Maria Kabisch
Dr. sc. hum.

Imputation of Missing Genotypes in Genetic Studies through Generalizations of the Basic Coalescent

Einrichtung: Deutsches Krebsforschungszentrum (DKFZ)
Doktormutter: Prof. Dr. rer. nat. Ute Hamann

In the present work, novel breast cancer susceptibility variants were identified by fine-mapping using standard methods for the imputation of missing genotypes. With the aim of improving the accuracy of genotype imputation in genetic studies, properties of the basic coalescent and two generalizations thereof were investigated.

Breast cancer is the major cause of morbidity and mortality in women worldwide. The identification of inherited variants that confer susceptibility to breast cancer is important to improve risk prediction and to target screening and disease prevention measures to those women most likely to benefit. Three genetic studies on breast cancer susceptibility were presented in this work investigating the relationship of inherited variants in *USPL1* and *UBC9* of the SUMO system and *INCENP* of the chromosomal passenger complex with breast cancer risk and specific tumor subtypes. Association analyses were carried out based on genotype and phenotype data from the German GENICA study and additional European studies participating in the Breast Cancer Association Consortium (BCAC). To enhance association fine-mapping, variants not directly genotyped in the study population were imputed using additional markers genotyped in an external reference population. It was found that the non-synonymous coding SNP rs7984952 in *USPL1* was associated with breast tumor grade in the GENICA population. Previous associations of rs7187167, which is located in the putative promoter region of *UBC9*, with breast tumor grade in the GENICA population were not replicated in validation analysis taking advantage of additional cases and controls from BCAC. Several variants located within and downstream of *INCENP* (top SNP rs144045115) were identified that associate with ER-negative breast cancer risk across different European study populations.

Standard methods for genotype imputation make use of the present linkage disequilibrium between variants to infer missing genotypes. However, this is accompanied by a number of limiting factors, which may result in loss of imputation accuracy. In the present work, coalescent-based methods for genotype imputation were investigated to improve imputation accuracy. To this end, the basic coalescent and two generalizations thereof were applied to simulated and real data in a practice-oriented framework to estimate the population genealogy underlying the study and the reference. Based on the resulting gene tree, imputation templates were identified via the coalescence time between study and reference haplotypes. Imputation accuracy of coalescent-based genotype imputation was evaluated as a function of population growth and structure and in comparison to standard imputation with IMPUTE2. Simulation experiments revealed that coalescent-based genotype imputation was of higher accuracy than IMPUTE2, in particular for the imputation of common variants. This effect was maintained in the presence of population growth and two subpopulations, but was attenuated when there were three or more subpopulations due to the increased number of rare variants. In the presence of population growth and/or structure, the genotype imputation based on the basic coalescent resulted in practically identical accuracy as the genotype imputation based on the generalized coalescent. Real data application using subpopulations from the 1000 Genomes Project as hypothetical study populations revealed that coalescent-based genotype imputation and IMPUTE2 were of similar accuracy when variants were imputed into regions of low recombination. In high recombination regions, IMPUTE2 was of higher accuracy than coalescent-based genotype imputation. To exploit the full potential of coalescent-based methods for the imputation of missing genotypes in genetic studies, advanced methods for modelling the gene genealogy for structured populations and for incorporating estimates of recombination are needed, as well as the implementation of these methods in user-friendly computer programs.