Barbara Peil
Dr. sc. hum.

Tailored reference panel selection using statistical depth to improve genotype imputation accuracy

Promotionsfach: Medizinische Biometrie und Informatik
Doktorvater: Priv.-Doz. Dr. sc. agr. Justo Lorenzo Bermejo

Genotypes that are not directly measured in a genetic study can be estimated ('imputed') using external genotype repositories, for example the HapMap or the 1000 Genomes Project. Several methodological studies have evaluated which data is most helpful for an accurate genotype imputation. The proposed techniques can be grouped into two categories. Initial approaches aimed to identify reference individuals that best reflect recombination patterns in the study population. More recent techniques aim to maximize the genetic diversity in the reference panel. It is well known that big reference panels are advantageous for genotype imputation.

In this thesis a novel strategy to select individuals for the reference panel relying on the identification of the ancestral kernel of the study population was developed and evaluated. The aim was the maximization of the accuracy of the subsequently imputed genotypes.

In a first set of simulations, subsets of individuals from an external data repository were selected based on several selection strategies. The subsets were then used as reference panels to impute variants in two different studies, and they were compared regarding the achieved accuracy of imputed genotypes. Classical selection strategies included random selection with fixed reference panel size, and the identification of the HapMap subpopulation genetically most similar to the study population. Genetic similarity was quantified by Wright's F_{ST} statistic and additionally by visual inspection after a principal component analysis. The novel strategy relied on a bivariate generalization of the boxplot, the bagplot, after a principal component analysis. Different reference panels with a fixed size and the largest univariate and multivariate depths, were also investigated. Genotype imputation accuracy was accessed by using leave-one-out cross-validation of the measured variants in the study.

In the two investigated studies, the size of the reference panel had a substantial influence on the imputation accuracy. When the size of the panel was fixed, simulations suggested that imputation benefits from a targeted selection.

It has recently been shown that in addition to genotypes from external data repositories, the integration of sequences from own-study individuals into the reference panel may improve the accuracy of imputed genotypes. In a second set of simulations, different strategies for the selection of own-study individuals for dense genotyping were compared regarding the imputation accuracy. The simulated study scenarios consisted of several combinations of subpopulations from HapMap phase III. HapMap individuals who did not belong to the study population constituted an external reference panel. This external panel was complemented with the sequences of study individuals selected according to different strategies. In addition to a random choice, individuals with the largest statistical depth for the first one, the first two and the first three genetic principal components were selected. In order to assess imputation accuracy, HapMap genotypes of study individuals not selected for sequencing which were not available on the Affymetrix SNP Chip 6.0 array were masked, and subsequently imputed relying on different reference panels.

The integration of sequences from a subset of the study population increased imputation accuracy in

all simulated study scenarios. The selection of sequenced individuals based on the univariate depth resulted in the highest mean IQS for the European-study scenario, whereas a random selection was the best strategy for the African-study scenario. The imputation of rare variants favored a selection based on the first three principal components.

The discrepancy between genetic diversity on the one hand, and similarity of linkage disequilibrium patterns in the study population and reference panel on the other hand, challenges the selection of study individuals for sequencing. Study-specific weights for both aspects need to be considered in order to optimize imputation accuracy. By masking and imputing measured variants, the optimal balance can be obtained in analogy to the simulations conducted in this work.