Week 4 Discussion Questions for "Genomic Approaches to the Study of Complex Diseases"

BIFX 504, Advanced Molecular Biology for Bioinformatics

Name: James Jedediah Smith

1. What is a complex trait?

Variations between people as a result of both genetic and environmental factors. Often associated multiple genes or variants that contribute to risk of a disease.

1. What is the general relationship between the effect size of a variant and its frequency in the population? In other words, are variants with strong effects or weak effects detectable in more people?

More common variants tend to have weaker effects, while more rare variants tend to have stronger effects. However, there are outliers to this general trend.

1. What is the difference between the “cohort” design and the “case/control” design for studying the genetic basis of complex disease in a population?

With cohorts, you enroll individuals regardless of health or disease and either assess them at a specific point or monitor them into the future. With case-control design, you specifically enroll based on disease status to get both cases and controls, then look back to see what happened in the cases prior to disease onset.

1. What are the benefits of each design? In other words, why might you use the cohort design instead of the case/control design and vice versa?

Cohort design allows you to study things into the future and provides a more broad view of a population. Case-control design allows you to look back and draw associations, but requires similar cases and controls to prevent population stratification. If a trait effect is seen in a broad population, a cohort design may be better. But if a trait effect appears in a certain subset of the population, case-control design may better.

1. Why is it necessary to control for the ancestry of individuals in a case/control study?

It is important to control for ancestry because certain subsets of a population may be more or less prone to developing a certain variant. As a result, if the case and controls have vastly different ancestry, it may lead to misleading conclusions or false positives.

1. How can you avoid or control for population stratification?

You could match cases with controls, restrict to one subgroup, adjust for genetic background, or use a family-based study design.

1. What are the two main benefits of using microarrays (as opposed to sequencing) to genotype individuals?

Arrays are much cheaper to do than sequencing. They can also be customized with specific variants based on the needs of a researcher.

1. How can you determine if a sample is poor quality and should be removed from a microarray-based genotyping study? What quality control checks should you perform?

There are four sample quality control checks used to identify bad samples that should be removed: poor quality samples, sample switches, unexpected relatedness, and differing ancestry. Poor quality samples are defined as having a sample success rate of < 95% or having excess heterozygous genotypes.

1. How can you determine if a SNP should be removed from a microarray-based genotyping study? What quality control checks should you perform?

There are SNP quality control checks used to identify bad SNPs that should be removed: low genotyping success of < 95%, differing genotypes in duplicate samples, expected genotype proportions are inconsistent with observed allele frequencies, non-mendelian inheritance in trios, and differential missingness in cases and controls.

1. What two major assumptions must be true in order to measure the effect size of a SNP using linear regression?

The assumptions are that the trait is normally distributed and the subjects are independent/unrelated to one another.

1. What does an odds ratio describe?

The risk of developing disease when a certain allele is possessed.

1. What is population substructure?

A subdivision of a population, such as ethnicity. It is important to adjust for this covariate or it could lead to population stratification.

1. Let's say that you performed an association study that looked at the genotype of only one SNP in relation to a trait, and you found that the p-value of the association 0.05. Is the association significant (within a 95% confidence interval)?

Yes! For a single test, as opposed to multiple tests, the standard p-value to use is 0.05.

1. Now let’s say that instead of looking at only one SNP you looked at 1,000,000 SNPs. After performing a multiple testing correction for the number of SNPs investigated, what p-value would your association need to have in order to be considered significant? In other words, what do we get when we perform a multiple testing correction for the p value 0.05 across 1,000,000 tests?

When performing multiple tests like this, you divide the 0.05 P-Value by the number of SNPs being tested. In this case the result would be: 0.05 / 1,000,000 = 5 x10-8.

1. What does imputation allow one to do?

It allows us analyze more variants more quickly by comparing variant sequences with more densely sequenced reference haplotypes to predict missing components.

1. What are the benefits of meta-analysis of GWAS data?

It combines data from multiple studies to give their predictions more statistical power than what would be observed in a single study. It can also investigate the consistency of effects across studies.

1. What does the effect size of a SNP tell you?

SNP effect size shows how associated it is with a trait or disease. Positive effect size means it is associated with the presence of said trait or disease. Negative effect size with the absence. Large numbers in either direction indicate a stronger association.

1. What does p value of a SNP tell you?

The p-value represents the likelihood of seeing this association randomly. Smaller p-value is good because it indicates a higher chance of non-random association.

1. List four approaches to identify the causal gene that is associated with a genetic locus.

The four approaches are: Look to other literature, see if gene has nonsynonymous SNPs, eQTL gene analysis, and pathway analysis.

1. What does a conditional association analysis tell you?

Whether or not SNP effect signals are independent of one another.

1. Why are rare variants typically identified by sequencing rather than array-based genotyping?

Assays require knowledge about the variants you are analyzing beforehand. With rare variants, we don’t usually know all the information up front. They are often too rare to associate with a single gene on their own. Some sequencing is necessary to help build a portfolio of associated variants. Once this information is gathered, arrays could be potentially used to help cut costs later down the line.

1. Describe three strategies that use sequencing to identify rare variants associated with disease by while minimizing the cost of the experiment.

Three strategies that minimize costs: only sequence selected individuals with extreme trait values, increase the number of individuals by either decreasing sequencing coverage or collecting rare variants onto a less expensive array, only sequence population isolates where the rare variants may appear with higher frequency.

1. How can combining information about variants from different studies that affect the same gene help to explain the biology of a disease? Give an example from the lecture.

Variants on their own may be too rare to build association. But multiple variants can be used to implicate genes. An example from lecture is the study that used multiple low frequency variants to show the association between the IFIH1 gene and diabetes.