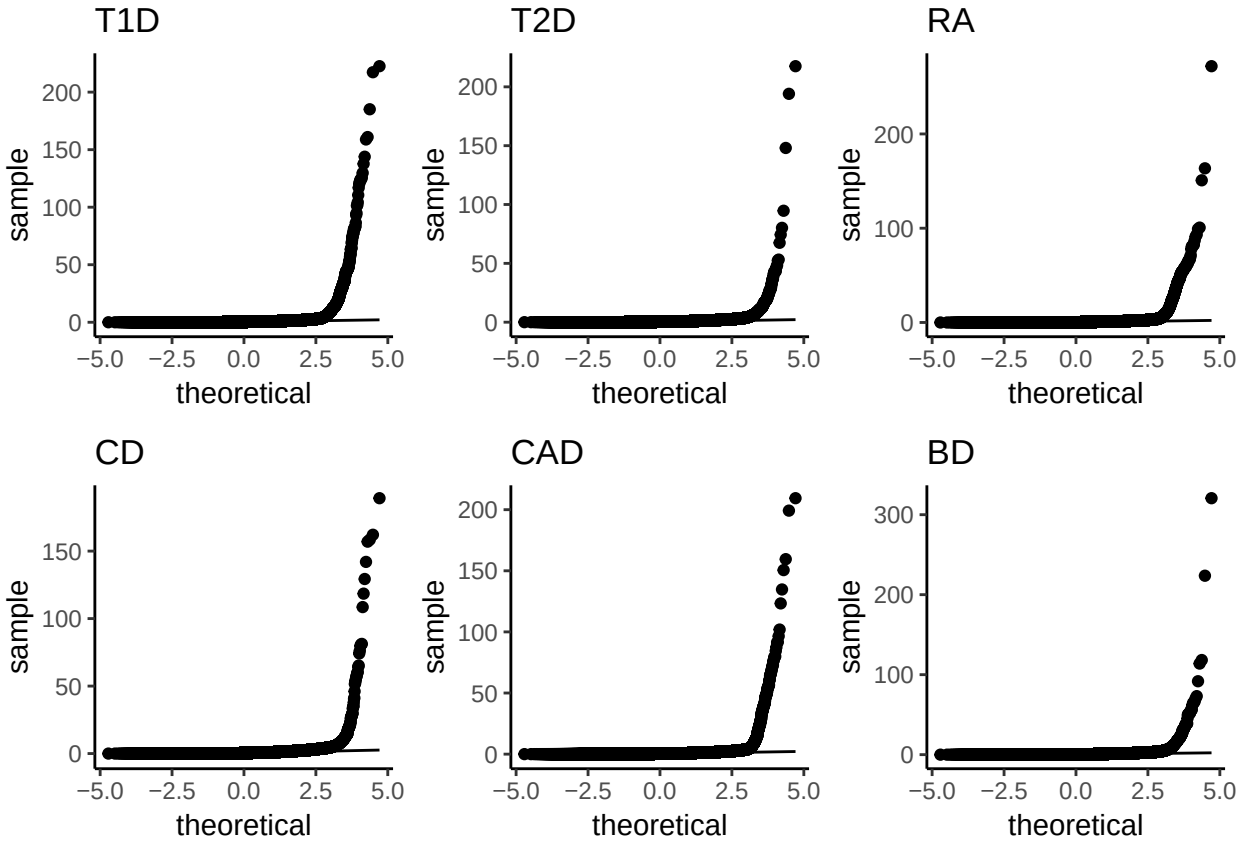
```
full_data <- list.files(here("NTR4_assignment1/data/"), pattern = ".csv", full.names = TRUE) %>%
  # removing the X chromosome file
  .[!grepl(., pattern = "X")] %>%
  .[!grepl(., pattern = "pgr")] %>%
  .[!grepl(., pattern = "T2D_lambda")] %>%
  .[!grepl(., pattern = "T2D.csv")] %>%
  map(., ~ .x %>% read_csv() %>% mutate(chr_n = as.character(chr_n),
                                        file = .x %>% str_split("/") %>%
                                          unlist() %>%
                                          .[length(.)] %>%
                                          str_split("_") %>%
                                          unlist() %>% .[1])) %>%
  bind_rows() %>%
  filter(hw_p.value > 0.05) %>%
  filter(!(minor_disease < 0.01 | minor_ctrl < 0.01))
```

QQ Plot

```
plot_qq <- function(df, pattern_in){
  p <- df %>%
    filter(file == pattern_in) %>%
    mutate(log_p = -log10(cd_p.value)) %>%
    ggplot(aes(sample = log_p)) +
    stat_qq() +
    stat_qq_line() +
    theme_classic() +
    labs(title = pattern_in)

  return(p)
}

library(patchwork)
plot_qq(full_data, pattern_in = "T1D") +
  plot_qq(full_data, pattern_in = "T2D") +
```

```r
  plot_qq(full_data, pattern_in = "RA") +
  plot_qq(full_data, pattern_in = "CD") +
  plot_qq(full_data, pattern_in = "CAD") +
  plot_qq(full_data, pattern_in = "BD") +
  plot_layout(ncol = 3) +
  ggsave("qq_group.png")
```

```
## Saving 6.5 x 4.5 in image
## Warning: Removed 1 rows containing non-finite values (stat_qq).
## Warning: Removed 1 rows containing non-finite values (stat_qq_line).
## Warning: Removed 2 rows containing non-finite values (stat_qq).
## Warning: Removed 2 rows containing non-finite values (stat_qq_line).
## Warning: Removed 1 rows containing non-finite values (stat_qq).
## Warning: Removed 1 rows containing non-finite values (stat_qq_line).
## Warning: Removed 1 rows containing non-finite values (stat_qq).
## Warning: Removed 1 rows containing non-finite values (stat_qq_line).
## Warning: Removed 2 rows containing non-finite values (stat_qq).
## Warning: Removed 2 rows containing non-finite values (stat_qq_line).
## Warning: Removed 1 rows containing non-finite values (stat_qq).
## Warning: Removed 1 rows containing non-finite values (stat_qq_line).
```



Ranking by number of significant genes

```
full_data %>%
  filter(cd_p.value < 1e-7) %>%
  group_by(file) %>%
  summarise(n = n()) %>%
  dplyr::select(file, n) %>%
  knitr::kable()
```

file	n
BD	266
CAD	267
CD	215
RA	475
T1D	713
T2D	195

2.) What disease, if any, diseases are associated with the the SNP rs6679677 (2 point bonus)?

```
full_data %>%
  filter(rs_id == "rs6679677") %>%
  filter(cd_p.value < 1e-7) %>%
  dplyr::select(file, everything()) %>%
  knitr::kable()
```

file	wtccc_id	hw_p.value	cd_p.value	cd_or	major	minor	major_ctrl	major_disease	minor_ct
RA	SNP_A-2267906	0.8235381	0	1.927912	C	A	0.9032956	0.8289145	0.096704
T1D	SNP_A-2267906	0.8235381	0	1.906399	C	A	0.9032956	0.8305000	0.096704

Both Rheumatoid Arthritis and Type 1 Diabetes are linked to the minor allele of this SNP. The SNP is located close to the transcription factor gene PHTF1.

3.) What are the potential biological meaning of a SNPs found in multiple diseases (3 point bonus)?

SNPs that are linked to multiple diseases can indicate a shared pathophysiology between them. In our case this might be due to a common regulatory process in which PHTF1 is involved. Similar to many autoimmune diseases, we can identify a shared set of SNPs in malignant diseases, too.

4.) Devise a method to correlate diseases on the basis of their association statistics (e.g., pvalues and/or odds ratios). What diseases are most correlated with each other? (10 point bonus).

We can estimate the degree of correlation by calculating the pearson correlation of odds-ratios for the two diseases. We can represent relationships in a graph.

As we can see below, the two diabetes forms are closely related. T2D and CAD (both metabolic) are linked. T1D and RA (both non-GI autoimmune) are linked as well.

```r
library(corrr)

full_corrr <- full_data %>%
  arrange(wtccc_id, file) %>%
  dplyr::select(cd_or, wtccc_id, file) %>%
  spread(file, cd_or) %>%
  dplyr::select(-wtccc_id)

full_corrr %>%
  correlate(method = "pearson") %>%
  network_plot(min_cor = .3)
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```