Challenges in whole exome sequencing to identify disease-causing variants in human rare diseases

Javier Santoyo Lopez✉
Medical Genome Project (MGP), Genomics & Bioinformatics Platform of Andalusia (GBPA), Seville, Spain

Next Generation Sequencing (NGS) Technologies have greatly improved our ability to mine variants out of the entire genome, and the extensive use of Whole Exome Sequencing (WES) has allowed the discovery of variants responsible of disease especially in monogenic diseases. However, finding disease-causing variants requires a primary and secondary analysis where several factors that can affect a successful outcome need to be taken into consideration.

Using Life Technologies SOLiD sequencers and Nimblegen SeqCapEZ exome capture library, we have systematically analyzed more than 300 exomes accounting for 20 different diseases and more than 200 exomes from control Spanish population within the Andalusian Medical Genome Project (MGP), a public-private partnership that aims to find mutations responsible of hereditary Rare Diseases and to establish first steps towards the implementation of personalized medicine.

Thus, the reliability of calling variants in a sample is highly related to the sequencing instrument used due to the sequencing chemistry and the intrinsic properties of each sequencing technology, so in the first place we have optimized a primary analysis pipeline to obtain high-quality variants from color-space. This pipeline provides high-quality variants with an extremely low rate of false positives as shown by Sanger sequencing validations. Furthermore, the analysis of whole exome data, shows more than 98% correlation of normalized coverage for library pairs and the distribution of coverage is consistent between samples, which shows a good reproducibility of the enrichment and sequencing technology.

After sample variant calling a secondary analysis pipeline filters those variants that are present in variant public databases and have a MAF that indicates that are SNPs. The filtering of SNPs from local population is a critical step, as there are many SNPs in our study population that are not represented in the public databases (e.g. dbSNP or 1000 Genomes), which is done by using MGP database for WES Spanish Healthy Control Individuals. Finally mutations are filtered based in family pedigree keeping only those mutations that segregate with the disease.

All this WES approach and subsequent variant analysis has allowed us to find the gene and the mutations responsible of disease in a great number of retinal dystrophies in particular for Retinitis Pigmentosa Autosomal Recessive (RPAD).