iNGS: a prototype tool for genome interpretation and annotation

Ismael Navas-Delgado✉, **Maria Jesús García Godoy, Fátima Arjona-Pulido, Trinidad Castillo-Castillo, Ana Isabel Ramos-Ostio, Sarai Infantes Díaz, Ana Medina García, José F. Aldana-Montes**

University of Málaga, Spain

Received 31 July 2013; **Accepted** 10 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Currently, clinical interpretation of whole-genome NGS genetic findings are very low-throughput because of a lack of computational tools/software. The current bottleneck of whole-genome and whole-exome sequencing projects is in structured data management and sophisticated computational analysis of experimental data. In this work, we have started designing a platform for integrating, in a first step, existing analysis tools and adding annotations from public databases to the findings of these tools. This platform can be used to produce tools for different kind of users. As a first experiment with this platform, we have developed a Web tools for running multiple analysis tasks, completing the findings with public data and producing a simple report similar to blood test reports.

Motivation and Objectives

Personalised medicine can be considered as a medical approach which proposes the customisation of healthcare involving medical decisions, treatments, etc., that are applied to a patient individually and tailored to that patient. This medical methodology is possible due to the rapid advances in technology in areas such as genomics, transcriptomics, proteomics, metabolomics, etc. In this context, it is important to mention that the development of sequencing approaches of personal human genomes and the detection of DNA variations by means of a reference human genome that was unveiled in 2003-2004 (Human Genome Sequencing Consortium International, 2004) are both huge contributions that should be integrated into the personal "omics" (in this case, genomics) of each patient.

Nonetheless, clinical interpretation of whole-genome and NGS (Next Generation Sequencing) genetic findings are currently very low-throughput because of a lack of computational tools/software to integrate all this information. In this sense, the reason for the current bottleneck of whole-genome and whole-exome sequencing projects is the management of structured data and sophisticated computational analysis of the experimental data obtained.

Therefore, we have started designing a platform for integrating, firstly, existing genome analysis tools and then adding more annotations than those currently provided from the findings of these tools in public databases. As a first experiment with this platform, we have developed iNGS[1], a Web tool for running multiple analysis tasks. All findings of these analysis tools are completed with public data to generate a simple report to complete information provided by genome interpretation and annotation tool results.

Method

In the context of tools for analysing the whole-genome, there are many available tools that are able to provide users with NGS genetic findings. Some of these tools, described in more detail below are Annovar (Wang et al., 2010), GATK (McKenna et al., 2010), SeatleSeq (Ng et al., 2009), VAAST (Yandell et al., 2011) and Galaxy (Goecks et al., 2010).

Annovar is a command-line tool widely used to annotate functional effects of variants with respect to genes, genomic region-based annotations (which refer to those regions that are different to genes) and compare variants with those variations stored in databases. Furthermore, an interesting feature of this tool is the detection of diseases associated with the regional annotations of GWAS[2] (Genome-Wide Association Studies) catalogue.

GATK is a Genome Analysis Toolkit that presents five main functionalities: 1) initial read mapping; 2) local realignment around INDELs (insertions or deletions); 3) base quality score recalibration; 4) SNP (single-nucleotide polymorphism) discovery and genotyping to find all potential variants; and 5) machine learning to separate true segregat-

1 http://khaos.uma.es/iNGS
2 http://www.genome.gov/gwastudies/

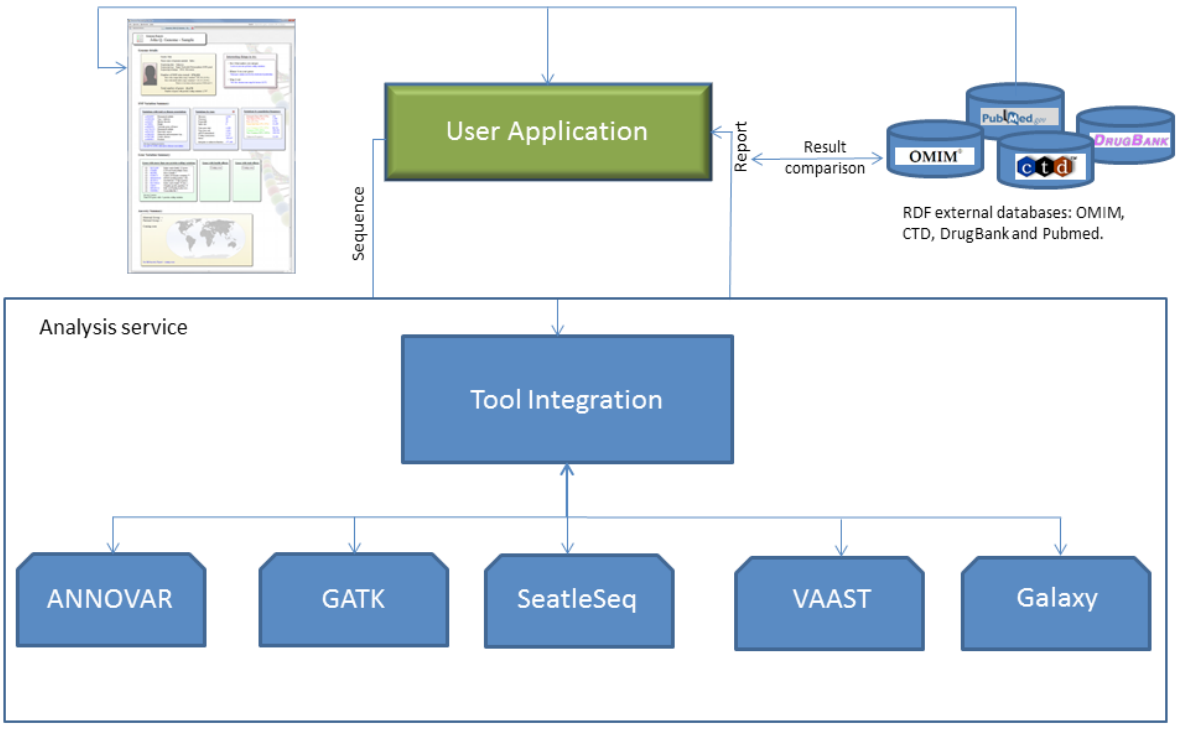

Figure 1. Platform and Pilot User Tool.

ing variation from machine artefacts common to NGS technologies.

SeatleSeq Annotation provides annotations of known and novel SNPs. The annotations include dbSNP (rs IDs), gene names and accession numbers, SNP functions, protein positions, amino-acid changes, conservation scores, HapMap frequencies, PolyPhen predictions and clinical associations. Furthermore, this tool has been tested to identify rare and common variants in over 300 megabases (Mb) of coding sequence using exomes from Freeman-Sheldon syndrome patients.

VAAST is a probabilistic tool that has been designed to identify damaged genes and their disease-causing variations in personal genome sequences. This tool has been benchmarked in multiples studies, which include 100 Mendelian conditions (Ng et al., 2009) and also the identification of genes responsible for common diseases (Lesage et al, 2002).

Galaxy is a framework that provides a set of web-based tools including different analysis variation tasks. This tool is used to look for disease SNPs in a full genome, to detect SNPs differing between populations, to look for disease-associated SNPs in a pedigree and for population structures and selective sweeps.

Besides the aforementioned tools, there are many databases that provide up-to-date findings that can help to provide interpretations for the genome analysis of these tools. Examples of such findings could be DNA variants related with diseases, genes, bibliography, drugs and drug targets. This information is publicly available in different data sources such as OMIM (Hamosh et al., 2005), CTD (Mattingly et al., 2003), DrugBank (Wishart et al., 2006) and PubMed (Roberts, 2001). Furthermore, as Figure 1 shows, the output files obtained contain information on DNA variants detected by the genotype analysis that can be completed with information from different repositories. To do this, we have selected a set

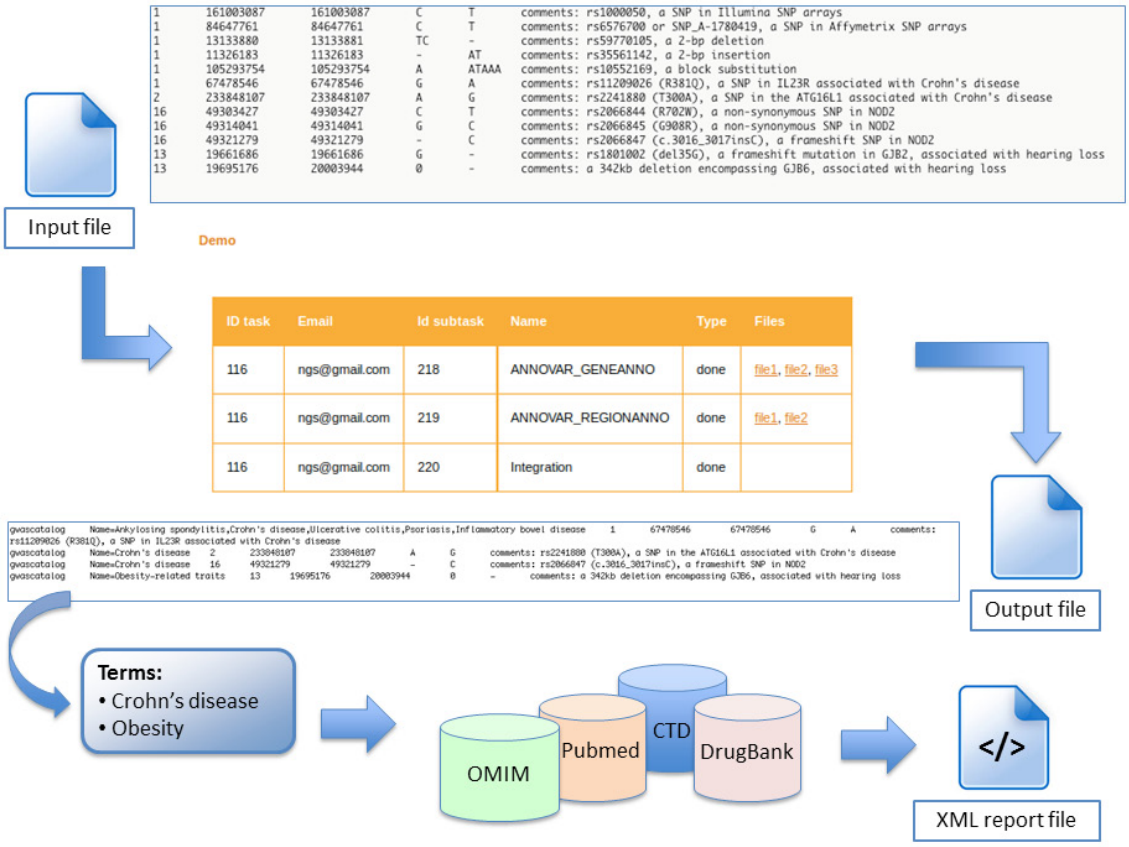

Figure 2. General scheme of how the genome-analysis tool platform works. The user introduces an input file (e.g. a whole-exome sequence of a patient). The analysis tools previously selected by the user process the file generating an output file. The interesting terms (e.g., disease and gene annotations) written in the output file are extracted automatically from the output file and included as parameters in a parameterized SPARQL query. The information retrieved is included in a report in XML format to complete the information obtained from the genomic analysis tools. The user with a biological or medical background can interpret this report.

of RDF (Resource Data Framework) data sources, information of which is stored following the principles of Open Linked Data.

OMIM endpoint provided by Bio2RDF server[3], contains data classified into different classes (such as allelic variant, clinical synopsis, gene, phenotypes gene-phenotypes etc.) that can be useful for phenotype-genotype analysis.

CTD endpoint[4] provides information on annotated associations between genes, diseases, proteins, bibliographic references and toxic agents.

DrugBank[5] is a Bio2RDF endpoint the data structure of which is divided into several interesting classes such as drug-drug, drug-target, drug-enzyme and drug-transporter interactions.

We have also included the PubMed repository[6] which is an important source of bibliographic data. Furthermore, it is worth mentioning that PubMed resources are mostly associated with other PubMed resources stored in OMIM, CTD and DrugBank endpoints.

The availability of these databases allows a platform to be designed that integrates analysis tools and public data, to be used by end-users without requiring them to install and configure complex software packages. Figure 1 depicts a general view of the proposed platform and how it can be used to develop a tool for combining analysis tools and public data. In this case, the tool presented aims to provide simple to understand reports. The analysis processes will run on

3 http://s4.semanticscience.org:16019/sparql
4 http://s4.semanticscience.org:16004/sparql
5 http://s4.semanticscience.org:16006/sparql
6 http://pubmed.bio2rdf.org/sparql

the served side, freeing the user from having to rely on computational resources.

The first prototype built using this platform will run several analysis tools in parallel, generating a set of DNA annotations results: element variants in genes (either exonic and intronic regions) and intergenic regions and their relationships with diseases according to GWAS studies, transcription binding site annotations, DNA variation annotations that fall within conserved genomic regions etc. Then, these results will be complemented with information retrieved from biological endpoints through a set of SPARQL queries. To do this, the outputs of the analysis tools (e.g., annotated diseases, gene names, segmental duplication annotations, etc.) will be used to retrieve information from the data sources to complete the analysis findings of the platform tools. Figure 2 shows how this platform works.

In this regard, this work attempts to ease the task of understanding DNA variants such as short and large INDELs, SNPs, especially disease-associated SNPs, large-scale rearrangements, CNVs (copy number variation), and their functional consequences in target genome sequence(s) by completing these findings with a report containing information on genotypes, phenotypes, diseases, bibliography, drugs and drug targets.

Results and Discussion

To test the functionality of the prototype tool, we have selected an input (CEU.low _ coverage.2010 _ 07.indel.sites) in VCF format from the 1000 genomes repository[7]. This input file was loaded into the analysis tool platform. The region-based annotation analysis using GWAS catalogue demonstrated that CEU population contains a 1bp insertion (C) in exonic region of NOD gene that is related with Crohn's disease (among other relevant DNA mutations). The NOD genes and Crohn's disease terms were extracted automatically from the output files and were included as parameters in a previous parameterized SPARQL query to retrieve information from GWAS, OMIM, DrugBank and Pubmed data in order to complete the information obtained by the analysis tools of iNGS. The data retrieved contain phenotype, genetic, bibliographic and pharmacogenomic information related with the extracted terms. Such information is included in an XML report to be visualised by the user such as that shown in the schema in Figure 2. This task represented a challenge due to the lack of standards in this context. Regarding automating the term to be extracted from each output file (e.g. disease or gene annotations), it was necessary to analyze how the annotations detected by each analysis tool are structured. This implies that the annotation structure of the outputs of each tool integrated in the iNGS platform had to be analysed according to the different output file formats (e.g., Annovar has its own output and input file formats).

The aforementioned use case indicates that this platform could be useful for the analysis of genome sequence inputs. The generated reports are completed with integrated information from different data sources. As mentioned in the Method section, such information is retrieved through these SPARQL queries. At this point it is important to mention that another important challenge was the design of this set of SPARQL queries. Initially, we have designed a set of queries that retrieves non-redundant and repetitive information. In this way, the information is represented in an XML report making it more comprehensible for users.

For future work, we are planning to integrate a query federation system in order to efficiently retrieve information from more than one data source. Moreover, we are considering integrating other DNA variation analysis tools that focus on the prediction of the impact of DNA variations at DNA level the use of which is much extended throughout the user community in the field of personalised medicine. Finally, once tested by users we hope to open up this project to involve more developers in the production of an open platform.

7 ftp://ftp.sanger.ac.uk/pub/1000genomes/PILOT/REL-1007/indels/

Acknowledgements

The Project Grant [TIN2011-25840) (Spanish Ministry of Education and Science) and P11-TIC-7529 (Innovation, Science and Enterprise Ministry of the regional government of the Junta de Andalucía).

References

Goecks J, Nekrutenko A, *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* **11**(8), R86. doi: 10.1186/gb-2010-11-8-r86

Hamosh A, Scott AF, *et al.* (2005) Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**(Database issue), D514-517. doi:10.1093/nar/gki033

Lesage S, Zouali H, *et al.* (2002) CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* **70**(4), 845-857. doi:10.1086/339432

Mattingly CJ, Colby GT, *et al.* (2003) The Comparative Toxicogenomics Database (CTD) *Environ Health Perspect.* **111**(6), 793-795. doi:10.1289/txg.6028

McKenna A, Hanna M, *et al.* (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**(9), 1297-1303. doi: 10.1101/gr.107524.110

Ng SB, Buckingham KJ, *et al.* (2009) Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics* **42**(1), 30-35. doi: 10.1038/ng.499

Ng SB, Turner EH, *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**(7261), 272-276. doi:10.1038/nature08250

Roberts RJ. (2001) PubMed central: The GenBank of the published literature. *Proc. Natl. Acad. Sci. U. S. A.* **98**(2), 381-382. doi: 10.1073/pnas.98.2.381

Wang K, Li M, *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**(16):e164. doi: 10.1093/nar/gkq603

Wishart DS, Knox C, *et al.* (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**(Database issue), D668-D672. doi: 10.1093/nar/gkj067

Yandell M, Huff C, *et al.* (2011) A probabilistic disease-gene finder for personal genomes. *Genome Res* **21**, 1529–1542. doi: 10.1101/gr.123158.111