GWAS3D: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications

Mulin Jun Li[1,2], Lily Yan Wang[1,2], Zhengyuan Xia[2,3], Pak Chung Sham[2,4,5,6] and Junwen Wang[1,2,4,*]

[1]Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China, [2]Shenzhen Institute of Research and Innovation, The University of Hong Kong, Shenzhen, Guangdong 518057, China, [3]Department of Anaesthesiology, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China, [4]Centre for Genomic Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China, [5]Department of Psychiatry LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China and [6]State Key Laboratory in Cognitive and Brain Sciences, The University of Hong Kong, Hong Kong SAR, China

ABSTRACT

Interpreting the genetic variants located in the regulatory regions, such as enhancers and promoters, is an indispensable step to understand molecular mechanism of complex traits. Recent studies show that genetic variants detected by genome-wide association study (GWAS) are significantly enriched in the regulatory regions. Therefore, detecting, annotating and prioritizing of genetic variants affecting gene regulation are critical to our understanding of genotype–phenotype relationships. Here, we developed a web server GWAS3D to systematically analyze the genetic variants that could affect regulatory elements, by integrating annotations from cell type-specific chromatin states, epigenetic modifications, sequence motifs and cross-species conservation. The regulatory elements are inferred from the genome-wide chromosome interaction data, chromatin marks in 16 different cell types and 73 regulatory factors motifs from the Encyclopedia of DNA Element project. Furthermore, we used these function elements, as well as risk haplotype, binding affinity, conservation and P-values reported from the original GWAS to reprioritize the genetic variants. Using studies from low-density lipoprotein cholesterol, we demonstrated that our reprioritizing approach was effective and cell type specific. In conclusion, GWAS3D provides a comprehensive annotation and visualization tool to help users interpreting their results. The web server is freely available at http://jjwanglab.org/gwas3d.

INTRODUCTION

Recent studies on human genetics, such as The International HapMap Project (1) and 1000 Genomes Project (2), have identified a large number of genetics variants in the human genome. Furthermore, genome-wide association studies (GWAS) (3) and exome sequencing (4,5) are extensively used to globally investigate the relationship between genetic variants and human diseases/traits. By looking at the genomic location of the associated variants detected in GWAS, a large portion (~88%) of them fall outside of coding regions, which are harder to interpret than the protein-coding variants (6). Therefore, elucidating the molecular function of genetic variants locating in the non-coding regions is critical to our full understanding of genetic disorders.

However, there are many difficulties and computational challenges in achieving this goal (7). One of the major difficulties comes from the unclear role of non-coding genetic variants in the relevant processes underlying disease/trait association. These variants could affect many biological activities including transcription, splicing, post-transcriptional regulation, translation initiation/elongation and post-translational modification (8). Previously, conservation information was frequently used to prioritize the functional importance of non-coding genetic variation (9,10). At the transcription regulation level, mutations in the promoter regions may impact the recruitment of RNA polymerase and other regulators, especially the binding of transcriptional

---

*To whom correspondence should be addressed. Tel: +852 2831 5075; Fax: +852 2855 1254; Email: junwen@hku.hk or junwen2u@gmail.com

factors (TFs) to the promoter region to initiate gene transcription. Tools such as is-rSNP (11), sTRAP (12) and regSNPs (13) have been successfully developed to evaluate the binding affinity affected by genetic variation. However, although algorithms that solely used TF motifs are effective in finding regulatory elements in the immediate promoter regions but may inevitably introduce a large number of false positives in the distal promoter/enhancer regions where the searching space becomes substantially larger. More and more studies showed that mutations within the distal regulatory elements, such as enhancer, insulator and silencer, could also disrupt or change the binding of TFs, nucleosome positioning signals and chromatin states. Furthermore, the locally changed chromosome conformation can block or create looping interaction between distal elements and promoter regions (14) and subsequently influence gene regulation. Unfortunately, few tools or resources have used such information to study genetic variants.

The Encyclopedia of DNA Elements (ENCODE) project has identified a comprehensive map of functional elements and active chromatin marks by advanced techniques such as ChIP-seq, DNase-seq, bisulfate sequencing, chromosome conformation capture and so forth. (15,16). Recent studies showed that disease-associated single-nucleotide polymorphisms (SNPs) detected by GWAS are significantly enriched in the regions that harbor functional elements, such as transcriptional factor binding sites (TFBSs), histone modification marked regions, DNase I hypersensitive sites (DHSs) and expression quantitative trait loci (16–19). Two recently published databases, HaploReg (20) and RegulomeDB (21), have used these regulatory signals and marks to annotate the genetic variants, which offer comprehensive resources on regulatory variation. On the other hand, different functional elements have been reported to function in a tissue/cell type-specific manner. SNPs associated with the same trait are likely to locate in active chromatin marks in the same/relevant cell type (22), implying the possibility of detecting regulatory signals using the chromatin marks of phenotypically relevant cell type. Computational algorithms including ChromHMM (23) and Seaway (24) have been successfully applied to scan different functional elements in the genome. Therefore, combinatory analysis of GWAS data and functional elements in a specific cell type to capture regulatory variants for a particular disease/trait are needed.

Here, we develop a web server GWAS3D (http://jjwanglab.org/gwas3d) to systematically analyze the probability of genetic variants affecting regulatory pathways and underlying disease/trait associations by integrating chromatin state, functional genomics, sequence motifs and cross-species conservation for a set of GWAS data or variant list. We first collected and curated genome-wide chromosome interaction (5C, Hi-C, ChIA-PET) data, enhancer/insulator/promoter marks [H3K4me1, H3K27ac, p300, CCCTC-binding factor (CTCF), DHS] and ChromHMM predicted functional elements in 16 different cell types. Using those regulatory regions, we mapped genetic variants to the reference genome and evaluated the binding affinity changes of regulatory

factors by scanning 73 ENCODE motifs. Finally, we combined original GWAS signal, risk haplotype, binding affinity significance and conservation information to prioritize the genetic variants. In addition, the system provides comprehensive annotation and visualization to help users to interpret the results. Comparing with existing software and databases, GWAS3D uses the latest information to build a one-stop web-based tool for clinicians and biologists to evaluate the deleteriousness of disease/trait-associated variants that affect transcription regulation on a broader spectrum, especially on non-coding genetic variation.

METHOD AND PIPELINE

Data collection and processing

GWAS3D integrates multiple genome-wide experimental data to connect genetic variants with underlying gene regulation mechanism through high-dimensional regulatory interactions. We first collected and curated the experimental results of long-range interactions, for 16 different cell types, measured by high-throughput chromosome conformation capture technologies (5C, ChIA-PET and Hi-C) from the ENCODE project, Gene Expression Omnibus (GEO) database and published resources (Supplementary Table S1). We directly used 5C and ChIA-PET interactions in the database and processed the Hi-C interactions by the iterative correction and eigenvector decomposition (ICE) algorithm (25), which can largely reduce the false positives and biases. Some chromatin marks have been reported and validated as the active signals of enhancers, including histone modifications of H3K4me1 and H3K27ac, DHSs and E1A-binding protein p300 (26,27), we therefore extracted the related ChIP-Seq peaks for the above 16 cell types from ENCODE. We also collected ChIP-Seq data of CTCF-binding sites, which imply transcription repression and chromatin insulation. For predicted elements, we downloaded the ChromHMM genome-wide maps of chromatin state annotations for supported cell types and extracted the promoter, enhancer and insulator elements whose signals are predicted as 'Strong'. We also merged the genomic profiles of three ENCODE tier 1 cell lines (GM12878, K562, H1 human embryonic stem cells) to support the 'no cell type restriction' option.

Genetic variants data sets were collected from dbSNP137 (28) and 1000 Genomes Project phase 1 release version 2 (29), which comprise 52 054 804 and 26 152 995 SNPs and Indels, respectively. We assigned reference allele and all alternative alleles to dbSNP137 variants and used biallelic variants for 1000 Genomes Project. The allele information was used to calculate the binding affinity of TFs. Linkage disequilibrium (LD) data for 11 populations were retrieved from the merged data of HapMap phases I + II + III. LD data for four 1000 Genomes Project super populations were computed and retrieved from MACH (30). Genomic coordinate of each locus was converted to GRCh37 hg19 by UCSC liftover tool. Variants with dbSNP ID were mapped to dbSNP137 using dbSNP merge file. In addition, annotations for genes
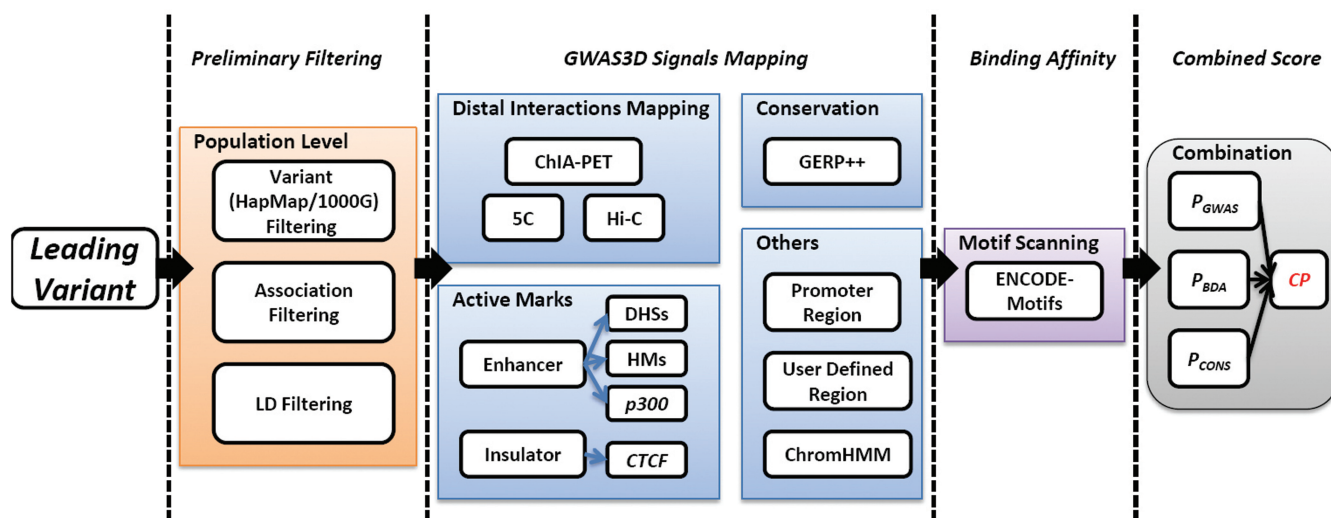
Figure 1. The workflow of GWAS3D (see the description of pipeline for details).

and other DNA elements were downloaded from the UCSC Genome Browser. Furthermore, GWAS3D used position frequency matrices of 73 transcription factors motifs grouped by family (ENCODE motifs) from ENCODE web site, which provides a comprehensive resource of 245 known motifs curated from TRANSFAC, Jaspar and protein-binding microarray experiments, and 293 novel motifs discovered by motif finding tools (including MEME, MDscan, Weeder, AlignACE) using large numbers of ChIP-Seq data. To consider the regions with evolutionary constraint, we also used the conservative elements by genomic evolutionary rate profiling (31) and used these signals to prioritize the suspected deleterious variants.

Pipeline

Given a set of GWAS data or a SNP list, GWAS3D can detect the variants' regulatory effects such as the assigned population haplotype, the experimentally derived genetic/epigenetic signals, the predicted change of transcription factor binding affinity on different alleles and sequence conservation in a particular cell type. The server further calculates the combined effect of each variant and prioritizes them based on the probability of affecting gene regulation. The overall workflow of GWAS3D is shown in Figure 1.

Preliminary data filtering

The system accepts inputs either from an association study or a SNP list. Many formats are supported, including the Plink-like (32) format, VCF-like format, single dbSNP ID and variant chromosome position. The input of association P-value is compulsory when the GWAS effect size is considered for prioritization. A user-defined P-value cut off is applied to filter out the less significant SNPs and to reduce data volume. SNPs or Indels will be checked and mapped to dbSNP137 or 1000 Genomes Project variants. Variants not using VCF-like format will be assigned respective alleles according to dbSNP137 then 1000

Genomes Project. The web server will filter the variant not mapped onto either dbSNP137 or 1000 Genomes Project unless VCF-like format is used. Then, it would fetch all variants in LD of each aforementioned leading variants by user-defined LD standard (HapMap or 1000 Genomes Project), population and r-square (r^2) cut off.

Identifying GWAS3D regulatory signals

Cell-type specific marks, including genome-wide long-range interactions, active promoter/enhancer/insulator marks, predicted ChromHMM maps, as well as a user-defined promoter region, are then used to identify the potential regulatory effects of the variants. We defined a variant mapped onto any of these regions as a 'GWAS3D signal', which implies a relevant regulatory function such as affecting distal interaction in high dimension or direct promoter activity of a target gene. The variants that are not mapped to any regulatory regions are filtered out at this stage.

Computing the binding affinity effect by ENCODE-motifs

To quantitatively measure the difference on the binding affinity caused by different alleles of candidate variants with GWAS3D signals and its significance, we used a comprehensive TF motif set to evaluate the possible reduced or enhanced binding. We first computed the position weight matrices (PWMs) from position frequency matrices of all ENCODE motifs by converting normalized frequency value to log-scale value using the method described previously (33,34). Given a variant (V) with GWAS3D signal, we took 30 bp of surrounding sequence and constructed the mutated sequences between the reference alleles (A_r) and the alternative alleles ($A_{a1}, \ldots A_{an}$). For user-selected motifs of TFs, we scanned these sequences using PWMSCAN (35) and fetched P-values represent the significance of each putative TF-binding site. We set a PWMSCAN P-value threshold (1E-3) to reduce the number of false positive bindings. We then measured the score of binding affinity

change by calculating the log-odds (LOD) of probabilities of paired binding sites for each motif (m):

$$LOD_m(V_{A_r A_{an}}) = log_{10} \frac{P_m(V_{A_r})}{P_m(V_{A_n})}$$

To estimate the statistical significance of binding affinity change, we performed permutations of on 52 054 804 SNPs in dbSNP137 and computed respective LOD of all ENCODE motifs to form the null distribution of binding affinity difference. We then used FastPval (36) to compute the *P*-value of each LOD from aforementioned empirical distribution.

Prioritization of regulatory variants

We combined the original GWAS signal, risk haplotype, binding affinity significance and conservation information to prioritize the leading variants (L) detected by GWAS. For each L, we first find all variants (V) with GWAS3D signal and in LD with L ($r^2 >$ user selected cutoff). We then calculated V's phenotypically associated effect (P_{GWAS}) by dividing the GWAS *P*-value of L with the r^2 between V and L in the user-specified populations. We then selected the most significant *P*-value of the LOD related to a specific TF motif (P_{BDA}) to represent the binding affinity effect of the variant. We further mapped the variant to genomic evolutionary rate profiling++ (37) constrained elements and calculated the corresponding conservation *P*-value (P_{CONS}). Using the *P*-values of aforementioned three independent measurements (GWAS, binding affinity, conservation), we performed Fisher's combined probability test to calculate a combined *P*-value, *CP*, for each V. We then assigned the most significant *CP* from all the variants Vs to the corresponding leading variant L. All the Ls are then re-ranked according to their new *CP* values, with special focus on their regulatory effects.

Plotting the GWAS3D regulatory variants

To visualize global chromosome interactions among putative regulatory variants and their associated loci, GWAS3D also provide informative circle plots of high-dimensional chromosome interactions. We selected top significant variants (defined by the user) detected by GWAS3D and mapped them to RefGene for gene names or Cytoband for chromosome locations. We generated an intuitive circle graph using VIZ-GRAIL (38) with some modifications.

WEB SERVER DESCRIPTION

Usage and interface

The system accepts four formats for variants including either GWAS format such as Plink-like format and single dbSNP ID, or NGS format such as VCF-like format and single variant chromosome position resulting from high-throughput sequencing. LD information of different populations for both HapMap and 1000 Genomes Project is well supported by GWAS3D, which also allows users to define the cutoffs of association *P*-value and LD.

More stringent settings will reduce the running time but some truly associated variants with moderate effect size might be lost. Also, GWAS3D provides information for 16 different cell types, which have been extensively investigated on chromatin states by recent ENCODE project, especially for long-range chromosome interactions. We recommend users to select the cell type that is relevant to the observed disease/trait in their GWAS/ NGS study. Furthermore, GWAS3D allows users to choose relevant TF families and related known/novel motifs, which benefit capturing the binding affinity changes for a set of specific TFs. User can define a specific *P*-value cut off for putative TF-binding site scanning. Other settings, such as user-defined genomic regions and visualization options, including promoter definition, allowed number of variants and distal intentions for plotting, are also adjustable by the users.

GWAS3D uses a series of user-friendly interfaces to display the results, which summarize the potential regulatory effects of these variants and facilitate the identification and selection of casual variants for follow-up experimental validation. The detected regulatory variants and their associated loci/interactions can be globally viewed from a circle plot in the left panel of 'GWAS3D PLOT' page (Figure 2). User can also check the overview of current GWAS3D run and download related information from right panel (Supplementary Figure S1). To query the detailed information of each variant with GWAS3D signals, we designed a variant prioritization table as well as comprehensive tab viewers in the 'GWAS3D INFO' page. In the prioritization table, the most significant regulatory variant in the LD of each leading variant is ranked by the significance of combined *P*-value. Also, variants with different type of GWAS3D signals are marked in different colors. For example, variants with significant TF-binding affinity change will be marked by purple stamp, and variants with active enhancer signal will be marked by green stamp. In the right tabs, user can identify the deleteriousness of selected variants by analyzing many annotations of its regulatory features in a dynamic manner (Figure 3).

Annotation

For each variant detected by GWAS3D, we provided six annotations, including variant summary, binding affinity, GWAS3D signals, genomic elements, LD signals and external annotation. Users can systematically analyze the regulatory properties of variants based on these annotations. First tab is about variant summary, showing the important attributes related to the selected variant and reports the information of previous GWAS result recorded in GWASdb (39) for this variant. In the tab of binding affinity, we listed top five significant affinity differences of TF motifs with detailed binding sites information. For the GWAS3D signals information, we offered a tab to show all of mapped functional elements used in GWAS3D and related marks information. To help user identify all putative regulatory variants in the LD of observed leading variant, we used one additional tab to list related information. Lastly, three useful external
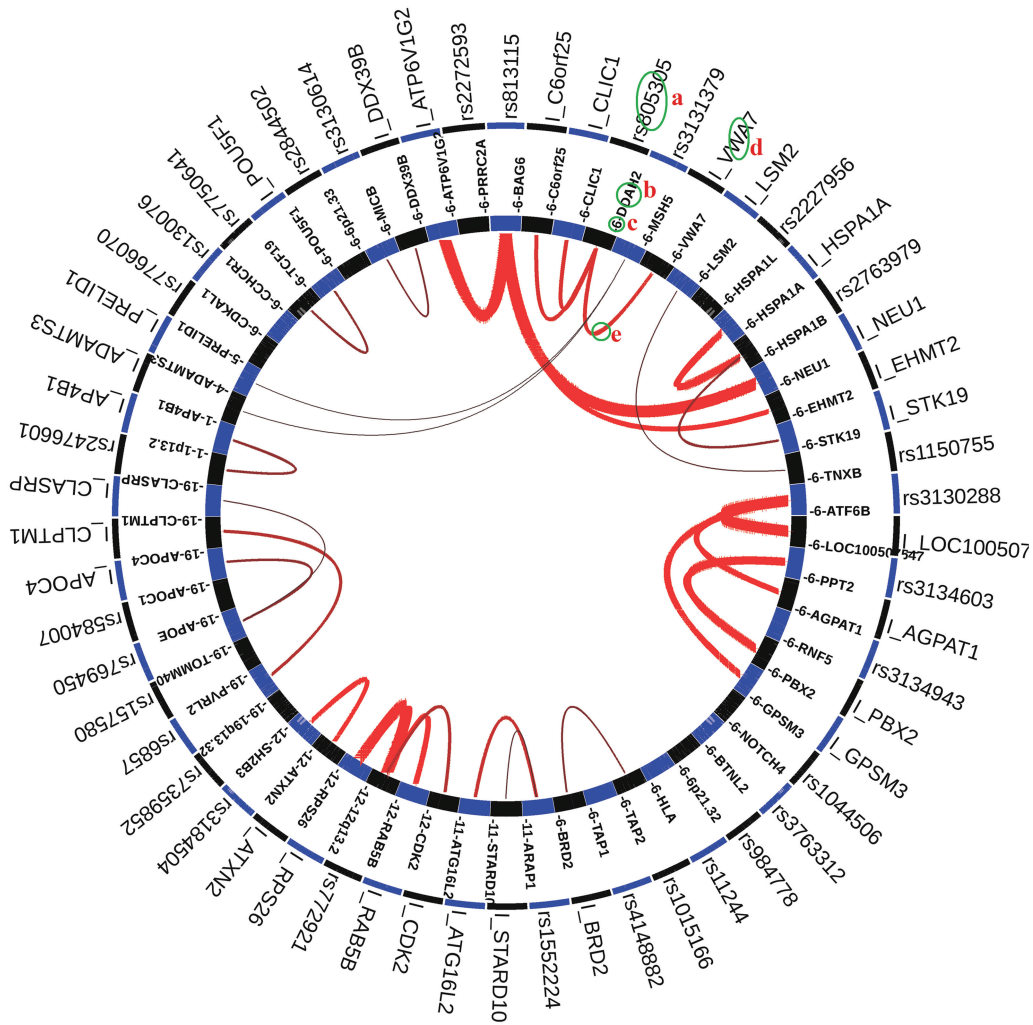
Figure 2. The circle plot of GWAS3D for all GWAS SNPs of diabetes mellitus from NHGRI GWAS Catalog based on the K562 cell line and CEU population. Significant GWAS3D results are presented by the circle plot. From the outer to inner, there are significant regulatory variants and distal interaction regions, genes and genomic loci, chromosome number and distal interaction indicators. For example, GWAS SNP rs805305 is detected as a significant regulatory variant by GWAS3D (**a**), this variant located on the intronic region of *DDAH2* (**b**) in chromosome 6 (**c**). One of the important regulatory features for this variant, which can be viewed from this plot, is that the region has a long-range interaction signal to another locus near *VWA7* (**d**), interactive elements with significant regulatory variant will start with 'I_'). The red line indicated this signal (**e**), and the intensity of interaction is represented by width.
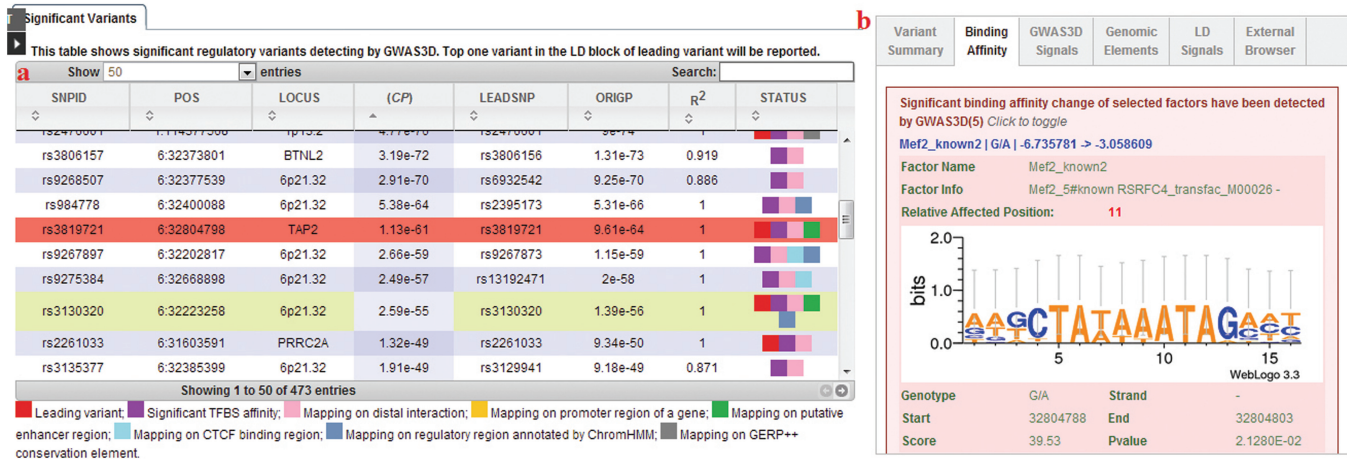

Figure 3. The 'GWAS3D INFO' page for detailed information of regulatory variant. The web page consists of two parts: (**a**) tabular viewer for significant variant detected by GWAS3D. (**b**) six annotation tabs of GWAS3D.

browsers were encapsulated into the system to give broad annotations and predictions including GWASrap (40), RegulomeDB and UCSC ENCODE genome browser. User can directly fetch the information in an internal window.

Evaluation

We first tested the web server using a well-studied locus known to be associated with plasma low-density lipoprotein cholesterol (LDL-C). We collected 17 associated SNPs in 1p13 region (Supplementary Table S2) genotyped in ~20 000 individuals of European descent and ~9000 African American individuals with LDL-C (41) and performed GWAS3D pipeline on those variants under HapMap CEU population and HepG2 cell type. We obtained five significant regulatory variants with distinguished GWAS3D signals and identified a leading variant rs12740374, as the top one in the prioritization table. The variant locates between genes *CELSR2* and *PSRC1*, and was scored a more significant *CP* (7.89E-46) than the original GWAS *P*-value (1.8E-42). GWAS3D reported that rs12740374 directly connected with two active enhancer marks (H3K27ac and DHS) and located in the conserved region. Importantly, binding affinity between allele G and risk allele T showed substantial difference for C/EBP transcription factors (Supplementary Figure S2). Those results were consistent with previous finding about the role of rs12740374 in the lipoprotein regulatory pathway (41). We also applied same associated SNPs set on YRI and CHB populations, besides the most significant leading SNP rs12740374, we further detected rs629301 that possibly disrupting the binding of transcription factor YY1, and other GWAS3D signals (Supplementary Figures S3 and S4). Another speculation is that this variant may influence the recognition and targeting of miR-199 in the 3′UTR of *CELSR2*. This variant was frequently reported as a highly associated signal in 1p13 region with LDL-C (42,43). Interestingly, we did not observe strong enhanced signals (active enhancer or promoter) at those GWAS-associated variants when we used non-liver cell type such as K562, H1-hESC and HeLa-S3. In contrast, CTCF-binding sites were observed around some of those associated variants, which may reflect the phenotypically cell type-specific association (Supplementary Figure S5) (22).

To further evaluate the ability of GWAS3D for detecting and prioritizing regulatory variants in a genome-wide manner, we collected 1370 associated SNPs with prostate cancer from NHGRI GWAS Catalog (6) and GWASdb (39). We detected 195 variants, which have GWAS3D regulatory signals (active promoter/enhancer/insulator marks, TFBS affinity changes, conserved elements), after applying the CEU population and RWPE1 cell type. Seven variants obtained more significant *CP* than the originals when analyzing top 20 putative regulatory variants in the prioritization table (Supplementary Table S3). Most of these significant variants exert the regulatory function of their associated loci by high linked LD variants other than leading SNPs. However, a noticeable result is a

leading SNP rs6983267 that harbors many GWAS3D signals has been successfully validated by many functional studies for affecting enhancer activity (44,45).

We then quantitatively evaluated the performance of our method. We first collected 118 known regulatory variants from OregAnno database (46). We randomly selected three data sets from dbSNP with same number of genetic variants in each of the regulatory regions (promoter, intergenic and genome-wide). For each of aformentioned four SNVs list, we performed GWAS3D pipeline without considering GWAS *P*-value, cell type restriction and population LD. Wilcoxon rank-sum test showed significant differences between OregAnno and each random sets, with *P*-values of 0.0344, 0.0011 and 0.0052 for promoter, intergenic and genome-wide data set, respectively, whereas there are no differences among the three random data sets. The experiment demonstrated that GWAS3D pipeline gives higher scores to regulatory variants and thus differentiates them from random variants.

The SNVs detected by GWAS are associated with the disease/trait, but may not be the one with function implications. GWAS3D can find functional SNVs from GWAS SNVs through LD and other information. To assess this capability, we used GWAS3D pipeline to find functional SNV for each of the selected GWAS SNVs (118 top GWAS significant variants in intergenic and promoter regions from GWAS Catalog database). We found there were no significant differences between the GWAS SNVs and any of the three random data sets (all *P* > 0.05, Wilcoxon rank-sum test). However, the functional SNVs found by GWAS3D are significantly highly scored with *P*-values of 2.966E-05, 9.591E-08 and 3.034E-07 compared with the three random sets (Wilcoxon rank-sum test), respectively. Those results further confirmed the capability of GWAS3D in identifying functional regulatory variants (Supplementary Figure S6).

Server design

We implemented the GWAS3D web server with a Perl-based web framework 'Catalyst', which provides a flexible programming interface on web development. Annotation information is stored in a back-end MySql database. We used Oracle Grid Engine as job management system for submitting tasks and offered three ways for users to retrieve their jobs: encrypted links, browser cookies and email notifications. jQuery and related UI components are used to construct dynamic web pages. GWAS3D is a one-stop framework with high usability and is freely available for academic use.

DISCUSSION

We have designed a web-based tool to detect, prioritize and annotate the regulatory genetic variations in combination with experimental data and computational predictions. Particularly, this tool takes advantage of recently generated ENCODE data, especially the experimental long-range interactions as well as the active marks of functional elements, to predict variants in the putative TFBSs

in distal and proximal promoter regions. GWAS3D is a tool dedicated to detect true functional variants that control gene regulation for genetic studies. Compared with recent software and databases such as VAAST, HaploReg and RegulomeDB, GWAS3D integrates more features and can be used in many scenarios. User can identify the most probable functional variant associated with interesting trait in one risk locus or prioritize the leading variants when given a full list of GWAS result or evaluate the deleteriousness of genetic variants affecting the gene regulation without any prior effect. GWAS3D also provides flexible configurations, such as human population, cell type specificity and TF family classification, for users to deal with different aspects of complex disease/trait. For example, user may select a matched cell type/tissue satisfying with a specific phenotype or manually define motifs of interested TFs used in following scanning when considering the tissue specificity of TFs. Recently, researchers found that the disease/trait-associated variants are highly related to active chromatin marks in relevant cell types (22). Therefore, these distinct features will greatly facilitate the discovery of regulatory variants under particular condition.

There are unbalanced genomic data of multiple domains for different cell types/tissues. For example, a lot of data were provided by ENCODE tier 1 cell lines (GM12878, K562, H1 human embryonic stem cells), whereas the data were few on tier 3 cell lines. It may potentially affect the quality of our annotation when applying GWAS3D pipeline to the cell lines having fewer data available. We therefore specially selected the 16 cell types, which included enough chromosomal looping data (5C or ChIA-PET or Hi-C) and important transcriptional markers data (H3K4me1, H3K27ac, DHSs, EP300 and CTCF). To cope with tissue/cell type limitation, we added a 'without tissue/cell type restriction' item in cell type selection option by merging the genomic profiles of three ENCODE tier 1 cell lines. Because the aforementioned three cell lines contain dynamic transcription signals from human normal adult cells, cancer cells and embryonic stem cells. In the future, we will continuously update the number of tissue/cell type when enough data are available for that cell line.

It was reported that many active chromatin marks are located in the intronic and exonic region of genes (47). Enhancers can also reside in intronic region of a gene to coordinate the looping with active promoter of another gene (48). Even for validated human fragments with enhancer activity (49), we found 30.42% of these fragments overlapped with coding region of genes. Thus, genetic variants not belonging to non-coding RNAs may also be associated with gene regulation. On the other hand, an exonic variant can associate with particular regulatory process by linking variants in the LD proxy. GWAS3D not only provides an efficient solution to interpret the regulatory role of genetic variation in the noncoding regions but also in other genic regions.

The computational process of our system is real-time, which is different from databases such as HaploReg and RegulomeDB, where the function annotations are pre-computed and stored in the database in advance.

Therefore, it can dynamically deal with the genetic variants input by users with maximum flexibility. Despite large computational burden in the background when LD is considered, our system can finish the job of a meta GWAS data set (thousands of variants with moderate GWAS significance, $P < 1.0 \times 10^{-5}$) within a few hours even with LD from the 1000 Genomes Project. It will be much quicker when using HapMap LD. To exploit the regulatory properties of personal genomics data, GWAS3D accepts VCF-like format and can evaluate the deleteriousness of rare/novel variation altering gene regulation associated with personalized trait.

Furthermore, our system provides visualization and instant annotation for detected variants. Using the circle plotting, important regulatory variants and its affected regions, as well as the intra/interchromosomal interactions related to variants, can be intuitively displayed. Although many tools, such as SeattleSeq, ANNOVAR (50) and ENSEMBL VEP (51), can help users retrieve sufficient variant annotations, the integrative function annotation of GWAS3D will benefit users for instant query and broader range of information. Therefore, besides the genomic mapping information of variant (such as information of gene and other genomic elements), we offered several direct links to the servers of GWASrap (40), RegulomeDB and UCSC ENCODE genome browser in the internal windows of GWAS3D.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3 and Supplementary Figures 1–6.

ACKNOWLEDGEMENTS

FUNDING

Conflict of interest statement. None declared.

REFERENCES

1. International HapMap 3 Consortium, Altshuler,D.M., Gibbs,R.A., Peltonen,L., Dermitzakis,E., Schaffner,S.F., Yu,F.L., Bonnen,P.E., de Bakker,P.I., Deloukas,P. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.
2. Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
3. McCarthy,M.I., Abecasis,G.R., Cardon,L.R., Goldstein,D.B., Little,J., Ioannidis,J.P. and Hirschhorn,J.N. (2008) Genome-wide

association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.

4. Bamshad,M.J., Ng,S.B., Bigham,A.W., Tabor,H.K., Emond,M.J., Nickerson,D.A. and Shendure,J. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.*, **12**, 745–755.

5. Do,R., Kathiresan,S. and Abecasis,G.R. (2012) Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum. Mol. Genet.*, **21**, R1–R9.

6. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.

7. Ward,L.D. and Kellis,M. (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.*, **30**, 1095–1106.

8. Sauna,Z.E. and Kimchi-Sarfaty,C. (2011) Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.*, **12**, 683–691.

9. Yandell,M., Huff,C., Hu,H., Singleton,M., Moore,B., Xing,J., Jorde,L.B. and Reese,M.G. (2011) A probabilistic disease-gene finder for personal genomes. *Genome Res.*, **21**, 1529–1542.

10. Lehmann,K.V. and Chen,T. (2013) Exploring functional variant discovery in non-coding regions with SInBaD. *Nucleic Acids Res.*, **41**, e7.

11. Macintyre,G., Bailey,J., Haviv,I. and Kowalczyk,A. (2010) Is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics*, **26**, i524–i530.

12. Manke,T., Heinig,M. and Vingron,M. (2010) Quantifying the effect of sequence variation on regulatory interactions. *Hum. Mutat.*, **31**, 477–483.

13. Teng,M., Ichikawa,S., Padgett,L.R., Wang,Y., Mort,M., Cooper,D.N., Koller,D.L., Foroud,T., Edenberg,H.J., Econs,M.J. *et al.* (2012) RegSNPs: a strategy for prioritizing regulatory single nucleotide substitutions. *Bioinformatics*, **28**, 1879–1886.

14. Sakabe,N.J., Savic,D. and Nobrega,M.A. (2012) Transcriptional enhancers in development and disease. *Genome Biol.*, **13**, 238.

15. Sanyal,A., Lajoie,B.R., Jain,G. and Dekker,J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.

16. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

17. Schaub,M.A., Boyle,A.P., Kundaje,A., Batzoglou,S. and Snyder,M. (2012) Linking disease associations with regulatory information in the human genome. *Genome Res.*, **22**, 1748–1759.

18. Hoffman,M.M., Ernst,J., Wilder,S.P., Kundaje,A., Harris,R.S., Libbrecht,M., Giardine,B., Ellenbogen,P.M., Bilmes,J.A., Birney,E. *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.*, **41**, 827–841.

19. Hardison,R.C. (2012) Genome-wide epigenetic data facilitate understanding of disease susceptibility association studies. *J. Biol. Chem.*, **287**, 30932–30940.

20. 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**, D930–D934.

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

22. Trynka,G., Sandor,C., Han,B., Xu,H., Stranger,B.E., Liu,X.S. and Raychaudhuri,S. (2013) Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.*, **45**, 124–130.

23. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.

24. Hoffman,M.M., Buske,O.J., Wang,J., Weng,Z.P., Bilmes,J.A. and Noble,W.S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.

25. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.

26. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.

27. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

28. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

29. Buchanan,C.C., Torstenson,E.S., Bush,W.S. and Ritchie,M.D. (2012) A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *J. Am. Med. Inform. Assoc.*, **19**, 289–294.

30. Li,Y., Willer,C.J., Ding,J., Scheet,P. and Abecasis,G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*, **34**, 816–834.

31. Cooper,G.M., Stone,E.A., Asimenos,G., Green,E.D., Batzoglou,S. and Sidow,A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.

32. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A., Bender,D., Maller,J., Sklar,P., de Bakker,P.I., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

33. Zhang,G.L., Chen,X.C., Chan,L., Zhang,M.X., Zhu,B.H., Wang,L.T., Zhu,X.Y., Zhang,J.Y., Zhou,B.P. and Wang,J.W. (2011) An SNP selection strategy identified IL-22 associating with susceptibility to tuberculosis in Chinese. *Sci. Rep.*, **1**, 20.

34. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.

35. Levy,S. and Hannenhalli,S. (2002) Identification of transcription factor binding sites in the human genome sequence. *Mamm. Genome*, **13**, 510–514.

36. Li,M.J., Sham,P.C. and Wang,J.W. (2010) FastPval: a fast and memory efficient program to calculate very low P-values from empirical distribution. *Bioinformatics*, **26**, 2897–2899.

37. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP plus. *PLoS Comput. Biol.*, **6**, e1001025.

38. Raychaudhuri,S. (2011) VIZ-GRAIL: visualizing functional connections across disease loci. *Bioinformatics*, **27**, 1589–1590.

39. Li,M.J., Wang,P., Liu,X., Lim,E.L., Wang,Z., Yeager,M., Wong,M.P., Sham,P.C., Chanock,S.J. and Wang,J. (2012) GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **40**, D1047–D1054.

40. Li,M.J., Sham,P.C. and Wang,J. (2012) Genetic variant representation, annotation and prioritization in the post-GWAS era. *Cell Res.*, **22**, 1505–1508.

41. Musunuru,K., Strong,A., Frank-Kamenetsky,M., Lee,N.E., Ahfeldt,T., Sachs,K.V., Li,X.Y., Li,H., Kuperwasser,N., Ruda,V.M. *et al.* (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, **466**, 714–719.

42. Teslovich,T.M., Musunuru,K., Smith,A.V., Edmondson,A.C., Stylianou,I.M., Koseki,M., Pirruccello,J.P., Ripatti,S., Chasman,D.I., Willer,C.J. *et al.* (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–713.

43. Aulchenko,Y.S., Ripatti,S., Lindqvist,I., Boomsma,D., Heid,I.M., Pramstaller,P.P., Penninx,B.W., Janssens,A.C., Wilson,J.F., Spector,T. *et al.* (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat. Genet.*, **41**, 47–55.

44. Pomerantz,M.M., Ahmadiyeh,N., Jia,L., Herman,P., Verzi,M.P., Doddapaneni,H., Beckwith,C.A., Chan,J.A., Hills,A., Davis,M. *et al.* (2009) The 8q24 cancer risk variant rs6983267 shows long-

range interaction with MYC in colorectal cancer. *Nat. Genet.*, **41**, 882–884.

45. Tuupanen,S., Turunen,M., Lehtonen,R., Hallikas,O., Vanharanta,S., Kivioja,T., Bjorklund,M., Wei,G., Yan,J., Niittymaki,I. *et al.* (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat. Genet.*, **41**, 885–890.

46. Griffith,O.L., Montgomery,S.B., Bernier,B., Chu,B., Kasaian,K., Aerts,S., Mahony,S., Sleumer,M.C., Bilenky,M., Haeussler,M. *et al.* (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.*, **36**, D107–D113.

47. Huff,J.T., Plocik,A.M., Guthrie,C. and Yamamoto,K.R. (2010) Reciprocal intronic and exonic histone modification regions in humans. *Nat. Struct. Mol. Biol.*, **17**, 1495–1499.

48. Ott,C.J., Blackledge,N.P., Kerschner,J.L., Leir,S.H., Crawford,G.E., Cotton,C.U. and Harris,A. (2009) Intronic enhancers coordinate epithelial-specific looping of the active CFTR locus. *Proc. Natl Acad. Sci. USA*, **106**, 19934–19939.

49. Visel,A., Minovitsky,S., Dubchak,I. and Pennacchio,L.A. (2007) VISTA Enhancer Browser-a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.

50. Wang,K., Li,M.Y. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, **38**, e164.

51. McLaren,W., Pritchard,B., Rios,D., Chen,Y.A., Flicek,P. and Cunningham,F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, **26**, 2069–2070.