OXFORD UNIVERSITY PRESS | Bioinformatics

VEXOR: an integrative environment for prioritization of functional variants in fine-mapping analysis

Journal:	*Bioinformatics*
Manuscript ID	BIOINF-2016-1435.R1
Category:	Applications Note
Date Submitted by the Author:	14-Dec-2016
Complete List of Authors:	Lemaçon, Audrey; Genomics Center, Centre Hospitalier Universitaire de Québec - Université Laval Research Center Joly-Beauparlant, Charles; Genomics Center, Centre Hospitalier Universitaire de Québec - Université Laval Research Center Soucy, Penny; Genomics Center, Centre Hospitalier Universitaire de Québec - Université Laval Research Center Allen, Jamie; Centre for Cancer Genetic Epidemiology, University of Cambridge, Public Health and Primary Care Easton, Douglas; Centre for Cancer Genetic Epidemiology, University of Cambridge, Oncology Kraft, Peter; Harvard School of Public Health, Epidemiology; Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health Simard, Jacques; Genomics Center, Centre Hospitalier Universitaire de Québec - Université Laval Research Center Droit, Arnaud; Genomics Center, Centre Hospitalier Universitaire de Québec - Université Laval Research Center
Keywords:	Bioinformatics, Genome analysis, Data integration, Genome annotation

BIOINFORMATICS

Vol. 00 no. 00 2016
Pages 1–3

VEXOR: an integrative environment for prioritization of functional variants in fine-mapping analysis

Audrey Lemaçon[1], Charles Joly Beauparlant[1], Penny Soucy[1], Jamie Allen[2], Douglas Easton[2,3], Peter Kraft[4,5], Jacques Simard[1] and Arnaud Droit[1,*]

[1]Genomics Center, Centre Hospitalier Universitaire de Québec - Université Laval Research Center, Quebec, Quebec, Canada, [2]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, [3]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK, [4]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA [5]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: The identification of the functional variants responsible for observed genome-wide association studies (GWAS) signals is one of the most challenging tasks of the post-GWAS research era. Several tools have been developed to annotate genetic variants by their genomic location and potential functional implications. Each of these tools has its own requirements and internal logic, which forces the user to become acquainted with each interface.

Results: From an awareness of the amount of work needed to analyze a single locus, we have built a flexible, versatile and easy-to-use web interface designed to help in prioritizing variants and predicting their potential functional implications. This interface acts as a single-point of entry linking association results with reference tools and relevant experiments.

Availability: VEXOR is an integrative web application implemented through the Shiny framework and available at: `http://romix.genome.ulaval.ca/vexor`.

Contact: arnaud.droit@crchuq.ulaval.ca

INTRODUCTION

Genome-wide association studies (GWAS) have identified thousands of robust and reproducible genetic relations for complex diseases. However, progress towards understanding pathological mechanisms has been limited by the difficulty in assigning molecular functions to the large majority of GWAS hits that do not affect protein-coding sequences. The latter fact implies that many susceptibility variants affect genes indirectly by disrupting their regulation. Therefore, multiple efforts have been invested in identifying functional non-coding elements. Despite the quality of resources such as the Encyclopedia of DNA elements (ENCODE) (ENCODE Project Consortium, 2012) and the National Institutes of Health (USA) Roadmap Epigenomics project (Bernstein et al., 2010), identifying functional variants remains a challenging task. Due to linkage disequilibrium, GWASs tend to identified large clusters of SNPs with similar levels of significance, making it difficult to differentiate causal SNPs from linked neutral variants (Farh and al, 2015). To gain a better insight into the mechanisms of disease, identifying the true functional variants underlying the observed GWAS signals is an essential step. The most common approach used in that respect consists first in assigning well-calibrated probabilities of causality to candidate variants, then selecting a set of likely causal variants using functional annotation, and ultimately, identifying target genes whose disruption leads to altered disease risk (Farh and al, 2015). Many tools and databases are available to assist fine-mapping analysis. However, these tools are scattered throughout different platforms (web interface, standalone programs and command-line tools), and they require specific inputs and, sometimes, computational skills. To obviate these limitations and assist the process of fine mapping, we developed a novel integrative environment, namely VEXOR (available at `http://romix.genome.ulaval.ca/vexor/`). VEXOR features are listed and compared to similar tools in the Table 1.

VEXOR significantly facilitates the visualization, exploration and interpretation of the outputs generated by genome-wide genotyping arrays and custom arrays. It also conveniently associates a selected set of variants with publicly available functional annotations. VEXOR clearly displays the overlaps between genetic variants and potential genomic predictors. Such overlaps emphasize potential functional variants, predict target genes or regulatory pathways, and allow the prioritization of variants for future functional assessment.

VEXOR SOFTWARE

Implementation

VEXOR from Variant EXplOreR is a Web interface coded in R (R Core Team, 2013) and based on the Shiny framework (Chang et al., 2016). The website and its server are hosted by the

*to whom correspondence should be addressed : arnaud.droit@crchuq.ulaval.ca

Audrey Lemaçon et al

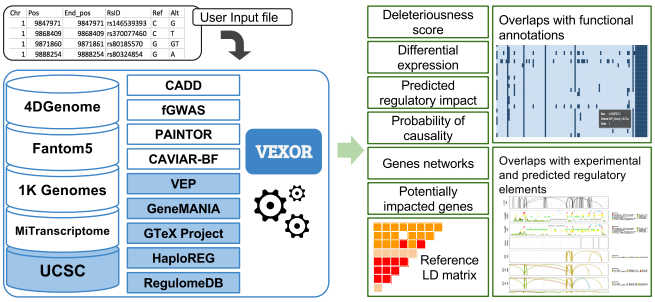

Fig. 1: VEXOR architecture, inputs and outputs. Blue rectangles and cylinders represent respectively tools and data sources ; filled shapes stand for linked resources whereas empty shapes stand for integrated resources.

Compute Canada infrastructure which provides a high-performance computing environment.

Input

VEXOR supports three input formats: comma-delimited, tabulation-delimited and excel (xls, xlsx). The input file must contain at least variants identification data (chromosome, position, rsId and alleles), but can contains various additional information such as epidemiological and statistical data. Uploaded file size is currently limited to 150 megaoctets (approximately 1 x 10^6 lines). For larger files, users will need to contact us to carry out data file entry, and private access to their data can be provided. This process has been used for fine-mapping of about 22 millions SNPs within the context of the Breast Cancer Association Consortium (BCAC; http://bcac.ccge.medschl.cam.ac.uk/). VEXOR and its resources are based on genome assembly GRCh37/hg19. All data that is uploaded in the current build (GRCh38/hg38) will be automatically converted to GRCh37 to ensure an effective integration.

Integrated Tools and Resources

VEXOR provides a single point-of-entry to several useful tools and resources such as VEP (McLaren et al., 2016), HaploReg (Ward and Kellis, 2012) and GTEx (The GTEx Consortium and al., 2015) (see Fig.1). No computational skills or file format editing between the tools is required. Users can execute all analyses needed within the VEXOR environment. Once the data has been successfully uploaded, the user can simultaneously study hundreds of variants of interest. We have implemented a tool that maps annotations (selected from a provided list or imported BED format file) against user-defined variants, and creates a display of overlapping features. Moreover, VEXOR can also link up to the UCSC (Rosenbloom and al, 2015) and Ensembl (Yates and al., 2015) Browser Session systems, where user-selected subsets of annotated data can be added as custom tracks for future display. VEXOR also includes minor-allele frequencies information coming form VEP REST service and a Haploview-like visualization of linkage disequilibrium created with the R package LDHeatmap (Shin and al, 2006). Deleteriousness and effect predictions can be obtained through VEXOR using VEP, HaploReg, RegulomeDB (Boyle et al., 2012) and CADD (Kircher et al., 2014) outputs. Results provided by each tool are processed by VEXOR and then presented as a summary table containing the main information as well as links to the relevant external websites for detailed data, when appropriate. GTEx summary results are also included to allow the study of correlations between genotype and tissue-specific gene expression levels. The impacted genes predicted through VEXOR can be automatically submitted to GeneMANIA (Warde-Farley and al, 2010) for pathway analysis and gene function prediction. A further useful feature offered by VEXOR is the implementation of an interface for statistical frameworks for variants prioritization. To fully exploit the annotation mapping functionality, we chose to implement three tools utilizing functional genomic information for multiple loci prioritization: fGWAS (Pickrell, 2014), fastPAINTOR (PAINTOR 3.0) (Kichaev et al., 2016) and CAVIAR-BF (Chen et al., 2015). These statistical frameworks are currently only available as a command-line program. Finally, taking advantage of R graphical abilities, we have included a section dedicated to the visualization of annotated features such as FANTOM5's enhancers (Lizio and FANTOM consortium, 2015), super-enhancers (Hnisz et al., 2015), computationally derived long poly-adenylated RNA transcripts referenced by MiTranscriptome catalog (Iyer and al., 2015), curated enhancer targets predicted by IM-PET as well as chromosomal interactions from the 4DGenome database (Teng et al., 2015) identified by various methods including Hi-C, chromosome conformation capture (3C), chromosome conformation capture carbon copy (5C) and chromatin interaction analysis by paired-end tag (ChIA-PET).

Conclusion

In conclusion, VEXOR is a versatile and scalable tool designed to help to characterize the functional context in fine-mapping analyses of complex traits. It represents a turnkey framework for predicting putative functional effects within both proximal and distal contexts for every variant. The user manual is available in the Supplementary material.

ACKNOWLEDGEMENT

The PERSPECTIVE project was supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research (grant GPH-129344), the Ministère de l'Économie, Science et Innovation du Québec through Genome Québec, and the Quebec Breast Cancer Foundation. We thank the BCAC for providing the impetus to create VEXOR, Compute Canada and particularly Jean-Philippe Dionne for his assistance in deploying the tool at the Compute Canada portal. We are also grateful to all the personnel of A.D.'s lab for their advice, and especially Frédéric Fournier and Mickaël Leclerq for their assistance in the preparation of this manuscript. We wish to extend our thanks to all current and future VEXOR users.

REFERENCES

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., and Thomson, J. A. (2010). The NIH

	Tools			
Features	DisGeNET	LocusExplorer	Enlight	VEXOR
Annotation mapping		✓	✓	✓
User data exploration system				✓
Minor allele frequency				✓
Linkage desequilibrium calculation			✓	✓
Manhattan plot		✓	✓	coming soon
Variant effect prediction	✓			✓
Variant visualization		✓	✓	✓
Variant scoring		✓		✓
Variant prioritization (stats)		✓		✓
Genomic context representation		✓	✓	✓
Disease association	✓			
Batch analysis	✓	✓	✓	✓
Platform	web interface	web interface + R library	web interface	web interface

Table 1. Features comparison between VEXOR, DisGeNET (Piñero et al., 2015), LocusExplorer (Dadaev et al., 2015) and Enlight (Guo et al., 2015).

roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, 28(10):1045–1048.

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M., and Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.*, 22(9):1790–1797.

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., and McPherson, J. (2016). shiny: Web application framework for R.

Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H., Poland, G. A., and Schaid, D. J. (2015). Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics*, 200(3):719–736.

Dadaev, T., Leongamornlert, D. A., Saunders, E. J., Eeles, R., and Kote-Jarai, Z. (2015). LocusExplorer: A user-friendly tool for integrated visualization of human genetic association data and biological annotations. *Bioinformatics*, 32(6):949–951.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Farh, K. K.-H. and al (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343.

Guo, Y., Conti, D. V., and Wang, K. (2015). Enlight: Web-based integration of GWAS results with biological annotations. *Bioinformatics*, 31(2):275–276.

Hnisz, D., Schuijers, J., Lin, C. Y., Weintraub, A. S., Abraham, B. J., Lee, T. I., Bradner, J. E., and Young, R. A. (2015). Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol. Cell*, 58(2):362–370.

Iyer, M. K. and al. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.*, 47(3):199–208.

Kichaev, G., Roytman, M., Johnson, R., Eskin, E., Lindstrom, S., Kraft, P., and Pasaniuc, B. (2016). Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics*.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, 46(3):310–315.

Lizio, Marina, a. and FANTOM consortium (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.*, 16:22.

McLaren, W., Gil, L., and al. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122.

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics*, 94(4):559–573.

Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., and Furlong, L. I. (2015). DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenbloom, K. R. and al (2015). The UCSC genome browser database: 2015 update. *Nucleic Acids Res.*, 43(Database issue):D670–81.

Shin, J.-H. and al (2006). LDheatmap : An R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J. Stat. Softw.*, 16(Code Snippet 3).

Teng, L., He, B., Wang, J., and Tan, K. (2015). 4DGenome: a comprehensive database of chromatin interactions. *Bioinformatics*, 31(15):2560–2564.

The GTEx Consortium and al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.

Ward, L. D. and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, 40(Database issue):D930–4.

Warde-Farley, D. and al (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, 38(Web Server issue):W214–20.

Yates, A. and al. (2015). Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710.