BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 2 2015, pages 275–276
doi:10.1093/bioinformatics/btu639

Genetics and population analysis

Advance Access publication September 26, 2014

Enlight: web-based integration of GWAS results with biological annotations

Yunfei Guo[1,2], David V Conti[1,2] and Kai Wang[1,2,3,*]

[1]Zilkha Neurogenetic Institute, University of Southern California, Los Angeles, CA 90033, [2]Department of Preventive Medicine, USC Keck School of Medicine, Los Angeles, CA 90032 and [3]Department of Psychiatry & Behavioral Sciences, USC Keck School of Medicine, Los Angeles, CA 90033, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: Identifying causal variants remains a key challenge in post-GWAS (genome-wide association study) era, as many GWAS single-nucleotide polymorphisms (SNPs) (including imputed ones) fall into non-coding regions, making it difficult to associate statistical significance with predicted functionality. Therefore, we created a web-based tool, Enlight, which overlays functional annotation information, such as histone modification states, methylation patterns, transcription factor binding sites, eQTL and higher-order chromosomal structure, to GWAS results.

Availability and implementation: Accessible by a Web browser at http://enlight.usc.edu.

Contact: kaiwang@usc.edu

Received on April 18, 2014; revised on September 9, 2014; accepted on September 22, 2014

1 INTRODUCTION

Genome-wide association studies (GWAS) enabled the identification of genetic variants associated with complex diseases and traits. To date, there are 14 012 genome-wide significant variants in the GWAS catalog (Hindorff *et al.*, 2009). However, most of the top hits are believed to be proxy markers for the underlying causal variants, and the majority of those variants lie within non-coding sequences (Maurano *et al.*, 2012). Despite substantial efforts in the post-GWAS era to identify causal variants, lack of knowledge on biological functions has become a major hurdle for differentiating causal variants from highly correlated tag single-nucleotide polymorphisms (SNPs).

A large amount of biological annotation is now available for non-coding sequences, given the community efforts devoted to elucidating functional significance of non-coding sequences in the past few years. For example, the ENCODE project has systematically mapped regions of active transcription, transcription factor (TF) association, chromatin structure and histone modification, and assigned at least one biochemical function to ~80% of the genome (Bernstein *et al.*, 2012). Recent studies showed that disease-related variants detected by GWAS are significantly enriched in regions where regulatory elements are inferred, such as DNase I hypersensitivity sites and TF binding sites (TFBS) (Maurano *et al.*, 2012; Schaub *et al.*, 2012). Proper integration of this biological information with GWAS results will likely shed light on prioritization and identification of candidate causal variants (McCarthy and Hirschhorn, 2008).

To that end, tools such as HaploReg (Ward and Kellis, 2012), RegulomeDB (Boyle *et al.*, 2012), GWAS3D (Li *et al.*, 2013) and FunciSNP (Coetzee *et al.*, 2012) were developed to leverage functional annotations to interrogate GWAS variants. However, users cannot visually recognize the annotation patterns.

The popularity of regional plots highlights the importance of visual inspection of the GWAS results. Although LocusZoom (Pruim *et al.*, 2010) generates high-quality images showing local linkage disequilibrium (LD) and association significance, it provides little information for interpreting non-coding variants. Therefore, we developed a web-based tool, Enlight, combining biological annotation and GWAS-related information into one single plot. It takes GWAS results (including imputed ones) as input, and returns a regional plot with a variety of epigenetic information, such as histone modification, methylation patterns, TFBS, eQTL and higher-order chromosomal structure.

2 IMPLEMENTATION

2.1 Architecture

The server backend is implemented in Perl CGI and MySQL. Regional plots are performed by LocusZoom (Pruim *et al.*, 2010); annotation plots are performed by custom R and Python code built on the basis of LocusZoom. Text annotation is performed by ANNOVAR (Wang *et al.*, 2010).

2.2 Input and output

The minimum input contains two columns: marker name (rsID) and *P*-value. Optionally, users can supply chromosome, start position, end position, reference allele and alternative allele in the first five columns; if they are absent, Enlight converts marker names to corresponding coordinates based on dbSNP137. On submission, Enlight will output a conventional regional plot containing association strength, LD information, gene position and chromosomal recombination rate. Meanwhile, Enlight plots categorical annotations like chromHMM classification as colored stripes, continuous annotation such as histone modification states, DNA methylation signal, TF binding strength, etc., as histograms, below the GWAS results panel (Fig. 1). Variants that appear in the GWAS catalog and eQTL databases from either GTEx project or eQTL browser of Pritchard Lab (http://eqtl.uchicago.edu/) can be shown by a XYplot. Hi-C

*To whom correspondence should be addressed.

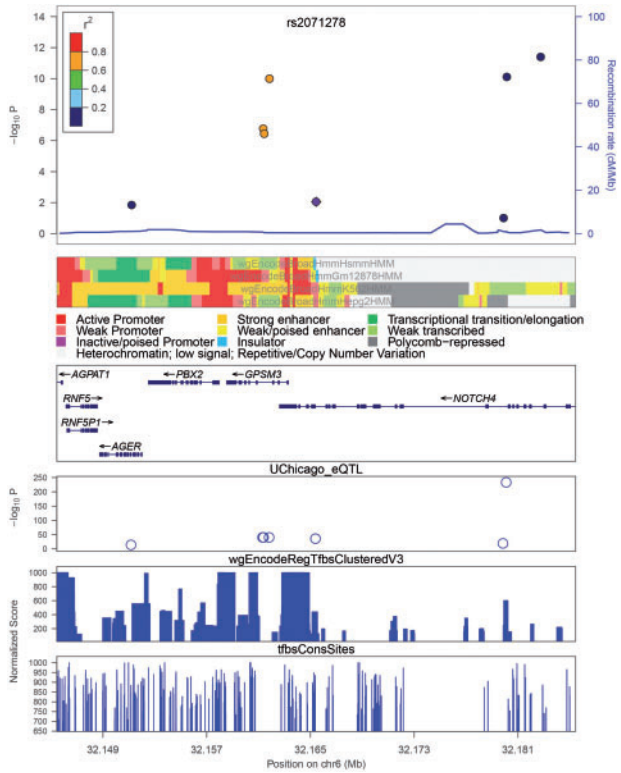

Fig. 1. The association signals for rs2071278 and nearby loci from the Wellcome Trust Case Control Consortium (WTCCC) Rheumatoid Arthritis study. The color stripe shows the chromHMM segmentation result, which combines multiple epigenetic chromatin marks (legend shown below). The three SNPs rs204991, rs204990 and rs204989 (all yellow), which are in high LD with the reference SNP, show up in a proposed enhancer region in blood cell lines, GM12878 and K562, but not in non-blood cell lines, HepG2 or HSMM. The UChicago_eQTL panel shows association signal based on data from eQTL browser at the University of Chicago. The wgEncodeRegTfbsClusteredV3 panel summarizes a large collection of TF ChIP-seq results, and tfbsConsSites panel describes TFBS conservation across human/mouse/rat. Hi-C interaction plot is not shown

interaction (Lieberman-Aiden, *et al.*, 2009) can be plotted as a heatmap. Users are able to select data tracks from 14 cell lines and 17 annotation types or plot with their own data in BED format.

In addition to visualization, Enlight generates text annotation for each variant using ANNOVAR, which can be used for downstream analyses, such as finding biofeature overlap, regression model building, etc.

Figure 1 shows a typical plot. rs2071278 is genome-wide significantly associated with serum complement C3 and C4 levels (Yang *et al.*, 2012), important measurements in the assessment of rheumatoid arthritis (Makinde *et al.*, 1989). Therefore, rs2071278 is possibly a SNP tagging nearby functional variants. For instance, rs204991, rs204990 and rs204989, all in high LD with rs2071278 (1000 genomes, European population, 2012), are shown to be significantly associated with expression of human leukocyte antigen genes by microarray (Zeller *et al.*, 2010). This finding partly explains their strong associations with rheumatoid arthritis in the WTCCC study (Wellcome Trust Case Control, 2007). When strong associations appear near known GWAS hits and are surrounded by interesting epigenetic and other functional features, they perhaps suggest functional importance.

3 CONCLUSIONS

We have created a user-friendly tool to generate regional plots of association results with biological annotation as well as text annotation for individual variant. Enlight makes it possible to visually connect epigenetic features or other functional annotations with the LD information, thus facilitating the identification of putative causal variants. Enlight can be accessed via a web-based interface.

ACKNOWLEDGEMENTS

The authors thank Dr. Jihong Kim, Mr. Zhuo Zhang and many users for valuable comments and helpful feedback.

Funding: National Institute of Health [R01 HG006465].

Conflict of interest: K.W. is a share holder and board member of Tute Genomics, a genome interpretation company.

REFERENCES

Bernstein,B.E. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

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

Coetzee,S.G. *et al.* (2012) FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Res.*, **40**, e139.

Hindorff,L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.

Li,M.J. *et al.* (2013) GWAS3D: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nucleic Acids Res.*, **41**, W150–W158.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Makinde,V.A. *et al.* (1989) Reflection of disease activity in rheumatoid arthritis by indices of activation of the classical complement pathway. *Ann. Rheum. Dis.*, **48**, 302–306.

Maurano,M.T. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.

McCarthy,M.I. and Hirschhorn,J.N. (2008) Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.*, **17**, R156–R165.

Pruim,R.J. *et al.* (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, **26**, 2336–2337.

Schaub,M.A. *et al.* (2012) Linking disease associations with regulatory information in the human genome. *Genome Res.*, **22**, 1748–1759.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Ward,L.D. and Kellis,M. (2012) HaploReg: a resource for exploring chromatin states, conservation: regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, **40**, D930–D934.

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.

Yang,X. *et al.* (2012) Genome-wide association study for serum complement C3 and C4 levels in healthy Chinese subjects. *PLoS Genet.*, **8**, e1002916.

Zeller,T. *et al.* (2010) Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PLoS One*, **5**, e10693.