Statgen final

Chi-Chun Liu

3/13/2020

Problem 1

a. A founding effect or a smaller effective population size together with limited gene-flows lead to lower haplotype diversity and enrichment of rare disease alleles. They also lead to less population structure and admixture that can serve as confounders. Moreover, environments in an isolated population are likely more homogeneous due to shared cultures and lifestyles, and can reduce potential confounding effects. These result in higher power in finding association between a locus and a phenotype.

b. Reduced effective population sizes as mentioned in (a) lead to lower haplotype diversity and long identity-by-descent tracts. So a smaller number of reference individuals sequenced by researchers can be representative enough to capture haplotype patterns in imputation targets formed by an isolated population. Longer shared haplotypes are easier to identify statistically as well.

c. When performing association analysis, we can use mixed effect models with a genetic relationship matrix to correct for cryptic relatedness and population structure.

d. The heritability from isolated populations is expected to be higher than that from unrealted subjects. This is because heritability is the phenotypic variation $V_p = V_G + V_E$ explained by the genetic variation V_G. The environmental variation V_E is lower due to more similar environments in an isolated population, and this ratio thus goes up.

Problem 2

a. Health condition and life-span is highly correlated, because an healthier individual is expected to live longer. We can thus investigate genetic contribution to longevity with certain health conditions. We can define our case cohort, which has achieved a critical age threshold without a set of pre-specified health complications and diseases. The cases will by definition be alive when the study is performed. We will select healthy young people as controls, since people without longevity born in the same birth cohort with our cases are not alive. We can't select people of similar age to the cases since we missed a lot of people who died young. However, we have to be very careful about confounders such as population stratification and environmental contributions. These confounders arise because our cases and controls were born in different times with different demographic compositions and environments. We will perform clustering analysis to match genome-wide genetic profile between our cases and controls, and use a mixed effect model to mitigate confounding effects.

b. We can use methods that integrate multiple tissue types. We will make use of transcriptome studies from publicly available databases such as GTEx.

c. We will again utilize utilize public available resources such as ENCODE as annotation.

d. One possibilty is that we have missed rare variants due to lower statistical power, and we can combine the rare-variants and perform SKAT-O tests. Another possibility is that there are many genes with modest effect sizes due to polygenicity. In this case we compute an aggregated test statistics for genes in a biological network, and test if the overall distribution of p-values follow a uniform distribution under the null.

e. We can test if there is significantly less disease risk alleles carried by the cases, the people who have achieved longevity compared to the controls. The disease risk alleles can be obtained from databases such as GWAS catalog and UK Biobank GWAS summary statistics. One caveat is that the risk alleles might have other protective effects for these cases. If the above test is not significant, we may lean toward the hypothesis that the cases have more protective alleles. If these variants are actually protective, they should increase in frequencies with respect to age. Moreover, if these variants are protective against another disease variant, we might expect an increase in frequencies of those risk variants in our cases.