# Reproducibility of Dressman JCO 2007

VJ Carey

July 20, 2010

In the light of recent high-level challenges to reproducibility of microarray studies (Ioannidis 2009 and others) the dispute between Baggerly and Coombes (BC) 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 BC 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 BC at `http://bioinformatics.mdanderson.org/Supplements/ReproRsch-Ovary/` to see if a simple characterization of the dispute, perhaps with resolution, might be possible.

The headline given to BC's letter was "Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer". This shows a focus on the confounding of array batch with gene expression measures and survival – see Figures 1(a) and (b) below, which respectively show a clear association, nonlinear in form, between mean RPS11 expression and run date, and a difference in survival distributions for early and late batches among platinum non-responders. The expression measures are the ones published by Dressman at the supplementary data web site for the 2007 paper, and thus employ the sparse factor regression corrections that are asserted by DPN to remove the confounding. Confounding happens. DPN are within their rights to hope that statistical modeling will remove it, but BC (and my Figure 1(a)) show that residual confounding was present after the sparse factor regression corrections. BC show that specific forms of adjusting for batch effects eliminate the significance of association between, e.g., Src activation and survival among platinum non-responsive patients; I will examine this more closely below. Because we are not sure what the absolutely right adjustment for confounding would be, the terminology of the title ("batch effects potentially compromise usefulness") may be called for.

BC have many more arguments, however, spread out over seven supplementary documents and hundreds of primary and derived data files. These arguments do not concern batch effects and confounding, but instead describe complete non-reproducibility of findings reported in Dressman's original paper. DPN assert that BC have no right to dispute reproducibility because BC do not "repeat" Dressman's original methods. The fact is that BC do "repeat" Dressman's original methods with one exception – derivation of coefficients for pathway activation scoring, which is such a simple task that neither Bild

nor Dressman troubled to publish their specific coefficients. Figure 1(c) shows that it is possible to obtain a very close approximation to Dressman's Figure 2B using the data from the supplemental web site, coupled with a simple linear algebraic operation to derive pathway scoring coefficients from Bild's cell-line archive.

When I obtained Figure 1(c) (BC have a similar display at p.5 of supplemental document ovca7.pdf, but based on RMA alone) I felt there might be some vindication at hand for Dressman's original paper. I then used the same methods in an attempt to reconstruct Figure 2C, involving the E2F3 pathway. Figure 1(d) shows no association between E2F3 and survival among platinum non-responders; Figure 1(e) shows an association but only among platinum responders. These are qualitatively similar to findings reported by BC, p51 of ovca7.pdf.

By using Dressman's posted quantifications, BC (and I) *do* meet DPN's standard for reproducing Dressman's original analyses. And part of Dressman's original publication is reproducible. But other parts are not. One important part that was reconstructed (Src pathway activation association with survival in platinum non-responders) appears to be affected by residual confounding. When a parametric (Weibull) model relating survival time to Src pathway activation is elaborated with a quadratic term (2 d.f.) for calendar time of array batch, the significance of the pathway effect is altered from an initial $p = 0.035$ to an adjusted $p = 0.47$. It may be of interest to note that in my reanalysis the effect of E2F3 activation remains significant ($p < 10^{-4}$) after the quadratic adjustment for calendar time of array batch, but this is only so for the platinum responders. In this group the test by Dressman was reported to be completely insignificant (Figure 2E of the original paper.)

There is no question that reasonable investigation of the data published in conjunction with the 2007 paper does *not* allow even approximate reconstruction of some of the key points of the paper. Clerical errors in the published archive have already been admitted by DPN and some may remain. If the findings of this letter are dependent on mistakes in the archive used, the investigations of this letter can be repeated and the non-reconstructibility claim can be reevaluated. However, any thorough re-evaluation of this study must consider both the problem of reconstructibility (really a computational concern regarding soundness of data representations and availability of analysis algorithms) and the problem of substantive reproducibility (an inferential condition satisfied by studies that will, when conducted independently in other settings, yield scientifically consistent interpretations). The confounding unearthed by BC is a special characteristic of the datasets generated by Dressman et al., and if batch effects are playing a role in the apparent association of survival and Src activation, the inferences are extremely likely to be non-reproducible in the inferential sense, because no validating experiment will be likely to possess the same relationships between batch, expression and survival as found in the Dressman data.

Dressman and colleagues are to be commended for making so much information related to their paper available in the supplemental data archives. Their harsh response to BC, which includes the claim that BC's analysis is "egregiously flawed", is completely

uncalled for. Many of the analyses conducted by BC were undertaken charitably, to find patterns that are consistent with the claims of the paper but that were not reconstructible in detail when the avenues suggested by the paper were first explored. Furthermore, much of the analytic effort expended by BC (and independently by me) relates to properly correcting the clerical errors that remain in the online supplements to this day.

How to improve the reliability of highly complex analyses of genome-scale data remains an open question. When certain choices of data representation and software interface are made, complex analyses involving multiple data sources can be expressed in a small number of highly platform-independent scripts. To illustrate this idea, this letter and all related data, calculations and graphics, can be obtained and concretely reconstructed from conventionally organized source codes using the Bioconductor package *dressCheck*, part of the Experimental data series at www.bioconductor.org. This approach exposes my work to scrutiny by external analysts, who can see and criticize, correct, or reuse every computation that leads to each graphic and $p$-value cited in this letter. BC exposed themselves to the same level of scrutiny in their supplemental archive. The work they have done to help clarify the role of pathway activation in platinum non-responsive tumors, and to show how to conduct careful and verifiable criticism of high-level publications involving genome-scale data analysis, is deserving of thanks, not approbation.

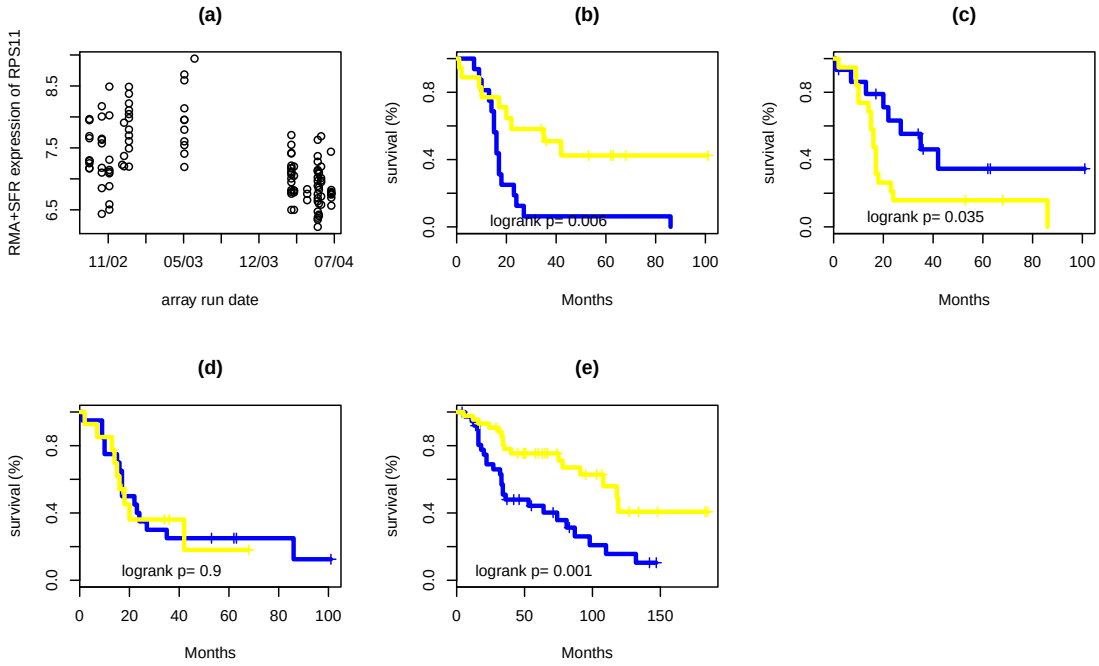


Figure 1: (a) Variation in expression of RPS11 over array preparation dates. (b) Survival distributions for early (blue) and later (yellow) array batches among platinum non-responders. (c) Association between Src pathway activation and survival among platinum non-responders. (d) Association between E2F3 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.