BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 1 2011, pages 144–146
doi:10.1093/bioinformatics/btq611

Data and text mining

ASSIMILATOR: a new tool to inform selection of associated genetic variants for functional studies

Paul Martin*, Anne Barton and Stephen Eyre

Arthritis Research UK Epidemiology Unit, Manchester Academic Health Science Centre, The University of Manchester, Stopford Building, Oxford Road, Manchester, M13 9PT, UK

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Fine-mapping experiments from genome-wide association studies (GWAS) are underway for many complex diseases. These are likely to identify a number of putative causal variants, which cannot be separated further in terms of strength of genetic association due to linkage disequilibrium. The challenge will be selecting which variant to prioritize for subsequent expensive functional studies. A wealth of functional information generated from wet lab experiments now exists but cannot be easily interrogated by the user. Here, we describe a program designed to quickly assimilate this data called ASSIMILATOR and validate the method by interrogating two regions to show its effectiveness.

Availability: http://www.medicine.manchester.ac.uk/musculoskeletal/research/arc/genetics/bioinformatics/assimilator/.

Contact: paul.martin-2@manchester.ac.uk

Received on June 4, 2010; revised on October 26, 2010; accepted on October 27, 2010

1 INTRODUCTION

Genome-wide association studies (GWAS) have been enormously successful in identifying regions associated with a variety of complex traits and diseases. Fine-mapping studies are underway for many of these disorders and are likely to identify a number of putative causal variants. The challenge then will be to prioritize which variants to select for the expensive functional studies required to fully translate how these variants affect risk. In many cases, it is expected that the likely causal variants will be single nucleotide polymorphism (SNP) markers that are in complete linkage disequilibrium and which cannot be prioritized further based on genetic evidence alone. SNPs within genes which affect the resulting protein or lie in a regulatory region would be obvious candidates for functional studies but, in many complex diseases, the causal SNPs identified to date map to intergenic, non-coding regions and it is more challenging to prioritize these based on likely function (Barton *et al.*, 2008; Thomson *et al.*, 2007; Wellcome Trust Case Control Consortium, 2007).

There is now a wealth of information available from the ENCyclopaedia Of DNA Elements (ENCODE) international consortium (Birney *et al.*, 2007; ENCODE Project Consortium 2004) hosted by the University of California Santa Cruz (UCSC) through their Genome Browser (Kent *et al.*, 2002). These data have been generated from wet lab experiments including Chromatin ImmunoPrecipitation Sequencing (ChIP-Seq), DNase hypersensitivity and histone modification studies, and thus may provide better evidence of putative function compared with predictive algorithms used previously to infer function at a locus. An enormous amount of data is available including studies in different cell lines and different cell compartments, but currently these sites cannot be easily interrogated by the user simultaneously. Other potential resources for prioritizing SNPs for functional studies are now becoming more widely available and include eQTL studies and programs which predict likely effects of non-synonymous polymorphisms. Here, we describe a program designed to quickly assimilate all available data for SNPs or locations entered by the user, called ASSIMILATOR. Importantly, the ability to enter SNPs using base pair position will allow the interrogation of novel variants identified, for example, by the 1000 Genomes project (http://www.1000genomes.org) even if an rs number has not yet been assigned. We also validate the method by interrogating SNPs in two regions: one associated with colorectal cancer (Pomerantz *et al.*, 2009) and one with type II diabetes (T2D) (Gaulton *et al.*, 2010). We show that, based on the information drawn together by ASSIMILATOR, we would have prioritized the subsequently confirmed causal SNPs for functional investigation from both previous studies.

2 METHODS

Written in Perl, ASSIMILATOR retrieves, queries and processes information for the desired SNPs from the UCSC Genome Browser's public MySQL database and displays this in a simplified, user-friendly manner. All available ENCODE tracks are queried in addition to predefined tracks, such as mRNAs, ESTs and CpG islands. In addition, eQTL data hosted by the Pritchard laboratories (http://eqtl.uchicago.edu), PolyPhen2 functional annotation (Adzhubei *et al.*, 2010) and SNP location relative to the gene are displayed. Multiple systems have been designed to improve the efficiency of data retrieval such as an XML-based track database, which minimizes the number of database queries and multi-threading support to query multiple SNPs simultaneously, reducing processing time with minimal reduction in individual performance.

The output can be viewed in a standard web browser and allows the user to quickly identify SNPs, which could be functionally important. To add extra functionality, the ability to view selected SNPs in NCBIs dbSNP (Sherry *et al.*, 2001) and in the UCSC Genome Browser has been incorporated into the output. To efficiently display features for a SNP in the UCSC Genome Browser, only tracks that contain features in the SNP region are displayed. The user interface has been designed to allow further mining of the output (Fig. 1) to display information from the multiple cell types and links to external data. This includes the ability to view the detailed experimental data thereby allowing users to assess the biological relevance of the results in the

*To whom correspondence should be addressed.

(a)

Results - Pomerantz et al.

Location Information			Standard					Expression					Regulation						
SNP ID	SNP Position	Information	Conservation	Human ESTs	Human mRNAs	Gencode Genes	Relative Location	Affy RNA Loc	Caltech RNA-seq	CSHL Sm RNA-seq	GIS PET Loc	RIKEN CAGE Loc	Broad Histone	Open Chromatin	HAIB TFBS	UW Histone	UW DNase DGF	UW DNaseI HS	Yale TFBS
rs6983267	chr8:128482487-128482487	snp130	Yes	Yes	Yes			Yes	Yes		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
rs10808556	chr8:128482329-128482329	snp130		Yes	Yes				Yes		Yes		Yes	Yes	Yes	Yes	Yes	Yes	Yes
rs4871788	chr8:128490967-128490967	snp130		Yes	Yes				Yes		Yes	Yes	Yes	Yes		Yes		Yes	
rs3847137	chr8:128483680-128483680	snp130											Yes	Yes		Yes		Yes	
rs2060776	chr8:128489299-128489299	snp130											Yes						
rs4276648	chr8:128496554-128496554	snp130											Yes	Yes		Yes			
rs4871022	chr8:128496902-128496902	snp130											Yes	Yes		Yes		Yes	
rs10956369	chr8:128492999-128492999	snp130											Yes	Yes		Yes		Yes	
rs7837644	chr8:128492580-128492580	snp130											Yes					Yes	
rs871135	chr8:128495575-128495575	snp130											Yes						
rs10505477	chr8:128476625-128476625	snp130											Yes					Yes	
rs10505474	chr8:128486686-128486686	snp130											Yes						
rs7837328	chr8:128492309-128492309	snp130											Yes			Yes			
rs10956368	chr8:128492832-128492832	snp130											Yes	Yes					
rs7837626	chr8:128492523-128492523	snp130		Yes	Yes				Yes	Yes	Yes		Yes	Yes				Yes	

wgEncodeRikenCage (rs6983267)

1. Track: ENCODE RIKEN CAGE Tags (PolyA- RNA in prostate whole cell)
Date Unrestricted: 2009-09-09
Name: 00003022001131200022003113
Position: chr8:128482488-128482513
Strand: -
Score: 200

2. Track: ENCODE RIKEN CAGE Tags (PolyA- RNA in prostate whole cell)
Date Unrestricted: 2009-09-09
Name: 00003022001131200022003113
Position: chr8:128482488-128482513
Strand: -
Score: 200

(b)

Results - Gaulton et al.

Location Information			Standard					Expression							Regulation								VarRep
SNP ID	SNP Position	Information	Conservation	Human ESTs	Human mRNAs	Gencode Genes	Relative Location	Affy RNA Loc	Caltech RNA-seq	GIS PET Loc	GIS RNA-seq	HudsonAlpha RNA-seq	RIKEN CAGE Loc	UW Affy Exon	Broad Histone	Open Chromatin	HAIB TFBS	SUNY RBP	UW Histone	UW DNase DGF	UW DNaseI HS	Yale TFBS	Common Cell CNV
	chr8:128482487		Yes	Yes	Yes			Yes	Yes	Yes			Yes		Yes	Yes	Yes		Yes	Yes	Yes	Yes	
rs7903146	chr10:114748339-114748339	snp130		Yes	Yes	Yes	Intronic		Yes	Yes	Yes	Yes		Yes	Yes	Yes		Yes			Yes		Yes

Fig. 1. Examples of ASSIMILATOR output showing results for (**a**) Pomerantz *et al*. with the causal SNP highlighted and (**b**) Gaulton *et al*. showing the evidence that the SNP is in a region of open chromatin. In addition, an example of results for a SNP without an rs number, as might be the case for novel SNPs identified via the 1000 Genomes project (http://www.1000genomes.org), is shown.

context of the thresholds and criteria used. ASSIMILATOR automatically queries any new tracks appearing from the ENCODE project on UCSC and includes these in the analysis. To further ensure ASSIMILATOR stays up to date, an option is available, which searches all UCSC database versions for ENCODE tracks and automatically uses the latest suitable version [currently March. 2006 (NCBI36/hg18)]. The ENCODE data release policy places restrictions on the publication of ENCODE data; therefore, the date at which the data becomes unrestricted is also displayed to aid the user.

To analyse the data, a hierarchical approach can be employed by the user, where isolated evidence for conservation across species, evidence of histone modification or mapping to a methylated region might be assigned a low weighting by the user; conversely, consistent evidence for a region being active, such as evidence for histone modification, DNase-1 hypersensitivity and open chromatin in the same cell line, coupled with evidence that a SNP lies within a transcription factor binding site (TFBS) would receive a higher weighting and could help to prioritize that SNP for functional work and may inform the design of such studies.

3 RESULTS

To verify the usefulness of ASSIMILATOR, we used information from a published study by Pomerantz *et al.* who found that an intergenic SNP, rs6983267, associated with colorectal cancer, showed functional evidence for interaction with the *MYC* gene (Pomerantz *et al.*, 2009). We used the SNP Annotation and Proxy Search (SNAP) tool (Johnson *et al.*, 2008) to generate a list of SNPs highly correlated with rs6983267 ($r^2 > 0.8$). This generated a list of 15 SNPs that were subsequently used as the input to ASSIMILATOR. The results are shown in Figure 1a clearly indicating that rs6983267 has the strongest a priori evidence of function. Not only is it in an active region of the genome, but also it is one of only two SNPs to lie in a TFBS. Additionally, ASSIMILATOR correctly identified the same TFBS as the published data.

Similarly, a recent study by Gaulton *et al.* (2010) looking at open chromatin across the genome identified a SNP associated with T2D in an open region. As a further proof of concept, supplying ASSIMILATOR with the same SNP revealed three lines of evidence showing bioinformatically that the SNP was in a region of open chromatin (Fig. 1b). This selection was achieved quickly and easily using our programme.

4 CONCLUSIONS

ASSIMILATOR provides a user-friendly interface with which to collate and assess the wealth of experimental evidence available for SNPs in order to prioritize efficiently for functional studies. ASSIMILATOR does not try to make assumptions about the likelihood of a SNP being functional and as such allows the user to make their own judgements about the candidacy of a SNP. ASSIMILATOR will also quickly and easily incorporate new data added to the ENCODE project ensuring that it maintains its relevance. With the wealth of information emerging from genome annotation studies, the task of manually mining the thousands of data points would be daunting. Here, we provide a one-stop solution that quickly and efficiently allows the user to view only relevant studies for their SNPs of interest and to mine that data with ease.

We have validated the program using published data and have shown that it allows the correct prioritization of a SNP subsequently shown to be the causal variant in a region associated with colorectal cancer. It thus provides an efficient portal to gather the essential information on which to base decisions regarding priorities for functional work. We have made ASSIMILATOR freely available through our web site as a download and we are also developing a web-based interface which will be found at the same location.

Funding: Arthritis Research UK (grant no. 17752).

Conflict of Interest: none declared.

REFERENCES

Adzhubei,I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Barton,A. *et al.* (2008) Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nat. Genet.*, **40**, 1156–1159.

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.

Gaulton,K.J. *et al.* (2010) A map of open chromatin in human pancreatic islets. *Nat. Genet.*, **42**, 255–259.

Johnson,A.D. *et al.* (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, **24**, 2938–2939.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Pomerantz,M.M. *et al.* (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.*, **41**, 882–884.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Thomson,W. *et al.* (2007) Rheumatoid arthritis association at 6q23. *Nat. Genet.*, **39**, 1431–1433.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.