Norbis GENESTAT course, 4-8 June 2018

Take-home project

Due date: 22 June 2018

Return to: hakon.gjessing@uib.no

Family recurrence

In the paper Nordtveit TI, Melve KK, Albrechtsen S, Skjaerven R. Maternal and paternal contribution to intergenerational recurrence of breech delivery: population based cohort study. BMJ 2008; 336: 872-6. (you'll find it under "take-home" on the course web page), data are provided for recurrence of breech presentation at delivery from father to offspring and from mother to offspring.

1. Read the paper and suggest an interpretation of possible genetic effect on the risk of breech delivery. Are there likely fetal or maternal genetic effects? What about the likelihood of parent of origin effects?

2. Discuss possible designs for a GWAS-study that may find the most likely genetic effects.

3. Discuss the added value of having a case-dyad or case triad design.

4. If you had a limited budget to spend on genotyping, what design and setup for genotyping would you choose?

Article comprehension check

Answer the following questions based on the article in PLoS Genetics: Kenny, E. E., *et al.* A genome-wide scan of ashkenazi jewish crohn's disease suggests novel susceptibility loci. *PLoS Genetics*, 8(3), **2012**. *(Do not forget to look at the Supplementary Material!)*

1. Figure 1 shows PCA — what was the reasoning behind this analysis? Why did they exclude certain ethnicities in Figure 1B compared to 1A?

2. Which steps of QC mentioned in the lecture were used in this study (list also `PLINK` options used)? Which were not used and why?

3. Did they use the data from individuals with any missing genotypes? If yes, how was the imputation performed? (provide a general answer, no software or settings details, please)

Quality control

Practical tasks

Perform the following tasks on the files `exam_data.ped` and `exam_data.map` from the course web page, using `PLINK`. The `.ped` file contains trio data. The answer should show your reasoning together with any commands used and output information.

1. **Check for Hardy-Weinberg equilibrium:**

   (a) create `.bed/.bin/.fam` files containing only autosomal chromosomes and with MAF > 0.01 (minor allele frequency) and check HWE on these files;

   (b) what does the `TEST` column contain and what does it mean?

   (c) how many and which markers have the calculated p-value lower than 10^{-4}?

   (d) how many and which markers have the calculated p-value lower than 10^{-3}?

2. **Check for Mendelian errors:**

   (a) perform the check on the original data files;

   (b) were there any Mendelian errors found? If yes, how many?

   (c) how many errors of each type were found?

We recommend using R for questions 1c, 1d and 2c. The commands presented on the first day of the course are sufficient to exctract the needed information.

Association analysis with trios

We will use the dataset `pres.data` the way it was loaded during the lectures.

1. How many individual lines of data are there in the file? How many family trios?

2. Check that parents have the right gender. Find the number of boys and girls among the children in the file.

   Hint: In the new format, this is not yet quite streamlined.
   In `pres.data$cov.data`, you see all covariate data. Each column has been recoded to 1, 2, 3, etc...., ordered alphabetically.
   `pres.data$aux$variables` is a list with one element for each column in `pres.data$cov.data`. Each element is a frequency table of the original values in the file. For instance,
   `pres.data$aux$variables[["sex.m"]]` shows that there are 559 mothers with `sex.m == 2`. In this way you can obtain the requested information.

3. Find the total number of SNPs, and how they are distributed on chromosomes. (Hint: Use the map file).

4. Locate the SNP rs666 in the map file. Run a standard `haplin` analysis on this SNP. Choose a multiplicative response model and make sure any missing data is begin imputed. Does the default setting include only boys, only girls, or a combination? Does the SNP have a significant effect on the risk of disease?

5. The default analysis assumes that the data come from a "pure" case-triad design, i.e. no independent controls. However, there is a case-control variable in the data file, which separates between case-triads and control-triads. Change the `design` argument and re-run the analysis so that `haplin` analyzes the data as a hybrid design, not only as pure case triads.

6. Extend the last analysis so that it incorporates not only rs666 but also one more SNP on each side of rs666. `haplin` will then find and analyze haplotypes over three SNPs.

7. Run a sequence of analyses, one for each SNP on the X-chromosome (i.e. `winlength = 1`), using the same settings as above. Use 3 cores in parallel.

8. Join the results into one large table, removing redundant rows (one line from each SNP).

9. Sort the table by overall p-value and find the top hit. Check allele frequencies and HWE test for this hit. Are there any Mendelian inconsitencies at that SNP?

10. Create a QQ-plot for all overall p-values on the X-chromosome that you have just calculated.

11. Run `haplin` on the top hit SNP together with the SNP to the left and right and look at haplotypes.

EWAS

1. You have just received a set of .idat files containing Illumina HumanMethylation450 beadchip data. Before analysis the data must be preprocessed. What would you prioritize during the preprocessing steps? And, perhaps, what not?

2. Assume that your EWAS dataset consists of methylomes from cord blod. After you have performed quality control (QC) and normalization on your dataset you would like to explore whether there are any effects of maternal alcohol intake on DNA methylation. How would you set up the linear regression equation? What additional covariates would you include in the model?

Hetionet

Based on the presentation about Hetionet and reference on the net, create the Cypher queries to answer the following questions. As answers, you need to give the cypher queries together with a short explanation and/or screenshot of the output figure.

Q1 Which protein family is overrepresented in the group of proteins/genes responsible for inducing the compounds that cause a side effect called "Cushingoid"?

Q2 Retrieve the genes that are associated with "multiple sclerosis", as shown by GWAS studies (hint: use a property of the relationship called "sources").

Power calculations

Perform the power/sample size calculations in the R file Power_excercises.R