Response to “Safikhani *et al.* reply”.

Paul Geeleher, Eric R. Gamazon, R. Stephanie Huang

Abstract:

Here, we respond to Haibe-Kains/Quackenbush's response to our Brief Communication Arising (BCA) “Consistency in large pharmacogenomic studies” (Nature 2016). Our BCA highlighted the key shortcoming of their manuscript “Inconsistency in Large Pharmacogenomic Studies” (Nature 2013). One of their main counter-arguments, which they have made repeatedly, was that in our BCA we have only shown consistency for nilotinib, and that nilotinib is an isolated example. Both of these claims are false and clearly contradictory to the contents of our manuscript (see “Response 1” below). They also argue that our ideas are “not new” and that they themselves have already discussed these ideas in a publication. This is very misleading. Indeed, the manuscript they reference to support this—a preprint that they posted to bioRxiv—was posted a year and a half after they had already seen our manuscript during the 2 and a half year peer-review process (see “Response 7” and “Timeline”). Below, we include a point-by-point response to their response to our BCA.

Note: While Haibe-Kains/Quackenbush first saw our paper on March 12th 2014 and were afforded multiple opportunities to comment on it during the peer review process, we were not given an opportunity to comment upon or critique their response to us. Thus, we are using this platform to discuss their response.

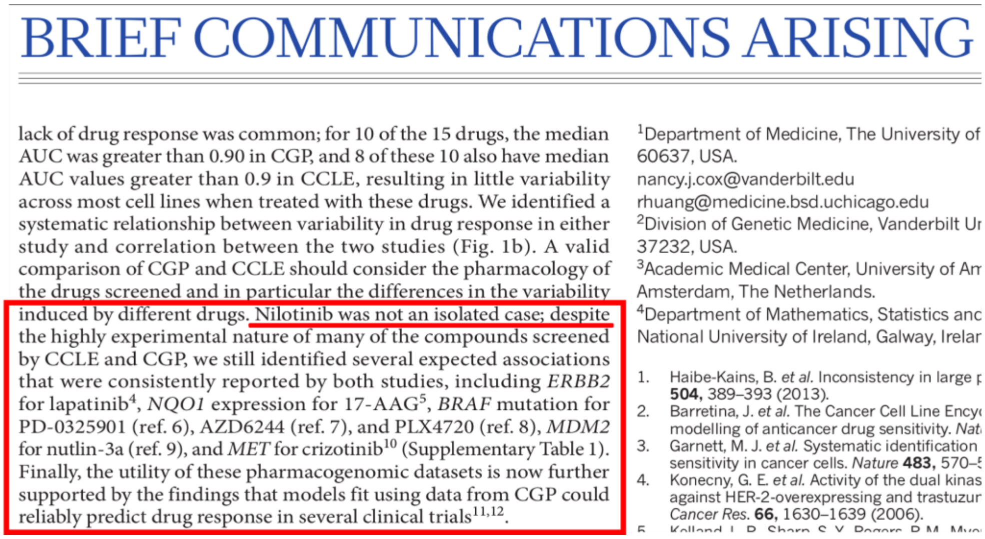
Haibe-Kains/Quackenbush 1:

In the accompanying Comment[1], Geeleher et al. claim to have discovered overall consistency between the Cancer Genome Project (CGP)[2] and the Cancer Cell Line Encyclopedia (CCLE)[3] by analysing the response of three cell lines containing the BCR-ABL1 fusion gene to the highly targeted drug nilotinib (Supplementary Fig. 1).

Response 1:

This is a false argument by the authors. Nowhere did we state that reproducibility for nilotinib (or “three cell lines”) implies general reproducibility between the two studies. In fact, we used the nilotinib example for the *sole* purpose of clearly demonstrating why the Spearman correlation metric employed by the authors in their original Nature publication is confounded by variability in drug sensitivity (Figure 1 in our BCA). The example has *nothing* to do with demonstrating overall consistency between the two studies.

As shown in our manuscript (Supplementary Table 1), a canonical drug target was identified in both CCLE and CGP for 9 of 15 drugs, and for 14 of 15 drugs in at least one study, indicating that the authors' flawed approach (as illustrated by the nilotinib example) is not an isolated case. Unfortunately, the authors still insist that nilotinib is an “isolated example”, when it is clearly demonstrated in our text that it is not (see Figure below), and given the fact that they were repeatedly made aware of this during the peer-review of our manuscript.



BRIEF COMMUNICATIONS ARISING

lack of drug response was common; for 10 of the 15 drugs, the median AUC was greater than 0.90 in CGP, and 8 of these 10 also have median AUC values greater than 0.9 in CCLE, resulting in little variability across most cell lines when treated with these drugs. We identified a systematic relationship between variability in drug response in either study and correlation between the two studies (Fig. 1b). A valid comparison of CGP and CCLE should consider the pharmacology of the drugs screened and in particular the differences in the variability induced by different drugs. Nilotinib was not an isolated case; despite the highly experimental nature of many of the compounds screened by CCLE and CGP, we still identified several expected associations that were consistently reported by both studies, including *ERBB2* for lapatinib[4], *NQO1* expression for 17-AAG[5], *BRAF* mutation for PD-0325901 (ref. 6), AZD6244 (ref. 7), and PLX4720 (ref. 8), *MDM2* for nutlin-3a (ref. 9), and *MET* for crizotinib[10] (Supplementary Table 1). Finally, the utility of these pharmacogenomic datasets is now further supported by the findings that models fit using data from CGP could reliably predict drug response in several clinical trials[11,12].

[1]Department of Medicine, The University of
60637, USA.
nancy.j.cox@vanderbilt.edu
rhuang@medicine.bsd.uchicago.edu
[2]Division of Genetic Medicine, Vanderbilt U
37232, USA.
[3]Academic Medical Center, University of Am
Amsterdam, The Netherlands.
[4]Department of Mathematics, Statistics and
National University of Ireland, Galway, Irela

1. Haibe-Kains, B. *et al.* Inconsistency in large p **504,** 389–393 (2013).
2. Barretina, J. *et al.* The Cancer Cell Line Ency modelling of anticancer drug sensitivity. *Nat*
3. Garnett, M. J. *et al.* Systematic identification sensitivity in cancer cells. *Nature* **483,** 570–5
4. Konecny, G. E. *et al.* Activity of the dual kinas against HER-2-overexpressing and trastuzu *Cancer Res.* **66,** 1630–1639 (2006).

<u>*Haibe-Kains/Quackenbush 2:*</u>

They use this example to argue that owing to the targeted nature of many of the 15 drugs screened in both CGP and CCLE (Supplementary Table 1), the rest of the drugs should show consistent sensitivity measurements if analysis is limited to the highly sensitive cell lines.

<u>Response 2:</u>

We did not make this argument (see Response 1). To reiterate, we used the nilotinib example only to illustrate why the authors' chosen metric for testing consistency between the two studies is fundamentally flawed. Because of the targeted nature of many of the drugs (which by design require specific molecular targets for response and which the authors' original paper did not take into account), the authors approach fails for these drugs as well. We did not, however, draw the conclusion, from this demonstration of the authors' fundamentally flawed approach, that the drugs *should* show consistent sensitivity measurements. Again, the results for the other drugs were included in Supplementary Table 1 (again, see Figure above).

<u>*Haibe-Kains/Quackenbush 3:*</u>

However, as we describe in more detail below, the consistency seen in the sensitivities of cell lines to nilotinib is an isolated example that does not generalize to other targeted drugs, supporting our initial finding[4] *of a broader inconsistency in reported phenotypes between CGP and CCLE*[4]*.*

<u>Response 3:</u>

Again, we did not claim that this is the case (see Responses 1 and 2).

<u>*Haibe-Kains/Quackenbush 4:*</u>

Geeleher and colleagues[1] *raise two potential issues with our published study*[4] *and we welcome the opportunity to address them here. First, in our initial study, we computed the correlation of gene expression and mutation profiles between cell lines to assess whether large transcriptomic changes and/or genetic drift might be the cause of the observed inconsistency in drug sensitivity data (which was calculated across cell lines)*[4]*. We agree with Geeleher et al.*[1] *that correlations 'across' and 'between' cell lines should be compared in a consistent manner, as was done in our recent re-analysis study*[5]*. We found that overall correlation across cell lines is lower than correlation between cell lines (Supplementary Fig. 2).*

<u>Response 4:</u>

While not the fatal flaw in Haibe-Kains *et al.*, the reason we highlighted this problem was to undermine the **ad-hoc** definitions of "consistency" used by the authors in their original study and to highlight the completely arbitrary nature of such qualitative definitions of statistical measures of concordance. We appreciate that this "mistake" has now been acknowledged. The "re-analysis" that is being referenced here in regard to "between" and "across" sample correlations was posted by the authors to the bioRxiv a year and a half after they first saw our manuscript as part of the ongoing peer review process of our manuscript. More details on this bioRxiv paper are below (see "Response 7" below)…

<u>*Haibe-Kains/Quackenbush 5:*</u>

However, gene expression data are significantly more concordant between studies than the drug response summary statistics (half-maximum inhibitory concentration (IC_{50}) and area under the curve (AUC)) values in all comparisons (Wilcoxon rank sum test $P < 0.002$).

<u>Response 5:</u>

This comparison, using a Wilcoxon ranksum test, is meaningless because the analysis is still ignoring the differences in variability between these two types of data, i.e. the fundamental flaw in the analysis being that highly targeted cancer drugs induce very little variability in drug response across a panel of cancer cell lines, which causes lower correlation. A key point is that even if correlations were vastly higher for gene expression than for drug sensitivity data (which they are not after correcting the "mistake"), it still could not be concluded that gene expression data are more concordant than drug sensitivity data.

Haibe-Kains/Quackenbush 6:

Consequently, our original conclusion that gene expression data are significantly more correlated than pharmacological response data still holds. In fact, the lower correlation of gene expression values across cell lines could suggest that there is even less consistency between the CCLE and CGP studies than we initially reported.

Response 6:

There is no basis to conclude from this that gene expression data are now less concordant than originally thought. Again it seem the authors have missed the main point of our argument. Correlation between two matrices can be calculated in two different directions, and in both cases correlations are higher between samples than across samples. The between and across sample correlations are exactly the same as they have always been, as are the levels of biological variability between and across samples. There are many examples in the literature where these correlations have been reported for repeated measures of gene expression data (e.g. PMID20220758) and between sample correlations are consistently much higher than across sample correlations, and in fact, the across samples correlations for CGP/CCLE are actually much higher than those reported in many previous studies (but it needs to be clear that because the differences in biological variability in the different datasets have not been taken into account, it still does not mean that the data are or are not more consistent).

Haibe-Kains/Quackenbush 7:

Second, the argument of Geeleher et al.[1] that the lack of variability in drug sensitivity measurements may complicate biologically meaningful assessment of concordance between pharmacogenomic datasets is not new, as it was already discussed by our group[5] and others[6, 7].

Response 7:

We are concerned about the authors' (mis)use of our paper during the peer review process without proper acknowledgment (such as in a YouTube presentation and in a posted bioRxiv preprint; see "TIMELINE" section at the end). We are doubly concerned about the subsequent use of the bioRxiv preprint in the authors' final response to our Brief Communication Arising (BCA) to undermine the novelty of our findings.

The bioRxiv paper that is being referenced here (reference 4) supporting the statement that our idea was "not new" was posted by Haibe-Kains/Quackenbush on September 6th 2015. As is procedure for a BCA, we were informed by Nature editor Barbara Marte that our manuscript had been sent to Haibe-Kains/Quackenbush on April 4th 2014 (email transcripts available) - less than 4 months after the publication of Haibe-Kains *et al.* (19th December 2013). Thus, the bioXiv paper that is being cited, supporting this claim that our idea is "not new", was posted **a year and half after the authors had already seen our paper**. This bioRxiv paper has also been cited by the authors to support their claim that they have since considered the difference between "between" and "across" sample correlation (see section "*Haibe-Kains/Quackenbush 4"* above). Again these were posted during the peer review process of our manuscript, and long after the authors had read our paper as part of the peer-review process.

Interestingly, during the original rounds of peer review of our manuscript, where Haibe-Kains/Quackebush corresponded with us and Nature about the content of our paper, they did not attempt to take credit for any of the ideas in our manuscript, the ideas they are now claiming by referencing in their later bioRxiv paper. During review, there was no mention of them already having discussed between/across sample correlations, or that our ideas about variability confounding correlation were “not new”, or any external references to support our ideas being “not new” (see “TIMELINE” below for details). Indeed, these two claims and reference to the bioRxiv paper never appeared anywhere during the review process and were only added to the very last draft of their response, which was emailed to us by the Nature editors on September 6th 2016. We had seen one previous draft, but were not given an opportunity to comment on it. If in fact our ideas were “not new”, it seems unusual that Haibe-Kains/Quackenbush would not have raised this point during the review of our manuscript, as that clearly would have meant that our manuscript did not satisfy the criteria for publication in Nature. We have kept copies of all correspondence between us, Haibe-Kains/Quackenbush and the Nature editors.

Furthermore, there is a YouTube video (https://www.youtube.com/watch?v=_1IKAld1YwY, posted February 27th 2015 - almost one year after seeing our paper) of Benjamin Haibe-Kains presenting work to a Toronto Bioinformatics User Group, which contains reference to (18 mins) between and across sample correlation (the first point we make in our paper) and how they had made a “mistake”. At 21 mins 25 seconds he states "Some people argue that correlation is a bad measure of concordance for these drugs...". We encourage Dr. Haibe-Kains to read the guidelines provided on the Nature website as regards work that is in peer review “Editors, authors and reviewers are required to keep confidential all details of the editorial and peer review process on submitted manuscripts.” (source: http://www.nature.com/authors/policies/confidentiality.html). We have downloaded a copy of this YouTube video and it is available upon request, if the link posted above should stop working.

Haibe-Kains/Quackenbush 8:

Geeleher et al.[1] focus on the sole example of nilotinib, for which there are three highly sensitive cell lines out of the 200 cell lines screened in both datasets.

Response 8:

Again: No we don’t (see Responses 1, 2 and 3).

Haibe-Kains/Quackenbush 9:

However, even among these three cell lines, the AUC values are not concordant; the least sensitive of the three cell lines in CGP is actually the most sensitive one in CCLE (Supplementary Fig. 1 and Supplementary Information).

Response 9:

This statement could be easily misinterpreted, so to clarify: we reported in our BCA that, of 189 cell lines treated with nilotinib, shared between CCLE and CGP, that the three cell lines harboring the BCR-ABL1 drug target are the three most sensitive samples in both studies. To clarify for the reader, Haibe-Kains have pointed out here that the cell line ranked most sensitive to nilotinib in CGP is 3rd most sensitive in CCLE and the cell line most sensitive in CCLE is 3rd most sensitive in CGP (out of 200 cell lines total). While the interpretation of concordance/discordance is subjective, we do not agree that this even comes close to representing discordance.

Haibe-Kains/Quackenbush 10:

Therefore, the only way to consider these results to be concordant is to classify these three cell lines as sensitive and the remainder as resistant, which cannot easily be done using the waterfall method described in the CCLE study[3, 4].

Response 10:

If the three (of 189) cell lines that contain the canonical drug target are the three most sensitive cell lines in both studies, then the results are obviously almost as concordant as they could ever possibly be under any drug sensitivity assay. There is no need for a “Waterfall approach” or to “classify these cell lines as sensitive…” or any similar approach to reach this obvious conclusion; indeed this conclusion could have been reached using data from either study and the standard methods employed by either study (e.g. MANOVA, ElasticNet Regression - see “Response 16”).

Haibe-Kains/Quackenbush 11:

The authors claim that nilotinib is not an isolated example[1], but do not seem to propose any statistic to quantify the consistency of the other drugs.

Response 11:

Again, we produced data and results (Supplementary Table 1 included in our BCA) to show that it is not an isolated example (again see Responses 1, 2, 3 and 8). Again, as stated in our paper, a canonical drug target with pre-existing literature support was identified for 14 of the 15 drugs in at least one study, and in both studies for 9 drugs, despite the experimental nature of over half of these compounds. So again, the association of nilotinib with its canonical target (BCR-ABL) is not an “isolated example”, despite the fact that the authors repeatedly make this false claim.

Haibe-Kains/Quackenbush 12:

We adapted the Matthews correlation coefficient8 (AMCC; see Supplementary Methods) to select the optimal cutoff for consistency between drug sensitivity calls, in which only a few cell lines may be sensitive, or between gene expression data, in which the gene of interest may be rarely expressed. As expected, nilotinib yielded an AMCC value of 1, which denotes perfect consistency between the two studies (Supplementary Fig. 3). However, the other drugs yielded much lower AMCC values, with only AZD0530, lapatinib and critozinib yielding AMCC values of around 0.65, with another five drugs (17-AAG, AZD6244, erlotinib, PD-0325901 and PLX4720) yielding moderate consistency ($0.5 \leq$ AMCC < 0.6), and the rest of the drugs yielding poor consistency (AMCC < 0.5). Notably, there was no systematic relationship between variability in drug sensitivity and AMCC estimates (Supplementary Figs 4 and 5), suggesting that AMCC values, although potentially overoptimistic, are a more appropriate statistic for consistency than Spearman’s or Pearson’s correlation coefficients. It should be noted that the inter-laboratory replicates of the measurements of camptothecin and AZD6482 sensitivity performed using the same experimental protocols at two different locations within CGP yielded AMCC values of only 0.55 and 0.41 (Supplementary Fig. 6), indicating a lack of reproducibility of drug phenotype measures between biological replicates. Consistent with our previous report, gene expression data yielded significantly larger AMCC than drug sensitivity data across cell lines (Wilcoxon rank sum test $P < 0.006$; Supplementary Fig. 7). This re-analysis confirms that nilotinib is an anecdotal case, and that drug sensitivity measurements for the rest of the drugs (cytotoxic or targeted) remain only poorly to moderately consistent.

Response 12:

Here, the authors have seemingly devised a new metric they call the Adapted Matthews Correlation Coefficient (AMCC) and decided that “poor consistency” is now defined as “AMCC < 0.5”. Why? There is no basis for defining the numbers calculated here as “consistent”, “inconsistent” or anything else. As in the original study, the definitions provided are arbitrary, thus these results are meaningless.

Haibe-Kains/Quackenbush 13:

Geeleher et al.[1] also state that their previous findings[9] support the utility of the data from the CCLE and CGP, and suggest that their findings provide evidence for a consistency between CGP and CCLE.

Response 13:

We cited two studies in clinical data (one by us, one from a completely independent group) as examples of the utility of the data, which had seemingly been undermined by the publication Haibe-Kains *et al.* The example has *nothing* to do with demonstrating consistency between CCLE and CGP.

Haibe-Kains/Quackenbush 14:

However, a true test of this assertion would be to train their models on the CGP and to use these to predict phenotypes reported by CCLE (and vice versa). If they could predict the drug response phenotype in the independent validation set with high accuracy, this would provide some quantitative evidence of a consistency between the two datasets in the context of their predictive models[1]. However, we[10] and others[11, 12, 13] have shown that such an analysis does not yield robust predictions for most drugs.

Response 14:

In their own earlier study published in JAMIA (http://www.ncbi.nlm.nih.gov/pubmed/23355484, cited above), Haibe-Kains and co-authors actually conclude (in the abstract) that “These results suggest that genomic predictors could be robustly validated for specific drugs.” Here, they cite this study here to support the opposite of this statement.

Furthermore, we suggest that comparing predicted drug sensitivity data, as the authors did in their JAMIA paper, will suffer from exactly the same biases we described when comparing measured drug sensitivity from different studies (as they have done in their Nature paper). Future studies using the approach employed in the JAMIA paper should also absolutely consider differences in variability induced by different types of drugs and in particular the lack of biological variability induced by highly targeted drugs. Our BCA was, however, about the flawed approach reported in the Nature paper (https://www.ncbi.nlm.nih.gov/pubmed/24284626, by some of the authors).

Haibe-Kains/Quackenbush 15:

In our original report[4], we found statistically significant non-zero correlations between phenotype measurements for almost all drugs, supporting the fact that there is biologically relevant signal in these datasets, albeit confounded by significant noise. Concurring with recent reports[6, 7, 14], we identified known gene–drug associations that are reproducible between CGP and CCLE (Supplementary Information); however, half of the known associations were inconsistent across datasets (significant in only one dataset; Supplementary Table 2). In a recent re-analysis of the updated CGP and CCLE datasets, we reported that the discovery of new, potentially weaker biomarkers, which was the main goal of the CGP and CCLE studies, was much more challenging owing to inconsistency in pharmacological phenotypes[5]. In our original conclusions[4], we argued that additional work is necessary to improve the consistency of phenotypic measures with the ultimate goal of making data from these large-scale projects more useful for development of robust predictors of drug response, and we believe that this conclusion still holds upon our re-analysis. We and others are actively working on identifying stable measures that could lead to improved consistency across inter-laboratory experiments.

Response 15:

Our interpretation is that the two studies are nowhere near as inconsistent as Haibe-Kains *et al.* have led readers to believe.

Haibe-Kains/Quackenbush 16:
Like Geeleher et al.[1], we originally hoped to use the CCLE and CGP to develop robust biomarkers that could predict responses to treatment. Although we could use methods to find some consistency in selected subsets of the data, we found no general methods that identified an overall consistency between the studies. Geeleher et al.[1] have shown that if the drug sensitivity data are appropriately discretized, consistency can be found for nilotinib—but not for the other compounds tested in both studies. Despite the single example of nilotinib, we conclude as we did originally, that sensitivity phenotypes lack consistency for most of the drugs screened both in CGP and CCLE.

Response 16:
Again, this statement is untrue. Again, we do not say anywhere (explicitly or even implicitly) that consistent results for nilotinib imply consistent results for the entire study and our paper clearly reports consistent results across a majority of drugs. Also, nowhere have we suggested dichotomizing or discretizing data as seems to be suggested here; these words (or any synonyms of them) do not even appear anywhere in our text. The results showing consistent findings for the majority of drugs for the two studies, included in Supplementary Table 1 of our BCA, were generated by CGP using MANOVA and those reported by CCLE using ElasticNet regression coupled with a bootstrap procedure; neither of which involve discretizing/dichotomizing data.

TIMELINE

Finally, because of the vast time lag between submission of our manuscript and its publication, we include here details of the review process and the timelines involved, as it may be of interest to some readers.

11/7/13 Haibe-Kains *et al.* released on Nature's website.

12/19/13 Publication date of Haibe-Kains *et al.*

3/12/14 We complete our manuscript and send to Haibe-Kains *et al.* for pre-Nature submission communication (as required by Nature). We received the acknowledgement of receipt from Haibe-Kains on the same day, but did not receive further correspondence.

3/28/14 We submitted to Nature.

4/11/14 We received request for code from Nature editor.

4/18/14 We emailed original code to Nature editor.

6/20/14 We received first round of response from Haibe-Kains (agreed to the majority of our criticism), including to agreeing that correlations were reported inconsistently for drug sensitivity and gene expression data, and that the differences in variability for different drugs was problematic. There was no mention of these ideas being “not new” or having been previously discussed by them or any other group (because they were new and had not been discussed by any other group).

7/17/14 We submitted the first round of revised manuscript to Nature.

10/12/14 We received email from Nature editor, stating they are awaiting for revised formal response from Haibe-Kains et al and once received, the exchange will be send for formal peer review.

11/12/14 We received email notice stating that Haibe-Kains had never provided a revised reply and made additional requests on our manuscript.

11/18/14 We submitted our second rounds of revision and response to all additional requests to Nature.

2/10/15 Change of Nature editor. Notified that Nature is awaiting for original Nature authors’ response.

3/28/15 One year has passed since our initial submission.

5/28/15 Our manuscript is sent out for external review.

7/14/15 External review is completed. Original authors submit their response as “confidential comments”.

8/7/15 We submitted our response to external review.

9/6/16 HK/Q post their “Revisiting inconsistency in large pharmacogenomic studies” to bioRxiv. The paper is tweeted many times. We notify the editors of Nature that the paper has been posted and that some of its contents bare an alarming similarity to the contents of our manuscript. We also find the YouTube video discussed above and include this in our email. We receive a reply stating that “Thank you for informing me about the potential disclosure of your information in these outside venues. That obviously should not be happening, and I will consult with the chief biology editor about what should be done. ”, however we never receive further communication on the matter.

12/4/2015 Nature notify us that they are “prepared to accept the BCA for publication”.

3/28/16 Two years have passed since our initial submission.

12/11/15 First time seeing HK/Q response to our paper when it appears on the Nature submission portal and we are able to download and view it. We are not informed of this fact or given any opportunity to challenge any of the points therein. In this version of their response, there is no mention of our ideas being “not new”, nor have they attempted to claim any of the ideas in our paper by citing their own bioRxiv paper.

9/6/16 Second time seeing HK/Q response from Nature editor prior to copy editing. **The sentences claiming that our ideas are “not new” and the accompanying citations to their bioRxiv paper, claiming our ideas, have now been included in their response.**

9/12/16 Nature formally accept our BCA, two and half years after its initial submission.

30/11/16 Our BCA is published online.