Genome Medicine

SOFTWARE Open Access

Annotate-it: a Swiss-knife approach to annotation, analysis and interpretation of single nucleotide variation in human disease

Alejandro Sifrim[1,2*], Jeroen KJ Van Houdt[3], Leon-Charles Tranchevent[1,2], Beata Nowakowska[3], Ryo Sakai[1,2], Georgios A Pavlopoulos[1,2], Koen Devriendt[3], Joris R Vermeesch[3], Yves Moreau[1,2] and Jan Aerts[1,2]

Abstract

The increasing size and complexity of exome/genome sequencing data requires new tools for clinical geneticists to discover disease-causing variants. Bottlenecks in identifying the causative variation include poor cross-sample querying, constantly changing functional annotation and not considering existing knowledge concerning the phenotype. We describe a methodology that facilitates exploration of patient sequencing data towards identification of causal variants under different genetic hypotheses. Annotate-it facilitates handling, analysis and interpretation of high-throughput single nucleotide variant data. We demonstrate our strategy using three case studies. Annotate-it is freely available and test data are accessible to all users at http://www.annotate-it.org.

Background

Context

With the advent of massively parallel high-throughput sequencing technologies and the increasing availability of reference genomes, new opportunities emerge for discovering genome-wide variation across individuals and populations, at both the large-scale level (deletions, duplications, and rearrangements) and the base-pair level (single nucleotide variants and small indels and repeats). As more and more sequence data are produced, accurate assessment of the frequency of variants in specific subpopulations (patients versus controls or any phenotypically different populations) is vital to the interpretation of how these variants segregate across populations. International projects, such as the 1000 Genomes Project [1] and the Hapmap Project [2], have been set up to assess the genetic variation across large groups of 'normal' human individuals. Full-genome [3], whole-exome [4-6] or targeted gene panel [7,8] sequencing studies of individuals struck by Mendelian disorders can aid in identifying the genetic cause for these diseases; for example, by leveraging publicly available data, under the assumption that these rare variants do not occur in the normal population. Furthermore, trio sequencing of an affected child and of his or her parents can identify *de novo* variants in sporadic cases of genetic disease [9,10].

Because of the large size and complexity of next-generation sequencing data sets, new computational and statistical methods for analyzing and interpreting the data are required to accurately find the variation of biological interest. To transform raw sequencing data into variation data, we need to undertake the following steps: (1) sequence alignment, (2) variant calling, (3) variant annotation, and (4) variant interpretation [11,12]. In this study, we mainly focus on the latter two steps. After sequence alignment and variant calling, we end up with a list of variants with their genomic coordinates and the variant alleles that differ from the reference sequence. Based on current knowledge of functional elements annotated in the human genome sequence, overlapping variants within the annotated features are found and the impact on RNA and protein sequence level is computed. Variant lists can be reduced further by applying functional impact prediction tools, such as Polyphen2 [13], SIFT [14], FoldX [15], and others [16]. Such tools are computationally intensive and require dedicated

* Correspondence: alejandro.sifrim@esat.kuleuven.be
[1]KU Leuven, Department of Electrical Engineering-ESAT, SCD-SISTA, Kasteelpark Arenberg 10, B-3001, Leuven, Belgium
Full list of author information is available at the end of the article