GWASdb2.0: a database for human genetic variants identified by genome-wide association studies

Mulin Jun Li[1,†], Zipeng Liu[1,2,†], Panwen Wang[1], Maria P Wong[3], Meredith Yeager[4], Pak Chung Sham[5,6], Stephen J Chanock[4], Zhengyuan Xia[2], Junwen Wang[1,6,*]
Departments of [1]Biochemistry, [2]Anaesthesiology, [3]Pathology, [5]Psychiatry, [6]Centre for Genomic Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.
[4]Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD, USA.
[*]Contact: Junwen Wang (Tel: +852 2831 5075; Fax: +852 2855 1254; Email: junwen@hku.hk)
[†]Both authors contributed equally to this work

INTRODUCTION

Genome-wide association study (GWAS) have produced large numbers of human genetic variants (GVs) associated with hundreds of medical traits and common diseases. Although databases such as NHGRI GWAS Catalog have attempted to collect significant trait/disease associated SNPs (TASs), comprehensive curation and function annotation of GVs, especially for those in the noncoding regulatory regions, are still lacking. Moreover, the inconsistent terminology of trait/diseases and populations among different GWASs prevents further comparison and integrative analysis of GWAS results. To address these issues we introduce a batch of new features in our newly update version of GWASdb[1]. **http://jjwanglab.org/gwasdb**

METHODS & MATERIALS

Data curation and collection

We manually selected TASs from full text and supplementary materials of published GWAS sources (Table 1) by using a moderate P-value of less than 1E-3.

Table 1. Data source

GWAS Source	GWAS catalog, HuGE, GRASP, PheGenI, GWASdb (curated by ourselves)
Collected Data	SNP ID, PubMed ID, P-value, Odds Ratio/beta, CI95, population, sample size, trait/disease, risk allele (and frequency), etc.

We grouped different populations into 8 ethnogeographic categories. (Table 2)

Table 2. Categories of 8 super populations

AFR - African	EUR - European/Caucasian
ASN - East Asian	HIS - Hispanic/Latino
SAN - South Asian	MEA - Middle Eastern
OCN – Oceania	AMR - Native American

Ontology mapping

We mapped various trait/disease descriptions to several well-defined ontology systems, including Disease Ontology (DO), Human Phenotype Ontology (HPO), Disease Ontology Lite (DOLite).

Variant annotation

We utilized over 40 different dataset and prediction tools to annotate all the TASs. (Table 3)

Table 3. Annotation items of GWASdb2.0

Summary information	• dbSNP • 1000 Genomes • HapMap project
Knowledge-based annotation	• GTEx, eQTL • Human Enhancer, Insulator • ENCODE functional elements, etc
Gene-based annotation	• Small RNA, Lnc RNAs • Ensemble Gene • RefGene, etc
Functional prediction annotation	• Transcription factor/miRNA-target binding affinity • Splicing site affection, phosphorylation effect • Synonymous/non-synonymous SNP, etc
Evolutionary annotation	• Conservative constraint • Positive selection • GERP++ elements, etc
Disease association	• OMIM, COSMIC • NCBI ClinVar • GAD, DGV, etc
External annotation	• dbPSHP, rSNPBase, • UCSC Genome Browser • Regulomedb, DMDM, etc

RESULTS

Database Statistic

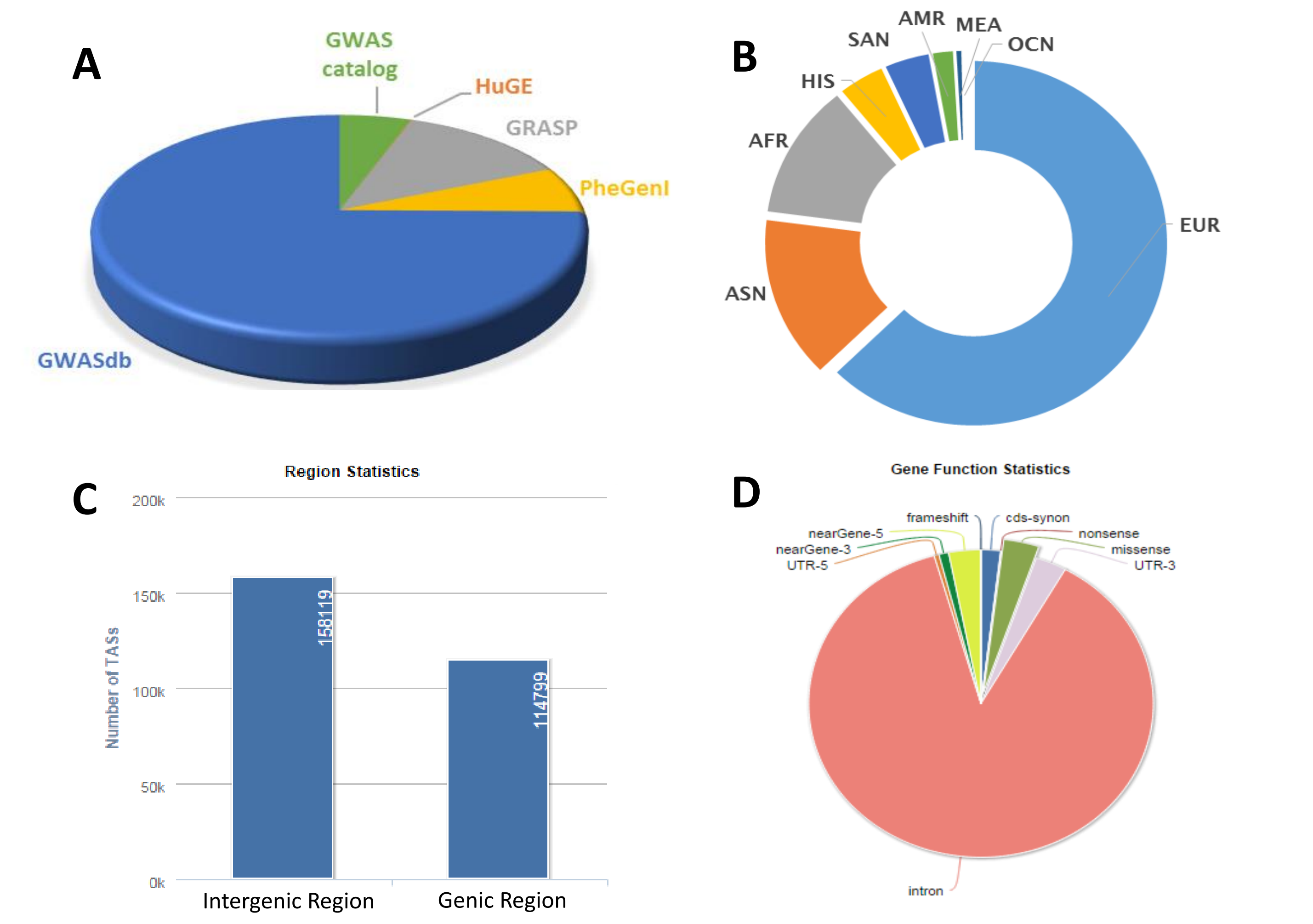


Figure 1. Update in Aug, 2014. (A) Composition of GWASdb2.0 by data source; (B) Data distribution by super populations; (C) TASs distribution in human genomic region. More than half of TASs are located in the intergenic region and (D) even for TASs in gene region, 87.3% of them come from intronic region, which indicates the potential regulatory role of these non-coding genetic variants.

Ontology mapping

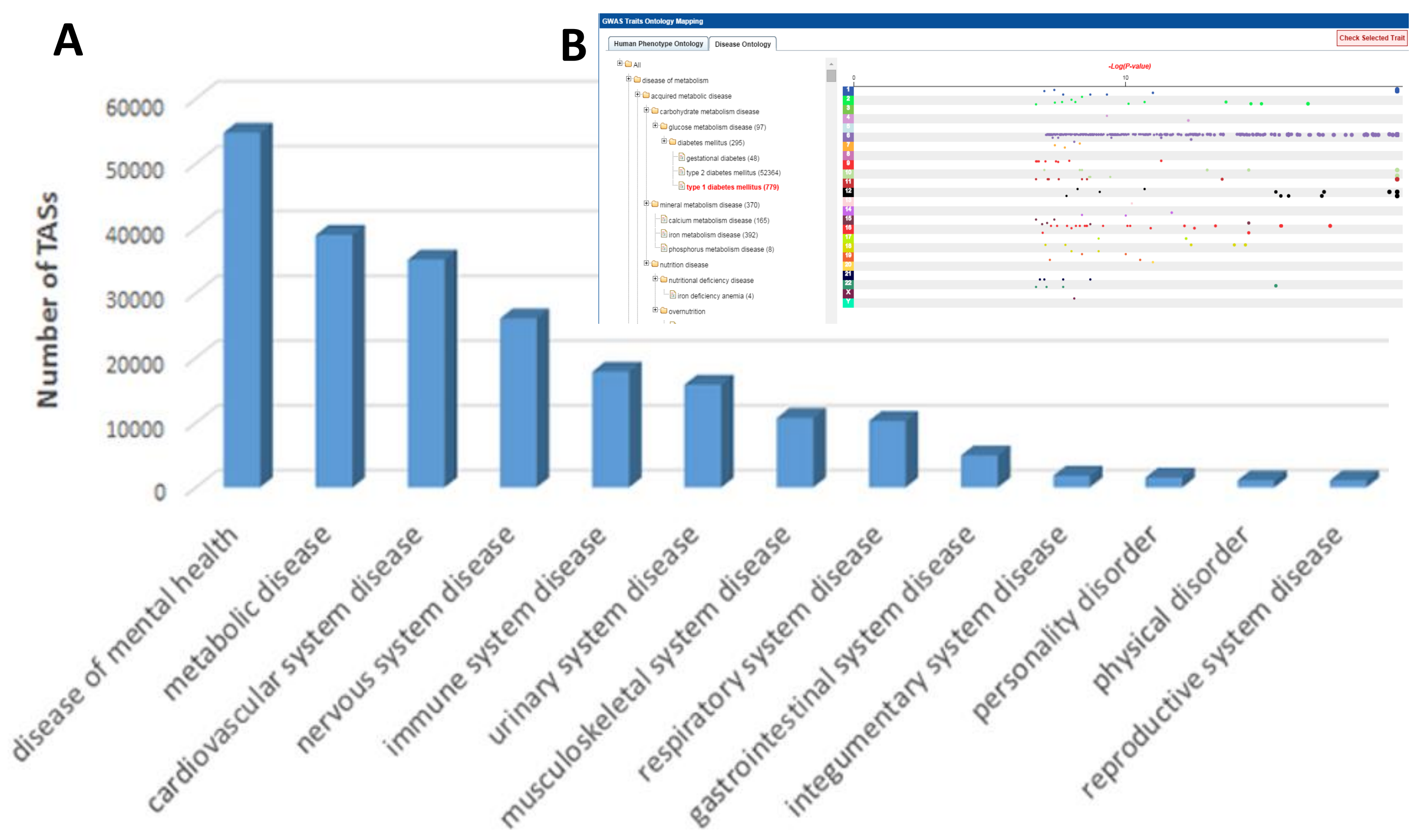
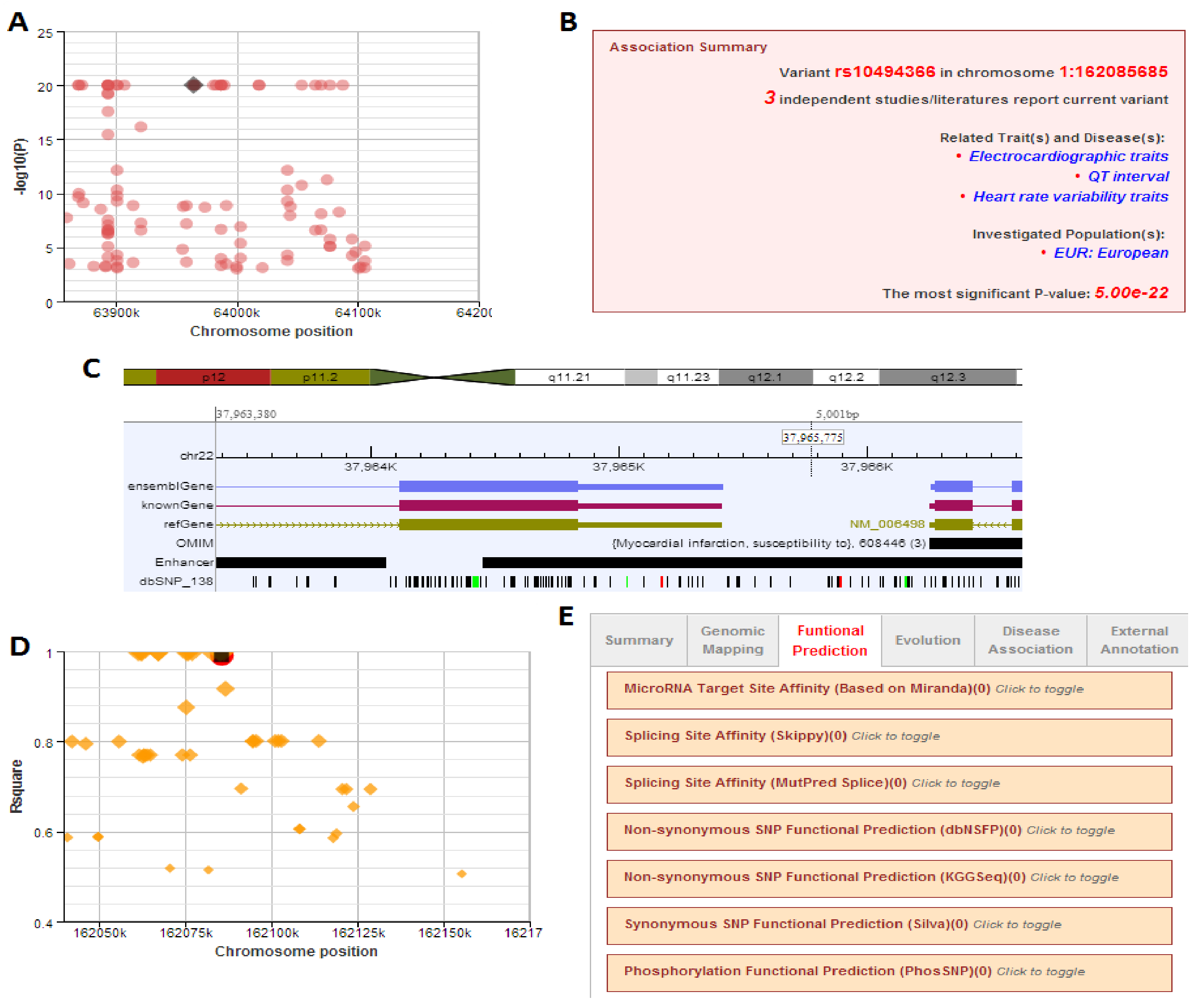


Figure 2. Distribution of TASs by mapping to Disease Ontology (DO). (A) Traits/diseases with more than 1000 TASs after mapping are shown; (B) Trait/diseases ontology mapping interface.

Annotation interface



Association Summary

Variant rs10494366 in chromosome 1:162085685
3 independent studies/literatures report current variant

Related Trait(s) and Disease(s):
- Electrocardiographic traits
- QT interval
- Heart rate variability traits

Investigated Population(s):
- EUR: European

The most significant P-value: 5.00e-22

MicroRNA Target Site Affinity (Based on Miranda)(0) Click to toggle
Splicing Site Affinity (Skippy)(0) Click to toggle
Splicing Site Affinity (MutPred Splice)(0) Click to toggle
Non-synonymous SNP Functional Prediction (dbNSFP)(0) Click to toggle
Non-synonymous SNP Functional Prediction (KGGSeq)(0) Click to toggle
Synonymous SNP Functional Prediction (Silva)(0) Click to toggle
Phosphorylation Functional Prediction (PhosSNP)(0) Click to toggle

Figure 3. Visualization and annotations of GWASdb2. (A) interactive Manhattan panel; (B) TAS summary information; (C) genome browser to show important functional elements; (D) interactive LD panel; (E) GWASdb annotation tabs.

ACKNOWLEDGEMENT

We thank Dr. Matthew R. Nelson and Dr. Philippe Sanseau in GSK to share their GWAS-MeSH mapping result.

REFERENCE

1. Li, M.J., et al., GWASdb: a database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res, 2012. 40(Database issue): p. D1047-54.