HHS Public Access

Author manuscript

Trends Genet. Author manuscript; available in PMC 2017 March 28.

Published in final edited form as:

Trends Genet. 2014 December ; 30(12): 513–514. doi:10.1016/j.tig.2014.10.001.

Pitfalls in the application of gene set analysis to genetics studies

Adriana Estela Sedeño Cortés[1] and **Paul Pavlidis**[2,3,*]

[1]Graduate Program in Bioinformatics, University of British Columbia, Vancouver, Canada. V6T 1Z4

[2]Center for High Throughput Biology, University of British Columbia, Vancouver, Canada. V6T 1Z4

[3]Department of Psychiatry, University of British Columbia, Vancouver, Canada. V6T 1Z4

Gene set analysis (GSA; “enrichment”) is a popular approach for the interpretation of genome-wide analyses. GSA is most commonly applied to the analysis of transcriptomes, but from the outset it has been considered useful for any study that provides rankings or “hit lists” of genes. The recent review by Mooney et al. [1] is a valuable resource for geneticists wishing to apply gene set analysis to the output of GWAS. Here we describe some additional points of practical importance if the methods are to be applied and interpreted soundly.

As described by Mooney et al., associating a gene with a SNP requires making some assumptions relating to relative location, unless the functional variant is known. But all the assignment methods described by Mooney et al. can result in the implication of more than one gene by a single variant. This is not problematic from a biological standpoint, but if those genes share any annotation used as input for GSA, the statistical significance of the shared annotation will be inflated. As described by Mooney et al., one aim of GSA is to try to capture the distributed nature of the heritability of the trait across multiple loci. Counting the same locus multiple times defeats this purpose. Put another way, the assumption (inherent in many GSA methods) of statistical independence of the genes can be violated in a particularly insidious way. Few of the methods and tools reviewed by Mooney et al. appear to address this problem.

This “multiple counting” problem has practical impact, leading to the recent retraction of a GWAS study of memory [2], in which the primary finding was the significance of the Gene Ontology (GO) term “synapse organization and biogenesis”. In this study a single SNP in the PCDHB cluster on chromosome 5 was assigned to at least eight PCDHB cluster members. Because those genes are very similar in their annotations, a GO term they shared reached statistical significance; without the duplication, it does not [2]. Based on our own experience and discussions with other genomics and genetics research groups, this is a common occurrence (protocadherins in particular seem especially problematic). The same issue crops up in genome-wide methylation studies (“EWAS”), in which CpGs are analyzed

*Corresponding author: Dr. Paul Pavlidis, paul@chibi.ubc.ca, (604) 827-4157, 177 Michael Smith Laboratories, 2185 East Mall, Vancouver BC, Canada. V6T 1Z4.

rather than SNPs. A remedy is to collapse the GO annotations for all genes assigned to a SNP or CpG to a single “meta-gene” analysis unit, rather than using the default gene-to-annotation mappings. Computationally intensive sample permutation methods should be considered [3], but the simple meta-gene approach will avoid much of the trouble.

The second issue surrounds the conceptual coherency of GSA and the interpretation of the results. For the most part, GSA results are treated as exploratory add-ons to primary findings. In such situations mistakes or problems in using GSA are not of major consequence. But there is a temptation for researchers to salvage negative or underpowered studies (in genetics, epigenetics or transcriptomics) by appealing to groups of genes. This was apparently the approach of Dixson et al., who had a sample size of a few hundred individuals, too small to yield SNPs reaching genome-wide significance. They are not alone, and enrichment results have been reported as a primary result in other studies [4,5]. But we must strongly stress the dangers. As Mooney et al. point out, there is no agreement on what gene sets to use, and sources differ dramatically even when they are attempting to describe the same concepts. Equally problematic, sources such as GO can change rapidly [6], which can lead to unstable results [7–10]. Dixson et al. used GO annotations dating from 2008 [11], and the GO group they discuss now has at least 59 genes, not 23 as reported. While the impact is unknown in this case, the incomplete, changeable, conflicting and partly arbitrary nature of gene annotations should be taken into account before treating them as units of analysis with biological meaning. Furthermore, one cannot easily defend assigning biological significance to specific gene set members without considering the strength of association at the gene level. Again referring to the Dixson et al. study, they expressed strong interest in genes in the “synaptic organization” set having nominal (uncorrected) p-values of 0.1 or higher. It seems risky to consider such genes of interest merely due to sharing an annotation with a locus that does have a signal. Finally, GSA is highly questionable if there is no evidence for any association signal at all (i.e., the SNP p-value distribution is uniform, as appears to be the case [12] for at least one of the disorders considered in [4]). For all these reasons, GSA should be used as a replacement for a variant-level analysis with trepidation.

Acknowledgments

We thank Paul Thomas (USC) and Jesse Gillis (CSHL) for discussion and comments on drafts of the manuscript. Supported by NIH grant GM076990 to PP.

References

1. Mooney MA, et al. Functional and genomic context in pathway analysis of GWAS data. Trends Genet TIG. 2014; doi: 10.1016/j.tig.2014.07.004
2. Retraction for Dixson et al., Identification of gene ontologies linked to prefrontal–hippocampal functional coupling in the human brain. Proc Natl Acad Sci. 2014; doi: 10.1073/pnas.1414905111
3. Holden M, et al. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics. 2008; 24:2784–2785. [PubMed: 18854360]
4. Torkamani A, et al. Pathway analysis of seven common diseases assessed by genome-wide association. Genomics. 2008; 92:265–272. [PubMed: 18722519]
5. Heck A, et al. Converging Genetic and Functional Brain Imaging Evidence Links Neuronal Excitability to Working Memory, Psychiatric Disease, and Brain Activity. Neuron. 2014; 81:1203–1213. [PubMed: 24529980]

6. Huntley RP, et al. Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. GigaScience. 2014; 3:4. [PubMed: 24641996]
7. Clarke EL, et al. A task-based approach for Gene Ontology evaluation. J Biomed Semant. 2013; 4:S4.
8. Gillis J, Pavlidis P. A methodology for the analysis of differential coexpression across the human lifespan. BMC Bioinformatics. 2009; 10:306. [PubMed: 19772654]
9. Groβ A, et al. Impact of ontology evolution on functional analyses. Bioinformatics. 2012; 28:2671–2677. [PubMed: 22954631]
10. Alam-Faruque Y, et al. The Impact of Focused Gene Ontology Curation of Specific Mammalian Systems. PLoS ONE. 2011; 6:e27541. [PubMed: 22174742]
11. Release Notes - GeneSetEnrichmentAnalysisWiki. [online]. Available: http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Release_Notes. [Accessed: 09-Sep-2014]
12. Burton PR, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]