Cell Research (2012) 22:1505-1508.

www.nature.com/cr

LETTER TO THE EDITOR

Genetic variant representation, annotation and prioritization in the post-GWAS era

Cell Research (2012) **22**:1505-1508. doi:10.1038/cr.2012.106; published online 17 July 2012

Dear Editor,

Since 2005 [1], genome-wide association studies (GWAS) and Next Generation Sequencing (NGS) have opened up new realms of investigation into the association between different diseases/traits and a large number of genetic loci. To date, more than 1 200 GWAS on over 250 traits have been published [2]. The advent of NGS and affordable whole genome and exome sequencing has accelerated the discovery of the genes underlying Mendelian diseases and has also enhanced our ability to detect rare variants. These variants may explain the missing heritability of common diseases and specific traits [3].

To discriminate the true trait/disease-associated SNPs (TASs) from the large number of loci discovered by GWAS and NGS studies, we need further downstream statistical and bioinformatics analyses [4]. Variant visualization, functional annotation and prioritization are vital for determining the true associations between genetic markers and diseases/traits, from the multiple signals many of which represent chance findings. In addition, an increasing number of requirements, such as clarity, diversity and interactivity, pose demands on data visualization [5, 6]. Thorough functional annotations based on genomic location and potential biological effects are needed, especially for markers with moderate effects located in regulatory regions (e.g., non-coding RNAs, enhancers and promoters, and those in evolutionarily conserved regions) [7-9], and for markers that have functional interactions with other TASs [10]. Comprehensive variant annotation will undoubtedly accelerate this process. Existing bioinformatics tools such as ANNOVAR [11] and VAAST [12] use genomic mapping to produce variant annotation, but few tools focus on the potential functional effects of TASs. Importantly, to properly distinguish the true association of TASs from the huge amount of GWAS signals, particularly for hidden TASs with moderate *P*-value, an annotation-based prioritizing process is required. Therefore, tools that can visualize, annotate and prioritize such data are urgently needed.

We have developed the GWASrap tool (http://jjwanglab.org/gwasrap) that systematically supports genetic variant representation, annotation and prioritization for data generated from GWAS and NGS (Figure 1 and Supplementary information, Data S1). Our web-based framework utilizes state-of-the-art web technologies to maximize user interaction and visualization of the results (Supplementary information, Figure S3). For a given SNP dataset with its *P*-values, GWASrap will first provide a Circos-style plot to visualize any genetic variants at either the genome or chromosome level (Supplementary information, Figures S5 and S6). The tool then combines different genomic features (SNP/CNV density, disease susceptibility loci, *etc*.) with comprehensive annotations that give the researcher an intuitive view of the functional significance of the different genomic regions (Supplementary information, Figure S4). The detailed statistics of the underlying study are also displayed on the web page, including variant distribution in different functional categories, classic Manhattan plot and QQ plot (Supplementary information, Figure S8). Users can perform interactive operations in the Manhattan panel, such as zooming in and out to search regions or markers of interest (Supplementary information, Figure S7). The system can also display a comprehensive range of relevant information from variant genetic attributes to nearby genomic elements, such as enhancers or non-coding RNAs. Furthermore, researchers can obtain extensive functional predictions for various features including transcription factor-binding sites, miRNA and miRNA target sites, and their predicted changes caused by the genetic variants (Supplementary information, Table S1). Our system can re-prioritize genetic variants by combining the original statistical value and variant prioritization score based on a simple additive effect equation (Supplementary information, Figures S1, S2 and Table S2). Researchers can also re-evaluate the significance of a TAS using the dynamic linkage disequilibrium (LD) panel (Supplementary information, Figure S8) or the tree-like network panel (Supplementary information, Figure S10). The GWASrap supports input variants in different formats, not only common variants with a dbSNP rs ID but also

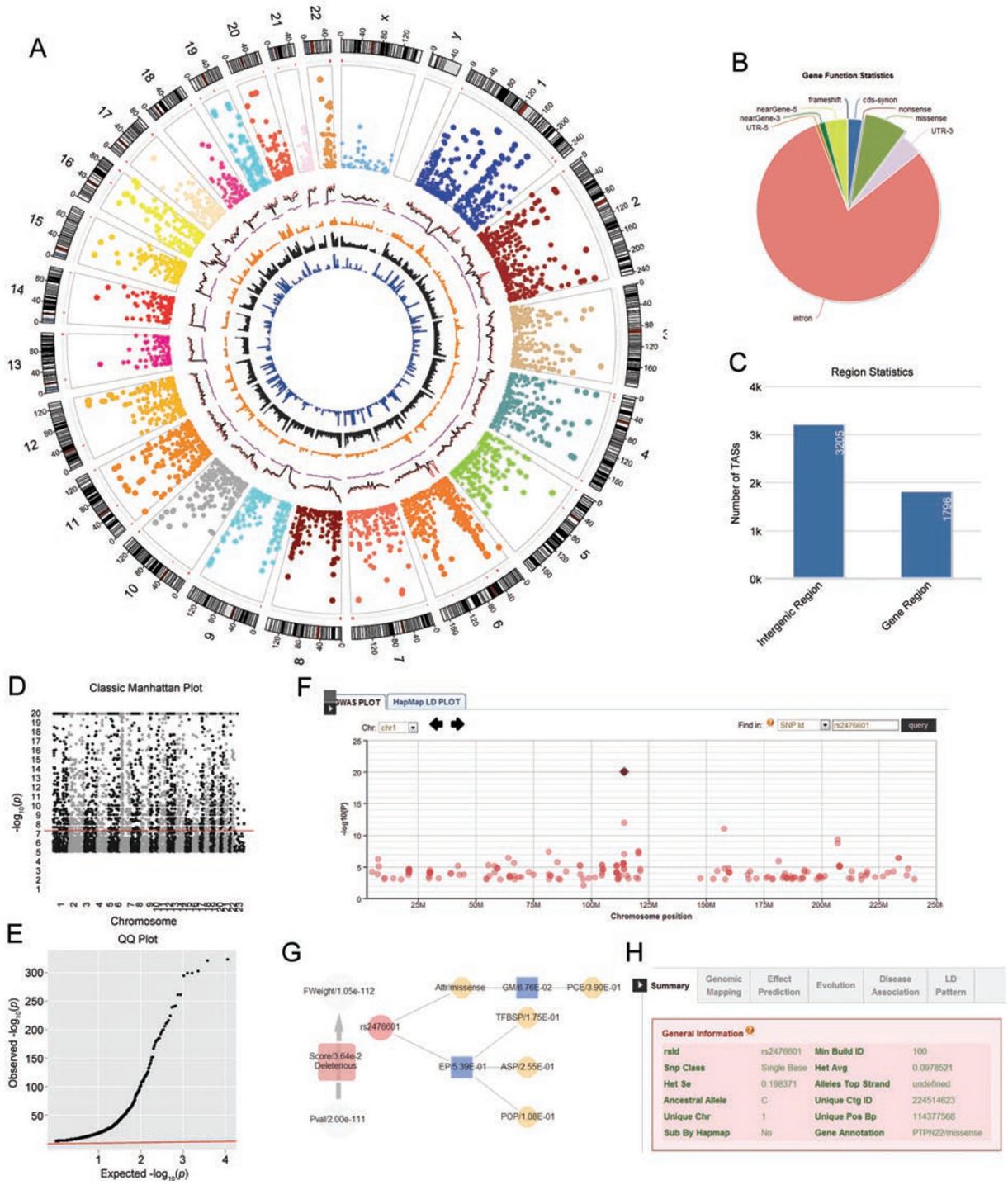

Figure 1 The functional components of GWASrap. **(A)** The GWAS representation of 7 340 significant TASs with *P*-value < 1.0×10^{-5} from NHGRI GWAS Catalog (up to February 2012). Displayed from the outer to the inner circle are the features or glyphs, the number of chromosome, the chromosome ideograms, copy number variation hotspots (red region), scatter plot for TASs with −logarithm (*P*-value), genome variant density (red: dbSNP, black: 1 000 genomes, purple: HapMap 3), OMIM gene distribution, copy number variation distribution and disease-susceptible region distribution. **(B)** The distribution of variants in different parts of the genes. **(C)** The distribution of variants in intergenic regions and gene regions. **(D)** Static Manhattan plot. **(E)** QQ plot. **(F)** An interactive Manhattan plot that can be queried and zoomed in and out. **(G)** A prioritization tree displays the functional annotations/predictions of a selected variant. Each node can be clicked to view detailed information of the representation. **(H)** Comprehensive annotation tabs for each item of the variant including variant summary, genomic mapping information, effective prediction information, evolution information, disease susceptibility information and LD ranking function.

rare variants from NGS data, which are represented by chromosome and locations.

We used a Circos-style [13] plot to represent all SNPs in the NHGRI GWAS Catalog (up to February 2012). Compared to the original graph and the traditional Manhattan plot, our plot provides a very broad horizontal area for sanity checking of the current status of GWAS data. For a given GWAS dataset for a specific disease/trait, researchers can easily locate the significant region by looking at the single chromosome plot or by interacting with the dynamic Manhattan panel. The surrounding features and glyphs contain sufficient genetic information to provide an intuitive overview of the GWAS results.

In the prioritization step, we first estimated the likelihood of disease associations for each type of SNP (missense, nonsence, synonymous, *etc*.) by mapping those data to the dataset of HapMap3. We then computed the variant's genomic mapping score and functional prediction score based on the variant's annotation information. The prioritization score was then computed from the product of above scores (Supplementary information, Data S1). We have tested our prioritization method on several different datasets and found it to be very reliable. We first applied our prioritization method to disease-causal SNPs from the OMIM database, and found that the prioritization scores were significantly different between disease-causal SNPs and randomly sampled background SNPs (Supplementary information, Figures S11-S13). Next, we re-scored the top 100 SNPs from the GWAS Catalog taking into account the synthetic associations of the LD proxy. The resulting scores were closer to that of the benchmarked 100 OMIM SNPs compared with the original scores and random background (Supplementary information, Figure S14). Finally, by applying the prioritization method to a bipolar disorder (BPD) GWAS study, we successfully selected the variants with the strongest effect that were confirmed in a separate study (Supplementary information, Tables S3 and S4). For example, SNP rs1042779 in *ITIH1* gene, with many deleterious attributes, obtained a progressively stronger signal after the prioritization step. The variant is highly associated with a non-synonymous variant rs11177 in *GNL3* gene and a synonymous variant rs2251219 in *PBRM1* gene, which had a very significant signal in a recent BPD GWAS (Supplementary information, Figure S15). In addition, an intronic variant rs420259, which was indicated as a high-risk marker in another GWAS, scored considerably higher after applying our prioritization method.

We have built a local database designed around a user-friendly web interface and web services to provide a rapid diagnostic tool for genetic variants. Our system accepts a submission query as either a dbSNP ID or a chromosomal location, and will quickly return the annotation information displayed on an interactive LD panel (Supplementary information, Figure S9). In the case of a rare variant or *de novo* variant without a dbSNP ID, the user can use genomic coordinates to obtain sufficient annotation. Our database also hosts well-structured and up-to-date repositories from all significant TASs found in popular diseases frequently investigated by GWAS. For each specific disease/trait, we have integrated significant SNPs for the disease/trait, which were annotated in the GWASdb database [14].

In summary, our system offers a universal web portal for GWAS/NGS representation, annotation and prioritization. The system will benefit users for data visualization and facilitate the functional annotation of genetic variants discovered by GWAS and NGS studies.

Acknowledgments

This study was supported by grants from the Research Grants Council (781511M, 778609M, N_HKU752/10, AoE M-04/04), Food and Health Bureau (10091262) of Hong Kong, and The University of Hong Kong Strategic Research Theme on Genomics.

Mulin Jun Li[1, 2], Pak Chung Sham[3, 4, 5], Junwen Wang[1, 2, 3]

[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]Centre for Genomic Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China; [4]Department of Psychiatry, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China; [5]The State Key Laboratory in Brain and Cognitive Sciences, The University of Hong Kong, Hong Kong SAR, China

Correspondence: Junwen Wang
Tel: +852 2831 5075; Fax: +852 2855 1254
E-mail: junwen@hku.hk

References

1 Klein RJ, Zeiss C, Chew EY, *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; **308**:385-389.
2 Hindorff LA, Sethupathy P, Junkins HA, *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009; **106**:9362-9367.
3 Bamshad MJ, Ng SB, Bigham AW, *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011; **12**:745-755.
4 Wang K, Li M, Hakonarson H. Analysing biological pathways

in genome-wide association studies. *Nat Rev Genet* 2010; **11**:843-854.
5 Robinson JT, Thorvaldsdottir H, Winckler W, *et al.* Integrative genomics viewer. *Nat Biotechnol* 2011; **29**:24-26.
6 Pruim RJ, Welch RP, Sanna S, *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010; **26**:2336-2337.
7 Wang J. A database of genetic variants in microRNA genes and their putative functional roles in gene regulation. *Hum Mutat* 2012; **33**:vii-vii.
8 Wang W, Wei Z, Lam TW, Wang JW. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Sci Rep* 2011; **1**:55.
9 Wei Z, Jensen ST. GAME: detecting cis-regulatory elements using a genetic algorithm. *Bioinformatics* 2006; **22**:1577-1584.
10 Hu X, Liu Q, Zhang Z, *et al.* SHEsisEpi, a GPU-enhanced genome-wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder. *Cell Res* 2010; **20**:854-857.
11 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010; **38**:e164
12 Yandell M, Huff C, Hu H, *et al.* A probabilistic disease-gene finder for personal genomes. *Genome Res* 2011; **21**:1529-1542.
13 Krzywinski M, Schein J, Birol I, *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* 2009; **19**:1639-1645.
14 Li MJ, Wang P, Liu X, *et al.* GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res* 2012; **40** (Database issue):D1047-D1054.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)