Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Data and Text mining

myVCF: a desktop application for high-throughput mutations data management

Alessandro Pietrelli[1,2]* and Luca Valenti[1,3]

[1]Internal Medicine and Metabolic Diseases, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy, [2]Istituto Nazionale Genetica Molecolare (INGM), "Romeo ed Enrica Invernizzi", Bioinformatic Group, Milano, Italy, [3]Department of Pathophysiology and Transplantation, Università degli Studi Milano, Milan, Italy

*To whom correspondence should be addressed.

Associate Editor: Dr. Jonathan Wren

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Next-generation sequencing technologies have become the most powerful tool to discover genetic variants associated with human diseases. Although the dramatic reductions in the costs facilitate the use in the wet-lab and clinics, the huge amount of data generated renders their management by non-expert researchers and physicians extremely difficult. Therefore, there is an urgent need of novel approaches and tools aimed at getting the "end-users" closer to the sequencing data, facilitating the access by non-bioinformaticians, and to speed-up the functional interpretation of genetic variants. We developed myVCF, a standalone, easy-to-use desktop application, which is based on a browser interface and is suitable for Windows, Mac and UNIX systems. myVCF is an efficient platform that is able to manage multiple sequencing projects created from VCF files within the system; stores genetic variants and samples genotypes from an annotated VCF files into a SQLite database; implements a flexible search engine for data exploration, allowing to query for chromosomal region, gene, single variant or dbSNP ID. Besides, myVCF generates a summary statistics report about mutations distribution across samples and across the genome/exome by aggregating the information within the VCF file. In summary, the myVCF platform allows end-users without strong programming and bioinformatics skills to explore, query, visualize and export mutations data in a simple and straightforward way.

Availability: https://apietrelli.github.io/myVCF/

Contact: pietrelli@ingm.org

Supplementary information: http://myvcf.readthedocs.io/

1 Introduction

The advent of next-generation sequencing (NGS) led undoubtedly to a revolution in the field of human genetics. During the last years, the dramatic reduction in the costs of this technology opened up new scenarios in sequencing entire exome/genomes of a large number of individuals, not only for research purposes, but also in the clinics as a standard diagnostic tools (Wang *et al.*, 2013; Smedley and Robinson, 2015; Gullapalli *et al.*, 2012). Although data generation is becoming increasingly easier and cheaper, worries arise due to the tremendous complexity of managing and handling large amount of data (Sboner *et al.*, 2011).

The Variant Call Format (Danecek *et al.*, 2011) (VCF) is the standard file to store mutations data derived from next-generation sequencing technologies. The VCF has been developed in the 1000 Genomes project (Auton *et al.*, 2015) and then adopted as the standard by most of the software that treat variant calls (McLaren *et al.*, 2016; McKenna *et al.*, 2010). Although it is a very useful and compatible format for bioinformatics tools, it lacks readability and is difficult to interrogate for scientists without programming skills.

The necessity to fill the gap between the raw data and the "end-user" urges to simple and clear tools with an intuitive graphical interface to explore and query variants and facilitate data interpretation, thereby fostering translational research in the field of genetics.

2 Implementation

The core system of myVCF application is implemented in Django (https://djangoproject.com), an open-source Python web framework that connects data models and relational databases using the object-relational mapping technique. The application, provided as stand-alone program, runs as a local web server and works entirely on a simple web browser allowing a cross-platform compatibility among the most used operating systems (Unix, Mac and Windows).

The data storing system implemented is a SQLite backend database that eases the use of the package without hard database configuration or system administration requirements.

The web page architecture, design and functionality, was developed using the Bootstrap framework together with several Javascript libraries (https://datatables.net/, http://www.highcharts.com, https://jquery.com). The export feature has been implemented libraries to easily save, process tables, results and plots using standard formats. These include PDF, Excel, CSV for tables, and JPG, PNG, TIFF for images and plots.

The myVCF application integrates and supports the annotations provided by both Annovar (Wang *et al.*, 2010) and VEP (McLaren *et al.*, 2016). The file must include i) at least one genotyped sample and ii) the variant information field for VCF annotation (i.e. CSQ field for VEP, ExonicFunc_ensGene for Annovar). The size of the VCF file and the hardware of the user machine can influence the speed of the storing process, which may vary from few seconds for targeted sequencing project to minutes for whole-exome sequencing projects. Of note, thanks to the database storing system, this procedure will be computed just once for each VCF.

The application is able to manage multiple VCF files that are reachable from the home page containing the list of the projects link.

The main function of myVCF application is the ability to browse the information contained in the VCF files. To perform queries, the application implements a flexible search engine with the possibility to type multiple terms such as: region/variant coordinates, gene name or dbSNP id. Based on the query, results are displayed with two alternative output page layouts:

1) Gene/region results page (Figure 1a)

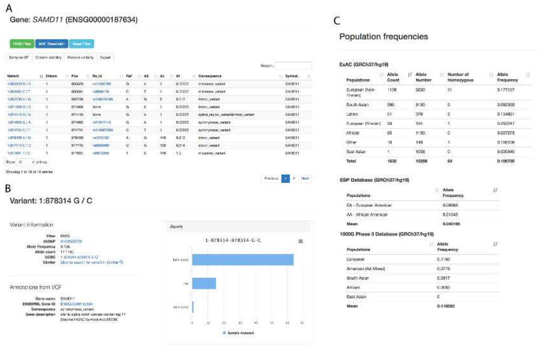
2) Variant results page (Figure 1b)

In the gene/region results page the mutations and related information are displayed in a tabular format. The user can use filters to select mutations by i) allele frequency threshold and ii) PASS filter quality field and visualize a set of informative columns (INFO fields in the VCF, Genotypes…) by clicking on the relative buttons.

The variant results page shows all the information relative to a single variant within the project. For each variant, different annotations are reported regarding the allele frequency, the functional consequence and the number of alleles found with zygosity distribution.

myVCF also provides links to external resources such as the ClinVar database (Landrum *et al.*, 2014) and UCSC. Moreover, myVCF integrates the information relative to the allele frequency present in public databases, such as the 1000Genomes project, Exome Sequencing Project (ESP) (Fu *et al.*, 2012), the Exome Aggregation Consortium (ExAC) (Lek *et al.*, 2016) (Figure 1c) together with in-silico predictor tools such as Polyphen2 (Andreasen *et al.*, 2013) and SIFT (Ng and Henikoff, 2003).

myVCF implements functions that perform both a global VCF metrics summary as quality control report, and a functional overview of the mutations distribution across samples, chromosomes, and gene feature (Supplementary Figure).



Gene: *SAMD11* (ENSG00000187634)

Variant	Chrom	Pos	Rs_id	Ref	Alt	Ac	Af	Consequence	Symbol
[illegible]	1	[illegible]	[illegible]	G	A	1	0.00625	missense_variant	SAMD11
[illegible]	1	[illegible]	[illegible]	C	T	1	0.00625	missense_variant	SAMD11
[illegible]	1	[illegible]	[illegible]	A	G	2	0.013	intron_variant	SAMD11
[illegible]	1	[illegible]	None	G	C	1	0.00625	intron_variant	SAMD11
[illegible]	1	[illegible]	None	G	A	1	0.00625	splice_region_variant&intron_variant	SAMD11
[illegible]	1	[illegible]	[illegible]	G	A	1	0.00625	synonymous_variant	SAMD11
[illegible]	1	[illegible]	[illegible]	C	T	1	0.00625	synonymous_variant	SAMD11
[illegible]	1	[illegible]	[illegible]	A	G	146	0.912	intron_variant	SAMD11
[illegible]	1	[illegible]	[illegible]	C	G	139	0.914	intron_variant	SAMD11
[illegible]	1	[illegible]	[illegible]	T	C	146	1.0	missense_variant	SAMD11

Show 10 entries

Showing 1 to 10 of 16 entries

Variant: 1:878314 G / C

Variant information

Filter	PASS
dbSNP	[illegible]
Allele Frequency	0.106
Allele count	17 / 160
UCSC	1-878314-878314-G-C
ClinVar	Click to search for variant in ClinVar

Annotations from VCF

Gene name	SAMD11
ENSEMBL Gene ID	ENSG00000187634
Consequence	synonymous_variant
Gene description	sterile alpha motif domain containing 11 [Source:HGNC Symbol;Acc:28706]

Population frequencies

ExAC (GRCh37/hg19)

Populations	Allele Count	Allele Number	Number of Homozygous	Allele Frequency
European (Non-Finnish)	1108	6260	61	0.177157
South Asian	386	6190	9	0.062359
Latino	51	378	2	0.134921
European (Finnish)	39	154	1	0.253247
African	30	1100	0	0.027273
Other	16	148	1	0.108108
East Asian	1	1058	0	0.000945
Total	**1632**	**15288**	**64**	**0.106750**

ESP Database (GRCh37/hg19)

Populations	Allele Frequency
EA - European American	0.08686
AA - African American	0.01345
Mean	**0.050155**

1000G Phase 3 Database (GRCh37/hg19)

Populations	Allele Frequency
European	0.1193
American (Ad Mixed)	0.0778
South Asian	0.0317
African	0.0083
East Asian	0
Mean	**0.118350**

Figure 1: Gene and Variant template example layouts. (A) Example of SAMD11 gene search results. Sample genotype and columns visualization preferences can be set by using the white background buttons. (B) Example of variant search results. The web page contains information retrieved from the INFO field of the VCF file and external resources links such as ClinVar and UCSC are available in the Variant Information section. Within the variant results page the most used population frequency databases are linked to display the allele frequency of the given variant in the available populations (C).

3 Conclusion

myVCF is a standalone, cross-platform and freely available tool to manage and browse VCF files in an efficient and easy way. It was developed to help end-users without strong bioinformatics skills to get in contact with mutation data and relative annotations in an informative, straightforward, intuitive and flexible fashion. The specific advantages of myVCF, as compared to already available applications, are represented by the simplicity of usage conjugated with a graphical interface, allowing the use by non-experts on almost every platform, together with the possibility to manage different projects and queries, and export the data in multiple formats. Moreover facilitate the mutation data interpretation in human sequencing projects (from target to whole-exome sequencing) by integrating different annotations according to the most used allele frequency population databases and in-silico prediction tools.

In conclusion, myVCF is the first tool that merges the functionality expressed in command-line tools, including multiple project management, exporting results and tables, with an easy and clear visualization through web pages in a simple browser.

Acknowledgements

We wish to thank all the people of the Bioinformatics facility group at INGM and those in the Valenti's laboratory at IRCCS Policlinico Cà Granda for their extensive application testing, suggestions and support.

Funding

This project was supported by the following grants: Fondazione Ca' Granda and INGM Molecular Medicine Grant 2015/6, and by the myFIRST AIRC [Grant n.16888 – EPIDEMIC] from the Italian Association for Cancer Research (AIRC).

Conflict of Interest: none declared.

References

Andreasen,C. *et al.* (2013) New population-based exome data are questioning the pathogenicity of previously cardiomyopathy-associated genetic variants. *Eur. J. Hum. Genet.*, **21**, 918–928.

Auton,A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Danecek,P. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Fu,W. *et al.* (2012) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**, 216–220.

Gullapalli,R.R. *et al.* (2012) Clinical integration of next-generation sequencing technology. *Clin. Lab. Med.*, **32**, 585–99.

Landrum,M.J. *et al.* (2014) ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**.

Lek,M. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.

McKenna,A. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–303.

McLaren,W. *et al.* (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, **17**, 122.

Ng,P.C. and Henikoff,S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–4.

Sboner,A. *et al.* (2011) The real cost of sequencing: higher than you think! *Genome Biol.*, **12**, 125.

Smedley,D. and Robinson,P.N. (2015) Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Med.*, **7**, 81.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Wang,Z. *et al.* (2013) The Role and Challenges of Exome Sequencing in Studies of Human Diseases. *Front. Genet.*, **4**.