2019

Varanto: variant enrichment analysis and annotation

Timonen, O
**Genome analysis**

**Varanto: variant enrichment analysis and annotation**

Oskari Timonen¹*, Mikko Särkkä², Tibor Fülöp², Anton Mattsson², Juha Kekäläinen² and Jussi Paananen²

¹Institute of Clinical Medicine, University of Eastern Finland, Kuopio, Finland, ²Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

**Abstract**

**Summary:** Genome-wide association studies (GWAS) aim to identify associations of genetic variations such as single-nucleotide polymorphisms (SNPs) to a specific trait or a disease. Identifying common themes such as pathways, biological processes and diseases associations is needed to further explore and interpret these results. Varanto is a novel web tool for annotating, visualizing and analyzing human genetic variations using diverse data sources. Varanto can be used to query a set of input variations, retrieve their associated variation and gene level annotations, perform annotation enrichment analysis and visualize the results.

**Availability and implementation:** Varanto web app is developed with R and implemented as Shiny app with PostgreSQL database and is freely available at http://bioinformatics.uef.fi/varanto. Source code for the tool is available at https://github.com/oqe/varanto.

**Contact:** oskari.timonen@uef.fi

**Supplementary information:** Supplementary data are available at Bioinformatics online.

---

1 Introduction

Genome-wide association studies (GWAS) are regularly used to identify genetic variations that contribute to normal and pathological traits (MacArthur et al., 2017). The result of a typical GWAS study is a list of variations and their association score for the trait in question. Finding more information about these variations requires additional analysis and retrieval of annotations from multiple data sources.

Enrichment analysis is used for studying enrichment of annotations within a set of genes in comparison to a background set (Huang et al., 2009). Annotations can be for example related signaling pathways, biological processes, molecular functions, disease associations or functional effects of a genetic variation. Enrichment determines which annotation terms are over- or underrepresented in the context of the study. Enriched terms can be used to further explore, interpret and validate the results from the study. Most existing genomic enrichment tools are centered around genes, and Varanto expands this approach by adding a more detailed layer of genetic variation level enrichment.

2 Methods

2.1 Implementation

Varanto web app is developed with R (R Core Team, 2017) and utilizes the Shiny framework (Chang et al., 2017) and PostgreSQL database. Publicly accessible server is hosted by University of Eastern Finland Bioinformatics Center. Data extraction, transformation and loading utilize BiomaRt (Durinck et al., 2005, 2009) and dplyr (Wickham et al., 2017) R-packages, Bash shell-scripting and Python. Source code and local deployment instructions are available at GitHub.

2.2 Resources

Varanto utilizes Ensembl (Yates et al., 2016), UCSC Genome Browser (Kent et al., 2002), Genome-Environment-Trait-Evidence (GET-Evidence) (Ball et al., 2012) and Molecular Signature Database
(MSigDB) (Subramanian et al., 2005) data resources imported into a local integrated database. Data from Ensembl Variation database is used as the backbone of Varanto database, and incorporates annotation data from for example dbSNP (Sherry, 2001), ClinVar (Landrum et al., 2016), PolyPhen (Adzhubei et al., 2013), SIFT (Kumar et al., 2009) and Gene Ontology (GO) (Ashburner et al., 2000; Carbon et al., 2017). Varanto has two levels of annotations 1) genes and 2) variations. Gene annotations such as GO terms or signaling pathways are linked to a gene level, and a variation is associated to these terms if the variation resides within the gene in question (therefore all variations within a gene have the same gene level annotations). The variation annotations include e.g. functional effect of the variation, variation specific disease associations, population frequency, and alleles. Genetic variation background sets are based on Affymetrix and Illumina SNP microarrays, and retrieved from UCSC Genome Browser. Content of the integrated database is updated regularly and data resource version information is available in the web application’s About tab.

2.3 Input

As an input, Varanto takes in a set of genetic variations in the form of dbSNP rs-identifiers, Ensembl variation identifiers and genomic locations. The set can be pasted to a textbox in the web app, or uploaded as a text file. Additionally, Varanto also contains example variation input set for quick testing.

Inputs can be filtered with a selection of a background variation set. This filter excludes all variations not present in the selected background set, removing bias caused by for example using results from a specific SNP microarray. To avoid the issue where your input set contain several closely located variations (e.g. within a single gene) resulting in an over-presentation of annotations linked to this loci, you can filter the input variation set by their genomic distance. The distance filtering enables thinning out of the input variation set based on genomic proximity. The distance filtering algorithm creates a window of a selected size (for instance 1 kilobase) starting from the first variation. All variations which are inside this window except the first one are excluded. The next window is created from the first variation which is outside of the previous window. This repeats until all variations are processed.

After input of variations and selection of possible background set and distance filter, user chooses variation and gene annotations from their respective menus. Multiple annotations from both categories can be selected. Pressing the Submit button validates the user input and submits the query. After a successful query, basic summary data for the query, as well as annotated data table is shown. Further analysis, results and visualization can be accessed from their respective tabs. Web architecture and user interface figure, example case study, database performance benchmark and comparison to existing software is available in Supplementary materials.

2.4 Enrichment analysis

Enrichment analysis can be used to identify over- and underrepresentation of annotations linked to variations in the input set. For each selected annotation term, odds ratio and statistical significance for over- and underrepresentation is listed in the enrichment results. Enrichment analysis is based on hypergeometric testing and the p-values are adjusted with Benjamini-Hochberg false discovery rate (FDR) method (Benjamini and Hochberg, 1995).

2.5 Visualization

Heatmap tab in the Varanto web app provides a graphical visualization of the binary matrix of the variations and their associated annotations. The variations and annotations are ordered using hierarchical clustering, enabling inspection of clustered variations and annotations. Interactivity of the heatmap is produced with plotly R-package (Siervert et al., 2017). Static heatmap and preliminary plot is generated with ggplot2 (Wickham, 2016). In the karyogram-tab genomic locations of the variations are visualized on the chromosome level with chromPlot R-package (Oróstica and Verdugo, 2016), enabling visual inspection of phenomena related to locational clustering or spread of the variations.

3 Conclusions

Varanto can be used to annotate, analyze and visualize sets of genetic variations, thus facilitating further exploration, interpretation and validation of results from high-throughput genetic studies. All data, results and source code are downloadable through Varanto website. Case study example, database benchmarking and comparisons to existing software is available in Supplementary materials.

Funding

This work has been supported by grants from Saastamoinen Foundation and Academy of Finland.

Conflict of Interest: none declared.

References


Kent, W.J. et al. (2002). The Human Genome Browser at UCSC. Genome Res. 12, 996–1006.


O. Timonen et al.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz046/5299998 by University of Eastern Finland user on 30 January 2019