# Development of pipeline for exome sequencing data analysis

**M.R. De Filippo[1] ✉, G. Giurato[1], C. Cantarella[1], F. Rizzo[1], F. Cirillo[1], A. Weisz[2]**

[1]Laboratory of Molecular Medicine and Genomics, Faculty of Medicine and Surgery, University of Salerno, IT, Naples, Italy
[2]Division of Molecular Pathology and Medical Genomics, 'SS Giovanni di Dio e Ruggi d'Aragona' Hospital, University of Salerno, Italy

## Motivations

Exome sequencing the targeted sequencing of the subset of the protein coding human genome is a powerful and cost-effective new tool for dissecting the genetic basis of diseases and traits that have proved to be intractable to conventional gene-discovery strategies. Until now many algorithms have been produced, each of them addressing a different task in the downstream analysis of next-generation sequencing (NGS) data. The aim of this work is to combine these algorithms into an analysis pipeline for the detection of SNP and deletion/insertion polymorphisms within DNA sequences obtained by whole exome sequencing. The pipeline tested with data obtained from SRA (http://www.ncbi.nlm.nih.gov/sra), will then be applied to studies undergoing in our laboratory.

## Methods

Starting from raw sequence data, we first performed quality statistics and filtering of sequence reads and then aligned them to a reference genome. To this end, BWA was used to align both single- and paired-end reads for its computational efficiency and multi-platform compatibility. Post-alignment analysis, including removal of duplicate reads and quality score recalibration, was carried out using GATK, which takes into account several covariates such as machine cycle and dinucleotide context. Next, SNP calling was done using GATK UnifiedGenotyper, that uses a Bayesian model to estimate the most likely genotypes and allele frequency in a population of N samples, giving an annotated VCF file as output. Subsequently, variant quality score was recalibrated to estimate the probability of each variant being a true polymorphism, rather than a sequencer, alignment or data processing artifact, and finally filtered to improve the accuracy of genotype and SNP calling.

## Results

The results obtained support the accuracy of our pipeline to identify SNP and short indels, to provide a global and quantitative catalog of nucleotide variants in the exome. The next step will be to apply this pipeline to samples sequenced in our laboratory.