GENE 210: Genomics and Personalized Medicine

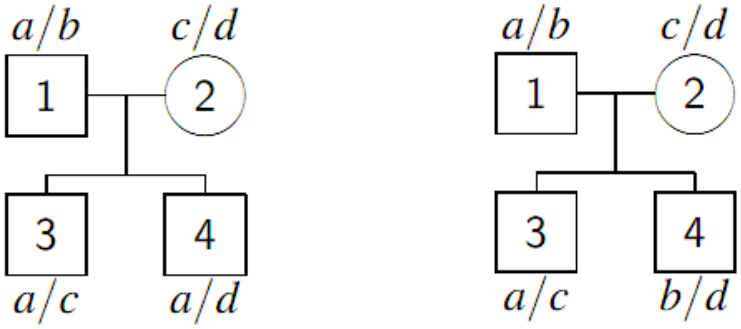
Genetic Similarity

Konrad Karczewski

Anyone with a personal genotype might be interested in how similar their genomes are to other individuals, both in their family and to other individuals in their or other populations. The methods to measure genetic similarity are fairly straightforward and easy to calculate for a single locus. Here, we will discuss two methods, identity by descent (IBD) and identity by state (IBS). The former is only applicable to family members, while the latter can be used to determine similarity between any two individuals.

Identity by Descent (IBD):

Identity by descent is useful for determining similarity between family members. IBD simply measures the amount of genetic material inherited by each offspring from each parent, and compares the two. For instance, we have the family below on the left, where at some locus, the mother has a/c alleles, and the father has b/d alleles. If we compare the offspring to the parents, we see that both of them have inherited an "a" from the father, but inherited different alleles from the mother. Their IBD is then 50% for these 2 loci. For another family (right), we may have no IBD between the two offspring if one child inherited one allele from each parent and the other child inherited the other allele from each.



Since we don't have family data, we won't be calculating this for ourselves, but this would be simple to calculate if data for both parents and a sibling are available.

Identity by State (IBS):

Identity by state is an easy, generally applicable method to measure similarity between unrelated individuals. IBS simply considers the similarity between genotypes at each locus and averages over all the loci of interest. So,

$$IBS(i,j) = \frac{1}{2n}\sum_{i=1}^{n} Sim(g_i, g_j)$$

In essence, we test the genotype similarity at each locus, assigning a 0 if the two genotypes are dissimilar, a 1 if one allele is shared, and a 2 if two alleles are shared. Then, we sum over all the loci of interest, and dividing by 2 times the possible loci (the maximum similarity if the two individuals had the same genotypes at all loci). For example, if we have two individuals, Konrad and Nick, we might try to measure their similarity across a few loci.

rsid	Konrad	Nick
rs4504908	AG	AA
rs478859	CT	CC
rs6577168	CC	CC
rs7415343	CT	CC
rs3176879	AA	AG
rs1922987	CT	CC
rs10493945	CC	CC
rs12041851	GG	AA
rs6541080	AG	GG
rs10493947	GG	AA

At the first locus (rs4504908), the two individuals share an A allele, but the other allele is different, and so, we'd assign a score of 1 for this locus. The 3rd locus (rs6577168) is CC in both individuals, so we'd assign a 2, and the last locus (rs10493947) is completely different between both individuals, so we'd assign a 0. If we summed over all 10 loci, we'd get a total score of 10. If the two individuals were identical at these loci, we'd have a possible score of 20, so the IBS for these loci is `10/20 = 50%`.

It should be noted that IBS is very dependent on which SNPs you consider. In this example, we only considered 10 random loci and got 50% similarity. If we were to take all loci on the genotyping arrays, we would find 74.29% similarity between these two individuals. In fact, as with any two humans, these two individuals are 99.97% similar, but here we are enriching for those loci that are already likely to vary between individuals.