I12. Data mining in the 1000 Genomes Project: A closer look at blood groups in Africa

Mattias Möller & Martin L. Olsson[1]

Division of Hematology and Transfusion Medicine, Department of Laboratory Medicine, Lund University, Lund, Sweden

Blood group genotyping has recently developed into a clinical tool to improve compatibility of blood transfusions and management of pregnancies. Next-generation sequencing (NGS) is rapidly moving toward routine practice for patient and donor typing and has the potential to remedy some of the limitations of currently used platforms. However, a large-scale investigation into the blood group genotypes obtained by NGS in a multiethnic cohort is lacking. The 1000 Genomes Project provides information on genome variation among 2504 individuals representing 26 populations worldwide. We extracted their NGS data for all 36 blood group systems to a custom-designed database. In total, 210 412 alleles from 43 blood group–related genes were imported and curated. Matching algorithms were developed to compare them to blood group variants identified to date. Of the 1241 non-synonymous variants identified in the coding regions, 241 are known blood group polymorphisms. Interestingly, 357 of the remaining 1000 variants are predicted to occur on extracellular portions of 31 different blood-group-carrying proteins and some may represent undiscovered or altered antigens. Of the alleles analyzed, 1504 were not previously described. Intriguingly, the *ABO/GBGT1/FUT2/FUT3* and *GYPB/GYPC* genes showed the highest degree of variation per kilobase coding sequence, and *ACKR1* variants had the most skewed distribution across 5 continental superpopulations in the dataset. These genes have all been implicated as susceptibility markers and pathogen receptors, e.g. in infectious diseases like malaria, HIV and gastroenteritis.

When looking closer at the *ABO* alleles in the dataset, we realized that a 5,821-bp deletion encompassing exons 5-7 was called in twenty 1000G individuals, predominantly Africans. This allele was confirmed and its exact deletion point defined by bioinformatic analyses and *in vitro* experiments. A PCR assay was developed, and screening of an independent cohort of African samples revealed three more donors heterozygous for this deletion, which was thereby phenotypically established as an *O* allele. Analysis of upstream genetic markers indicated a presumed ancestral origin from *ABO*O.01.02*. A second deletion was called in nine individuals and showed greater diversity regarding ethnic/geographic distribution and allelic background. It could neither be confirmed by *in silico* nor *in vitro* experiments. Thus, a previously uncharacterized *ABO* deletion among Africans was comprehensively mapped and a genotyping strategy devised. The false prediction of another deletion emphasizes the need for cautious interpretation of NGS data and calls for strict validation routines.

In summary, blood group gene results from the 1000 Genomes project were exported to an online search engine, www.erythrogene.com, which presents data according to the allele nomenclature developed for clinical reporting by the International Society of Blood Transfusion (ISBT). This database deepens our knowledge on blood group polymorphism globally and provides a long-sought platform for future research.