OPTIMAL POOLING FOR THE DETECTION OF SINGLE NUCLEOTIDE POLYMORPHISMS USING NEW GENERATION GENOME SEQUENCERS

ANDREAS FUTSCHIK & DAVID M. RAMSEY

The genome consists of sequences made of 4 nucleotides. At a majority of the sites along these sequences, each individual in a species will have the same nucleotide. A site where there is variation within a population is called a single nucleotide polymorphism (SNP). At virtually all such sites, just two of the four nucleotides appear. These variants are called alleles, the most common (rare) allele is termed the major allele (minor allele, respectively).

Genome sequencers read these sequences by utilizing DNA from an individual (or pool) placed within a lane. Each lane gives a random number of reads for a particular site (modelled using the Poisson distribution). Consider a pool of size m. If the same amount of genetic material is taken from each individual, we may assume that given the number of reads for a site in a lane is r, then the number of reads from an individual has a binomial distribution with parameters r and $1/m$. For ease of presentation, by individuals we understand individual chromosomes rather than biological individuals. However, the analysis presented here can be generalized. With some small probability the nucleotide present at a site is read incorrectly.

We consider the optimal pooling of DNA to detect sites at which a population shows variation. Since any reasonable test will detect alleles of relatively large frequency with power close to 1, we concentrate on the detection of low frequency alleles (against the null hypothesis of no variation). Given a fixed number of lanes, pooling individuals allows us to increase the probability that a rare allele actually appears in the sample. However, as the pool size increases, the mean number of reads from an individual decreases, making it harder to distinguish reads of an individual allele from errors.

One could use the following simple test for the presence of a minor allele: accept that there is a minor allele if in any lane the number of reads for an allele that is not the major allele exceeds a given threshold. We specify the value of this threshold given the parameters of the genome sequencer and significance level for the test. We derive a function which estimates the power of such a test when the minor allele has a low frequency. This can be used to define the optimal pool size for detecting low frequency alleles. Simulations indicate that the estimates of the power of this test are robust to deviations from the assumptions and that such a test is conservative. This simple test is compared to an appropriately defined likelihood ratio test.