Questions on statistical analysis of **Genetic Case-control** and **Population Association** studies

1. **Hardy-Weinberg equilibrium (HWE) – Bhavana / Dongmin / Will**
   a. What is Hardy-Weinberg equilibrium? What is the basic principle of probability that HWE assumes?
   b. What factors can give rise to deviations from HWE? Which one are researchers typically interested in for association studies?
   c. What are two basic goodness-of-fit tests for deviation from HWE, and how do they differ?
   d. Why might it be of interest to discard SNP loci that deviate significantly from HWE among controls?
   e. Briefly outline how a departure from HWE would be tested for a single diallelic locus.
   f. What are QQ-plots of P-values used for? Why are log values used? What's being plotted exactly, and why?

2. **Imputation of missing data - Emma / Julie / Xiaoai**
   a. Sometimes genotype data is incomplete due to factors such as variation in sequencing depth. What are some factors that can lead to **bias** in sampling different genotypes from a population?
   b. How can the potential for sampling bias of missing genotypes in case vs. control groups be tested?
   c. What are some methods used to impute missing data? What data / assumptions are these based on?
   d. What is the "common-disease common variant" (CDCV) hypothesis and why is it an important factor in association studies? How prevalent does a minor allele need to be in the population in order to be considered "common"?
   e. List three factors that will affect the power of association studies to identify minor alleles that are associated with disease.

3. **Linkage disequilibrium (LD) - Lauren / Maria**
   a. What is LD?
   b. What is one way that LD between typed SNPs and causal variants can be used to the advantage of association studies?
   c. In what way can LD interfere with association studies? For GWA, how are SNPs 'tagged' in order to maximize the chance of identifying common causal variants?
   d. One measure of LD is r^2. How are r^2 and sample size related to the ability to detect LD? What statistical measure does it relate to?
   e. Another measure of LD is ***D'***. Why do **LD maps** use an exponential decay function to estimate local LD?

4. **Models and measures of association - Maxim / Natalia / Yuya**
   a. What is a **risk factor**? What is **disease penetrance**?
   b. What are the four common models for disease penetrance?
   c. What is the **relative risk** (RR)? How does it relate genotype to phenotype?
   d. How does the **odds ratio** (OR) differ from the RR? Why do case-control studies use the OR instead of the RR?
   e. What is the mathematical relationship between the RR and the allelic OR? When are these approximately the same?
   f. What type of statistical models can combine the OR with other kinds of data, and why?

5. **Tests of association – Nuha / Peien / Raya**
   a. Describe the basic approach to test for association using a Chi-squared test.
   b. For genotype association, how are dominant and recessive models of penetrance handled?
   c. How does the treatment of a multiplicative model differ from that for an additive model?
   d. Under what conditions is an allelic association test more powerful than a genotypic test?
   e. When can the Cochran-Armitage trend test be applied? What is the advantage?
   f. Why might the Fisher exact test be preferred?
   g. An alternative test of association is the likelihood ratio (LR). What is it? Can you figure out why, for large samples, the Chi-square and LR are equivalent under the null hypothesis? (This is mentioned but not explained in the papers.)
   h. Logistic regression models can incorporate additional risk factors, confounding factors, and interactions between loci. What is the response variable? What are the explanatory variables, and how are they modeled? What do the coefficients for the predictors represent?

6. **Multiple hypothesis testing – Selena / Setiembre / Sydney**
   a. Why is multiple testing important in GWA studies?
   b. What is the type I error? What does the significance level represent?
   c. What is the tradeoff that must be made in multiple testing?
   d. What is a family-wise error rate (FWER)? What two correction methods are used to control the FWER? Why are they considered conservative?
   e. What is the FDR? What is the limitation of using FDR in GWAS?
   f. What approach can be used to overcome these limitations, and how does it work?
   g. How are log QQ p value plots useful here?
   h. One of the papers mentions (without further explanation) that "Bayes factors have also been proposed for the measurement of significance." Can you imagine how this might work?

7. **Epistatic and gene-environment effects – Theo / Titir**

a. Genetic interactions between multiple loci and with environmental factors both contribute to complex diseases. How can such interactions be incorporated into association models?

b. What are two ways that association studies can improve the chances of identifying such interactions?

8. **Confounding - Theo / Titir**

a. What factors can contribute to spurious associations?

b. How can **PCA** be used to help minimize the effect of confounding factors?

c. How can the technique of **Genomic Control** (GC) be used to help control for confounding factors? What additional typing data are specifically required in order to make GC work?