Journal of Biology

BioMed Central

Comment

'Validation' in genome-scale research

Timothy R Hughes

Address: Banting and Best Department of Medical Research, Department of Molecular Genetics, and Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON M5S 3E1, Canada. Email: t.hughes@utoronto.ca

Published: 26 January 2009

Journal of Biology 2009, **8**:3 (doi:10.1186/jbiol104)

The electronic version of this article is the complete one and can be found online at http://jbiol.com/content8/1/3

Abstract

The individual 'validation' experiments typically included in papers reporting genome-scale studies often do not reflect the overall merits of the work.

Following the advent of genome sequencing, the past decade has seen an explosion in genome-scale research projects. Major goals of this type of work include gaining an overview of how biological systems work, generation of useful reagents and reference datasets, and demonstration of the efficacy of new techniques. The typical structure of these studies, and of the resulting manuscripts, is similar to that of a traditional genetic screen. The major steps often include development of reagents and/or an assay, systematic implementation of the assay, and analysis and interpretation of the resulting data. The analyses are usually centered on identifying patterns or groups in the data, which can lead to predictions regarding previously unknown or unanticipated properties of individual genes or proteins.

So that the work is not purely descriptive - anathema in the molecular biology literature - there is frequently some follow-up or 'validation', for example, application of independent assays to confirm the initial data, an illustration of how the results obtained apply to some specific cellular process, or the testing of some predicted gene functions. As the first few display items are often schematics, example data, clustering diagrams, networks, tables of *P*-values and the like, these validation experiments usually appear *circa* Figure 5 or 6 in a longer-format paper. This format is sufficiently predominant that my colleague Charlie Boone refers to it as "applying the formula". I have successfully used the formula myself for many papers.

My motivation for writing this opinion piece is that, in my own experience, as both an author and a reviewer, the focal point of the review process - and of the editorial decision - seems too often to rest on the quality of the validation, which is usually not what the papers are really about. While it is customary for authors to complain about the review process in general (and for reviewers to complain about the papers they review), as a reader of such papers and a user of the datasets, I do think there are several legitimate reasons why our preoccupation with validation in genomic studies deserves reconsideration.

First, single-gene experiments are a poor demonstration that a large-scale assay is accurate. To show that an assay is consistent with previous results requires testing a sufficiently large collection of gold-standard examples to be able to assess standard measures such as sensitivity, false-positive rate and false-discovery rate. A decade ago, there were many fewer tools and resources available; for example, Gene Ontology (GO) did not exist before the year 2000 [1], and many of the data analysis techniques now in common use were unfamiliar to most biologists. Proving that one could make accurate predictions actually required doing the laboratory analyses. But today, many tools are in place to make the same arguments by cross-validation, which produces all of the standard statistics. It is also (gradually) becoming less fashionable for molecular biologists to be statistical Luddites.

Second, and similarly, single-gene experiments, or illustrations relating to a specific process, do not describe the general utility of a dataset. Many studies have shown (even if they did not emphasize) that specific data types and reagents are more valuable for the study of some things than others. Validation experiments tend to focus on the low-hanging fruit, for instance, functional categories that seem to be yielding the best examples, and the largest numbers. To minimize the ire of my colleagues, I will give an example from my own work. Our first efforts at systematically predicting yeast gene functions from gene-expression data [2] resulted in more predictions relating to RNA processing than to any other category, and Northern blots are something even my lab can do, so these were the ones we tested. Although we would like to think that the success at validating predictions from other processes will also be as high as our cross-validation predicted, laboratory validation of predictions from only one category does not show that. Moreover, if one is engaged in high-throughput data collection, it is possible to perform a large number of validations, and show only those that work. It is also possible to choose the validation experiments from other screens already in progress, or already done, or even from other labs. I suspect this practice may be widespread.

A third issue is that focus on the validation is often at the expense of a thorough evaluation of the key points of the remainder of the paper. I may be further ruffling the fur of my colleagues here, but I think it is fair to say that a hallmark of the functional genomics/systems biology/network analysis literature is an emphasis on artwork and *P*-values, and perhaps not enough consideration of questions such as the positive predictive value of the large-scale data. David Botstein has described certain findings as "significant, but not important" - if one is making millions of measurements, an astronomically significant statistical relationship can be obtained between two variables that barely correlate, and an overlap of only one or a few percent in a Venn diagram can be very significant by the widely used hypergeometric test. A good yarn seems to distract us from a thorough assessment of whether statistical significance equates to biological significance, and even whether the main dataset actually contains everything that is claimed.

I'm writing for an issue of *Journal of Biology* that is about how to make the peer review process easier, but I do believe that papers in our field would be better if referees were allowed and expected (and given time) to look at the primary data, have a copy of the software, use the same annotation indices, and so on, and see whether they can verify the claims and be confident in conclusions that are reached from computational analyses. Even simple reality checks such as comparing replicates (when there are some) are often ignored by both authors and reviewers. I bring this up because one of the major frustrations expressed by a group of around 30 participants at the Computational and Statistical Genomics workshop I attended at the Banff International Research Station last June was the difficulty of reproducing computational analyses in the functional genomics literature. Often, the trail from the primary data to the published dataset is untraceable, let alone the downstream analyses.

Fourth, and finally, the individual validation experiments may not garner much attention, unless they are mentioned in the title, or have appropriate keywords in the abstract. They are rarely as useful as they would be in a paper in which they were explored in more depth and in which the individual hypothesis-driven experiments could be summarized. For instance, a paper we published in *Journal of Biology* in 2004 [3] described an atlas of gene expression in 55 mouse tissues and cell types. Using SVM (Support Vector Machine) cross-validation scores, we found that, for many GO annotation categories, it was possible to predict which genes were in the category, to a degree that is orders of magnitude better than random guessing, although usually still far from perfect. The most interesting aspect of the study to me was the observation that there is a quantitative relationship between gene expression and gene function; not that this was completely unexpected, but it is nice to have experimental evidence to support the generality of one's assumptions. The SVM scores were used mainly to prove the general point, and whether any individual predictions were correct was not the key finding - we knew ahead of time (from the cross-validation results) that most of the individual predictions would not be correct; this is the nature of the business. Nonetheless, final acceptance of the manuscript hinged on our being able to show that the predictions are accurate, so at the request of reviewers and editors, we showed that Pwp1 is involved in rRNA biogenesis, as predicted. According to Google Scholar, this paper now has 139 citations, and my perusal of all of them suggests that neither Pwp1 nor ribosome biogenesis is the topic of any of the citing papers. The vast majority of citations are bioinformatics analyses, reviews, and other genomics and proteomics papers, many of them concerning tissue-specific gene expression. Thus, the initial impact appears primarily to have been the proof-of-principle demonstration of the relationship between gene function and gene expression across organs and cell types, and the microarray data themselves. It is the use of genome-scale data and cross-validation that proves the point, not the individual follow-up experiments.

A small survey of my colleagues suggests that many such examples would be found in a more extensive analysis of the literature in functional genomics and systems biology.

For instance, Jason Moffat explained that in the reviews of his 2006 *Cell* paper describing the RNAi Consortium lentivirus collection [4], which already contained a screen for alteration of the mitotic index in cultured cells, a major objection was that more work was needed to validate the reagents by demonstrating that the screen would also work in primary cell cultures - which may be true, but so far, even the mitotic index screen seems to have served primarily as an example of what one can do with the collection. The paper has clearly had a major impact: it has 161 citations according to Google Scholar, the vast majority of which relate to use of the RNAi reagents, not any of the individual findings in this paper.

To conclude, I would propose that, as authors, reviewers and editors, we should re-evaluate our notion of what parts of genome-scale studies really are interesting to a general audience, and consider carefully which parts of papers prove the points that are being made. It is, of course, important that papers are interesting to read, have some level of independent validation, and a clear connection to biology. But it seems likely that pioneering reagent and data collections, technological advances, and studies proving or refuting common perceptions will continue to be influential and of general interest, judging by citation rates. As erroneous data or poorly founded conclusions could have a proportionally detrimental influence, we should be making an effort to scrutinize more deeply what is really in the primary data, rather than waiting to work with it once it is published. Conversely, the individual 'validation' studies that occupy the nethermost figures, although contributing some human interest, may be a poor investment of resources, making papers unnecessarily long, delaying the entry of valuable reagents and datasets into the public domain, and possibly distracting from the main message of the manuscript.

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
2. Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ: **Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters.** *Nat Genet* 2002, **31:**255-265.
3. Zhang W, Morris QD, Chang R, Shai O, Bakowski MA, Mitsakakis N, Mohammad N, Robinson MD, Zirngibl R, Somogyi E, Laurin N, Eftekharpour E, Sat E, Grigull J, Pan Q, Peng WT, Krogan N, Greenblatt J, Fehlings M, van der Kooy D, Aubin J, Bruneau BG, Rossant J, Blencowe BJ, Frey BJ, Hughes TR: **The functional landscape of mouse gene expression.** *J Biol* 2004, **3:**21.
4. Moffat J, Grueneberg DA, Yang X, Kim SY, Kloepfer AM, Hinkle G, Piqani B, Eisenhaure TM, Luo B, Grenier JK, Carpenter AE, Foo SY, Stewart SA, Stockwell BR, Hacohen N, Hahn WC, Lander ES, Sabatini DM, Root DE: **A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen.** *Cell* 2006, **124:**1283-1298.