frontiers in
GENETICS

REVIEW ARTICLE
published: 09 December 2013
doi: 10.3389/fgene.2013.00280

A review of post-GWAS prioritization approaches

Lin Hou and Hongyu Zhao*

Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA

Edited by:
Shuang Wang, Columbia University, USA

Reviewed by:
Yun Li, University of North Carolina, USA
Jinming Li, Southern Medical University, China

****Correspondence:***
Hongyu Zhao, Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06510, USA
e-mail: hongyu.zhao@yale.edu

In the recent decade, high-throughput genotyping and next-generation sequencing platforms have enabled genome-wide association studies (GWAS) of many complex human diseases. These studies have discovered many disease susceptible loci, and unveiled unexpected disease mechanisms. Despite these successes, these identified variants only explain a small proportion of the genetic contributions to these diseases and many more remain to be found. This is largely due to the small effect sizes of most disease-associated variants and limited sample size. As a result, it is critical to leverage other information to more effectively prioritize GWAS signals to increase replication rates and better understand disease mechanisms. In this review, we introduce the biological/genomic features that have been found to be informative for post-GWAS prioritization, and discuss available tools to utilize these features for prioritization

Keywords: genome-wide association studies, prioritization, eQTL, DNase I hypersensitive site, non-coding

INTRODUCTION

With the developments of affordable and reliable high-throughput genotyping and next-generation sequencing platforms, many genome-wide association studies (GWAS) have been successfully conducted to identify DNA variants associated with many complex human diseases and traits, such as cancer, autoimmune diseases, height, blood pressure, body mass index, among others. As of 11/16/13, there were 11,907 single nucleotide polymorphisms (SNPs), 940 traits with 15,052 associations documented in the GWAS catalog, maintained by the National Human Genome Research Institute (Hindorff et al., 2009). These studies have uncovered many novel genes and implicated unexpected pathways associated with disease mechanisms, leading to great insights on disease etiology.

In spite of these accomplishments, many challenges remain in GWAS design and analysis. The first limitation is the limited statistical power to identify all disease-associated loci. Although many susceptible loci have been identified, they only explain a small fraction of the overall heritability, with the majority of heritability remaining unexplained. One possible reason is that the missing heritability is due to the lack of coverage of genetic variations on the genotyping platforms, such as those rare or even private variations. Another explanation is that most disease-associated variants have small effect sizes that are not likely detected due to low statistical power, even with thousands of subjects. To better identify these variants, more powerful and cost-effective designs and statistical methods are desired. Several approaches have proved cost-effective to enrich signals and increase statistical power. For example, a number of customized genotyping platforms have been used to target certain genomic regions with high density to fine map disease-associated variants. The successful examples include the use of the ImmunoChip (Trynka et al., 2011) to fine map 186 distinct loci associated with 12 autoimmune diseases, and the use of the MetaboChip (Voight et al., 2012) to fine map established trait-associated loci. As for rare variant analysis, a number of whole exome sequencing studies have enjoyed success for diseases like autism and schizophrenia (Xu et al., 2011; Sanders et al., 2012). It has also been found that studies focusing individuals with extreme phenotypes can increase statistical power because of the enriched genetic signals in the study subjects (Lin et al., 2013). In addition to improved platforms and study designs, many statistical methods have also been developed to increase statistical power. Meta-analysis is commonly applied to leverage all the information from separate studies to increase statistical power to identify disease-associated loci. A number of methods aim to investigate the combinatorial effects of a group of SNPs, including both marginal and interaction effects. These are accomplished by explicitly modeling the interactions of two or more SNPs; joint analysis of all the SNPs in a gene or defined region; and joint consideration of the SNPs in proximity of all the genes annotated in a specific pathway/network. The advances in study designs and statistical methods have led to many novel discoveries. For example, a recent study of the inflammatory bowel disease (IBD; Jostins et al., 2012), which is a meta-analysis of both ImmunoChip and GWAS datasets, increased the number of IBD loci to 163, where these loci account for 13.6% of the genetic risk of Crohn's disease (CD) and 7.5% of ulcerative colitis. See Cantor et al. (2010) for a comprehensive review of some of the topics.

The second limitation is the difficulty to interpret the biological relevance of susceptible loci and link them with the disease etiology. Because the ultimate goal of association studies is to understand disease etiology and develop effective strategies to prevent and treat diseases, it would be desirable that GWAS results can implicate functional variants and disease pathways that can be experimentally studied in follow-up functional studies. However, a large proportion of disease-associated variants fall into non-coding regions of the genome, with 88% of the associated SNPs in GWAS catalog non-coding (Hindorff et al., 2009), making it difficult to form testable hypothesis. Even when the variants in the coding region, it is often not clear whether they are functional due to the presence of several closely linked variants. To address this problem, many statistical methods have been proposed to

prioritize GWAS signals by incorporating diverse functional evidence, so that variants with small effect sizes but possessing functional features may be prioritized over variants with similar effect sizes but less likely to be functional. GWAS signals can be prioritized at both the SNP level and gene level, depending on the biological features considered and the input signals. Statistical approaches that prioritize at the SNP level are especially helpful in pinpointing the causal variants with sequence data, where essentially all the variations in the genome can be identified. This is contrast to earlier GWAS that only interrogated a subset of SNPs, such as tag SNPs, in genotyping platforms. One benefit of such approaches is that the functional evidence provides paths to derive plausible and testable hypotheses for the prioritized genes or loci. Moreover, with the incorporation of other data informative about disease association, the prioritized genes/loci are more likely to be truly associated with disease. For example, it has been observed that trait-associated loci are more concentrated in regions with certain genomic features, such as protein coding regions and expression quantitative trait loci (eQTL). In this review, we will review (1) biological/genomic features that are informative for prioritizing GWAS results; and (2) computational methods and tools that prioritize disease-associated SNPs by integrating these biological/genomic features.

BIOLOGICAL FEATURES USED IN PRIORITIZATION AND THEIR JUSTIFICATIONS

The first step, and sometimes the only step of many SNP prioritization approaches is to annotate the candidate SNPs by intersecting GWAS signals with desired genomic features, such as eQTLs, transcription factor binding sites, DNase hypersensitive sites, histone modifications, and others. For CD, Fransen et al. (2010) showed that *cis*-eQTL SNPs were enriched in known CD-associated SNPs. Based on this observation, the authors proposed to select a subset of SNPs to follow-up a public CD GWAS dataset, to intersect the top 500 GWAS hits with *cis*-eQTLs in an eQTL database. The SNPs thus selected, *cis*-eQTLs of the genes *UBE2L3* and *BCL3*, were replicable in two independent replication cohorts. This represents a successful application of the annotation-based prioritization methods.

The genomic features may implicate functional roles of the prioritized SNPs to disease etiology, and these hypotheses can be formally tested through molecular studies. These include variants both in coding and non-coding regions. Through these filtering/intersecting methods, researchers can focus on a much smaller number of SNPs in follow-up studies. Although the proximity of a SNP with a documented genomic feature may suggest a functional role of the SNP, it may not necessarily increase the probability that it will affect the phenotype of interest, nor the probability that the locus is truly susceptible. In general, the genomic features discussed above are considered biologically plausible and extensively used to prioritize SNPs, but whether each feature is informative on a SNP's functional relevance is disease and context dependent. In a recent study, Minelli et al. (2013) tried to identify features that are important in selecting SNPs for follow-up studies by surveys in experts. They sent questionnaires to ten experts who conducted GWAS studies, and asked their opinion on the importance of a set of selected features. The features included relative position of the SNP to a nearby transcript, whether the SNP causes an amino acid change, etc. (see Table 2 in their paper). The result was not surprising, as experts considered gene level evidence more important, such as the SNP in a gene that is previously associated with the phenotype, or that encodes a protein in a phenotype related pathway, or that has gene/protein interaction relevant to the phenotype. Experts opinions are valuable, however, they might be biased toward existing knowledge and also expertise in specific diseases. Nevertheless, this paper highlighted the need for understanding what features should be considered in prioritization. In the following, we review these features and statistical methods to use these features, to inform human geneticists in their applications of the annotation-based approaches.

EXPRESSION QUANTITATIVE TRAIT LOCI

By contrasting the SNPs documented in the GWAS catalog (Hindorff et al., 2009) with those randomly sampled SNPs with matching minor allele frequency distribution, Nicolae et al. (2010) showed that complex trait-associated SNPs are more likely to be eQTLs. The conclusion remained valid for a linkage disequilibrium (LD)-pruned subset of SNPs in the GWAS catalog. Since the eQTL annotation considered by Nicolae et al. (2010) was derived from an expression dataset of lymphoblastoid cell lines, it was of interest to investigate whether cell line-specific eQTLs showed different levels of enrichments across diseases with different focal tissues, including cancer, neurological/psychiatric disorders, and autoimmune disorders. By tissue of relevance, the lymphoblastoid cell lines should be a good proxy for autoimmune disorders, and relatively poor for cancer and neurological/psychiatric disorders. As expected, there was greater enrichment of eQTLs in the group of autoimmune disorders, while only moderate enrichment for the other two groups of diseases. Furthermore, in the examination of the results in the Wellcome Trust Case Control Consortium (WTCCC) GWAS dataset of CD, eQTLs were enriched in SNPs with association p-value less than 0.01, but not for the missense SNPs, indicating potential loss of information if non-coding SNPs are ignored.

PROTEIN DELETERIOUSNESS PREDICTIONS

Polymorphisms in the coding region may have different effects on protein function. Synonymous SNPs do not change the corresponding protein sequence; non-synonymous SNPs change amino acid composition, or truncate the protein sequence by causing an early codon; indels can change protein sequence with varying consequence depending on whether the indel is in-frame or frame-shifting; SNPs and indels can also interrupt splicing sites, thus change the mRNA isoform. In other words, mutations in the coding region may be benign or deleterious to protein function, with deleterious mutations more likely to have a phenotypic effect. Many computational tools have been developed to predict "deleteriousness" of SNPs and indels (Ng and Henikoff, 2003; Adzhubei et al., 2010). These methods generally take features like biochemical property of the altered amino acid, conservation and sequence homology as input, and use machine learning technique to train a classifier. These methods have been comprehensively

reviewed by Cooper and Shendure (2011) and Ng and Henikoff (2006).

The most extreme case of protein function interruption is the loss of function mutations. However, genome sequencing studies found that all human carry loss of function mutations without obvious phenotypic effect, and common loss of function variants were depleted in polymorphisms associated with complex disease like CD and rheumatoid arthritis (MacArthur et al., 2012). The results indicate that the "deleteriousness" feature should be interpreted with caution, since disruption of protein function does not necessarily have a phenotypic effect. In this regard, the "residual variance intolerance score" has been defined quantitatively measure the tolerance of a protein to mutations (Petrovski et al., 2013). The number of missense and non-sense variants found in each gene in the cohort of the National Heart, Lung, and Blood Institute (NHLBI) exome sequencing project was compared to the number of functionally neutral variants, and genes with variants fewer than expected are assigned a negative score, indicating they are less tolerant to variations.

DIFFERENTIAL GENE EXPRESSIONS

Gene expression microarrays and RNA-seq are commonly used to study gene expression profiles in disease cases and matched controls, and differentially expressed genes thus identified may suggest disease mechanisms and potential biomarkers that can be further explored in follow-up studies. Chen et al. (2008) analyzed 476 expression datasets in the Gene Expression Omnibus (GEO), and calculated the frequency that a gene was differentially expressed in these datasets, which they called "differential expression ratio." They found that differential expression ratio is positively correlated with the likelihood that a gene harbors disease-associated variants, where the list of disease-associated genes was created by combining information from Genetic Association Database (GAD; Becker et al., 2004) and Human Gene Mutation Database (HGMD; Stenson et al., 2003). In addition, they found that among the genes discovered in the initial scan of the WTCCC type 1 diabetes mellitus GWAS dataset, the differential expression ratio was higher in genes that were replicable than those not replicable in follow-up studies. These authors have developed an online server, FitSNPs, to incorporate this feature (see **Table 1**).

DNase I HYPERSENSITIVE SITES

DNase I hypersensitive sites (DHSs) are markers of accessible chromatin, which indicate regulatory roles in the transcription process. DHS have been mapped in 349 cell and tissue samples genome-wide by next-generation sequencing (Thurman et al., 2012). Enrichment analysis showed that trait-associated SNPs in the GWAS catalog (Hindorff et al., 2009) are more concentrated within DHS regions, excluding confounding factors such as allele frequency and distance from the nearest transcriptional start site (Maurano et al., 2012).

Table 1 | A list of online SNP prioritization tools.

Name	Website	Reference
FASTSNP	http://fastsnp.ibms.sinica.edu.tw	Yuan et al. (2006)
FitSNPs	http://fitsnps.stanford.edu/fitSNPs.php	Chen et al. (2008)
SNPranker 2.0	http://www.itb.cnr.it/snpranker	Merelli et al. (2013)
SPOT	http://spot.cgsmd.isi.edu	Saccone et al. (2010)

OTHERS

There are many more genomic features collected and annotated in large community projects, such as the Encyclopedia of DNA Elements (ENCODE; Consortium, 2011), which are potentially valuable for SNP prioritization. Kindt et al. (2013) examined enrichment or depletion of trait-associated SNPs in 58 genomic features. The features investigated covered genic and regulatory features, conservation features, and chromatin state features (see Table 1 in Kindt et al., 2013). Among those features, genomic regions annotated as "heterochromatin" and "low expression signals" are depleted of trait-associated SNPs, while eQTLs and "strong enhancer" showed the highest level of enrichment.

The biological features discussed so far are measured/inferred from laboratory cell lines and the sequence and annotation of the human genome, which do not provide trait-specific information. However, trait-relevant features are intrinsically helpful for prioritization. For example, a DNase-seq experiment in intestine tissues and immune cells of CD patients would be more informative for prioritizing variants associated with CD than those measured in brain tissues. Maunakea et al. (2010) and Portela and Esteller (2010) reviewed recent progress in mapping the epigenome (including DNA methylation and histone modification), showing that epigenetic modifications play important roles in human diseases, including cancer, neurodevelopmental disorder, neurodegenerative disease, neurological disease, and autoimmune diseases. Thus, epigenome data in disease states is valuable for understanding disease and prioritize disease susceptible loci. However, efforts in disease-specific epigenome mapping and systematic database to document such data are still lacking and greatly needed. For DNA methylation alone, a database, DiseaseMeth, has incorporated methylation data for 72 human diseases (Lv et al., 2012).

SNP PRIORITIZATION APPROACHES AND AVAILABLE WEB SERVERS

Here we review methods and tools that prioritize GWAS signals at single SNP level. There are mainly two steps in these methods. The first step applies annotations or filters based on whether or not the candidate SNP has the desirable features and the second step scores the candidate SNPs by integrating evidence from multiple sources. Saccone et al. (2008, 2010) developed an online prioritization tool, SPOT, which systematically combines multiple biological databases to prioritize SNPs by the genomic information network (GIN) model (see **Table 1**). In their model, each SNP is assigned a prioritization score, which is a linear sum of scores derived from pathway information, comparative genomics, linkage scan, and results of other independent GWAS studies. The weights are decided by the strength of the link between the SNP and the annotations. For example, for a SNP that is in LD with a non-synonymous SNP of a susceptible gene, the assigned weight will be less than that of SNPs physically in the gene. The methodology prioritized SNPs with increased biological relevance

in a GWAS study of nicotine dependence. Thompson et al. (2013) incorporated biological features in a Bayesian framework, where the prior probability that a SNP is associated with the phenotype is determined by its annotations. They first curated a training set, including SNPs that were confirmed replicable as the positive set, and 1,000 randomly selected SNPs as control set. For a selected set of features, a logistic regression model was fit for each disease. Thus, the log odds ratio that a test SNP is associated with the disease can be estimated through the model.

There are also web servers that perform SNP prioritization in an annotation fashion. They annotate the candidate SNPs by single or multiple features, but do not combine the results. They differ by the features and strategies they use in prioritization. A list of SNP prioritization resources are provided in **Table 1**. Most of these tools are only applicable to SNPs, and tools that can prioritize indels are still lacking.

FASTSNP uses a decision tree framework to assign different risk level to SNPs by considering the genomic location and functional effect of the SNPs (Yuan et al., 2006).

FitSNPs calculated a differential expression ratio for all genes in the genome, and prioritize SNPs by the differential expression ratio of their associated genes (Chen et al., 2008).

SNPranker 2.0 first annotates the SNPs with different features, and then user a user interactive way to integrate features (Chen et al., 2008). Users can specify the features they want to include and the weight of each feature, which would give the users an opportunity to enforce their biological priors. But they also provide an optimal set of weights by default. The optimal weights were determined by a cross-validation approach.

TOOLS FOR VARIANT ANNOTATION

Besides the SNP prioritization tools, there are also many web servers and software for variant annotation (**Table 2**), which could provide useful information for prioritization. Basically, these tools take a list of query variants as input, and annotate them with their in-house databases. Among these, HaploReg (Ward and Kellis, 2012a) and RegulomeDB (Boyle et al., 2012) provide annotation for variations in non-coding regions. HaploReg annotates variations by their chromatin state, conservation across mammals, and computationally predicted transcription factor binding sites. Besides, by utilizing LD information from the 1000 Genomes Project, HaploReg automatically reports, and annotates all variations within a user-specified LD threshold of the query variant. RegulomeDB has incorporated many data sources, including the ENCODE project, available transcription factor ChIP-seq data, and eQTL datasets. The other tools are designed for variations in the whole genome. They annotate the query variations by dbSNP ID, allele frequency in different ethnic groups, position in a transcript (intron, exon, 5′ UTR, etc.,), and the resultant amino acid change if any. SeattleSeq (Ng et al., 2009) and Variant Effect Predictor (VEP; McLaren et al., 2010) has convenient web interface, suitable for users who are not familiar with scripts and programming languages, while ANNOVAR (Wang et al., 2010) and Snpeff (Cingolani et al., 2012) are stand-alone software packages, so that they can be easily incorporated into variant analysis pipelines. Discussions on variant annotation tools can also be found in Ward and Kellis (2012b).

DISCUSSION

In this review, we have focused on biological and genomic features that are informative for SNP prioritization. In the second phase of association studies, researchers can use these databases or tools to choose SNPs in follow-up studies. The observation that eQTLs and open chromatin regions are enriched of trait-associated SNPs highlights the potential rich information in the non-coding regions of the genome (Nicolae et al., 2010). Nonetheless, the gene-centric approaches may be helpful in disease gene discovery, and there are many approaches that perform prioritization at the gene level. A review of the methods and tools for gene-based prioritization can be found in Tranchevent et al. (2011).

The prioritization methods discussed here take as input a list of candidate SNPs, which is usually derived by taking all the SNPs achieving a specified significance level in GWAS, e.g., all SNP with p-values less than 0.01. The SNPs are treated equally regardless of the association p-values. However, the p-value, the effect size, and other statistics that summarize the association level of individual SNP could be informative for SNP selection. A computational framework that incorporates the significance level with the biological/genomic features discussed above might improve the performance of the prioritization scheme. A discussion of different signal measures of association was given by Strömberg et al. (2009).

Although there are many web servers and databases for SNP prioritization, most of them provide annotations of different types of features, but do not rank these SNPs through integrating GWAS and annotation information. Also, these methods do not employ disease-specific information. There is still a great need for statistical methods that select and integrate multiple annotations in

Table 2 | A list of tools for variant annotation.

Name	Website	Reference
ANNOVAR	http://www.openbioinformatics.org/annovar/	Wang et al. (2010)
HaploReg	http://www.broadinstitute.org/mammals/haploreg/haploreg.php	Ward and Kellis (2012a)
RegulomeDB	http://regulome.stanford.edu/	Boyle et al. (2012)
SeattleSeq	http://snp.gs.washington.edu/SeattleSeqAnnotation137/	Ng et al. (2009)
Snpeff	http://snpeff.sourceforge.net	Cingolani et al. (2012)
VEP	http://useast.ensembl.org/info/docs/variation/vep/index.html	McLaren et al. (2010)

a disease-specific manner, and re-rank SNPs under a coherent statistical framework.

ACKNOWLEDGMENTS

Supported in part by the NIH grants R01 GM59507, the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development, and the Clinical and Translational Science Award UL1 RR024139 from the National Center for Research Resources, National Institutes of Health.

AUTHOR CONTRIBUTIONS

Lin Hou and Hongyu Zhao conceived and wrote the paper.

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248

Becker, K. G., Barnes, K. C., Bright, T. J., and Wang, S. A. (2004). The Genetic Association Database. *Nat. Genet.* 36, 431–432. doi: 10.1038/ng0504-431

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1797. doi: 10.1101/gr.137323.112

Cantor, R. M., Lange, K., and Sinsheimer, J. S. (2010). Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* 86, 6–22. doi: 10.1016/j.ajhg.2009.11.017

Chen, R., Morgan, A., Dudley, J., Deshpande, T., Li, L., Kodama, K., et al. (2008). FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease. *Genome Biol.* 9, R170. doi: 10.1186/gb-2008-9-12-r170

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6, 80–92. doi: 10.4161/fly.19695

Consortium, T. E. P. (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol.* 9:e1001046. doi: 10.1371/journal.pbio.1001046

Cooper, G. M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* 12, 628–640. doi: 10.1038/nrg3046

Fransen, K., Visschedijk, M. C., Van Sommeren, S., Fu, J. Y., Franke, L., Festen, E. A. M., et al. (2010). Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. *Hum. Mol. Genet.* 19, 3482–3488. doi: 10.1093/hmg/ddq264

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9362–9367. doi: 10.1073/pnas.0903103106

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., Mcgovern, D. P., Hui, K. Y., et al. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491, 119–124. doi: 10.1038/nature11582

Kindt, A. S. D., Navarro, P., Semple, C. A. M., and Haley, C. (2013). The genomic signature of trait-associated variants. *BMC Genomics* 14:108. doi: 10.1186/1471-2164-14-108

Lin, D.-Y., Zeng, D., and Tang, Z.-Z. (2013). Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proc. Natl. Acad. Sci. U.S.A.* 110, 12247–12252. doi: 10.1073/pnas.1221713110

Lv, J., Liu, H., Su, J., Wu, X., Liu, H., Li, B., et al. (2012). DiseaseMeth: a human disease methylation database. *Nucleic Acids Res.* 40, D1030–D1035. doi: 10.1093/nar/gkr1169

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335, 823–828. doi: 10.1126/science.1215040

Maunakea, A. K., Chepelev, I., and Zhao, K. (2010). Epigenome mapping in normal and disease states. *Circ. Res.* 107, 327–339. doi: 10.1161/CIRCRESAHA.110.222463

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195. doi: 10.1126/science.1222794

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26, 2069–2070. doi: 10.1093/bioinformatics/btq330

Merelli, I., Calabria, A., Cozzi, P., Viti, F., Mosca, E., and Milanesi, L. (2013). SNPranker 2.0: a gene-centric data mining tool for diseases associated SNP prioritization in GWAS. *BMC Bioinformatics* 14:S9. doi: 10.1186/1471-2105-14-S1-S9

Minelli, C., De Grandi, A., Weichenberger, C. X., Gögele, M., Modenese, M., Attia, J., et al. (2013). Importance of different types of prior knowledge in selecting genome-wide findings for follow-up. *Genet. Epidemiol.* 37, 205–213. doi: 10.1002/gepi.21705

Ng, P. C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509

Ng, P. C., and Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* 7, 61–80. doi: 10.1146/annurev.genom.7.080505.115630

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461, 272–276. doi: 10.1038/nature08250

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6:e1000888. doi: 10.1371/journal.pgen.1000888

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., and Goldstein, D. B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* 9:e1003709. doi: 10.1371/journal.pgen.1003709

Portela, A., and Esteller, M. (2010). Epigenetic modifications and human disease. *Nat. Biotech.* 28, 1057–1068. doi: 10.1038/nbt.1685

Saccone, S. F., Bolze, R., Thomas, P., Quan, J., Mehta, G., Deelman, E., et al. (2010). SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. *Nucleic Acids Res.* 38, W201–W209. doi: 10.1093/nar/gkq513

Saccone, S. F., Saccone, N. L., Swan, G. E., Madden, P. A. F., Goate, A. M., Rice, J. P., et al. (2008). Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence. *Bioinformatics* 24, 1805–1811. doi: 10.1093/bioinformatics/btn315

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485, 237–241. doi: 10.1038/nature10945

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S., et al. (2003). Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* 21, 577–581. doi: 10.1002/humu.10212

Strömberg, U., Björk, J., Vineis, P., Broberg, K., and Zeggini, E. (2009). Ranking of genome-wide association scan signals by different measures. *Int. J. Epidemiol.* 38, 1364–1373. doi: 10.1093/ije/dyp285

Thompson, J. R., Gögele, M., Weichenberger, C. X., Modenese, M., Attia, J., Barrett, J. H., et al. (2013). SNP prioritization using a bayesian probability of association. *Genet. Epidemiol.* 37, 214–221. doi: 10.1002/gepi.21704

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., et al. (2012). The accessible chromatin landscape of the human genome. *Nature* 489, 75–82. doi: 10.1038/nature11232

Tranchevent, L.-C., Capdevila, F. B., Nitsch, D., De Moor, B., De Causmaecker, P., and Moreau, Y. (2011). A guide to web tools to prioritize candidate genes. *Brief. Bioinform.* 12, 22–32. doi: 10.1093/bib/bbq007

Trynka, G., Hunt, K. A., Bockett, N. A., Romanos, J., Mistry, V., Szperl, A., et al. (2011). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* 43, 1193–1201. doi: 10.1038/ng.998

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* 8:e1002793. doi: 10.1371/journal.pgen.1002793

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164. doi: 10.1093/nar/gkq603

Ward, L. D., and Kellis, M. (2012a). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40, D930–D934. doi: 10.1093/nar/gkr917

Ward, L. D., and Kellis, M. (2012b). Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotech.* 30, 1095–1106. doi: 10.1038/nbt.2422

Xu, B., Roos, J. L., Dexheimer, P., Boone, B., Plummer, B., Levy, S., et al. (2011). Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat. Genet.* 43, 864–868. doi: 10.1038/ng.902

Yuan, H.-Y., Chiou, J.-J., Tseng, W.-H., Liu, C.-H., Liu, C.-K., Lin, Y.-J., et al. (2006). FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Res.* 34, W635–W641. doi: 10.1093/nar/gkl236

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 07 October 2013; paper pending published: 04 November 2013; accepted: 23 November 2013; published online: 09 December 2013.

*Citation: Hou L and Zhao H (2013) A review of post-GWAS prioritization approaches. Front. Genet. **4**:280. doi: 10.3389/fgene.2013.00280*

This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics.

Copyright © 2013 Hou and Zhao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.