Improving Reproducibility and Candidate Selection in Transcriptomics Using Meta-analysis

Journal of Experimental Neuroscience
Volume 12: 1–3
© The Author(s) 2018
DOI: 10.1177/1179069518756296
SAGE

Laurence A Brown and Stuart N Peirson

Sleep and Circadian Neuroscience Institute (SCNi), Nuffield Department of Clinical Neurosciences, Sir William Dunn School of Pathology, University of Oxford, Oxford, UK.

ABSTRACT: Transcriptomic experiments are often used in neuroscience to identify candidate genes of interest for further study. However, the lists of genes identified from comparable transcriptomic studies often show limited overlap. One approach to addressing this issue of reproducibility is to combine data from multiple studies in the form of a meta-analysis. Here, we discuss recent work in the field of circadian biology, where transcriptomic meta-analyses have been used to improve candidate gene selection. With the increasing availability of microarray and RNA-Seq data due to deposition in public databases, combined with freely available tools and code, transcriptomic meta-analysis provides an ideal example of how open data can benefit neuroscience research.

KEYWORDS: Meta-analysis, circadian, open science, transcriptomics, reproducibility

RECEIVED: January 8, 2018. **ACCEPTED:** January 8, 2018.

TYPE: Commentary

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHORS: Laurence A Brown, Sleep and Circadian Neuroscience Institute (SCNi), Nuffield Department of Clinical Neurosciences, Sir William Dunn School of Pathology, University of Oxford, Oxford OX1 3RE, UK.
Email: laurence.brown@ndcn.ox.ac.uk;

Stuart N Peirson, Sleep and Circadian Neuroscience Institute (SCNi), Nuffield Department of Clinical Neurosciences, Sir William Dunn School of Pathology, University of Oxford, Oxford OX1 3RE, UK. Email: stuart.peirson@eye.ox.ac.uk

COMMENT ON: Brown LA, Williams J, Taylor L, Thomson RJ, Nolan PM, Foster RG, Peirson SN. Meta-analysis of transcriptomic datasets identifies genes enriched in the mammalian circadian pacemaker. Nucleic Acids Res. 2017 Sep 29;45(17):9860-9873. doi: 10.1093/nar/gkx714. PubMed PMID: 28973476.

Background

A well-planned experiment allows the collection of relevant data while minimising unwanted or unexplained variation. Variation will always exist in complex biological systems, as will a degree of variation introduced by technical issues associated with the method of data collection. With the use of new technologies, cost can also be a limiting factor in the scale of an experiment, which can lead to underpowered individual studies. The acknowledgement that no study will be ideal does not mean that the data contained within small studies are without value. Issues of reproducibility between studies have been raised in many fields of science, including neuroscience.[1] As such, methods to combine the data from multiple studies - meta-analysis - provide a critical way of providing a scientific consensus. Meta-analysis is widely used in clinical medicine and forms one of the cornerstones of modern evidence-based medicine, on which clinical guidance and policy decisions are based. Building upon the pioneering work of epidemiologist Archie Cochrane, the Cochrane Collaboration is a good example of where routine meta-analysis is used to provide and update guidance for a range of potential medical interventions.[2,3]

Preclinical medical sciences and fundamental biological research have been slower to embrace meta-analysis. In some fields, the type of experimental techniques may vary too much for direct comparison. Moreover, data are typically not deposited on publication, making further analysis challenging. However, one field where data deposition is commonplace is in transcriptomic studies.[4] Transcriptomics poses a number of challenging statistical issues – in particular, the high false-positive rates that arise due to the simultaneous assessment of changes in very large numbers of transcripts. Conversely, when stringent methods are used to account for false positives, transcripts that are really changing may be excluded (false negatives).[5] Transcriptomic meta-analysis provides a valuable future avenue for addressing these problems.

Vote Counting and Beyond

The simplest way to look for consistent changes in gene expression across multiple transcriptomic studies is look for common transcripts among genes identified in each study. This is typically represented by the ubiquitous Venn diagram, showing overlap between studies (Figure 1). However, not all studies are equal. Some may contain larger sample sizes or less biological noise. As a result, any candidate gene list will contain false positives and will be missing false negatives that do not reach the chosen significance level for inclusion. As a result, similar to many other areas of modern life, vote counting alone may not always lead to the best-informed decisions.

By contrast with simple vote counting, meta-analysis involves applying a weighting to each study to account for factors such as sample size and biological variation. For a transcriptomic study, a simple approach to this is the use of inverse variance weighting, whereby increasing confidence is given to studies with lower variation. This provides a simple metric by which an effect size can be determined for each transcript across each study. Studies can then be compared to give a combined effect size.

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (http://www.creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

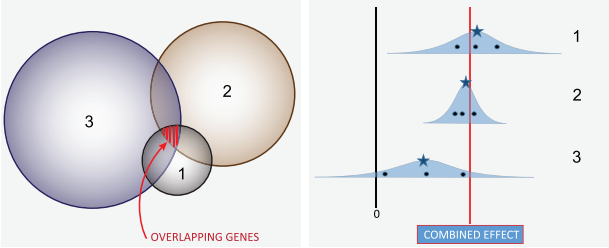

Figure 1. Possible approaches to meta-analysis; bringing together data from 3 different transcriptomic studies. An example of vote counting, where although the coverage of transcripts in more modern studies has improved, any agreement will be limited by those transcripts also present in earliest or lower-powered studies (left panel). Bringing together the effect sizes found in each study, placing greater confidence in those studies with lower variation (right panel).

Enrichment of Transcripts in the Suprachiasmatic Nuclei

We have recently applied transcriptomic meta-analysis to identify genes involved in circadian rhythms.[6] Circadian rhythms are ~24-hour changes in physiology and behaviour that are found in virtually all organisms. The disruption of such rhythmicity has been associated with a wide range of disorders, ranging from changes in metabolism, cardiovascular function, mental health, and even cancer.[7] In mammals, the primary circadian pacemaker is located in the suprachiasmatic nuclei (SCN), a paired structure of 20 000 neurons in the hypothalamus.[8] The SCN receives light input via the retinohypothalamic tract, enabling endogenous biological time to be set by the external light/dark cycle.[9] Circadian rhythms are generated by an intracellular transcriptional-translational feedback loop comprising a number of so-called 'clock genes'.[10] As many of these genes show 24-hour variations in expression, a number of studies have investigated cycling transcripts in the SCN as well as peripheral tissues, to identify genes important for circadian function.[11–13] However, the overlap between transcripts identified as rhythmic in different studies is often limited. For example, in the SCN, 2 studies identified 101 and 337 rhythmic genes, respectively.[11,13] Of these, only 23 were common to both.[14] A similar picture emerged in *Drosophila* microarray studies. To address this issue, Keegan et al[15] used transcriptomic meta-analysis to reduce the false discovery rate (FDR), identifying 81 previously identified cycling transcripts, as well as 133 transcripts that were not reported in any of the previous studies.

Not all genes important for circadian physiology are expected to be rhythmic, and many may play key roles in post-transcriptional or post-translational circadian mechanisms. Furthermore, other genes may be involved in the development or intercellular communication within the SCN. Can we therefore identify genes important for circadian function via their selective enrichment in the SCN? This question led to our recent meta-analysis, based on 79 microarray and 17 RNA-Seq data sets. Our data showed enrichment of many genes known to play roles in the generation of circadian rhythms[16,17] and previously identified markers[18,19] but also a host of other transcripts with little or no known role in circadian biology. The limitations of such an approach must also be acknowledged. We used data from tissue from different mouse strains collected at different times across the circadian cycle. Although this may be expected to increase variance in both SCN and whole brain samples, the increased sample size still identified 426 transcripts with a combined effect size of >3 and an FDR-corrected *P* value of <.01. Such robust changes are rarely detected in individual transcriptomic studies.

Confidence in Your Candidate: Tackling the Ignorome

The process of identifying candidate genes or transcripts that play important roles in biological processes or disease is central to modern biomedical research. In broad terms, 2 approaches have been applied to address this issue. Forward genetics, based on identifying interesting phenotypes and determining the causative gene(s), has played a critical role in many areas of research. However, over the past 2 decades, reverse genetics, moving from a gene of interest to associated phenotypes, has become much more common, primarily due to the remarkable genomic resources now available.

Transcriptomic studies provide a way of identifying potential candidate genes for subsequent study in relation to a particular biological process. However, following up a candidate gene can take months or often even years. As such, the risks of pursuing the wrong candidate have major time and financial considerations and can be especially damaging to the career of the individual researcher who is tasked with such a project. This raises a transcriptomics dilemma: how can we be sure it is worthwhile following up a specific candidate gene?

A major problem in selecting a candidate for further study is that for many genes, little is known about their function. As a result, researchers tend to focus on genes which have already been well-characterised. This is well illustrated by the power-law relationship in biomedical publication and funding, showing that researchers favour the study of relatively few genes.[20] This is further emphasised by neuroscience studies showing that the top 5% of genes account for a remarkable 70% of the research literature. By contrast, around 20% of genes have effectively no neuroscientific literature. Such genes for which we lack any detailed biological understanding have been collectively termed the 'ignorome'.[21] The use of transcriptomic meta-analysis provides increasing confidence that changes in a candidate gene are real and justify further study. This in turn should encourage more research on the uncharted regions of the genome.

The Future: Wide Open Data?

The number of scientific papers published each year featuring transcriptomic data has been rising steadily (Figure 2), as well

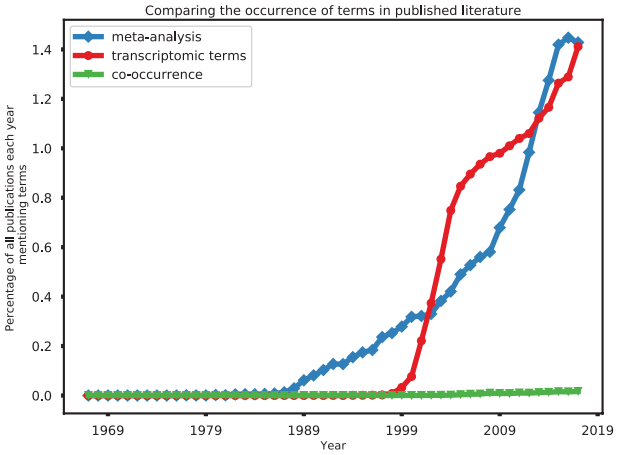

Figure 2. The increasing prevalence of meta-analyses and transcriptomic data in the scientific literature: even when controlling for a general increase in the volume of scientific literature, in the past 30 years, the number of scientific publications mentioning 'meta-analysis' each year is increasing. More recently, there has been an increasing proportion of publications that contain the terms 'microarray', 'RNA-Seq', or 'transcriptomics'. However, transcriptomic data are the focus of less than 2% of the publications involving meta-analysis in any given year. Terms searched using https://www.ncbi.nlm.nih.gov/pubmed, on January 2, 2018 and taken as percentages of all publications in the database for said year.

as the volume of data related to these publications (quadrillions of bases and counting).[22,23] As a result, if a question can be addressed now using available public data, it may be possible to get a better answer next year, or even next month. Automated periodic checks for new data may also allow refinements or additional questions to be asked of the data, in the form of subgroup analysis.

It should also be the case that the methods of analysis are clear but not immutable, with incremental refinements being made over time. To aid transparency, we published our full analysis as interactive notebooks, to allow others to follow our methods. We may have even made some errors but feel it is far better to be open about this possibility than to hide our analysis from the research community. Others are taking this approach further with innovations such as the Jupyter project (https://jupyter.org/) and Binder (https://mybinder.org/)[24] allowing for full analyses to be re-run in a Web browser. Reproducible research and open science are rightly being demanded by funding bodies and research communities alike.[25] Luckily, the tools to facilitate these processes are improving rapidly.

Finally, it is worth stressing that transcriptomic meta-analysis is only possible because of the requirement for data deposition on acceptance of a paper for publication, along with minimal meta-data about the experimental design[4] (although the more methodological detail the better).[23] By contrast, widespread deposition of data in other fields does not always occur and is the first hurdle to overcome before meta-analysis becomes possible in other research areas. As well as improving candidate gene selection in transcriptomic studies, more broadly, meta-analysis provides an ideal way of improving the reproducibility and transparency of neuroscience research.

Author Contributions

LAB and SNP contributed equally to this article.

ORCID iDs

Laurence A Brown https://orcid.org/0000-0001-8116-4100
Stuart N Peirson https://orcid.org/0000-0003-3653-834X

REFERENCES

1. Button KS, Ioannidis JP, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14:365–376.
2. Chalmers I. The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Ann N Y Acad Sci*. 1993;703:156–163 (discussion 163-165).
3. Cochrane AL. *Effectiveness and Efficiency: Random Reflections on Health Services*. London, England: The Nuffield Provincial Hospitals Trust; 1972.
4. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat Genet*. 2001;29:365–371.
5. Nadon R, Shoemaker J. Statistical issues with microarrays: processing and analysis. *Trends Genet*. 2002;18:265–271.
6. Brown LA, Williams J, Taylor L, et al. Meta-analysis of transcriptomic datasets identifies genes enriched in the mammalian circadian pacemaker. *Nucleic Acids Res*. 2017;45:9860–9873.
7. Maywood ES, O'Neill J, Wong GK, Reddy AB, Hastings MH. Circadian timing in health and disease. *Prog Brain Res*. 2006;153:253–269.
8. Weaver DR. The suprachiasmatic nucleus: a 25-year retrospective. *J Biol Rhythms*. 1998;13:100–112.
9. Hughes S, Jagannath A, Hankins MW, Foster RG, Peirson SN. Photic regulation of clock systems. *Methods Enzymol*. 2015;552:125–143.
10. Reppert SM, Weaver DR. Coordination of circadian timing in mammals. *Nature*. 2002;418:935–941.
11. Panda S, Antoch M, Miller B, et al. Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell*. 2002;109:307–320.
12. Storch KF, Lipan O, Leykin I, et al. Extensive and divergent circadian gene expression in liver and heart. *Nature*. 2002;417:78–83.
13. Ueda HR, Chen W, Adachi A, et al. A transcription factor response element for gene expression during circadian night. *Nature*. 2002;418:534–539.
14. Duffield GE. DNA microarray analyses of circadian timing: the genomic basis of biological time. *J Neuroendocrinol*. 2003;15:991–1002.
15. Keegan KP, Pradhan S, Wang JP, Allada R. Meta-analysis of Drosophila circadian microarray studies identifies a novel set of rhythmically expressed genes. *PLoS Comput Biol*. 2007;3:e208.
16. Doi M, Ishida A, Miyake A, et al. Circadian regulation of intracellular G-protein signalling mediates intercellular synchrony and rhythmicity in the suprachiasmatic nucleus. *Nat Commun*. 2011;2:327.
17. Harmar AJ, Marston HM, Shen S, et al. The VPAC(2) receptor is essential for circadian function in the mouse suprachiasmatic nuclei. *Cell*. 2002;109:497–508.
18. Hong HK, Chong JL, Song W, et al. Inducible and reversible Clock gene expression in brain using the tTA system for the study of circadian behavior. *PLoS Genet*. 2007;3:e33.
19. Husse J, Zhou X, Shostak A, Oster H, Eichele G. Synaptotagmin10-Cre, a driver to disrupt clock genes in the SCN. *J Biol Rhythms*. 2011;26:379–389.
20. Su AI, Hogenesch JB. Power-law-like distributions in biomedical publications and research funding. *Genome Biol*. 2007;8:404.
21. Pandey AK, Lu L, Wang X, Homayouni R, Williams RW. Functionally enigmatic genes: a case study of the brain ignorome. *PLoS ONE*. 2014;9:e88889.
22. Kodama Y, Shumway M, Leinonen R. International nucleotide sequence database C. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res*. 2012;40:D54–D56.
23. Lima MS, Smith DR. Don't just dump your data and run. *Embo Rep*. 2017;18:2087–2089.
24. eLife. Blog_post. 2018; https://elifesciences.org/labs/8653a61d/introducing-binder-2-0-share-your-interactive-research-environment. Accessed January 4, 2018.
25. Eglen S. Open access: sharing your data is easier than you think. *Nature*. 2014;510:340.