GAMES: a new tool for genomic annotation of next generation sequencing data

Sana ME[1], Galasso M[1], Volinia S[1]

Motivation

Next-Generation Sequencing (NGS) methods are rapidly revolutionizing the landscape of biomedical science, but at the same time pose the problem of the analysis of the huge amounts of data. There are many commercial and public software packages for the analysis of NGS data. However, the outputs of these tools often seem to be poorly annotated and of difficult functional interpretation. Detection of single nucleotide polymorphisms (SNPs), insertions and deletions (InDels) and other genetic rearrangements are among the major aims of processing ultra high-throughput sequencing data. In this work, we present GAMES (Genomic Analysis of Mutations Extracted by Sequencing), a new tool that, from the alignment of the reads to a reference genome, performs mutational analysis and integrates the data with genome annotations. For each SNP, the program extracts the information about the coordinates in genome browser, the genomic location and associated protein/s, the effect of the mutation in the translated codons, the conservation in placental mammal phylogenetic tree, and the association with known SNPs in HapMap and dbSNP (NCBI). GAMES aims to be a middleman between hard data and the interpretation by the investigator.

Methods

We applied GAMES to reads obtained by Genome Analyzer (Illumina) after SureSelect Target Enrichment of 36 genes involved in hypertrophy cardiomyopathy. The data were processed using BWA (Burrows-Wheeler Alignment) to map the reads to the human genome (hg18) and to extract the alignment in SAM (Sequence Alignment/Map) format. As a first step, for each mismatch in the alignment, GAMES extracts the position and the respective base in the reference sequence, consensus quality score, per-base sequence coverage, counts and frequencies, and the repetitivity, defined as the number of reads that can be uniquely mapped to cover this location. The script evaluates heterozygotes, defined as the two best calls for each position in the reads. The implemented quality parameter is the standard Phred score, logarithmically linked to error probabilities, that determinates the accuracy of consensus sequence and of the alignment. The list of SNPs and InDels is filtered by a set of threshold parameters: quality, minimum coverage, minimum count and repetitivity (non-unique elements

[1] DAMA, Data Mining for Analysis of Microarrays, Department of Morphology and Embryology, University of Ferrara, Italy

are filtered out). A second stage of the script queries various databases to extract, together with genomic information, the effect of the mutation on the primary structure of the protein (if in coding regions), the known isoforms (if any), the possible known SNPs, the placental mammal conservation scores (PhyloP). This kind of analysis allows to underline the biological relevance of a SNP in cds, of a non-synonymous mutation or of an insertion/deletion in the coding regions. GAMES offers as processed output different useful files: -the annotations for each mutated nucleotide (text tab delimited file); -the selected protein alterations with links to major databases, NCBI, UCSC, dbSNP, HAPMAP, (html file); -the coordinates and coverage of reads (BED files); -the mutated positions for a sample aligned on Genome Browser as a track (MAF file); -the multiple alignment with the mutated bases for each of a number of sequenced genomes (MAF composite file).

Results

We proposed GAMES, a new application for mining functional SNPs and InDels from NGS data. Its sensitivity and specificity and the accuracy of SNP calling are guaranteed by taking into account different parameter, such as read length, Phred quality score, the number of reads and sequence covering a position, the repetitivity and the conservation of the mutated nucleotide during the evolution. GAMES provides various output files: MAF and BED files are useful for visualization (for example, UCSC genome browser and IGV) and for the evaluation of the experimental design. The main purpose of our script is aimed to gain biological insight for the mutation using a output immediately comprehensible to whom works in life science, even if without bioinformatic expertise. GAMES aims to be not only a tool for SNPs and InDels detection, but also one with a direct impact on the understanding of the significance of NGS data.

Contact e-mail

mariaelena.sana@unife.it