Reviewer's report

Title: Novel pharmacogenomic markers associated with paclitaxel response in cancer

Version: 1 Date: 12 October 2010

Reviewer: Daniel Hertz

Reviewer's report:

Major Compulsory Revisions:
1) Top paragraph on page 7 in the statistical methods says that cell lines missing 90% of their genotype data will be excluded. This is a very low cutoff. Typically if you were genotyping and had less than 80% successful genotype calls for a sample you would throw it away. In this case you have set your cutoff as 10% Successful calls. Why was this threshold chosen? How many cell lines would be thrown away if you changed to 80% genotype calls? Is there an explanation for why call rates would be less than 50-60% for a cell line? If so, can we still believe the other 50% that were successfully called? Finally, the 4 cell lines that are labeled as "N" for genotype data in additional table 1, which are not included in any analyses, are those lines with greater than 90% missing data or were those lines not genotyped at all?

2) There is no mention of gene expression data normalization, which is especially concerning because the sensitive lines had higher expression in every significant comparison. Since there are only 8 resistant lines this could be due to differences in experimental conditions if the data were never normalized before analysis. I assume they were, but it should be mentioned explicitly.

3) In every case the A1 allele is found more commonly in the resistant (F_A) than the sensitive (F_U) cell lines. Is this a consequence of the system that was used to determine which allele would be A1 and which would be A2? Otherwise, is there an explanation for this?

4) On page 19 you briefly mention a concern that requires much more inspection. DCT is a marker of melanoma and all the melanoma cell lines are in the sensitive group. This is also the case with all of the CNS cancer, leukemias, and prostate cancers. If any of the genetic markers are related to these cancers you will have the same issue. Additionally, the resistant cancers have an overrepresentation of NSCL and renal cancer. Is it possible that there is a confounding variable (such as race) that is being ignored? For instance, if there is a racial predisposition to renal cancer, and the renal cancer cells are inherently more resistant to paclitaxel, then variants that are more common in that ethnic group would be identified in this analysis. These two (related) concerns need to be addressed in a more up-front way in the discussion section, unless there was consideration of this built into the analysis and I didn’t see it.

Minor Essential Revisions:
1) Many times in this manuscript the phrase “therapeutic response” is used or the phrasing indicates that there is clinical/therapeutic response data. This should be avoided, particularly in the title. This study does not include “paclitaxel response in cancer” data, so it should not be titled to indicate that it does. Other instances where a change is warranted are the abstract (“therapeutic response”), pg 5 (“therapeutic response”), pg 13 (“paclitaxel response”) and pg 21 (“therapeutic response”). These could be changed to “paclitaxel sensitivity,” “cellular response,” “in vitro response” or other more accurate phrases. The last sentence of the manuscript (pg 21) states that these could be used for future mechanistic studies on “paclitaxel response” when they should probably be looked at as candidates for future studies into understanding paclitaxel mechanism.

2) The FDR chosen (<0.005) is far more stringent than that typically employed (FDR<0.1 or <0.2 I see regularly). I commend the authors for doing so, but would like to know if this decision was made a priori, and if so why was this level chosen, or if it was chosen after the investigators saw how many SNPs met this criteria.

3) The abstract conclusion states that the haplotype analyses highlight the role of SNP-SNP interactions. There were no SNP-SNP interactions tested for, so this claim is very speculative. The sentence should reflect this speculation and/or include alternative hypotheses such as the true causative genetic factor not having been identified.

4) The first sentence in the background (pg 4), the FDA is the Food and Drug Administration.

5) Second sentence of second paragraph of background is incorrect (“However…communication”). There are two taxanes (paclitaxel and docetaxel) and both cause microtubule stabilization. There are other microtubule-targeting drugs (such as Vincas) which cause microtubule instability. The sentence could be corrected by changing “other taxanes” to “other microtubule-targeting drugs” unless the author can cite a paper that demonstrates docetaxel causing microtubule instability.

6) The gene expression methods section states that the p value used was 0.05 even though there were 20 separate comparisons (8 independent comparisons since multiple probes in one gene are not independent). Regardless, the methods should state that this is a confirmatory or exploratory analysis so multiple comparisons correction was not applied. The haplotype analysis should also mention the exploratory nature of this process since many haplotypes were generated and tested to see if they outperformed the single SNP association.

7) The X-axis label of Figure 1 should be changed. The label implies that the value was measured at the 10^-6.0 M dose, when the value was actually extrapolated from an experiment that started at this dose and tested a range of doses. It would be more accurate to just label it “Normalized GI50 values.” Similarly, on page 14 in the “relevance to paclitaxel…” section it should say that the data was extrapolated from a range of doses which started at 10^-6 M, instead of “The GI50 response data at a 10^-6 M dose.”
8) Figure 1 (I think) and the additional file 1 (definitely) include 62 cell lines. You should state that the NCI60 panel actually includes 62 lines, if that’s the case. Also, the methods should make clear that the kernel density test was done on the entire 62 line population, and the cell lines without genotyping data were removed after the classification procedure. In the text for the tables and figures you should state that all 62 lines are included when they were.

9) Was there any follow-up on the 33 non-protein-coding SNPs that were significant in the GWAS. For example, did you look for LD with the 10 SNPs that were in protein-coding genes, did you look at whether they were proximal to other genes that were important in the IPA analysis etc. One of them (2836880) was found in the LD block with PTPRD SNP 2836025. This should be mentioned in the discussion and used as an example of a SNP that by GWAS analysis alone it was unclear how it would be related to paclitaxel sensitivity, but looking at haplotypes explains its relationship with the system (through PTPRD).

10) In the “impact of significant variants…” section (pg 14) you may want to mention that the FastSNP program correctly predicted no effect for 2 SNPs as well as correctly predicting the effect for the 4 expression enhancers. I was quite impressed by the performance of the algorithm from what you presented.

11) Page 20 you should mention if there was any overlap between the findings of Park et al, and if not, why your results didn’t confirm their work.

Discretionary Revisions:

1) Typically when referencing a program such as SAS you should include the name of the developer (SAS Inc) and their location (Cary, NC) instead of just their website, as was done here. This may also be applicable to Ingenuity Pathway Analysis, Affymetrix etc.

2) The last sentence in the conclusion should probably be flipped around. The way it is written it implies that the authors found biomarkers and are extrapolating their findings to claim these biomarkers may have a role in cellular response to paclitaxel. In fact, this is the opposite. It would be better if it read “These genetic variants may play a significant role in the cellular response to paclitaxel, and represent potential biomarkers for predicting paclitaxel response.”

3) The first sentence in the third background paragraph (“despite…patients”) you may want to change the wording. The response rates don’t vary among patients, because each patient is only treated once, it varies among groups of patients (ethnic groups, cancer types, tumor subclassifications etc).

4) Two sentences later, the sentence “one approach…increased toxicity” should also be modified. Increasing doses of chemotherapy is not an approach; the doses used have been found to be the maximally tolerated doses in humans. Increasing beyond the MTD is not an available approach. You can say that “increasing beyond the typical dose may increase the efficacy, but this is not an available approach due to the existence of dose-limiting toxicities such as…”

5) First sentence on page 5 (“although…patients”) should change “can” to “may.” At this time we “can” explain observed variation by age or compliance, these are validated relationships. The associations with genetic profiles “may” be another
factor that explains the variation, but until we validate these associations that is hypothetical.

6) The numbers of SNPs don’t match up (pg 7). You start with “over 124000 SNP” and exclude 40,690 (20514+20176) which should leave ‘over 83,810’ but your final number is 79622. Where did the other 4,000 SNPs go? You should recheck these numbers if you want to report them as exact numbers. Also, please include commas when reporting numbers larger than 999 (eg 1,000) as it is much easier to read.

7) In table 1 the headers F_A and F_U should be changed to something more straightforward such as AF_R and AF_S (ie Allele-frequency resistant and sensitive) unless the F_A designation means something that I didn’t understand.

8) Pg 11 typo: “were predicted to having no exonic” I think it should be “were predicted to not have”

9) (top of pg 12) At the end of IPA results you should probably list which protein-coding gene SNPs are related to P53/B-catenin axis and which are related to cellular microtubules, or reference a table that makes this differentiation. The details should remain in the discussion, but it seems that it would be a good idea just to show the breakdown in the results section since you make the claim that all the genes can be classified into these two categories.

10) In the gene expression analysis results section (pg 12) mention that there were 20 total probes in 8 genes analyzed. This could also be stated in the table 2 caption.

11) Pg 13 typo: I think it should be “transcriptomic” instead of “transcriptonomic”

12) Pg 15 typo: “or in case of drug response in general.” I do not understand that statement, it should probably be reworded for clarity.

13) Pg 16 typo: “due to the in vitro nature of study” should probably have “our” between “of” and “study.” One of the “only”s should be deleted from “it only identifies only genes”

14) In vivo and in vitro should probably be italicized throughout the manuscript

15) Pg 18 typo: “GRIK1 is a kainite receptor used interneuronal communication” should probably have “for” between “used” and “interneuronal”

16) Pg 21 typo: “our use cell lines” should probably have “of” between “use” and “cell”

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Needs some language corrections before being published

Statistical review: No, the manuscript does not need to be seen by a statistician.
Declaration of competing interests:

I declare that I have no competing interests