IMGT/HighV-QUEST 2011

Eltaf ALAMYAR[1], Véronique GIUDICELLI[1], Patrice DUROUX[1] and Marie-Paule LEFRANC[1]

[1] IGH, UPR CNRS 1142, 141, rue de la Cardonille, 34396, Montpellier, Cedex 05, France
{eltaf.alamyar,giudicel,patrice.duroux,marie-paule.lefranc}@igh.cnrs.fr

Keywords Next-Generation Sequencing, NGS, immunoglobulin, antibody, T cell receptor, expressed immune repertoires, IMGT-ONTOLOGY.

The analysis of expressed repertoires of antigen receptors - immunoglobulins (IG) or antibodies and T cell receptors (TR) - represents a huge challenge for the study of the adaptive immune response in normal and disease-related situations, such as viral infections. To answer that need, IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) has developed IMGT/HighV-QUEST [1]. IMGT/HighV-QUEST is devoted to the analysis of large repertoires of IG and TR sequences that result from Next Generation Sequencing technologies. IMGT/HighV-QUEST, a high throughput version of IMGT/V-QUEST, analyses up to 150.000 sequences per run. It identifies the IG and TR variable (V), diversity (D) and joining (J) genes and alleles by alignment with the germline IG and TR gene and allele sequences of the IMGT reference directory. It describes the V-REGION mutations and identifies the hot spot positions in the closest germline V gene. It integrates IMGT/JunctionAnalysis for a detailed analysis of the V-J and V-D-J junctions, and IMGT/Automat for a full V-J- and V-D-J-REGION annotation. The analysis is based on the IMGT-ONTOLOGY concepts of description, classification and numerotation.

IMGT/HighV-QUEST uses two different systems of HPC resources at CINES, and a local computational server, in order to perform the analysis of submitted sequences by a standalone version of IMGT/V-QUEST. Since the management of many analyses of thousands of sequences is a challenging task, IMGT/HighV-QUEST manipulates a local database for the local analysis queue, and in order to manage the jobs, the tasks are split into three independent layers. The web service layer is responsible for providing user interaction facilities and adding new analyses in the local queue. The scheduled tasks layer (also called background layer) includes all core logical functionalities of IMGT/HighV-QUEST. Background operations such as selection of a resource, dispatching of analyses, monitoring the running jobs, preparation of results of the completed analyses are performed in this layer. Finally the computational resources layer is where the real analysis of user sequences is performed. The analysis results of IMGT/HighV-QUEST comprise a set of text files which include 11 files in CSV format equivalent to the eleven sheets of the 'Excel files' of IMGT/V-QUEST and, for each analysed sequence, the 'Detailed view' that allows one to visualize the individual detailed results. These result files are archived in a single ZIP file that is downloaded by the user.

Since its availability in October 2010, more than 41 million sequences have been submitted and 118 users have registered to IMGT/HighV-QUEST (11/05/11). The jobs required 17,000 computational hours of resources and generated about one terabyte of results data. More than three quarters of the sequences were submitted by users from USA, the others being submitted by users from EU for most, but also from China, Japan, Australia, Canada, Korea and Venezuela. New functionalities have been developed that comprise the introduction of statistical analysis on the results of the batch to help the user in estimating the reliability of the results. Statistics are performed on results selected as ‘1 copy’ (redundancies are enregistered but not treated), and with quality criteria (identification of a single gene/allele, known functionality, REGION length, absence of IMGT/V-QUEST warnings regarding the CDR1-IMGT and CDR2-IMGT lengths or the percentage of identity). These statistics include the frequency of gene expression and of CDR3-IMGT length; they also report the number of identical CDR3-IMGT sequences and the number of sets of CDR3-IMGT with identical nucleotides (nt) and amino acid (AA) sequences. These functionalities, which have been set up in a first step for the human TR, are particularly useful to evaluate the significance of the results of a batch.

Acknowledgments: this work was granted access to the HPC resources of CINES under the allocation 2010-036029 made by GENCI (Grand Equipement National de Calcul Intensif).

[1] E. Alamyar, V. Giudicelli, P. Duroux and M.-P. Lefranc, IMGT/HighV-QUEST: A high-throughput system and web portal for the analysis of rearranged nucleotide sequences of antigen receptors - High-throughput version of IMGT/V-QUEST, *Proceedings of the 11th Journées ouvertes en Biologie, Informatique et Mathématiques (JOBIM)*, Montpellier, P27 pp. 156, 2010.