Norbis GENESTAT course, 30 May - 3 June 2022

Take-home project

Due date: 15 June 2022

Return to: hakon.gjessing@uib.no

Please NOTE:

- If there are any technical problems with the exam, please send me an email about it.
- It is OK to discuss/collaborate (a bit) when working on the exam, but the solutions/answers you hand in should be your own. No copy-paste.
- Most of the questions should be possible to answer based on the material provided in the lectures. Do your best; there is a lot to do, but perfection is absolutely NOT needed.
- The papers, data files etc. are available from the course web page.

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

Read the provided article and answer the questions below

> Moreno Uribe, L. M., Fomina, T., Munger, R. G., Romitti, P. A., Jenkins, M. M., Gjessing, H. K., Gjerdevik, M., Christensen, K., Wilcox, A. J., Murray, J. C., Lie, R. T., & Wehby, G. L. **(2017)**. A Population-Based Study of Effects of Genetic Loci on Orofacial Clefts. *Journal of Dental Research*, 96(11), 1322–1329. https://doi.org/10.1177/0022034517716914

1. The study focuses on individuals with orofacial clefts. What are the subtypes of clefts investigated here?
2. What is the study design and what was the model used in statistical analysis?
3. What was (approximately) the power to detect relative risk (RR) of 1.3, when including only complete dyads?
4. The study used data from several projects in different countries. What did the researchers do to avoid spurious associations due to possible differences between the data subsets?
5. What kind of effects were calculated? Could these be calculated in a case-control study?
6. Fig.1 - why not all the SNPs that had significant effect are shown?
7. Does this study support the usual merging of the two sub-types of clefts, CLO and CLP?

Bioinformatics services

You are provided the data (zip of csv under takehome on the course web page) that gives the differentially methylated CpGs separating the two different stages of differentiating cells. Adapt the instructions from the exercises in DAY5 to answer the questions below:

1. Read the data into R
   (a) how many CpGs are in the dataset?
   (b) what is the range of the p-values?
2. Calculate q-values.
3. Narrow the dataset to the CpGs that have a q-value < 0.01 and name this subset `diff_methyl_signif`
   (a) How many CpGs from this subset are located in chromosome 13?
4. Write out the files that give the regions
5. Go to ensembl BioMart and check which genes are located in the vicinity of these significant CpGs, write those to a file `mart_export.txt` and read it into R *(be careful with the naming of the columns - adjust according to the structure of your file!)*
   (a) Tabulate the genes according to which chromosome they are situated on.
6. *At this point, you can write out the names of the genes and input them in STRING db and examine the protein interaction network, as we did during the exercise, but this network is not that interesting, so you can just skip this part.*

NOTE: **the following two questions were not covered in the presentation, and are optional:**

7. Go to ensembl BioMart again and check which *regulatory regions* are located in the vicinity of these significant CpGs, write those to a file `mart_export_regulatory_feat.txt` and read it into R (be careful with the naming of the columns - adjust according to the structure of your file!)

   (a) How many items were found?

   (b) How many items *of each regulatory type* were found?

8. Examine one of the found enhancers by going to Gene Cards and checking which genes are possibly regulated by this region.

   (a) Is there anything in common these genes have?

Quality control

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?

Power calculations

Power calculations, parent-of-origin (PoO) effects

1. Execute the command below. What is the power to detect the given PoO effect applying 600 case-parent triads?

```
power_PoO <- hapPowerAsymp(cases = c(mfc=600),
  haplo.freq = c(0.2,0.8), RRcm = c(1.5,1), RRcf = c(1,1))
```

2. What is the power to detect the PoO effect if we set RRcf = c(1.5,1)? Comment on the result.

3. You receive extra money for your project on PoO effects, and you are wondering whether to genotype additional case triads or control triads. Run relevant power calculations and comment on the results.