GWAS Explorer (GWAX) Formulae: Gene-Trait Association Scoring

Jeremy Yang

August 2019

1 RCRAS = Relative Citation Ratio (RCR) Aggregated Score

The purpose is to evaluate the evidence for a gene-trait association, by aggregating multiple studies and their corresponding publications. The iCite[1] RCR[1] is itself a statistic designed to evaluate the evolving empirical impact of a publication (in contrast to the non-empirical impact factor). Hence by aggregating RCRs we seek an corresponding measure of scientific community impact.

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

Where

$$\begin{aligned} study &= \text{GWAS (study\_accession)} \\ gc &= \text{gene count (in study)} \\ pub &= \text{publication (PubMed ID)} \\ sc &= \text{study count (in pub)} \end{aligned}$$

RCR $median = 2.0$ and $90\%ile = 8.5$. The $log_2()$ function is used with the belief that additional publications add diminishing evidence. Division by *spp* 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 papers and studies may associate to few or many findings and reported genes.

From NHGRI-EBI GWAS Catalog[2] and iCite PubMed statistics.

This approach informed by bibliometric methodology, including fractional counting, of Jensen

[1] https://icite.od.nih.gov/

[2] https://www.ebi.ac.uk/gwas/

et al. as employed in DISEASES[2] and TIN-X[3].

As with loss or objective functions, the absolute value is less important than the gradient, which solely determines ranking. For example, ten supporting publications may not be twice the evidence than five, but is certainly more.

2 Association SNP-gene distance weighting

Mapping genomic variation of single nucleotides (SNPs) to genes is a challenging area of active research. In this project we seek not to advance this field but simply to employ current best practices in scoring and ranking mappings between GWAS SNPs and protein coding genes. Linkage disequilibrium issues are not directly addressed, since this application does not seek to identify causal SNPs, but rather to identify druggable genes. SNP to gene mappings are scored and ranked based on these understandings:

- SNPs *within* a gene are more strongly associated than *upstream* or *downstream*.
- SNPs *upstream* are more strongly associated than *downstream* for equal distance.
- Strength of association decreases with distance with no obvious probabilistic relationship. Hence we need to employ some ad hoc scoring function, informed by genomic theory. At present we propose an inverse exponential form. The other functions are plotted for comparison.

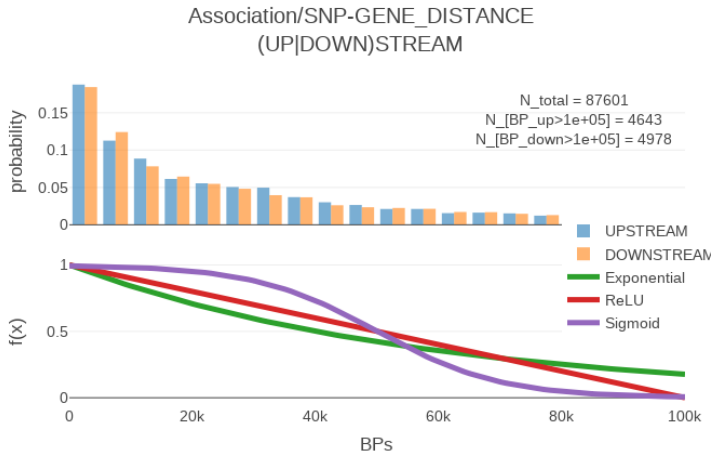

Figure 1: SNP-gene distances and weighting function candidates

$$F_{Exp}(d) = 2^{-d/k}$$
where $d =$ *distance in base pairs*
and $k =$ *"half-life distance"*

3 Multivariate, non-parametric ranking with μ scores

Multivariate ranking is a well studied problem which needs to be addressed for ranking GWAS associations. A weighted sum approach could be justified but requires ad hoc judgement based weights which could be fundamentally flawed and misleading. The shorthand version of the problem is how can we compare apples and oranges? In baseball, is a double worth two singles, or what is the right multiplier? From GWAS, how should effect size be weighted relative to p-value, or to the RCRAS described above? A rigorous approach exists and has been described well by Wittkowski[4]. 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 also define a Pareto-boundary or Pareto-front, which also is closely related to convex hull approaches. A μ score is defined simply as the number of lower cases minus the number of higher, but the ranking is the useful result. The ranking rule between case k and case k' may be formalized thus:

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

Or simply put, case k' is higher than case k if it is higher in some variable(s) and lower in none.

Variables of merit available, for evaluating relevance and confidence of disease-gene associations, by aggregating metadata and summaries from the GWAS Catalog:

- Odds ratio (OR) (median of trait-SNP ORs)
- pValue (median of trait-SNP pValues)
- N_study (# studies supporting association)
- N_sample (median of sample sizes)
- 1/N_trait (# traits associated with this gene)
- 1/N_gene (# genes associated with this trait)
- N_snp (# SNPs implicating gene)
- RCRAS (RCR aggregated score, described above)

References

[1] BI Hutchins, X Yuan, JM Anderson, GM Santangelo. *Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level*. PLoS Biol 14(9): e1002541, 2016, `https://doi.org/10.1371/journal.pbio.1002541`.

[2] Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ, *DISEASES: text mining and data integration of disease-gene associations*. Methods. 2015 Mar;74:83-9. `doi:10.1016/j.ymeth.2014.11.020`. Epub 2014 Dec 5.

[3] DC Cannon, JJ Yang, SL Mathias, O Ursu, S Mani, A Waller, SC Schürer, LJ Jensen, LA Sklar, CG Bologa, TI Oprea, Bioinformatics, 2017, btx200, *TIN-X: Target Importance and Novelty Explorer*. doi: 10.1093/bioinformatics/btx200.

[4] KM Wittkowski, T Song, K Anderson, JE Daniels, Journal of Quantitative Analysis in Sports, 4(3), 2008, *U-Scores for Multivariate Data in Sports*. doi:10.2202/1559-0410.1129.