Statistics and population genetics of haplotype sharing as a tool for fine-mapping of disease gene loci.

Nolte, Ilja Maria

This thesis discusses the Haplotype Sharing Statistic (HSS). HSS has been developed to map disease gene loci, in particular those involved in so-called complex genetic diseases. By analyzing haplotypes, i.e. combinations of alleles at associated genetic markers on the same chromosome, we believe that HSS contributes substantially to the existing statistical methodology for reasons described in this thesis. The hypothesis of HSS is that disease-associated haplotypes have descended from a small number of common ancestors at whose chromosomes a disease mutation occurred. Based on this assumption, current affected haplotypes will be identical for some small region spanning the disease mutation. HSS aims to identify the disease susceptibility locus by evaluating overlap of haplotypes, i.e. haplotype sharing, among patients and among controls. The marker that gives the maximal contrast in haplotype sharing is considered to be the best estimate of the location of the disease-susceptibility locus. In **Chapter 1**, current multilocus association methods based on the principle of present haplotypes being identical by descent for a segment spanning the disease mutation have been reviewed. Each method is evaluated by discussing how it is dealing with aspects from population genetics, statistics and design that are likely critical for its efficiency and usefulness to fine-map disease genes. As it is expected that no single method will be most powerful in every population and for each genetic model, it was not our intention to suggest which method is best or to give guidelines when to use a specific method. We do, however, list for each of the methods its advantages and disadvantages with respect to the issues mentioned above, leaving it to the reader to decide the usefulness of a method for the situation that he/she wants to investigate.

Chapter 2 is dedicated to the fundamental population genetics of haplotype methods. As mentioned before, haplotype methods are based on the assumption that patients have descended from common ancestors. Haplotype segments spanning the disease mutation that are shared by patients who have an ancestor in common for the disease mutation are therefore expected to be identical by descent (IBD). The question is whether identity of marker alleles on a haplotype can be considered as an indicator of IBD status. We have derived an analytical formula for the probability that two similar haplotypes are IBD. This enables evaluation of the influence of population genetic parameters on haplotype sharing. It appears that the probability that two haplotypes with identical markers are IBD is inversely proportional to the frequency of the haplotype in the population. The calculations also show that the probability of IBD increases when the number of markers with identical alleles increases and that the probability of IBD is higher in populations with a recent rapid expansion than in stable populations. Identical intermediate markers have a major influence on the probability of IBD, especially when the first two markers are far apart. This implies that length of haplotype sharing measured as the number of consecutive intervals with identical alleles may be an adequate measure to use in fine-mapping of disease gene loci.

The choice to use microsatellite markers or single nucleotide polymorphisms (SNPs) to create haplotypes also influences the probability that two haplotypes are IBD. Calculations show that haplotypes must consist of at least five microsatellite markers with identical alleles located within 1 Mb (~1 cM) to have a substantial probability to be IBD. This holds only if the population from which the haplotypes have been sampled is younger than 1,000 generations, e.g. the European population. On the other hand, haplotypes of three SNPs within 100 kb (~0.1 cM) are IBD with a high probability even in populations of 10,000 generations of age, i.e. the world population. Ancestral haplotypes consisting of

microsatellites may, therefore, be the same in different European subpopulations, but likely not in populations that separated earlier than European ancestry. For SNP haplotypes hold that they can be preserved all over the world and ancestral haplotypes may be the same worldwide.

In **Chapter 3**, several randomization methods for estimating the variance of the mean haplotype sharing have been compared. Repeated sampling without replacement provides separate variance estimates for haplotype sharing among patients and among controls, allowing patient and control samples to be drawn from heterogeneous (sub)populations in contrast to the reallocation method. The decision to sample without replacement is based on the fact that by sampling with replacement, i.e. the bootstrap method, a single haplotype may be drawn twice and sharing in a pair of the same haplotypes is infinite. This property is not desirable as infinite haplotype sharing is not representative of haplotype sharing in the entire population. Simulation results proved that indeed repeated sampling without replacement provides better variance estimates than the bootstrap and the reallocation methods, but is worse than the first-order jackknife procedure. Both the first-order jackknife and the repeated sampling without replacement procedure provide estimates of the variance of the mean haplotype sharing that are slightly too high. This results in a conservative test when evaluating the difference in haplotype sharing between patients and controls. The bias in the variance estimation decreases when the sample size is increased. Furthermore, as the number of randomizations increases, the distribution of the estimates converges meaning that, for 100, 1000 and 10,000 randomizations, the variance of the estimates relative to the variance of the true analytical estimates decreases from almost 6 to 1.57 and 1.12, respectively. Simulations showed that repeated sampling without replacement is robust with respect to the fraction that is sampled and to the number of markers comprised in the haplotypes. In order to expedite the randomization procedure, it is not required to sample a fraction of 0.5, but a smaller fraction, e.g. 0.1, may suffice. A drawback of the repeated sampling without replacement is, however, that a correction must be made for sampling a fraction. Deriving the correction term can probably only be accomplished when a formula for the variance is known. It may be difficult or even impossible to obtain such a formula for other complex statistics. In the case that was presented in **Chapter 3**, Appendix I, and in the general case of scalars, the correction term for a fraction of 0.5 equals 1, suggesting that repeated sampling without replacement of 50% of the data has a more general applicability.

In **Chapter 3**, an analytical sample estimate of the variance of the mean haplotype sharing is derived (Appendix II). This estimate can be used to compare the performance of the different randomization methods relative to the true variance. More importantly, it has allowed us to construct a version of HSS that is not based on randomization. The advantage of this analytical version of HSS is not only that, in this way, a statistically valid test is obtained, but also that the computational time to calculate the HSS in a dataset is much lower. This enables us to perform more extensive simulation studies to evaluate the power and mapping accuracy of HSS as shown in **Chapter 4**. Results show firstly that HSS appears to be more powerful and more accurate than single locus association analyses and the transmission/disequilibrium test (TDT) when different evolutionary histories underlie the patient and the control samples, in particular when patients coalesce faster to common ancestors than controls. If this latter is not the case, HSS does not detect excess in sharing among patients as compared to controls, even though strong association exists. Secondly,

simulation results show that HSS extracts different information from the data than association and TDT analyses do, since the results of HSS and association or TDT are not correlated under the null-hypothesis and only poorly when the alternative hypothesis is true. It is, therefore, suggested to apply both association or TDT analysis and HSS to evaluate empirical datasets, as combining such tests will probably improve power and mapping accuracy for fine-mapping of disease gene loci. Finally, the efficiency of HSS has been illustrated by applying it to two published datasets on hemochromatosis. In both datasets, the most significant HSS results have been obtained at the locus closest to the identified disease gene (i.e. HFE).

Chapters 5 and **6** show further simulation results of HSS. In **Chapter 5**, HSS has been applied together with non-parametric linkage (NPL) analysis. Simulations have been performed mimicking a genetically complex disorder in a partially isolated population of 10 generations. Multiplex families are selected for the NPL analysis. Haplotypes transmitted and non-transmitted to affected individuals from the last generation are selected as case haplotypes and control ones, respectively, for HSS. The difference in ascertainment schemes implies independence of the results and, therefore, allows combination of the results. It appears that HSS had power to map the simulated disease gene locus with great precision if the identity-by-descent status of the alleles is known (uncoded alleles). However, when alleles are downcoded to resemble microsatellites, HSS cannot detect excess of haplotype sharing among patients, where NPL still can, though with low mapping accuracy. It should be noted that the NPL results are probably inflated because full genotype information is assumed known for all affected individuals from all 10 generations. Reasons for not detecting a difference with HSS may be that a too sparse marker map was used (2-5 cM) such that preserved haplotypes are not present over many markers or in a substantial number. It is likely that for LD mapping higher marker densities are required to approximate full IBD information.

In the simulated dataset from the Genetic Analysis Workshop No. 12, HSS is able to locate the major gene involved in a complex disease in an isolated population (**Chapter 6**). No significant difference between patients and controls has been observed in the general population, implying either that the general population is too heterogeneous, or that the mutations in this population are too old such that haplotypes are not preserved over a substantial number of markers. It can also be that absence of a difference is an artifact due to too similar coalescence histories in the sample of patients and controls, as we saw before. Linkage disequilibrium analysis reveals that LD is one order of magnitude larger in the isolated population than in the general population, providing evidence for the second explanation. These results demonstrate that it is important to carefully choose the study population that is used for fine-mapping of disease gene loci involved in complex diseases.

In **Chapters 7** and **8**, HSS has been applied to empirical datasets. Among 124 multiple sclerosis (MS) patients, two frequent haplotypes in the human leukocyte antigen (HLA) region have been identified that are significantly more frequent and longer than among controls (**Chapter 7**). HLA has already long before been acknowledged to harbor genes involved in MS, but association results were inconsistent and contradictory. HSS, along with association analysis and the transmission disequilibrium test (TDT), has pinpointed the disease locus to an interval of only 51 kb between markers G511525 and D6S1666 in the HLA class II region. The only known gene in this region is DQB1.

In **Chapter 8**, HSS has beenapplied to a genome screen on Hutterite families in order to detect genes underlying asthma and related phenotypes, as provided by the Genetic Analysis Workshop No. 12. Results consistent with previous evidence for linkage are observed on chromosomes 7 and 12. The most significant result of HSS, however, indicates a possible susceptibility locus on chromosome 21, which has not been earlier identified as a candidate region for asthma.

In **Chapter 9**, we discuss that, although HSS has already been extensively tested on both simulated and empirical data, the true value of HSS still needs to be proven. At this moment, influences on the performance of HSS of missing and phase-ambiguous data, genotyping errors, errors in the marker order, use of distances instead of marker intervals to measure the length of haplotype sharing and the informativity of the markers are still unclear. Furthermore, we need to pay more attention to already developed, but not yet fully tested, methods related to HSS that we expect to make a substantial contribution to future research in fine-mapping of disease gene loci. One of these methods is the CROSS test, that measures haplotype sharing between a patient and a control haplotype in order to better account for association information that is not picked up by HSS. At the time of appearance of this dissertation, we have not succeeded in deriving the correct distribution of the CROSS statistic, like we have for HSS. Another test is the directional analysis that analyzes haplotype sharing from a marker in telomeric and centromeric direction separately. This test results in the identification of a single marker interval to contain the disease gene locus. Whether the identified interval can indeed be used as a confidence interval still needs to be evaluated. Additional future research should focus on testing HSS when SNPs are used instead of microsatellites, adapting HSS to make use of family data and adjusting HSS to analyze quantitative data. However, it will be important above all to apply HSS to empirical datasets because, especially then, irregularities arise that may influence the usefulness and efficiency of HSS. The true value of HSS will be demonstrated when a disease mutation is detected in an interval that was identified by HSS.