Mining of 141, 456 high-quality human exomes and genomes reveals the presence of 10, 909 putative immunoglobulin heavy chain IGHV variants

Tiago Mendes, Liza Figueiredo Felicori Vilela, Lucas Alves de Melo Pontes, Fábio Martins

UNIVERSIDADE FEDERAL DE MINAS GERAIS, Ufv, UFMG

Abstract

The correct identification of alleles can assist in the study of several human diseases associated with the antibody repertoire and in the development of new therapies using antibody engineering techniques. The advent of next-generation sequencing of human genomes and antibody repertoires enabled the development of several tools for the mapping and identification of new immunoglobulin (Ig) alleles. Some of these tools use 1, 000 Genomes (G1K) data for new Ig alleles discovery. However, genome data from G1K present low coverage and variant call problems. For these reasons, we used in this work, data from the Genome Aggregation Database (gnomAD), the largest high-quality catalogue of variation from 125, 748 exomes and 15, 708 human genomes. The methodology developed in this work identified 10, 909 putative immunoglobulin heavy chain variable region gene (IGHV) alleles, in which 10, 828 of them are new and 2, 024 appear at least in 6 different alleles. IGHV2-70 was the IGHV gene segment with the largest number of variants described. The majority of the variants were found in the framework 3 and most of them are missense. Interestingly, a large number of variants were found to be population exclusive. A database integrated with a web platform was created (YGL-DB) to store and make accessible the likely new variants found. This database is the largest human putative IGHV alleles repository to date. This available data can help the scientific community to validate new IGHV variants through the design of new primers (specific or not to a given population) or even to validate new variants found from AIRR-seq data.

Funding:

Link to Video: