Metaanalysis of flawed expression profiling data leading to erroneous Parkinson's biomarker identification

Santiago and Potashkin (1) propose that two RNAs in blood, hepatocyte nuclear factor 4 alpha (*HNF4A*) and polypyrimidine tract binding protein 1 (*PTBP1*), might be clinically useful biomarkers for diagnosing and tracking the progression of Parkinson's disease (PD), even speculating that they are better than neurological examination. Many have proposed biomarkers based on gene expression in blood, including the same authors (2, 3), but none of the studies propose the same biomarkers as each other, and to our knowledge such studies have not led to any substantial inroads in changing PD diagnosis methods. Given this backdrop, the bar to propose new blood-based biomarkers for PD should be quite high. In our opinion, this bar was not reached in this case.

Santiago and Potashkin's central claim is that a metaanalysis of four previous gene-expression studies identifies over 1,000 candidate biomarkers, leading them to eventually select *HNF4A* and *PTBP1* for follow-up. Our analysis reveals the dataset with Gene Expression Omnibus identifier GSE22491 is an outlier, with a very large amount of differential expression (affecting an estimated 50% of genes). The distinct nature of this dataset is apparent in the heat map shown in figure 1 of ref. 1. Such pervasive changes in blood gene expression might be expected in infection or leukemia, but for PD they seem surprising. Unfortunately, there is a simple explanation: GSE22491 suffers from an experimental design flaw in which all of the controls were run in one batch and all of the cases were run in another batch on a different date. This confound raises the strong possibility that any differential expression in GSE22491 is a batch effect, not the result of any biological difference between cases and controls.

When this dataset is excluded, and other issues with the input datasets are addressed, the prominence of *HNF4A* and *PTBP1* is eliminated (Fisher's combined probability test $P > 0.05$ before any multiple test correction, placing both genes at rank ∼1,400 among all genes tested). Thus, we conclude that the identification of *HNF4A* and *PTBP1* as biomarkers in the microarray metaanalysis is an artifact. We provide details of our analysis leading to this conclusion, as well as some additional points, in a separate document.*

How could this have been avoided? The use of a more robust metaanalysis method, as suggested by the developers of the metaanalysis tool Santiago and Potashkin used (4), might have helped. However, much of the problem may have been the uncritical use of previously published microarray datasets. It should not be assumed that because a dataset has been published in the peer-reviewed literature that it is of high quality. It is unknown to us why the problems such as batch confounds were not noted by the original study authors, and it is very possible the conclusions in those papers were adversely affected. Unfortunately, these types of problems are common, and the onus is on the data reuser to identify and address them. This case is also a wake-up call for developers of reanalysis tools (like ourselves) to assist users in maintaining vigilance.

Lilah Toker and Paul Pavlidis[1]

Department of Psychiatry and Centre for High-Throughput Biology, Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada V6T1Z4

1 Santiago JA, Potashkin JA (2015) Network-based metaanalysis identifies HNF4A and PTBP1 as longitudinally dynamic biomarkers for Parkinson's disease. *Proc Natl Acad Sci USA* 112(7):2257–2262.

2 Potashkin JA, Santiago JA, Ravina BM, Watts A, Leontovich AA (2012) Biosignatures for Parkinson's disease and atypical parkinsonian disorders patients. *PLoS ONE* 7(8):e43595.

3 Santiago JA, Scherzer CR, Potashkin JA (2014) Network analysis identifies SOD2 mRNA as a potential biomarker for Parkinson's disease. *PLoS ONE* 9(10):e109042.

4 Xia J, et al. (2013) INMEX—A web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Res* 41(Web Server issue, W1):W63–W70.

Author contributions: L.T. and P.P. analyzed data and wrote the paper.

The authors declare no conflict of interest.

[1] To whom correspondence should be addressed. Email: paul@chibi.ubc.ca.

*figshare.com/articles/Supplementary_information_and_analysis/1434042.

www.pnas.org/cgi/doi/10.1073/pnas.1507563112