TIGA: Target illumination GWAS analytics

Jeremy J Yang[1,2], Dhouha Grissa[3], Christophe G Lambert[1], Cristian G Bologa[1,4], Stephen L Mathias[1], Anna Waller[5], David J Wild[2], Lars Juhl Jensen[3] and Tudor I Oprea[1,3,4,6*]

1. Translational Informatics Division, Department of Internal Medicine, University of New Mexico Health Sciences Center, Albuquerque, NM 87131, USA
2. School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN 47408, USA.
3. Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark.
4. UNM Comprehensive Cancer Center, University of New Mexico Health Sciences Center, Albuquerque, NM 87131, USA
5. UNM Center for Molecular Discovery, University of New Mexico Health Sciences Center, Albuquerque, NM 87131, USA
6. Department of Rheumatology and Inflammation Research, Institute of Medicine, Sahlgrenska Academy at University of Gothenburg, 40530 Gothenburg, Sweden

** To whom correspondence should be addressed. Tel: +1 505 925 7529; Fax: +1 505 925 7625; Email: toprea@salud.unm.edu*

Abstract

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. This method, intended for drug target hypothesis generation, scoring and ranking, has been implemented as an analytical pipeline, available as open source, with public datasets of results, and a web application designed for usability by drug discovery scientists, at https://unmtid-shinyapps.net/tiga/.

Keywords

GWAS, data science, drug discovery, drug target, druggable genome

Introduction

Over the two decades since the first draft human genome was published, dramatic progress has been achieved in foundational biology with translational benefits to medicine and human health. Genome wide association studies (GWAS) contribute to this progress by inferring associations between genomic variations and phenotypic traits (Bossé and Amos, 2018; Rusu *et al.*, 2017). These associations are correlations which may or may not be causal. While GWAS can reveal important genotype–phenotype associations, data quality and interpretability must be addressed (Lambert and Black, 2012; Visscher *et al.*, 2017; Marigorta *et al.*, 2018; Gallagher and Chen-Plotkin, 2018). For drug discovery scientists seeking to prioritize targets based on evidence from multiple studies, quality and interpretability issues are broader than for GWAS specialists. For this use case, GWAS are one of several evidence sources to be explored and considered, and interpretability must be in terms of genes corresponding to plausible targets, and traits corresponding to diseases of interest.

Single nucleotide variants (SNV) are the fundamental unit of genomic variation, and the term single nucleotide polymorphism (SNP) refers to SNVs identified as common sites of variation relative to a reference genome, and measured by microarray or sequencing technologies. The NHGRI-EBI GWAS Catalog (Buniello *et al.*, 2019) -- hereafter "Catalog" -- curates associations between SNPs and traits from GWAS publications, shares metadata and summary data, standardizes heterogeneous submissions, maps formats and harmonizes content, mitigating widespread data and meta-data issues according to FAIR (Findable, Accessible, Interoperable and Reusable) principles (Wilkinson *et al.*, 2016). These challenges are exacerbated by rapid advances in experimental and computational methodology. As *de facto* GWAS registrar, the Catalog interacts directly with investigators and accepts submissions of summary statistic data in advance of publication. Proposing and maintaining metadata standards the Catalog advocates and advances FAIRness in GWAS, for the benefit of the community. The Catalog addresses many difficulties due to content and format heterogeneity, but there are continuing difficulties and limitations both from lack of reporting standards and the variability of experimental methodology and diagnostic criteria.

Other GWAS data collections include the Genome-Wide Repository of Associations between SNPs and Phenotypes, GRASP (Eicher *et al.*, 2015) and The Framingham Heart Study, which employs non-standard phenotypes and some content from the Catalog (not updated since 2015). GWASdb (Li *et al.*,

2016) integrates over 40 data sources in addition to the Catalog, includes less significant variants to address a variety of use cases, and has been maintained continually since 2011. GWAS Central, continually updated through 2019, includes less significant associations and provides tools for a variety of exploration modes based on Catalog data, but is not freely available for download. PheGenI (Ramos *et al.*, 2014) integrates Catalog data with other NCBI datasets and tools. Others integrate GWAS with additional data (e.g. pathways, expression, linkage disequilibrium) to associate traits or diseases with genes (Greene *et al.*, 2015; Shen *et al.*, 2017; Wainberg *et al.*, 2019; Li *et al.*, 2018; Pallejà *et al.*, 2012). Each of these resources offers unique value and features. For this use case, the Catalog is the logical choice, given its applicability and commitment to expert curation, data standards, support and maintenance.

Here we describe TIGA (Target Illumination GWAS Analytics), an application for illuminating understudied drug targets. TIGA enables ranking, filtering and interpretation of inferred gene-trait associations aggregated across studies from the Catalog. Each inferred gene-to-trait association is evaluated for confidence, with scores derived solely from evidence aggregated across studies, linking a phenotypic trait and protein-coding gene, mapped from single nucleotide polymorphism (SNP) variation. TIGA uses the Relative Citation Ratio, RCR (Hutchins *et al.*, 2016), a bibliometric statistic from iCite (Hutchins *et al.*, 2019). TIGA does not index the full corpus of GWAS associations, but focuses on the strongest associations at the protein-coding gene level instead, filtered by disease areas that are relevant to drug discovery. For instance, GWAS for highly polygenic traits are considered less likely to illuminate druggable genes. Here, we describe the web application and its interpretability for non-GWAS specialists. We discuss TIGA as an application of data science for scientific consensus and interpretability, including statistical and semantical challenges. Code and data are available under BSD-2-Clause license from https://github.com/unmtransinfo/tiga-gwas-explorer.

Methods

NHGRI-EBI GWAS Catalog preprocessing

The 2020-07-15 release of the Catalog references 8935 studies and 4628 PubMed IDs. The curated associations include 7433 studies and 2194 EFO-mapped traits. After filtering studies to require (i) mapped trait, (ii) p-value below 5e-8, (iii) reported effect size (odds-ratio or beta), and (iv) mapped protein-coding gene, we found 3930 studies, 1452 traits, and 12158 genes. For consistency, only genes mapped by the Ensembl pipeline (https://www.ebi.ac.uk/gwas/docs/faq) for genomics annotations were considered (not author-reported). Figures 1 and 2 illustrate the growth of GWAS research as measured by counts of studies and subjects.

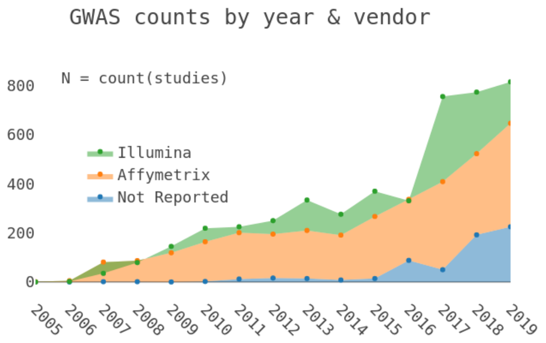


Fig 1: GWAS counts by year and vendor, indicating growth and platform trends.

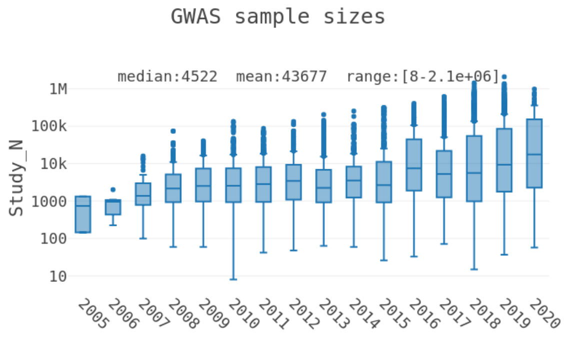


Fig 2: GWAS sample size distributions by year, on log scale, indicating variance in statistical power.

RCRAS = Relative Citation Ratio (RCR) Aggregated Score

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 statistic designed to evaluate the evolving empirical impact of a publication, in contrast to the non-empirical impact factor. By aggregating RCRs we seek to capture scientific community impact.

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

Where:

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

The $\log_2$() function is used with the assertion that differences of evidence depend on relative, rather than absolute differences in RCR. Division by sc effects a partial count for papers associated with multiple studies. Since RCR≥0, $\log_2$(RCR+1)≥ 0 and intuitively, when RCR= 1 and sc= 1, $\log_2$(RCR+ 1) = 1. Similarly division by gc reflects a partial count since studies may implicate multiple genes. This approach is informed by bibliometric methodology described elsewhere (Cannon *et al.*, 2017). For recent publications lacking RCR, we used the global median as an estimated prior. Computed thus, RCRAS extends RCR with similar logic, providing a rational bibliometric measure of evidence for scoring and ranking gene-trait associations.

Association weighting by SNP–gene distance

Mapping genomic variation of single nucleotides (SNPs) to genes is a challenging area of active research (Liu *et al.*, 2010; Mishra and Macgregor, 2015; Lamparter *et al.*, 2016). While TIGA does not contribute to mapping methodology, it does employ mappings provided by the Catalog between GWAS SNPs and genes, generated by the POSTGAP (https://github.com/Ensembl/postgap) Ensembl pipeline, which is based on STOPGAP (Shen *et al.*, 2017)). TIGA aggregates SNP-trait associations, assessing evidence for gene-trait associations, based on these understandings:

- SNPs within a gene are more strongly associated than SNPs upstream or downstream.
- Strength of association decreases with distance, or more rigorously stated, the probability of linkage disequilibrium (LD) between a SNP and protein coding gene decreases with genomic physical distance. Accordingly, we employ an inverse exponential scoring function, consistent with LD measure (Δ) and coefficient of decay (β) by Wang and coworkers (Wang *et al.*, 2006).

This function, used to weight N_snp to compute a distance-weighted SNP count N_snpw, is plotted together with the observed frequencies of mapped gene distances in supplementary Fig. 1, to illustrate how the extant evidence is weighted.

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

where d=distance in base pairs
and k = ”half-life distance” (50k)

Multivariate ranking

Multivariate ranking is a well studied problem which needs to be addressed for ranking GWAS associations. We evaluated two approaches, namely non-parametric μ scores (Wittkowski, 2008) and meanRank, and chose the latter based on benchmark test performance. **meanRank** aggregates ranks instead of variables directly, avoiding the need for *ad hoc* parameters. Variable-ties imply rank-ties, with missing data ranked last. We normalize scoring to (0,100] defining **meanRankScore** as follows.

Variables of merit used for scoring and ranking gene-trait associations:

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

Variables of merit and interest not currently used for ranking:

- OR: median(odds ratio, inverted if <1) supporting gene-trait association.
- N_beta: simple count of beta values with 95% confidence intervals supporting gene-trait association.
- N_snp: SNPs involved with gene-trait association.
- N_study: studies supporting gene-trait association.
- study_N: mean(SAMPLE_SIZE) supporting gene-trait association.
- geneNtrait: total traits associated with the gene.
- traitNgene: total genes associated with the trait.

From the variables selected via benchmark testing the meanRankScore is computed thus:

$$meanRankScore = 100 - Percentile(meanRank) = 100 - Percentile(\frac{1}{N}\sum_{i}^{N} rank_i)$$

$$\text{where } rank_i = rank\ of\ ith\ variable$$
$$\text{and } N = number\ of\ variables\ considered$$

Mu (μ) scores were implemented via the muStat (Wittkowski and Song, 2012) R package. Vectors of ordinal variables represent each case, and non-dominated solutions are cases, which are not inferior to any other case at any variable. The set of all non-dominated solutions defines a Pareto-boundary. The μ score is defined as the number of lower cases minus the number of higher cases, with ranking as the useful result. The ranking rule between case k and case k′ may be formalized thus:

$$\{x_k < x_{k'}\} \Leftrightarrow \{\forall_{l=1...L} x_{kl} \leq x_{k'l} \wedge \exists_{l=1...L} x_{kl} < x_{k'l}\}$$

Simply put, case k’ is higher than case k if it is higher in some variable(s) and lower in none.

Benchmark against gold standard

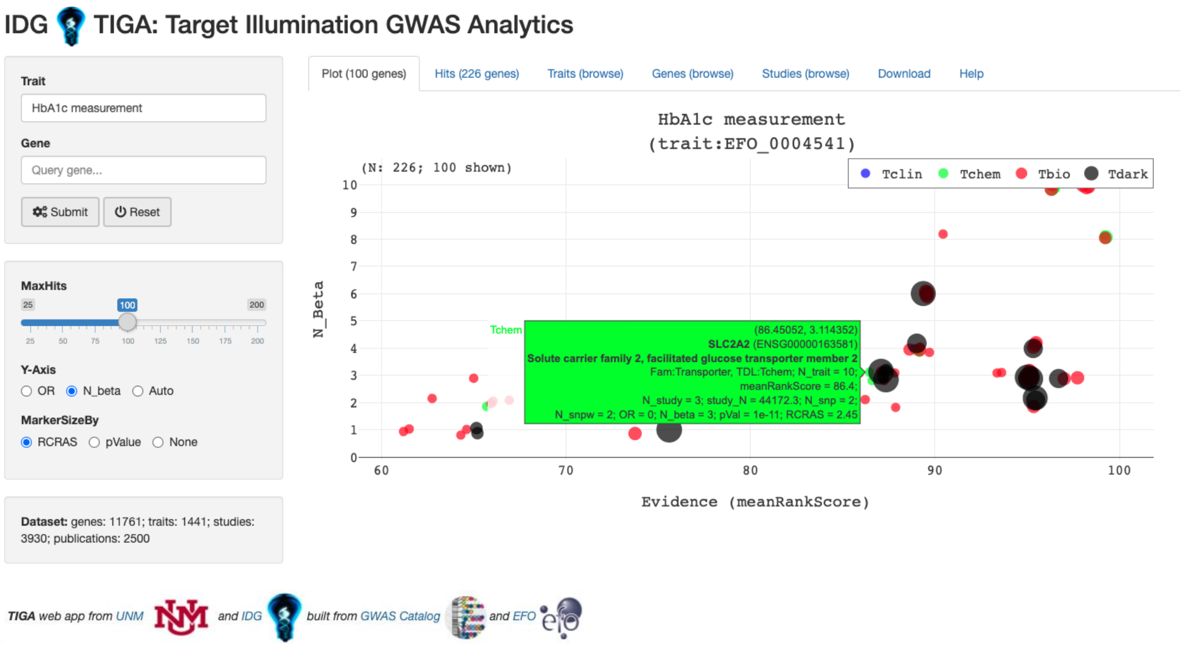
Lacking a suitable gold standard set of gene–trait associations in general, we instead relied on established gene–disease associations from the Genetics Home Reference, GHR (Fomous *et al.*, 2006) and UniProtKB (UniProt Consortium, 2018) databases. This gold stand set was built following a previously described approach (Pletscher-Frankild *et al.*, 2015). It consists of 5,366 manually curated associations (positive examples) between 3,495 genes and 709 diseases. All other (2,472,589) possible pairings of these genes and diseases were considered negative examples.

To assess the quality of the TIGA gene–trait associations, we mapped the Ensembl gene IDs to STRING v11 identifiers using the STRING alias file (Szklarczyk *et al.*, 2019) and the EFO terms to

Disease Ontology identifiers (Schriml *et al.*, 2019) based on ontology cross-references and the EMBL-EBI Ontology Xref Service. We then benchmark any individual-variable or multivariate ranking of the associations by constructing the receiver operating characteristic (ROC) curve by counting the agreement with the gold standard.

Results

The TIGA web application



IDG TIGA: Target Illumination GWAS Analytics

Trait: HbA1c measurement

Gene: Query gene...

MaxHits: 100

Y-Axis
○ OR ● N_beta ○ Auto

MarkerSizeBy: ● RCRAS ○ pValue ○ None

HbA1c measurement (trait:EFO_0004541)

(N: 226; 100 shown)

(86.45052, 3.114352)
SLC2A2 (ENSG00000163581)
Solute carrier family 2, facilitated glucose transporter member 2
Fam:Transporter, TDL:Tchem; N_trait = 10;
meanRankScore = 86.4;
N_study = 3; study_N = 44172.3; N_snp = 2;
N_snpw = 2; OR = 0; N_beta = 3; pVal = 1e-11; RCRAS = 2.45

Dataset: genes: 11761; traits: 1441; studies: 3930; publications: 2500

TIGA web app from UNM and IDG built from GWAS Catalog and EFO

Fig 3: TIGA web application (http://unmtid-shinyapps.net/tiga/), displaying a plot of genes associated with trait "HbA1c measurement" (EFO_0004541).

TIGA facilitates drug target illumination by currently scoring and ranking 101,762 associations between protein-coding genes and GWAS traits. While not capturing the entire Catalog, the TIGA app can aggregate and filter GWAS findings for actionable intelligence, e.g., to enrich target prioritization via interactive plots and hitlists (Fig 3), allowing users to identify the strongest associations supported by evidence.

Hits are ranked by meanRankScore described in Methods. Scatterplot axes are Effect (OR or N_beta) vs. Evidence as measured by meanRankScore. Plot markers may be sized by N_study or RCRAS. This app accepts "trait" and "gene" query parameters via URL, e.g. ?trait=EFO_0004541, ?gene=ENSG00000075073, ?trait=EFO_0004541&gene=ENSG00000075073. Gene markers are colored by Target Development Level (TDL)(Oprea *et al.*, 2018). TDL is a knowledge-based classification that bins human proteins into four categories: **Tclin,** mechanism-of-action designated targets via which approved drugs act (Santos *et al.*, 2017; Ursu *et al.*, 2019; Avram *et al.*, 2020); **Tchem** are proteins known to bind small molecules with high potency; **Tbio** includes proteins that have Gene Ontology (Ashburner *et al.*, 2000) “leaf” (lowest level) experimental terms; or meet two of these conditions: A fractional publication count (Pafilis *et al.*, 2013) above 5, three or more Gene “Reference Into Function” annotations (Mitchell *et al.*, 2003), or 50 or more commercial antibodies in Antibodypedia (Björling and Uhlén, 2008); **Tdark** are manually curated UniProtKB proteins that fail to place in any of the previous categories.

Benchmark against gold-standard disease–gene associations

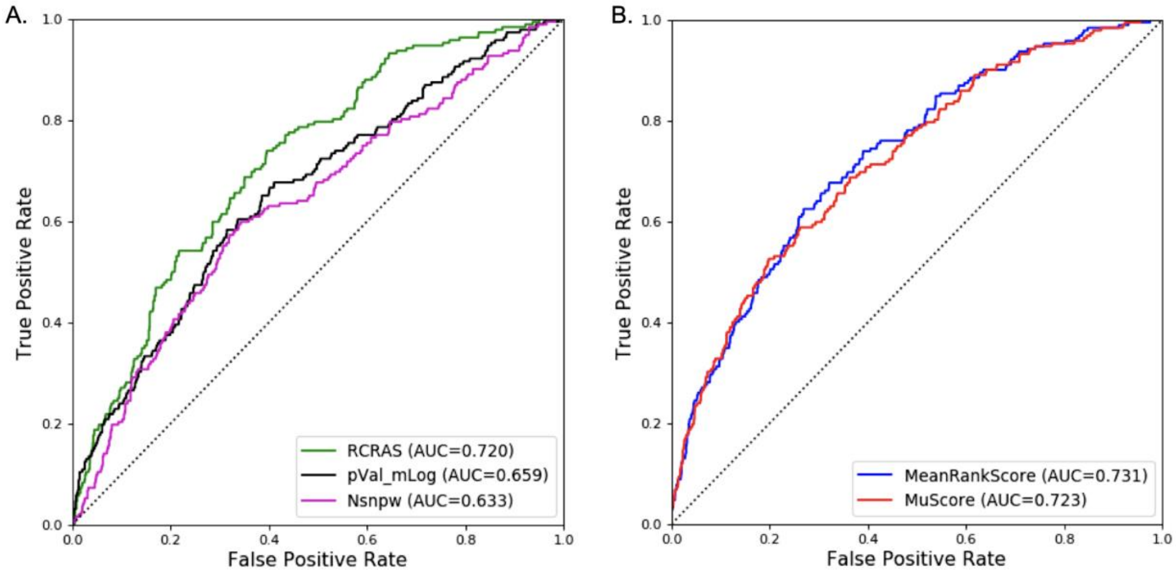


Fig. 4: **Performance evaluation.** The performance of TIGA on the gold standard of gene-disease associations. A) Results for the top-3 individual variables of merit. B) Results for the multivariate ranking by meanRankScore and μ score.

To benchmark the quality of the GWAS associations in TIGA, we focused on the 383 EFO terms that could be mapped to diseases and their 20,458 associations with genes. We evaluated the performance of each variable of merit individually against the manually curated gold standard gene–disease associations. The resulting ROC curves showed that the three best performing variables are RCRAS, pVal_mLog, and N_snpw, which have areas under the curve (AUC) of 0.72, 0.65, and 0.63, respectively (Fig. 4A). These variables are complementary, having a maximal pairwise Spearman correlation of 0.34 and evaluating different aspects of the associations. Based on these, we calculated two multivariate rankings, μ score and the meanRankScore. We benchmarked both rankings the same way as the individual variables and found that meanRankScore performs marginally better than μ score (Fig. 4B). As the meanRankScore is also more than five orders of magnitude faster to calculate, we selected it as the final ranking in TIGA.

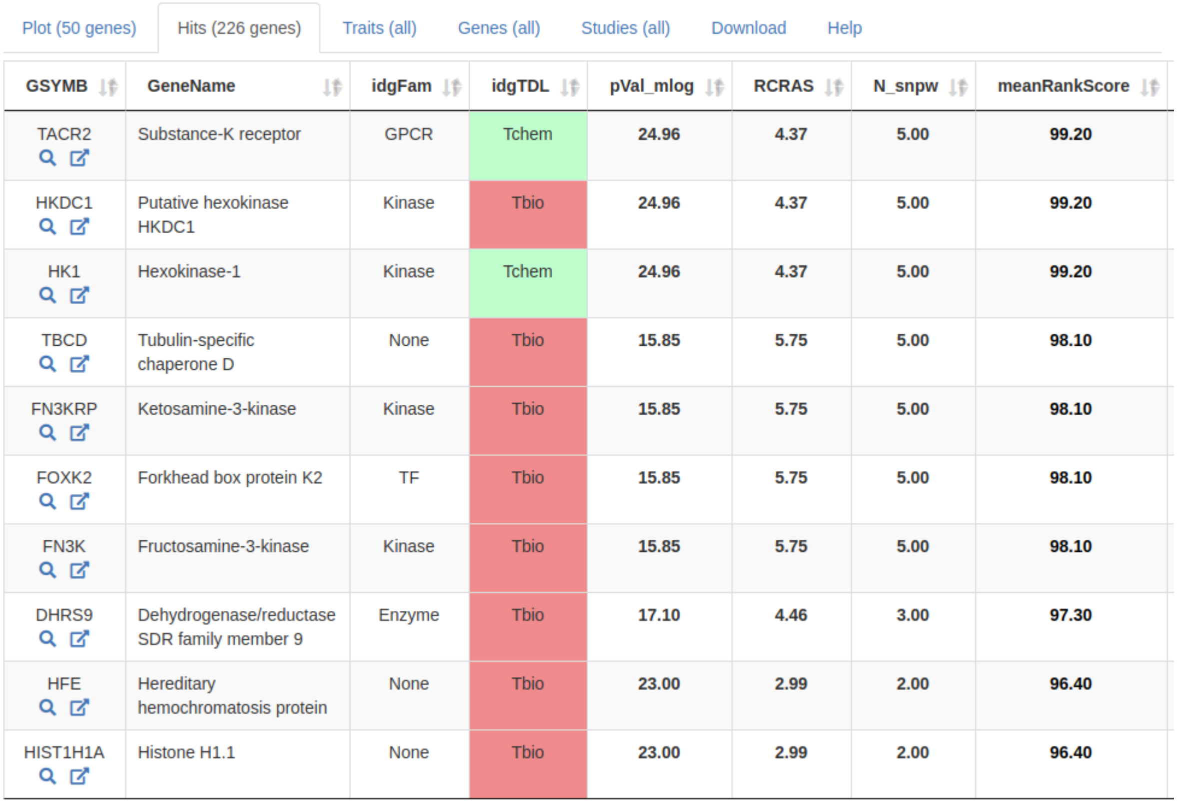
Using TIGA for drug target illumination

The main motivation of developing TIGA is to capture GWAS data when illuminating drug targets. Table 1 shows how many targets from each protein family and TDL are covered with associated traits in TIGA, with families as defined by Drug Target Ontology(Lin *et al.*, 2017) (DTO) Level 2. Noteworthy is the coverage for 2469 **Tdark** (understudied) proteins (Oprea *et al.*, 2018). The associations for other TDLs are also providing unique evidence, especially for **Tbio** proteins that are biologically characterized but have not before been clinically validated.

Figures 3 and 5 illustrate a typical use case, the plot and gene list for trait "HbA1c measurement" (glycated hemoglobin, signifying prolonged hyperglycemia), highly relevant to the management of type 2 diabetes mellitus. Figure 6 shows the provenance for one of the associated genes, SLC25A44 "Solute carrier family 25 member 44" with the scores and studies for this gene-trait association, including links to the Catalog and PubMed. SLC25A44 is an understudied (**Tdark**) transporter for branched-chain amino acids that acts as metabolic filter in brown adipose tissue, contributing to metabolic health (Yoneshiro *et al.*, 2019).

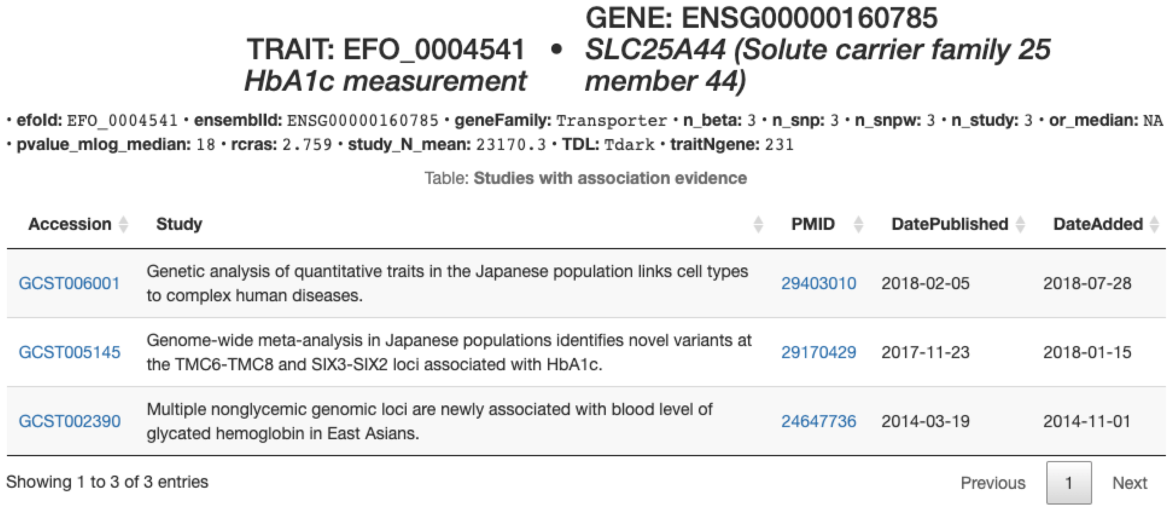
Table 1. TIGA mapped target (protein) counts by IDG Target Development Level (TDL) and Drug Target Ontology (DTO) level 2 gene family.

Family \ TDL	**Tclin**	**Tchem**	**Tbio**	**Tdark**	**Total**
G-protein coupled receptor	73 / 101	78 / 143	73 / 129	110 / 407	334 / 780
Ion channel	97 / 127	59 / 89	72 / 116	12 / 20	240 / 352
Kinase	57 / 66	278 / 360	97 / 133	12 / 20	444 / 579
Calcium-binding protein	3 / 5	1 / 3	58 / 93	8 / 11	70 / 112
Cell-cell junction	0 / 0	0 / 0	22 / 49	8 / 12	30 / 61
Cell adhesion	0 / 1	0 / 2	23 / 52	6 / 15	29 / 70
Cellular structure	4 / 10	5 / 11	244 / 323	44 / 86	297 / 430
Chaperone	0 / 1	8 / 9	27 / 46	6 / 8	41 / 64
Enzyme modulator	4 / 5	25 / 44	376 / 532	50 / 101	455 / 682
Enzyme	69 / 104	277 / 387	1022 / 1553	177 / 332	1545 / 2376
Epigenetic regulator	9 / 13	41 / 55	16 / 22	0 / 1	66 / 91
Extracellular structure	0 / 1	0 / 1	50 / 57	8 / 9	58 / 68
Immune response	0 / 1	0 / 2	13 / 41	4 / 6	17 / 50
Nuclear receptor	16 / 18	16 / 19	8 / 11	0 / 0	40 / 48
Nucleic acid binding	0 / 1	13 / 19	354 / 603	67 / 131	434 / 754
Transcription factor	1 / 2	12 / 16	385 / 557	73 / 163	471 / 738
Transporter	31 / 37	63 / 82	405 / 605	105 / 160	604 / 884
Receptor	20 / 24	6 / 12	157 / 225	27 / 55	210 / 316
Signaling	13 / 24	24 / 32	245 / 338	17 / 34	299 / 428
Storage	0 / 1	0 / 1	2 / 7	1 / 2	3 / 11
Surfactant	0 / 0	0 / 0	3 / 5	0 / 0	3 / 5
Other	95 / 134	233 / 337	3973 / 6131	1734 / 3416	6035 / 10018
Total	492 / 676	1139 / 1624	7625 / 11628	2469 / 4989	11725 / 18917



GSYMB	GeneName	idgFam	idgTDL	pVal_mlog	RCRAS	N_snpw	meanRankScore
TACR2	Substance-K receptor	GPCR	Tchem	24.96	4.37	5.00	99.20
HKDC1	Putative hexokinase HKDC1	Kinase	Tbio	24.96	4.37	5.00	99.20
HK1	Hexokinase-1	Kinase	Tchem	24.96	4.37	5.00	99.20
TBCD	Tubulin-specific chaperone D	None	Tbio	15.85	5.75	5.00	98.10
FN3KRP	Ketosamine-3-kinase	Kinase	Tbio	15.85	5.75	5.00	98.10
FOXK2	Forkhead box protein K2	TF	Tbio	15.85	5.75	5.00	98.10
FN3K	Fructosamine-3-kinase	Kinase	Tbio	15.85	5.75	5.00	98.10
DHRS9	Dehydrogenase/reductase SDR family member 9	Enzyme	Tbio	17.10	4.46	3.00	97.30
HFE	Hereditary hemochromatosis protein	None	Tbio	23.00	2.99	2.00	96.40
HIST1H1A	Histone H1.1	None	Tbio	23.00	2.99	2.00	96.40

Fig. 5: TIGA hit list of genes for trait "HbA1c measurement".



TRAIT: EFO_0004541 *HbA1c measurement* • GENE: ENSG00000160785 *SLC25A44 (Solute carrier family 25 member 44)*

• **efoId:** EFO_0004541 • **ensemblId:** ENSG00000160785 • **geneFamily:** Transporter • **n_beta:** 3 • **n_snp:** 3 • **n_snpw:** 3 • **n_study:** 3 • **or_median:** NA • **pvalue_mlog_median:** 18 • **rcras:** 2.759 • **study_N_mean:** 23170.3 • **TDL:** Tdark • **traitNgene:** 231

Table: **Studies with association evidence**

Accession	Study	PMID	DatePublished	DateAdded
GCST006001	Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases.	29403010	2018-02-05	2018-07-28
GCST005145	Genome-wide meta-analysis in Japanese populations identifies novel variants at the TMC6-TMC8 and SIX3-SIX2 loci associated with HbA1c.	29170429	2017-11-23	2018-01-15
GCST002390	Multiple nonglycemic genomic loci are newly associated with blood level of glycated hemoglobin in East Asians.	24647736	2014-03-19	2014-11-01

Showing 1 to 3 of 3 entries

Fig. 6: Provenance for association between gene SLC25A44 "Solute carrier family 25 member 44" and trait "HbA1c measurement".

Discussion

Target illumination

The explicit goal of the NIH Illuminating the Druggable Genome (IDG) program (Oprea *et al.*, 2018) is to "map the knowledge gaps around proteins encoded by the human genome." TIGA is fully aligned with this goal, as it evaluates the GWAS evidence for disease (trait) – gene associations. TIGA generates GWAS-centric trait–gene association dataset using an automated, sustainable workflow amenable for integration into the Pharos portal (Nguyen *et al.*, 2017; Sheils *et al.*, 2021). The OpenTargets platform (Koscielny *et al.*, 2017) uses Catalog data and other sources to identify and validate therapeutic targets by aggregating and scoring disease–gene associations for "practicing biological scientists in the pharmaceutical industry and in academia." In contrast, TIGA is a GWAS Catalog-only application that takes into account cited articles in a simple, interpretable manner.

From information to useful knowledge

In data-intensive fields such as genomics, specialized tools facilitate knowledge discovery, yet interpretation and integration can be problematic for non-specialists. Accordingly, this unmet need for integration and interpretation requires certain layers of abstraction and aggregation, which depend on specific use cases and objectives. Our target audience is drug discovery scientists for whom the aggregated findings of GWAS, appropriately interpreted, can provide additional value as they seek to prioritize targets. This clear purpose serves to focus and simplify all aspects of its design. Our approach for evidence aggregation is simple, easily comprehensible, and based on what may be regarded as axiomatic in science and rational inductive learning: First and foremost, evidence is measured by counting independent confirmatory results.

Interpretability concerns exist throughout science, but GWAS is understood to present particular challenges (Lambert and Black, 2012; Visscher *et al.*, 2017; Marigorta *et al.*, 2018; Gallagher and Chen-Plotkin, 2018). The main premise of GWAS is that genotype-phenotype correlations reveal underlying molecular mechanisms. While correlation does not imply causation, it contributes to plausibility of causation. Genomic dataset size adds difficulty. The standard GWAS p-value significance threshold is 5e-8, based on overall p-value 0.05 and Bonferroni multiple testing adjustment for 1-10 million tests/SNPs (Marigorta *et al.*, 2018). The statistical interpretation is that the family-wise error ratem FWER, or overall probability of a type-1 error, is 5%, but associations to mapped genes require additional interpretation. Motivated by, and despite these difficulties, it is our belief that GWAS data can be rationally interpreted and used by non-specialists, if suitably aggregated. Accordingly, TIGA is a rational way to suggest and rank research hypotheses, with the caveat that the identified signals may be accompanied by experimental noise and systematic uncertainty.

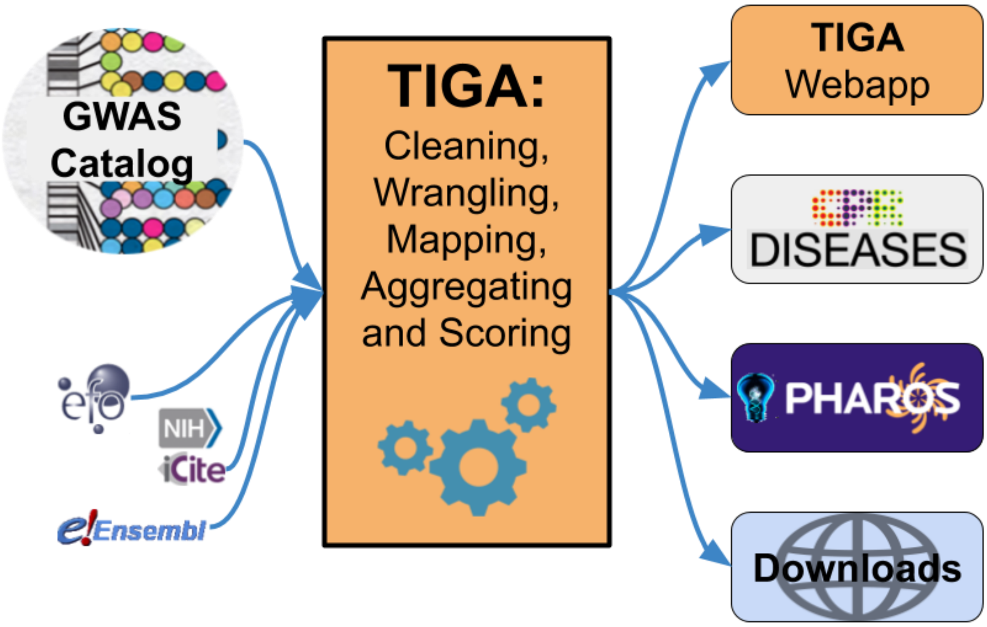


Fig 7: TIGA data sources and interfaces. TIGA integrates GWAS data from the Catalog and several other sources to rank gene-disease associations. These associations can be accessed through the TIGA webapp and are integrated into the DISEASES (Pletscher-Frankild *et al.*, 2015) and Pharos platforms. Bulk download is also available.

Designing for downstream integration

Biomedical knowledge discovery depends on integration of sources and data types which are heterogeneous in the extreme, reflecting the underlying complexity of biomedical science. These challenges are increasingly understood and addressed by improving data science methodology. However, provenance, interpretability and confidence aspects are underappreciated and rarely discussed. As in all signal propagation, errors and uncertainty accrue and confidence decays. Here, we proposed the use of simple, transparent, and comprehensible metrics to assess the relative confidence of disease-gene associations, via the unbiased meanRank scores. Figure 7, summarizing TIGA sources and interfaces, illustrates its well-defined role. Continuous confidence scores support algorithmic weighting and filtering. Standard identifiers and semantics support rigorous integration. Limiting provenance to the Catalog and its linked publications, semantic interpretability is enhanced.

Conclusions

We agree with Visscher et al. that: "the paradigm of 'one gene, one function, one trait' is the wrong way to view genetic variation"(Visscher *et al.*, 2017). Yet in the real world of biomedical science, progress often requires simplifying assumptions. Findings must be interpreted in context for an audience and application. Mindful of these concerns and limitations, TIGA provides a directly interpretable window into GWAS data, specifically for drug target hypothesis generation and elucidation. As interest in "interpretable machine learning" and "explainable artificial intelligence" (Gilpin *et al.*, 2018) grows, TIGA summarizes gene-trait associations derived solely and transparently from GWAS summary- and meta-data, with rational and intuitive evidence metrics and a robust, open-source pipeline designed for continual updates and improvements. Whether in stand-alone mode, or integrated into other resources, TIGA can contribute to drug target identification and prioritization.

Acknowledgements

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

Conflicts of Interest

CGL has a financial interest in Golden Helix Inc., a company which sells GWAS and other bioinformatics software. LJJ is one of the owners and Scientific Advisory Board members of Intomics A/S. TIO has received honoraria or consulted for Abbott, AstraZeneca, Chiron, Genentech, Infinity Pharmaceuticals, Merz Pharmaceuticals, Merck Darmstadt, Mitsubishi Tanabe, Novartis, Ono Pharmaceuticals, Pfizer, Roche, Sanofi and Wyeth. He is on the Scientific Advisory Board of ChemDiv Inc. and InSilico Medicine.

References

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Avram,S. *et al.* (2020) Novel drug targets in 2019. *Nat. Rev. Drug Discov.*, **19**, 300.

Björling,E. and Uhlén,M. (2008) Antibodypedia, a portal for sharing antibody and antigen validation data. *Mol. Cell. Proteomics*, **7**, 2028–2037.

Bossé,Y. and Amos,C.I. (2018) A Decade of GWAS Results in Lung Cancer. *Cancer Epidemiol. Biomarkers Prev.*, **27**, 363–379.

Buniello,A. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.

Cannon,D.C. *et al.* (2017) TIN-X: target importance and novelty explorer. *Bioinformatics*, **33**, 2601–2603.

Eicher,J.D. *et al.* (2015) GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. *Nucleic Acids Res.*, **43**, D799–804.

Fomous,C. *et al.* (2006) ‘Genetics home reference’: helping patients understand the role of genetics in health and disease. *Community Genet.*, **9**, 274–278.

Gallagher,M.D. and Chen-Plotkin,A.S. (2018) The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.*, **102**, 717–730.

Gilpin,L.H. *et al.* (2018) Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*.

Greene,C.S. *et al.* (2015) Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.*, **47**, 569–576.

Hutchins,B.I. *et al.* (2016) Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. *PLoS Biol.*, **14**, e1002541.

Hutchins,B.I. *et al.* (2019) The NIH Open Citation Collection: A public access, broad coverage resource. *PLoS Biol.*, **17**, e3000385.

Koscielny,G. *et al.* (2017) Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.*, **45**, D985–D994.

Lambert,C.G. and Black,L.J. (2012) Learning from our GWAS mistakes: from experimental design to scientific method. *Biostatistics*, **13**, 195–203.

Lamparter,D. *et al.* (2016) Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput. Biol.*, **12**, e1004714.

Li,M.J. *et al.* (2016) GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **44**, D869–76.

Lin,Y. *et al.* (2017) Drug target ontology to classify and integrate drug discovery data. *J. Biomed. Semantics*, **8**, 50.

Li,T. *et al.* (2018) GeNets: a unified web platform for network-based genomic analyses. *Nat. Methods*, **15**, 543–546.

Liu,J.Z. *et al.* (2010) A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.*, **87**, 139–145.

Marigorta,U.M. *et al.* (2018) Replicability and Prediction: Lessons and Challenges from GWAS. *Trends Genet.*, **34**, 504–517.

Mishra,A. and Macgregor,S. (2015) VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Res. Hum. Genet.*, **18**, 86–91.

Mitchell,J.A. *et al.* (2003) Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA Annu. Symp. Proc.*, 460–464.

Nguyen,D.-T. *et al.* (2017) Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.*, **45**, D995–D1002.

Oprea,T.I. *et al.* (2018) Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.*, **17**, 377.

Pafilis,E. *et al.* (2013) The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PLoS One*, **8**, e65390.

Pallejà,A. *et al.* (2012) DistiLD Database: diseases and traits in linkage disequilibrium blocks. *Nucleic Acids Res.*, **40**, D1036–40.

Pletscher-Frankild,S. *et al.* (2015) DISEASES: Text mining and data integration of disease–gene associations. *Methods*, **74**, 83–89.

Ramos,E.M. *et al.* (2014) Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.*, **22**, 144–147.

Rusu,V. *et al.* (2017) Type 2 Diabetes Variants Disrupt Function of SLC16A11 through Two Distinct Mechanisms. *Cell*, **170**, 199–212.e20.

Santos,R. *et al.* (2017) A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.*, **16**, 19–34.

Schriml,L.M. *et al.* (2019) Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.*, **47**, D955–D962.

Sheils,T.K. *et al.* (2021) TCRD and Pharos 2021: Mining the Human Proteome for Disease Biology. *Nucleic Acids Res.*

Shen,J. *et al.* (2017) STOPGAP: a database for systematic target opportunity assessment by genetic association predictions. *Bioinformatics*, **33**, 2784–2786.

Szklarczyk,D. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, **47**, D607–D613.

UniProt Consortium,T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, 2699.

Ursu,O. *et al.* (2019) Novel drug targets in 2018. *Nat. Rev. Drug Discov.*

Visscher,P.M. *et al.* (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.*, **101**, 5–22.

Wainberg,M. *et al.* (2019) Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.*, **51**, 592–599.

Wang,Y. *et al.* (2006) A Fine-Scale Linkage-Disequilibrium Measure Based on Length of Haplotype Sharing. *The American Journal of Human Genetics*, **78**, 615–628.

Wilkinson,M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, **3**, 160018.

Wittkowski,K.M. and Song,T. (2012) muStat: Prentice rank sum test and McNemar test. *R package version*, **1**.

Yoneshiro,T. *et al.* (2019) BCAA catabolism in brown fat controls energy homeostasis through SLC25A44. *Nature*, **572**, 614–619.

Definitions

Common terms used in GWAS and related fields can vary in their definitions and connotations depending on context. Therefore for clarity and rigor the following definitions are provided, which we consider consistent with best practices in the GWAS and drug discovery communities.

genotype An organism has one genotype, comprised of a germ line genome and multiple somatic genomes. Statistical models may assume a population distribution hence a population genotype.

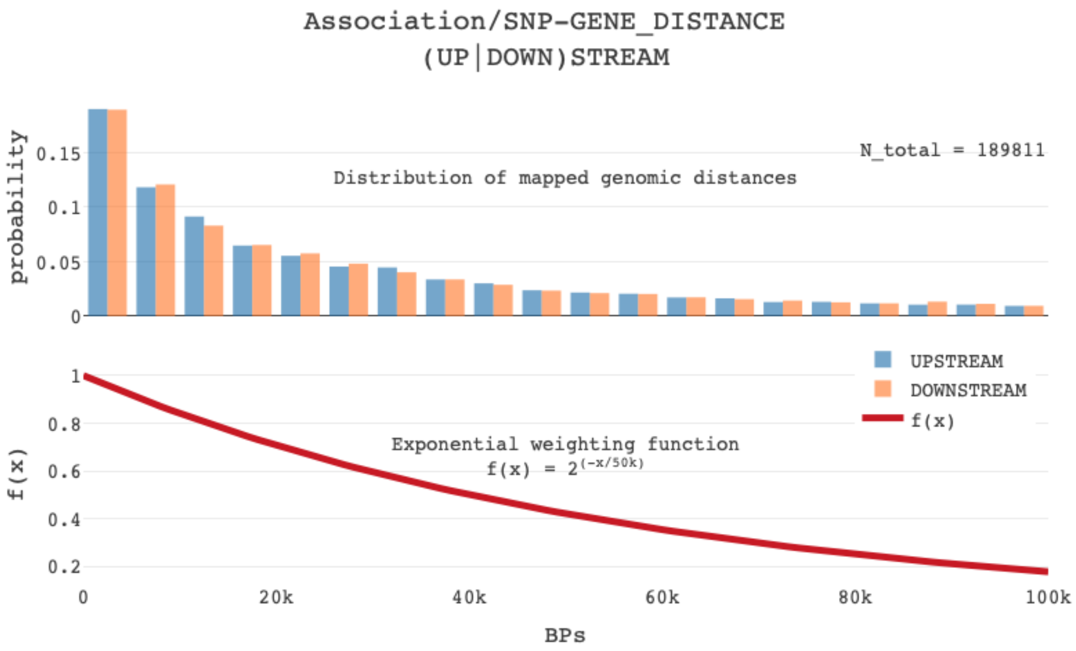
phenotype An organism has one phenotype, comprised of (potentially) all non-genomic observable characteristics, a.k.a. phenotypic traits.

gene Genomic unit responsible for an expression product. Protein coding genes are a subset of this definition.

trait Single non-genomic, observable characteristic.

drug target Biomolecular entity involved in the mechanism of action of a drug. The IDG project is human protein-centric; hence in this context, all drug targets are human proteins.

Supplementary material



Supplementary Fig 1: SNP-gene distances for (up|down)stream genes and TIGA weighting function

$$W_{Exp}(d) = 2^{-d/k}$$

where d=distance in base pairs
and k = ”half-life distance” (50k)