Reproducibility of Dressman JCO 2007

VJ Carey

October 26, 2023

In the light of recent high-level challenges to reproducibility of microarray studies (Ioannidis 2009 and others) the dispute between Baggerly, Coombes and Neeley (BCN) and Dressman, Potti and Nevins (DPN) in J Clin Oncology, 2008; 26(7):1186-1187, is of broad interest. But it seems that neither the editors of JCO nor the rebuttalists read the arguments of BCN with much care. In preparing an invited chapter on reproducible research in a forthcoming monograph on cancer bioinformatics, I decided to look closely at the archive generated by BCN at `http://bioinformatics.mdanderson.org/Supplements/ReproRsch-Ovary/` to see if a simple characterization of the dispute, perhaps with resolution, might be possible.

Briefly, the following points need to be understood by those interested in pathway activation and platinum responsiveness in ovarian cancer.

1. As of April 4 2009, the 'corrected RMA' archive at `http://data.cgt.duke.edu/platinum.php` has incorrect ID labeling of most of 119 samples. BCN show that a presumably correct relabeling can be established for 116 samples by finding maximum correlation between the corrected RMA samples and uncorrected RMA samples derivable from the CEL files also posted at the platinum.php site. Note that Dressman's 'corrected RMA' terminology appears to refer to the fact that RMA-based quantifications were corrected for batch effects using sparse factor regression, and not to correction of the published mislabeling problem.

2. BCN show how array run dates can be extracted from CEL files; standard Bioconductor tools facilitate this. Figures 1(a) and 1(b) below, computed independently of the source code of BCN, and based solely on the relabeled 'corrected RMA' quantifications, provide evidence that batch effect-related confounding is present. RPS11, a gene in the Src pathway signature, and survival among platinum nonresponders, have distributions that are systematically related to array batch date. The pattern seen in RPS11 indicates that the sparse factor regression corrections do not completely remove the batch effect.

3. Figure 1(c) is a close approximation to Dressman et al.'s 2007 Figure 2B, and is computed using only the quantifications and survival data published at the platinum.php web site. The only aspect of Figure 1(c) that does not use, and

thereby ‘repeat’, Dressman’s original analysis, is the scoring of pathway activation state of tumors. Neither Bild (2006) nor Dressman trouble to publish their scoring coefficients, a version of which may be easily derived by applying singular value decomposition to Bild’s cell-line data. Figure 1(c) is a strong suggestion that the data and methods used by me and by BCN in their reevaluation of the 2007 article *can* reproduce an important aspect of the results. Indeed, when I computed Figure 1(c) I felt that a vindication of Dressman’s 2007 article might be at hand.

4. Figure 1(d) shows the result of performing the analysis pattern that yielded Figure 1(c) to the E2F3 pathway. Figure 1(d) should yield the association seen in Dressman’s Figure 2C, but it does not. Among platinum non-responsive patients, there is no association between E2F3 activation and survival. However, among platinum-responsive patients, a significant association is seen, as in Figure 1(e). These observations were also made by BCN in their supplementary ovca7.pdf, but with different data sources.

5. By extracting run dates from the original CEL files, strata can be formed to adjust for date-related confounding. Using a parsimonious quadratic model for the effect of run date, the test for a Src pathway effect on survival among platinum non-responders, corrected for confounding, has $p = 0.47$. On the other hand, the same correction in the E2F3 setting, among platinum *responders*, does not substantially alter the association between pathway activation and survival: $p < 10^{-4}$ after adjustment. Similar findings were reported by BCN in ovca7.pdf.

In summary, the published data archives cannot be used to reconstruct key findings in Dressman’s 2007 paper, even when methods are confined strictly to those employed by Dressman et al. The standard for reproduction articulated by DPN in their scathing rebuttal to BCN is readily met (with the exception of pathway activation scoring) for any reanalysis based on ‘corrected RMA’ quantifications, because these quantifications enjoy the sparse factor regression adjustments unique to the Dressman methodology. Thus either the published data archive or the 2007 paper need substantial revision.

Three final remarks. 1) It is important to distinguish between reconstrucibility of quantitative analyses and reproducibility of research findings. Figure 1(c) shows that Dressman’s Figure 2B is *reconstructible*, which is in itself a good thing. The associated inference is probably not *reproducible*, however, because of the confounding: any experiment with similar observational resources possessing different biologically irrelevant relationships among batch, expression, and survival would yield results that are almost surely qualitatively different. It is customary for epidemiologists to check carefully for patterns indicative of confounding in their observational datasets; it must become similarly routine for analysts in genomics. 2) All the data, computations, and graphics on which this letter depends are available in the Bioconductor experimental data archive *dressCheck*, and the Sweave code for this letter is present there as ‘short.Rnw’. Any individual with a copy of R can regenerate, criticize, or reuse any programming underlying this letter. If the problem of reconstruction are due primarily to flaws in the published data archive, analyses underlying this letter can be regenerated with one command, once the data images are revised. 3) Dressman and colleagues are to be commended for making publicly available so much of the data underlying their report. Their analyses are extremely interesting, but it appears technical errors have led – at least – to confusions of subgroup labels. It is clear that the level of scrutiny to which their analysis has been subjected by BCN and by me has not been applied to the vast majority of publications based on genome-scale data analysis, and thus there is a kind of unfairness visited upon those who a) have very interesting findings worthy of further exploration and b) make their data archives available for reanalysis. How to make genome-scale data analysis more reliable and verifiable for the primary investigators is an open question. Increased reliability and verifiability will become necessary as complexity of assays and annotation schemes grows. Baggerly and colleagues show how the criticism of a complex high-level publication can be made transparent and thorough; unfortunately the work involves seven supplementary documents and hundreds of associated primary and derived data files. The *dressCheck* package on which this letter is based establishes less but does so in a more concise manner – Figure 1(c) for example is computed from the primary data with 11 lines of R code.

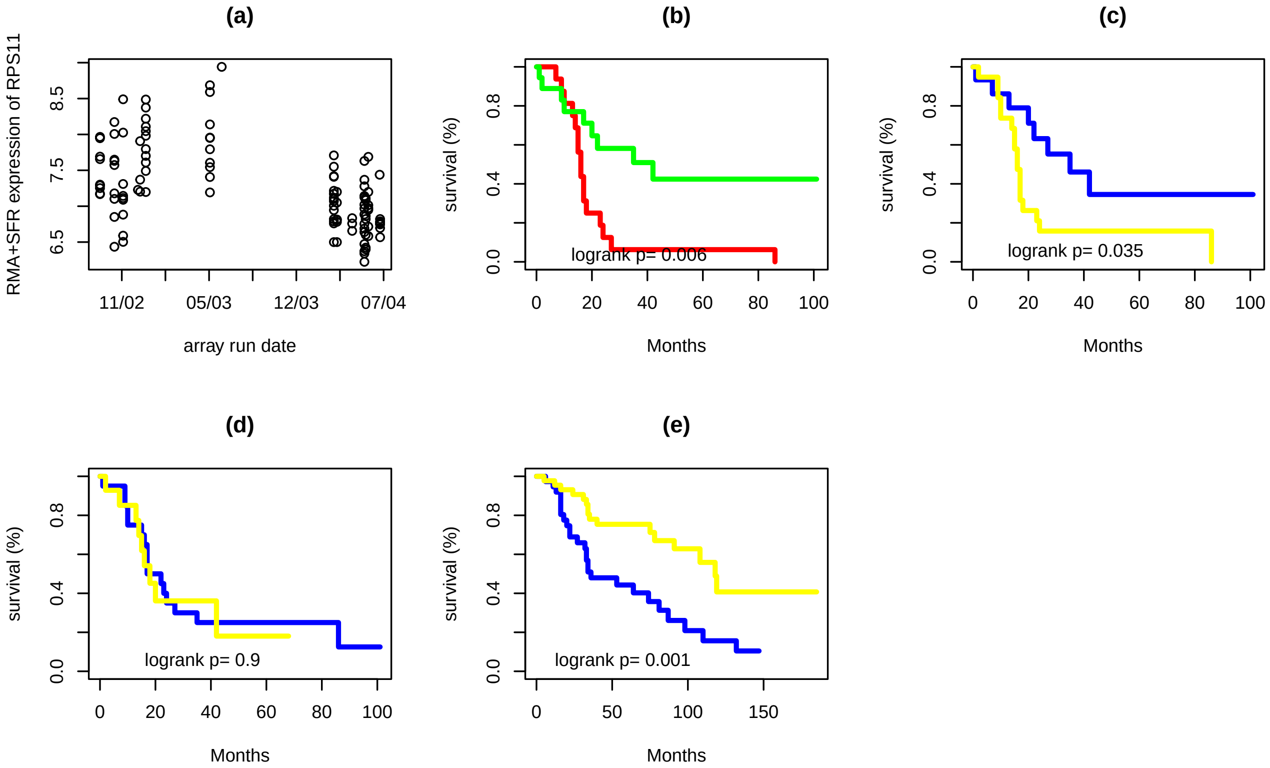

Figure 1: (a) Variation in expression of RPS11 over array preparation dates. (b) Survival distributions for early (red) and later (green) array batches among platinum non-responders. (c) Association between Src pathway activation and survival among platinum non-responders. (d) Association between E2F3 pathway activation and survival among platinum non-responders. (e) As (d) but for platinum responders. For Kaplan-Meier graphs (c-e), blue line is for low pathway activation score, yellow line for high.