Bioinformatics, 32(1), 2016, 151–153
doi: 10.1093/bioinformatics/btv511
Advance Access Publication Date: 12 September 2015
Applications Note

OXFORD

Genetics and population analysis

DIVAS: a centralized genetic variant repository representing 150 000 individuals from multiple disease cohorts

Wei-Yi Cheng†, Jörg Hakenberg, Shuyu Dan Li and Rong Chen*

Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

*To whom correspondence should be addressed.
†Present address: Roche Translational & Clinical Research Center, New York, NY 10016, USA
Associate Editor: Alfonso Valencia

Received on April 27, 2015; revised on July 28, 2015; accepted on August 25, 2015

Abstract

Motivation: A plethora of sequenced and genotyped disease cohorts is available to the biomedical research community, spread across many portals and represented in various formats.
Results: We have gathered several large studies, including GERA and GRU, and computed population- and disease-specific genetic variant frequencies. In total, our portal provides fast access to genetic variants observed in 84 928 individuals from 39 disease populations. We also include 66 335 controls, such as the 1000 Genomes and Scripps Wellderly.
Conclusion: Combining multiple studies helps validate disease-associated variants in each underlying data set, detect potential false positives using frequencies of control populations, and identify novel candidate disease-causing alterations in known or suspected genes.
Availability and implementation: https://rvs.u.hpc.mssm.edu/divas
Contact: rong.chen@mssm.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

DNA sequencing and genotyping data from populations of various demographic backgrounds are becoming available to the biomedical research community at an ever increasing pace. Individually, targeted studies have provided insights into the genetic underpinnings of diseases or confirm previously identified causal alterations: for congenital heart disease (Turki *et al.*, 2014), coloboma (Rainger *et al.*, 2014), schizophrenia (Rees *et al.*, 2014) and numerous others.

The power to discover novel, disease-causing or protective alleles potentially becomes even larger when combining data from different studies, thereby increasing the number of controls and cases for related phenotypes, and helping adjust for the mutational spectrum for individuals of diverse ethnic backgrounds. Projects such as the 1000 Genomes Project (1000 Genomes Project Consortium, 2012) or the NHLBI Exome Sequencing Project (www.evs.gs.washington.edu/EVS/) included large cohorts from various ethnicities, and have been frequently used as a control population for filtering out potential benign variants observed in a disease cohort. However, few efforts compile the information from multiple large cohorts into a centralized portal to obtain the distribution of variants observed across multiple studies.

The first two large portals providing joint access to several sequencing studies went live earlier this year: the Exome Aggregation Consortium (ExAC, 2015) (www.exac.broadinstitute.org) and the European Variation Archive (EVA, 2015) (www.ebi.ac.uk/eva/). ExAC compiled summarized variant information from more than 63 000 exomes by normalizing data from individual studies through an identical pipeline. EVA provides a query interface for obtaining summarized variant information from several large studies. Although both portals provide variant frequencies observed in numerous control cohorts of multiple ethnic groups, they do not include disease-specific variant frequencies from pure disease cohorts,

which can be a positive indication that a variant may be a candidate for causing the disease.

We have compiled a centralized genetic variant repository (Disease Variant Store, DIVAS) that includes genotype data from eight large-scale studies, including 1000 Genomes, ExAC, dbGaP GRU, GERA (Hoffmann *et al.*, 2011) and UK10K ALSPAC/TWINS (see Supplementary Information), consisting of 150 000 individuals from seven ethnic groups. Among this population, 84 928 individuals were annotated with 39 disease phenotypes, and 66 335 as control cohorts. We have computed the disease-specific, ethnicity-specific and control variant frequencies as well as genotype frequencies for all observed variants, and then visualized these summarized information through a public web interface. The broad spectrum of disease phenotypes and ethnic groups in DIVAS make it a simple and comprehensive tool to validate known pathogenic variants, or facilitate in the discovery of novel disease-causing variants.

2 Usage

The DIVAS web interface provides several ways to query for variants, including genes and coordinates. Results are presented in a table including information on effects from snpEff, variant frequencies observed in selected DIVAS cohorts, predicted functional impact (such as SIFT, MutationAssessor) and known disease associations from ClinVar, OMIM, SwissVar and HGMD (the latter with restricted access). Once the results are shown in tabular format, users can select one of those four annotation categories from the dropdown in the upper left. The frequencies of each variant are generated dynamically through bar charts. The bar charts were implemented in D3.js and thus allow the user to filter the frequencies by population, or to sort the frequencies based on various criteria (conditioned on disease/control and population).

One immediate use of DIVAS is to validate known disease variant associations in public databases such as ClinVar. For instance, the variant GATA4:p.Ala346Val (rs115372595) was reported in Rajagopal *et al.* (2007) to be observed only in a proband with endocardial cushion defect (ECD); it is annotated as pathogenic in ClinVar. In DIVAS, we observe that this variant has a frequency more than 2-fold higher in a congenital heart defect population than in any other disease and control cohort (Fig. 1). This observation is consistent with the original study that this variant may contribute to the risk of ECD. We provided other examples in the Supplementary Material.

We also provide RESTful API access to query DIVAS for allele frequencies, diseases, effects and predicted functional impacts by specifying genes, dbSNP IDs or variant keys. See Supplementary Methods and the DIVAS website for details.

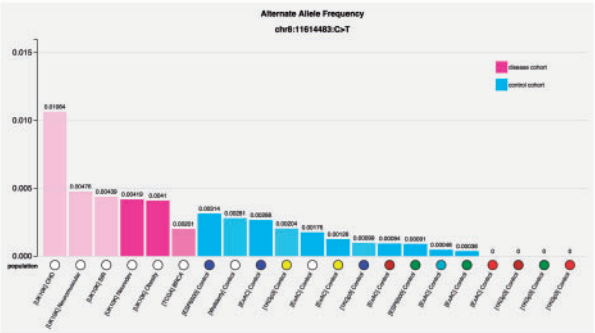

Fig. 1. Frequency bar chart showing the variant frequencies across all DIVAS disease and control cohorts for GATA4:p.Ala346Val. Opacity of bars indicates the sample size of that cohort

3 Discussion

Combining multiple studies helps to validate disease-associated variants in each underlying dataset, detect potential false positives using frequencies of other populations and identify novel candidate disease-causing alterations in known or suspected genes.

It has to be noted that several datasets represented in DIVAS do not provide detailed phenotypes. Disease annotations across different studies are not necessarily compatible; we plan to address this issue by structuring phenotypes using UMLS or HPO (Köhler *et al.*, 2014; Thorn *et al.*, 2007). Moreover, some studies are assumed to be healthy (1000 Genomes) or phenotypes are not published on an individual level (ExAC). Individuals of the former group may still develop a disease at a later time point due to a genetic component.

Since DIVAS variants are derived from a variety of sequencing and genotyping platforms, some variants are not profiled and missing data require imputation. Different platforms also impose different quality criteria; in DIVAS, we are currently using the quality metrics and filter criteria provided by each individual study. For that reason, DIVAS provides allele frequencies before and after filtering out non-passing sites.

A long-term goal of the service is to integrate the DIVAS infrastructure with several other large data access portals (for instance, dbGap and UK10K). This can help prospective data applicants get a clear idea whether their genes and variants of interest have a frequency distribution in any relevant study supporting their hypothesis.

4 Implementation

Integrating variants from heterogeneous sources requires left-aligned normalization on all genetic variants, which in our case was performed using an algorithm proposed by Tan *et al.* (2015). For variants containing multiple alternate alleles, we separated each alternate allele as one unique variant and normalized them independently. We have devised a unique, reversible, compressed variant key to represent all possible single nucleotide variations and deletions using a $\leq$15-byte string representation. This representation can also handle insertions of up to 2958 bases or multi-nucleotide variations using at most 1000 bytes (see Supplementary Methods). By indexing variants and disease-variant association databases in the same way, we are able to quickly retrieve detailed annotations from ClinVar (Landrum *et al.*, 2014), SwissVar, HGMD (Stenson *et al.*, 2014), OMIM and dbSNP. We compute gene- and protein-level annotations using snpEff (Cingolani *et al.*, 2012) and include pre-computed functional predictions from dbNSFP (Liu *et al.*, 2013).

For each dataset, we calculated frequencies for the annotated diseases and ethnicities when possible. We then calculated variant frequencies in each subset. We also calculated frequencies exclusively using variants that have passed the quality control employed by the original data publisher. Frequency calculation is implemented using Apache Pig and results were exported to a MySQL database backing the web query interface. Data visualization is based on D3.js, enabling the web interface to dynamically render charts of variant frequencies based on individual filter/sort criteria chosen by the user.

Funding

This work was supported in part through the computational resources provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai. We thank Benjamin Glicksberg for proofreading.

Conflict of Interest: none declared.

References

1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1 092 human genomes. *Nature*, **491**, 56–65.

Cingolani,P. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, **6**, 80–92.

European Variation Archive (EVA), Hinxton, UK. http://www.ebi.ac.uk/eva/, Last accessed July 2015.

Exome Aggregation Consortium (ExAC), Cambridge, MA. http://www.exac.broadinstitute.org, Last accessed July 2015.

Hoffmann,T.J. *et al.* (2011) Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics*, **98**, 79.

Köhler,S. *et al.* (2014) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.*, **42**(D1), D966–D974.

Landrum,M. *et al.* (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**(Database issue), D980–D985.

Liu,X. *et al.* (2013) dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.*, **34**, E2393–E2402.

Rainger,J. *et al.* (2014) Monoallelic and biallelic mutations in MAB21L2 cause a spectrum of major eye malformations. *Am. J. Hum. Genet.*, **94**, 915–923.

Rajagopal,S.K. *et al.* (2007) Spectrum of heart disease associated with murine and human GATA4 mutation. *J. Mol. Cell. Cardiol.*, **43**, 677–685.

Rees,E. *et al.* (2014) CNV analysis in a large schizophrenia sample implicates deletions at 16p12.1 and SLC1A1 and duplications at 1p36.33 and CGNL1. *Hum. Mol. Genet.*, **23**, 1669–1676.

Stenson,P. *et al.* (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, **133**, 1–9.

Tan,A. *et al.* (2015) Unified representation of genetic variants. *Bioinformatics*, **31**, 2202–2204.

Thorn,K.E. *et al.* (2007) The UMLS Knowledge Source Server: an experience in Web 2.0 technologies. *AMIA Annu. Symp. Proc.*, **11**, 721–725.

Turki,A. *et al.* (2014) Rare variants in NR2F2 cause congenital heart defects in humans. *Am. J. Hum. Genet.*, **94**, 574–585.