David M. Ramsey
Department of Mathematics and Statistics, University of Limerick
Andreas Futschik
Department of Statistics and Operations Research, University of Vienna

Statistical Tests for the Detection of Single Nucleotide Polymorphisms Using Genome Sequencers

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 (can be modelled using the Poisson or negative binomial distribution). 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. Due to the large number of sites on the genome it is necessary to adopt an appropriate multiple testing procedure (e.g. the Benjamini-Hochberg procedure). This simple test is compared to a likelihood ratio tests. Adaptations of these tests to real data are also considered.