TIGA: Target Illumination GWAS Analytics

Aggregating and assessing experimental evidence for interpretable, explainable, accountable gene-trait associations.

Jeremy J Yang[1,3], Dhouha Grissa[2], Cristian G Bologa[1], Stephen L Mathias[1], Anna Waller[1], Christophe G Lambert[1], David J Wild[3], Lars Juhl Jensen[2] and Tudor I Oprea[1]

[1]University of New Mexico, Albuquerque, NM, USA; [2]Novo Nordisk Foundation Center for Protein Research, Copenhagen, Denmark; [3]Indiana University, Bloomington, IN, USA

Goal: Interpretable, useful knowledge from complex and noisy GWAS data

Genome wide association studies (GWAS) can reveal important genotype–phenotype associations, however, data quality and interpretability issues must be addressed. For drug discovery scientists seeking to prioritize targets based on the available evidence, these issues go beyond the single study. Here, we describe rational ranking, filtering and interpretation of inferred gene–trait associations and data aggregation across studies by leveraging existing curation and harmonization efforts. Each gene–trait association is evaluated for confidence, with scores derived solely from aggregated statistics, linking a protein-coding gene and phenotype. We propose a method for assessing confidence in gene–trait associations from evidence aggregated across studies, including a bibliometric assessment of scientific consensus based on the iCite Relative Citation Ratio, and meanRank scores, to aggregate multivariate evidence. TIGA is intended for drug target hypothesis generation, scoring and ranking, via the TIGA web app, and integration by IDG TCRD+Pharos, and the JensenLab DISEASES resource.

Aggregate statistics for gene-trait association

pValue max of SNP pValues

OR median of SNP ORs

N_beta count of significant betas

N_snp SNPs

N_snpw N_snp weighted by genomic distance

N_study total studies with association

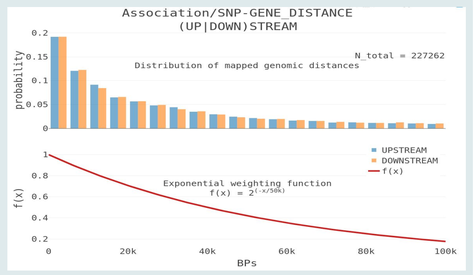
study_N mean(N_sample)

RCRAS RCR Aggregated Score

N_trait total traits associated with gene

N_gene total genes associated with trait

$$N\_snpw = \sum_{i}^{N\_snp} 2^{-d_i/k}$$



$$RCRAS_{gt} = \sum_{study}\left(\frac{1}{gc}\sum_{pub}\frac{log_2(RCR+1)}{sc}\right)$$

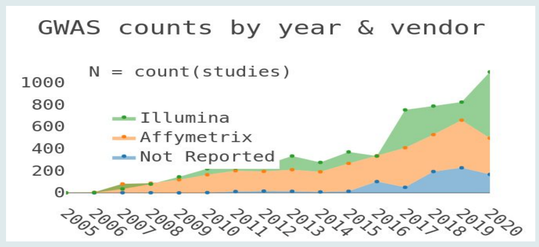
$study$ = GWAS (study_accession)
gc = gene count (in study)
pub = publication (PubMed ID)
sc = study count (in pub)

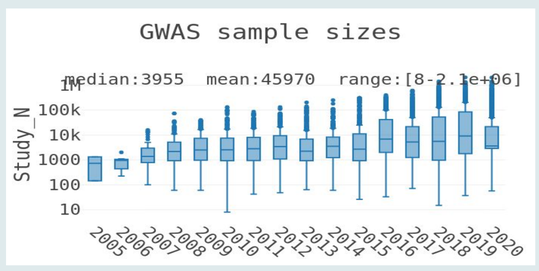
GWAS Catalog, powerful discovery platform

Since 2008 the NHGRI-EBI GWAS Catalog has provided a valuable and popular service by curating GWAS publications and effectively sharing GWAS metadata and summary data, addressing many difficulties of standardizing heterogeneous submissions, mapping formats and harmonizing content, promoting data standards according to FAIR principles, and sharing effectively via UI, API and downloads.

GWAS progress and problems

Despite dramatic progress in genomics, and the success of the GWAS Catalog, major problems and unmet needs remain. One reason is the rapid evolution of sequencing and analysis methodology, and consequent lack of standards. Problematic study designs may be due to practical clinical, recruitment challenges, sample size and statistical power, or traits poorly suited to genomic analysis, due to polygenicity or confounded etiology.

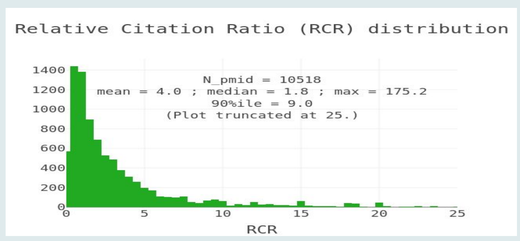


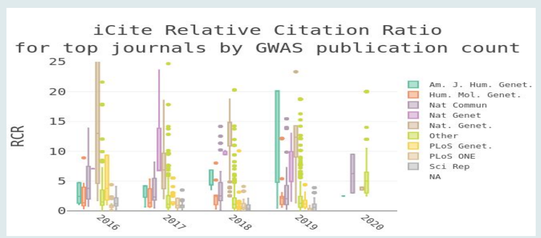


MAPPED_TRAIT	N_genes
self reported educational attainment	887
mathematical ability	875
heel bone mineral density	836
body mass index	741
type II diabetes mellitus	719
body height	616
high density lipoprotein cholesterol measurement	612
triglyceride measurement	582
sex hormone-binding globulin measurement	582
schizophrenia	570
intelligence	562
smoking status measurement	536
testosterone measurement	528
low density lipoprotein cholesterol measurement	483

RCRAS, bibliometric for consensus

The purpose of TIGA is to evaluate the evidence for a gene-trait association, by aggregating multiple studies and their corresponding publications. The iCite RCR (Hutchins et al., 2016) is a bibliometric statistic designed to evaluate the impact of an individual publication (in contrast to the journal impact factor). By field- and time-normalizing per-publication citation counts, the RCR measures evolving impact, in effect a proxy for scientific consensus. Hence by aggregating RCRs we seek a corresponding measure of scientific consensus for associations.

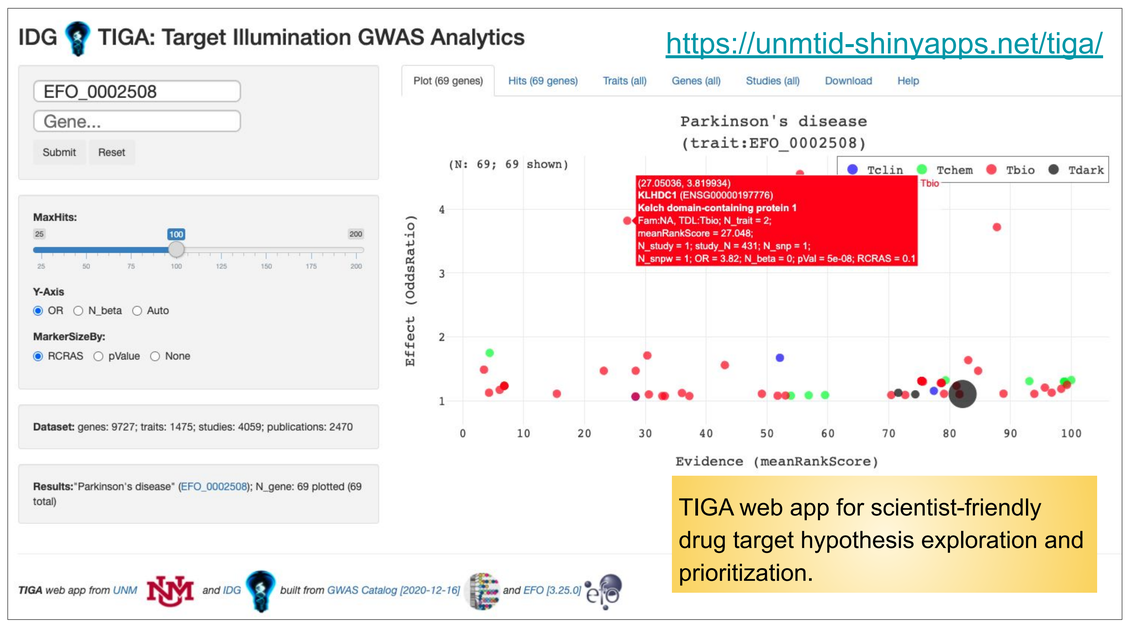




meanRank for unbiased multivariate scoring

meanRank aggregates ranks instead of variables directly, avoiding the need for ad hoc weighting parameters. Variable-ties imply rank-ties, with missing data ranked last. meanRankScore normalizes to (0,100] with variables of merit selected by benchmark against a gold-standard gene-disease association set:

- N_snpw: N_snp weighted by distance inverse exponential.
- pVal_mLog: max(-Log(pValue)) supporting gene-trait association.
- RCRAS: Relative Citation Ratio (RCR) Aggregated Score (iCite-RCR-based).



TIGA web app for scientist-friendly drug target hypothesis exploration and prioritization.

Biology matters.

As molecular biology, and even fundamental concepts such as "gene", evolve profoundly beyond the Crick-ian "central dogma", GWAS interpretation and utility will depend on these advances. However, the IDG mission simplifies and rationalizes prioritization, with a focus on protein-coding genes.

biotype	N_gene	%
protein_coding	15923	49.85
lncRNA	9097	28.48
processed_pseudogene	3521	11.02
unprocessed_pseudogene	805	2.52
OTHER	2599	8.14
TOTAL	31945	

Aggregation is hard. GWAS, like life, is un-FAIR.

The word "aggregation" encompasses a plethora of challenges. E.g., even a simple count of studies, since publications may report a meta-analysis using some or all of multiple datasets. Effect size beta units lack experimental and reporting standards. Expert curation is precious, as provided by GWAS Catalog and others, but ideally a community supported registry would promote standards.

Traits, phenotypes, diseases, semantics:

Interpretation of GWAS and related domains depend on semantics and ontology, i.e., the rigorous logic enabled by precise and accurate language. Imperfect semantics degrades downstream tasks, including data aggregation and validation of findings as testable clinical or animal-model hypotheses. Progress in medical science is required for progress in nosology, diagnostic criteria, and clinical informatics, all essential for improvements to disease models which frame much of our knowledge.

trait_id	trait_name	subclass_id	subclass_name	trait_N_study	subclass_N_study
EFO_0004247	mood disorder	EFO_0000289	bipolar disorder	18	96
EFO_0000289	bipolar disorder	EFO_1000650	bipolar I disorder	96	2
EFO_0000289	bipolar disorder	EFO_0009964	bipolar II disorder	96	1

Subclass traits for "mood disorder", EFO_0004247 with nonzero study counts.

Serving the IDG community

TIGA is fully aligned with IDG to evaluate evidence for disease-gene associations, focusing on protein-coding genes and drug target illumination, for scientists for whom GWAS, interpreted appropriately, can add value for exploring and prioritizing research opportunities. TIGA scores have been integrated into IDG db **TCRD** (v6.8.4) for display by portal **Pharos**, and JensenLab resource **DISEASES** (Nov 2020). With advances in target based, rational drug discovery, novel targets are an increasing unmet need, hence efforts to illuminate and prioritize target hypotheses.

TDL	N_gene (GWAS)
Tclin	685
Tchem	1684
Tbio	12495
Tdark	5499

TIGA manuscript under review; BioRxiv preprint:
https://www.biorxiv.org/content/10.1101/2020.11.11.378596v1

This work was supported by US National Institutes of Health (grants U54 CA189205 and U24 224370), Illuminating the Druggable Genome Knowledge Management Center (IDG KMC), and by the Novo Nordisk Foundation (grant number NNF14CC0001).