Published in final edited form as:

Trends Genet. 2014 December ; 30(12): 514–515. doi:10.1016/j.tig.2014.10.002.

‘Pitfalls in the application of gene set analysis to genetics studies’: a response

Michael A. Mooney^{1,2}, Joel T. Nigg³, Shannon K. McWeeney^{1,3,4}, and Beth Wilmot^{1,3,4}

¹Division of Bioinformatics & Computational Biology, Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, BICC 3181 S.W. Sam Jackson Park Rd., Portland, OR 97239-3098, USA

²Oregon Health & Science University Knight Cancer Institute, CR145, 3303 SW Bond Ave, Portland, OR 97239, USA

³Division of Psychology, Department of Psychiatry, Oregon Health & Science University, 3314 SW US Veteran’s Hospital Road DC7P, Portland, OR 97239, USA

⁴Oregon Clinical and Translational Research Institute, MQ280Y 3181 S.W. Sam Jackson Park Rd., Portland, Oregon 97239-3098, USA

The recent letter by Cortés and Pavlidis [1] brings attention to two important points regarding the interpretation of results from gene set analyses (GSA). In particular, the authors provided a clear example of the problems that can occur when correlation between genes is not taken into account [e.g., when the same single nucleotide polymorphism (SNP) is mapped to multiple genes] [2]. As we discussed in our recent review [3], the choice of strategy used to map SNPs to genes can have important consequences for the results obtained from a GSA, and we appreciate Cortés and Pavlidis’ clarification of the issue.

Cortés and Pavlidis correctly point out that few GSA methods directly address the problem of ‘multiple counting’ the same SNP when it is mapped to multiple genes. However, the use of functional predictions (the predicted consequence of the polymorphism) to assign SNPs in regions with overlapping genes has been suggested [4,5]. Another proposed method is to clean gene sets to remove genes in high linkage disequilibrium with others in the set [6].

In addition, Cortés and Pavlidis stress that extreme care should be taken when assigning biological significance to GSA results. We agree with this sentiment. The current state of GSAs makes it difficult to assign importance to biological processes without first scrutinizing the contributions of individual gene or SNP effects.

Although GSA is a promising approach for identifying the cumulative effect of multiple genes, the results obtained can differ significantly based on the GSA method used and the source of the gene set annotations (even for the same biological concept). Given our nascent understanding of the genetic architecture of complex diseases, it is difficult to judge the

validity of one method over another. Therefore, we agree with the caution expressed by Cortés and Pavlidis, and believe that current methods for GSA should be view as exploratory techniques for hypothesis generation.

Acknowledgments

Funding for this work was provided by the following grants: NIMH (1R01MH099064), NIH/NCI (5P30CA069533), NIH/NCATS (5UL1RR024140).

References

1. Cortés AES, Pavlidis P. Pitfalls in the application of gene set analysis to genetics studies. *Trends Genet.* 2014; 30:513–514. [PubMed: 25459301]
2. Dixon, et al. Retraction for Dixon *et al.*, Identification of gene ontologies linked to prefrontal–hippocampal functional coupling in the human brain. *Proc Natl Acad Sci USA.* 2014. <http://dx.doi.org/10.1073/pnas.1414905111>
3. Mooney, MA., et al. Functional and genomic context in pathway analysis of GWAS data. *Trends Genet.* 2014. <http://dx.doi.org/10.1016/j.tig.2014.07.004>
4. Bakir-Gungor B, Sezerman OU. A new methodology to associate SNPs with human diseases according to their pathway related context. *PLoS ONE.* 2011; 6:e26277. [PubMed: 22046267]
5. Saccone SF, et al. SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. *Nucleic Acids Res.* 2010; 38:W201–W209. [PubMed: 20529875]
6. Hong MG, et al. Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Hum Genet.* 2009; 126:289–301. [PubMed: 19408013]