Bioinformatics Advance Access published November 26, 2014

Databases and ontologies

SNiPA: an interactive, genetic variant-centered annotation browser

Matthias Arnold[1,†], Johannes Raffler[1,†], Arne Pfeufer[1], Karsten Suhre[1,2] and Gabi Kastenmüller[1,*]

[1] Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany, [2] Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Education City, Qatar Foundation, Doha, Qatar

Associate Editor: Dr. Jonathan Wren

ABSTRACT

Motivation: Linking genes and functional information to genetic variants identified by association studies remains difficult. Resources containing extensive genomic annotations are available but often not fully utilized due to heterogeneous data formats. To enhance their accessibility, we integrated many annotation datasets into a user-friendly webserver.

Availability and implementation: http://www.snipa.org/

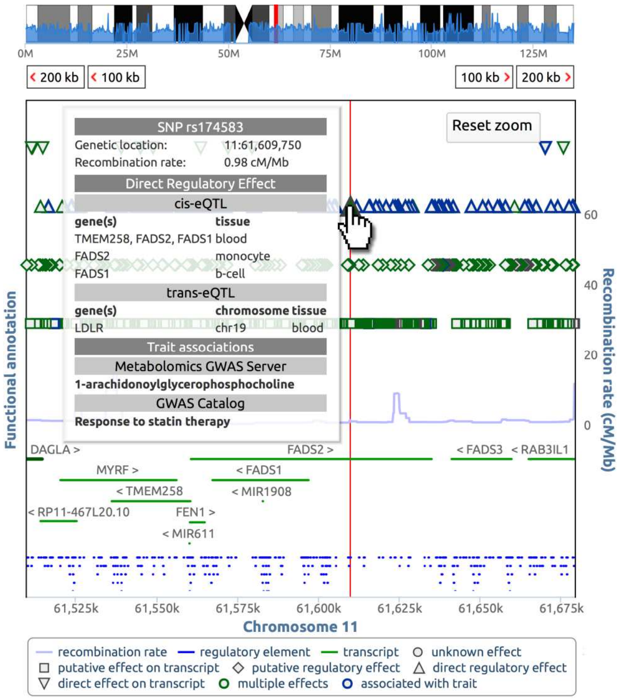
Contact: g.kastenmueller@helmholtz-muenchen.de

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

1 INTRODUCTION

Genome-wide association studies (GWAS) and next-generation sequencing (NGS) are performed routinely to identify genetic variants and novel genes implicated in both common and rare human diseases. A key step in translating results from such studies into a better understanding of molecular disease mechanisms and, ultimately, into clinical applications, is the prioritization of potentially functional variants that may be active *in vivo*. To this end, comprehensive collection and evaluation of existing functional annotation from genetic, informatics and experimental resources is essential (MacArthur, et al., 2014). This comprises the integration of data and knowledge across multiple levels including the variant, the gene, and the chromatin level.

Several large resources (Ensembl, UCSC, NCBI, etc.) aim at providing genome-wide genome-level annotation tracks from an extensive set of resources. However, retrieving statistical and functional annotation relevant at the single nucleotide level remains difficult. For instance, common genome browsers often display single nucleotide variants (SNVs) as thin bars that trail away in the wealth of other annotation tracks and are even less prepared to display statistics such as linkage disequilibrium (LD) relationships between variants. This limits visual distinction of relevant variants from those without relevant annotations and leaves the complex task of aggregating position-based data to the researcher. Variant-centered resources, on the other hand, typically concentrate on specific types of data such as amino acid changes (Adzhubei, et al., 2010; Kumar, et al., 2009), expression quantitative trait loci (eQTLs) (GTEx Consortium, 2013; Xia, et al., 2012), trait associations (Beck, et al., 2014; Hindorff, et al., 2009), or regulatory effect predictions (Boyle, et al., 2012). Moreover, these annotations are often presented in resource-specific data structures.



SNP rs174583

Genetic location: 11:61,609,750
Recombination rate: 0.98 cM/Mb

Direct Regulatory Effect

cis-eQTL

gene(s)	tissue
TMEM258, FADS2, FADS1	blood
FADS2	monocyte
FADS1	b-cell

trans-eQTL

gene(s)	chromosome	tissue
LDLR	chr19	blood

Trait associations

Metabolomics GWAS Server

1-arachidonoylglycerophosphocholine

GWAS Catalog

Response to statin therapy

Fig. 1. The *SNiPA* Variant Browser shows variants (top), genes (center in green), and regulatory regions (bottom in blue). Top-level information is available in mouse-over tooltips for all plot elements as shown here for the query SNP *rs174583*. The example highlights the value of variant-centered accumulation of annotations: *rs174583* is associated with the concentration of a lipid metabolite as well as with the expression levels of two genes encoding enzymes involved in lipid metabolism (*FADS1/2*) and the gene coding for *LDL receptor*, a major regulator of cholesterol homeostasis. Furthermore, the variant was linked to the response to lipid lowering drugs (statins), which target *HMG-CoA reductase* regulated by the *LDL receptor*.

For individual inspection of single variants, both resource types are extremely valuable. However, for simultaneous processing of

[*] To whom correspondence should be addressed.

[†] These authors made an equal contribution to this work.

larger variant sets, collection and examination of annotations from different data sources quickly becomes cumbersome. This presents a major bottleneck in genome-wide scans of genetic influences on human traits since the collection of such evidences is the key to understanding the effects of phenotype-linked genetic variants.

Here we propose *SNiPA*, a web service offering variant-centered genome browsing and interactive visualization tools tailored for easy inspection of many variants in their locus context (Figure 1).

2 DATA AND FEATURES

SNiPA includes a wide range of genome-level datasets contained in the Ensembl database (Flicek, et al., 2014) as an established backbone of annotations for the human genome. We combine this backbone with numerous variant-specific annotations taken from published datasets. Thus, *SNiPA* covers information ranging from regulatory elements, over gene annotations to variant annotations and associations (Table 1 and Supplementary Text 1). *SNiPA* contains annotations for all bi-allelic variants in phase 3 version 5 of the 1000 genomes project (1000 Genomes Project Consortium, et al., 2012) and provides pre-calculated LD-data for $r^2 \geq 0.1$ for all super-populations (African, American, South and East Asian, European). We use the Ensembl VEP tool (McLaren, et al., 2010) for primary effect prediction of SNVs. Additional position-based data is included in the VEP prediction as custom annotation files. For other annotations, we wrote a Perl module to extend the output provided by VEP (Table 1, Supplementary Text 1).

Table 1. Annotation data compiled in *SNiPA*

Entity type	Data type	N_{Entries}[a]	N_{Sources}[b]
Variant	*cis*-eQTL associations	919,860	8
	trans-eQTL associations	17,891	6
	Trait associations	245,333	9
	Conservation & deleteriousness scores	genome-wide	4
Gene	Trait annotations	3,752	3
Regulatory elements	microRNA target sites	606,408	5
	Promoters	106,169	2
	Enhancers	455,800	2
	ENCODE feature clusters	406,632	1

[a] Entries are unified w.r.t. the entities given in the first column, i.e. numbers listed are counts of annotated entities (e.g. variants). [b] Details and references for all included datasets are described in Supplementary Text 1.

SNiPA provides user-friendly starting points for annotating individual SNVs as well as sets of SNVs, LD blocks or genetic regions of interest. We have implemented several entry points to access the data: (i) a variant-centered implementation of a genome browser ("Variant Browser"); (ii) "Association Maps" for browsing through GWAS results; (iii) an interface for batch retrieval of variant annotations via ID-list, gene ID, or genomic coordinates ("Variant Annotation"); (iv) a combined listing of annotations across a set of variants within LD blocks or chromosomal regions ("Block Annotation"); (v) "Regional Association Plot" and "Linkage Disequilibrium Plot" (Diabetes Genetics Initiative of Broad Institute of Harvard, et al., 2007) that combine publication-ready plotting of association results and LD values, respectively, with the interactive interface of the "Variant Browser"; (vi) "Proxy Search" and "Pairwise LD" that allow querying pre-calculated LD values augmented with variant annotations. *SNiPA* enables the user to download condensed annotation data in tabular format for further off-line processing. Detailed descriptions of *SNiPA* modules are available in the online documentation and Supplementary Text 1.

The complex information contained in *SNiPA* is organized in a clear, comprehensive, and informative structure extending effect categories contained in the Sequence Ontology (Eilbeck, et al., 2005) (Supplementary Text 1). For instance, variant annotations are presented as "*SNiPA*cards" grouping information into semantic sections. All annotations are linked to their primary sources and to the Ensembl genome browser.

3 CONCLUSION

Mechanistic characterization of variants identified by genetic studies is the key to understanding molecular disease mechanisms. *SNiPA* combines a comprehensive set of genomic annotations with a genetic variant-based genome browser to simplify the task of variant annotation. *SNiPA* as well as all underlying data is freely available to the scientific community (commercial use may be limited by third-party constraints) and will be automatically updated following the Ensembl releases.

Funding: This work was supported by the Helmholtz Portfolio theme 'Metabolic Dysfunction and common disease' and by the research project Greifswald Approach to Individualized Medicine (GANI_MED) (BMBF: 03IS2061A). K.S. is supported by Biomedical Research Program funds at Weill Cornell Medical College in Qatar, a program funded by the Qatar Foundation.

Conflict of interest: none declared.

REFERENCES

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

Adzhubei, I.A., et al. (2010) A method and server for predicting damaging missense mutations, Nature methods, 7, 248-249.

Beck, T., et al. (2014) GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies, European journal of human genetics : EJHG, 22, 949-952.

Boyle, A.P., et al. (2012) Annotation of functional variation in personal genomes using RegulomeDB, Genome research, 22, 1790-1797.

Diabetes Genetics Initiative of Broad Institute of Harvard, M.I.T., Lund University, and Novartis Institutes of BioMedical Research,, et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels, Science, 316, 1331-1336.

Eilbeck, K., et al. (2005) The Sequence Ontology: a tool for the unification of genome annotations, Genome biology, 6, R44.

Flicek, P., et al. (2014) Ensembl 2014, Nucleic acids research, 42, D749-755.

GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project, Nature genetics, 45, 580-585.

Hindorff, L.A., et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, Proceedings of the National Academy of Sciences of the United States of America, 106, 9362-9367.

Kumar, P., Henikoff, S. and Ng, P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, Nature protocols, 4, 1073-1081.

MacArthur, D.G., et al. (2014) Guidelines for investigating causality of sequence variants in human disease, Nature, 508, 469-476.

McLaren, W., et al. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor, Bioinformatics, 26, 2069-2070.

Xia, K., et al. (2012) seeQTL: a searchable database for human eQTLs, Bioinformatics, 28, 451-452.