Overview

Table of Contents:

Motivation

Quite recently NGS techniques started to be used for the identification of trait-causing genomics variants. Interestingly, a sheer wealth of candidate causal mutations can be found in any human genome (~5 million per human genome, ~150.000 per human exome) many of which may provide a compelling story about how the variant may influence the trait; the so-called narrative potential of human genomes. The first step for these data to be translated into effective knowledge (e.g. clinical knowledge) is through the integration of reference annotation data that enables their filtering and provides a context for accurate interpretation.

The ability to obtain the list of genomic variants in any human genome has paved the way for the development of massive reference projects such as "The 1000 Genomes Project", which has sequenced the genome of more than 3000 individuals, The NHLBI Exome Sequencing Project, which has sequenced more than 6500 exomes or "The Genome of the Netherlands" who sequenced genomes from more than 700 individuals, among others. These projects are generating a genomic variation knowledgebase which sets an essential base to identifying disease-causing genotypes. These data are heavy, dispersed in different repositories, lack normalization and are delivered in different tastes of VCF-like formats. Accessing to all these data, putting information from the different projects into relation and being able to integrate those data in the analysis process is often extremely painful.

Features

- Most important high-quality variant studies normalised and integrated in just one server database
- Rich variant annotation performed, including HPO terms, consequence types, substitution effect prediction scores, Gene Ontology terms, etc.
- Population frequencies calculated, including populations and super-populations
- Interactive web-based data mining tool based on IVA
- Clients in Python, Java, JavaScript for fast programmatic access

Resource

The Human Genomic Variation Archive (HGVA) is an open access genetic variation resource that integrates all variants from key world-wide reference projects, but also added-value information such as basic variant annotation, population frequencies, protein effect predictions, variant-associated phenotyes, etc. HGVA currently hosts about 300GB of data from 13 different studies describing more than 200 million variants. HGVA is not a mere data archive, but a big data provider that enables users to efficiently query, filter and retrieve relevant information from its knowledge-base, either from a visual web-interface or programatically.