Author's response to reviews

Title: Bioinformatic analyses identifies novel protein-coding pharmacogenomic markers associated with paclitaxel sensitivity in NCI60 cancer cell lines

Authors:

Lawson Eng (lawson.eng@utoronto.ca)
Irada Ibrahim-zada (irada@lunenfeld.ca)
Hamdi Jarjanazi (hamdi@lunenfeld.ca)
Sevtap Savas (savas@mun.ca)
Mehran Meschian (meschian@lunenfeld.ca)
Kathleen I Pritchard (kathy.pritchard@sunnybrook.ca)
Hilmi Ozcelik (ozcelik@lunenfeld.ca)

Version: 2 Date: 5 January 2011

Author's response to reviews: see over
January-5-2011

Dear Editor,

Thank you for your email on October 28, 2010 concerning the comments and suggestions on our manuscript. This letter is to serve as a response to the comments provided by the reviewers in a point-by-point format. Each point by each reviewer is addressed below:

Reviewer 1

1) The major concern is all the results and data described in this manuscript were collected from bioinformatics tools analysis without further biological experiments validation. For example, mRNA of five genes (DCT, SNTG1, CFTR, GRIK1, and SGCD) showed significantly different expression levels between sensitive and resistant cell lines through computational analysis; any possibilities these genes can be validated by Real-time PCR or Northern blot analysis?

We agree with the reviewers concern that the results and data described in this manuscript was collected from bioinformatics tools analysis. Validation of these results by Real-time PCR or northern blot analysis would be valuable for this study. However, in silico analysis was the purpose of the primary analysis of the paper and hence we have changed the title to reflect this. As well, our primary objective in this paper was to use innovative bioinformatic tools to identify novel SNP markers and that the mRNA data only served as adjunctive support for our findings. Hence we felt that the microarray data was sufficient to support our findings and we edited the ending in the discussion section “Impact of significant variants on respective genes and gene expression” to indicate that the mRNA expression data was exploratory and the predicted changes from FastSNP along with the mRNA expression data will need to be further validated.

2) It will be interesting if the author could include some SNPs, mRNA gene expression array data from clinical tumor tissue samples in patients with or without paclitaxel treatment.

We definitely agree with the reviewer that this would be a valid extension in our investigation and would help further bridge the results of this study to the clinical management. However, in the clinical setting, paclitaxel is often combined with platinum-based regimes (i.e., in ovarian and lung cancers) or used sequentially with other chemotherapeutic agents (i.e., in breast cancer) and it would be difficult to isolate clinical outcomes solely based on paclitaxel. In addition, when such cases are possible (i.e., paclitaxel is used as a sole agent), the clinical setting is typically metastatic second or third line therapy, where the original tumor was either left intact or was removed lines of therapies earlier and the biology of the cancer has changed. Furthermore, the connection between biology of the tumor naive to therapy and the biology of the tumor at the time of paclitaxel monotherapy is often very weak. Hence it was difficult to find such comparable datasets and we chose used the NCI60 panel instead.

3) Page 8 and Page 9, only 8 cell lines (14%) were in the resistant group while there were 50 cell lines in sensitive group based on the arbitrary z-scores (1.2). As the NCI 60 cancer cell lines
were initially established from patient samples, this distribution does not reflecting the response rates in clinic. As the authors described in the manuscript, the response rates observed from clinical studies of breast cancer patients treated with paclitaxel vary from 21-86%. Similarly in ovarian cancer, the response rate varied from 20-65% and in non-small lung cancer, 30-56%. The distribution of the resistant and sensitive groups can be changes by using a different z-score. In addition, there are several paclitaxel resistant breast and cancer cell lines (with their mRNA gene array data) available and these cell line pairs could be included in this study.

We have chosen this z-score threshold to determine the resistant and sensitive cell lines based on a non-parametric kernel estimation approach. While it is true that changing the z-score can change the groups of resistant and sensitive cell lines and possibly the results obtained in this study, and to clarify that, we have revised the first sentence in the results section as “The non-parametric kernel density estimation of the NCI60 GI50 response data helped to identify the optimum cut-off z-score, thus categorizing the cell lines into a resistant group (n=8) if their z-scores were greater than 1.2 \(10^{-6.807}\) M) and a sensitive group (n=50) if they had z-scores less than 1.2.

Although inclusion of other paclitaxel-resistant cell lines is a great idea, we felt that the addition of resistant cell lines might not be fully comparable to the DTP data due to differences in experimental variables (such as the dose of paclitaxel used, the SNPs genotyped, timing).

3) The authors mentioned that “In all cases, these genes were found to have significantly increased mRNA expression in sensitive cell lines for all probes available, supporting their predicted effect by FastSNP”. In order to confirm the SNP predict genes such as DCT, SNTG1, CFTR, GRIK1, and SGCD play a role in paclitaxel response, some functional studies of these genes may be necessary. For example, inhibit gene expression in sensitive cell line by siRNA or overexpress gene in resistant cell line by transfection and then evaluated paclitaxel sensitivity.

Similar to point 1 addressed above, we agree that it would be valuable to perform these studies to confirm our results and demonstrate independently that these genes play a role in paclitaxel response. However, as mentioned our focus of this study was to use in silicio analysis and the primary objective was to use innovative bioinformatic tools to identify novel SNP markers. Despite this, we have emphasized in the manuscript’s discussion that our results need to be confirmed by further studies before they can be translated into clinical management of patients.

Reviewer 2

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?

Thank you for pointing this out. We initially selected this low cut-off in order to include all the cell lines available due to the exploratory nature of the project and wanting to obtain as many potential SNP markers as possible. However, after you have pointed this out, we have revisited this initial analysis and conducted the GWAS analysis with much higher thresholds ranging from 75% to 90%. When using a threshold of 75% successful calls, we were able to include all the cell lines which had genotype data available (58). However, upon raising this threshold to 80%, only 54 cell lines passed this criterion. At this level though, many of our SNP markers are lost in the analysis (For FDR < 0.005, we have 17 SNPs associated (with only 13 out of 43 of our original SNPs found in association and only one protein coding gene (SNTG1) from our original set of genes found), and at FDR < 0.01, we have 55 SNPs associated (with 41 out of 43 of our original SNPs found in association and 9 of our 10 original protein coding genes (excluding LPHN2) found). Once we have set the threshold to 90%, only 5 cell lines (with all 5 being sensitive cell lines) were found to make the criteria causing a significant loss in the analysis. Hence we have decided to keep the threshold set for 75% as it is an exploratory study and that when some more stringent criteria is applied (i.e., 80% cutoff), somewhat similar results are obtained. We are unaware why certain cell lines did not have as high of a call rate versus others.

Regarding those 4 lines that were not included in then analysis, those 4 cell lines did not have any genotype data available for them and hence were excluded from the GWAS.

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.

We have contacted the database – BioGPS with respect to this issue. Their response was that the data was normalized following the protocol (per-gene normalization to the median) described in this manuscript (Lattin JE, Schroder K, Su AI, Walker JR, Zhang J, Wiltshire T, Saijo K, Glass CK, Hume DA, Kellie S, Sweet MJ (2008) Expression analysis of G Protein-Coupled Receptors in mouse macrophages. *Immunome Res.* 4:5.). The following text has been incorporated into the manuscript “mRNA expression data for the NCI60 panel, as measured on the Affymetrix U133A Chip and normalized using the per-gene normalization to the median approach, was obtained from the online database, BioGPS (biogps.gnf.org) [19]”. Relevant citations have also been corrected for this.

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?

There is no real explanation for it but was a result of the consequence of the system used to determine which allele would be A1 and which would be A2. We tried to switch the case
(resistant) and control (sensitive) groups in the PLINK input files and still arrived at the same results with A2 in that being more common in that case than A1.

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.

We agree with the reviewer’s comments about these issues. Both issues (tumor specific markers and race-specific markers or population stratification) brought up by the reviewer are inherent to the use of any series of cell lines such as the NCI60 panel. As explain in our response to reviewer 1, point 2, it is very difficult to find appropriate human clinical dataset which would have potentially resolve some of these issues. We are aware of a number of groups who have started to utilize primary (non-cell lines derived) xenografted lines for pharmacogenomic testing, but their results are not available yet in publically available dataset. As such, we are restricted to using cell lines and the NCI60 panel was a well-characterized, well-recognized panel. Nonetheless we have added few sentences (provided below) that discusses these two potential confounders upfront in the “Impact of genes on drug response” section in the discussion.

“The NCI60 panel comprises of a limited number of human cancer-derived cell lines, and as such, there may be confounding by either race or tumor type in our identification of pharmacogenomic markers. For instance, DCT is also a marker for melanoma and is melanocyte specific [65, 67]. The melanoma cell lines in this case were all found in the sensitive group and this may serve as a confounding variable. Validation of our results in non-cell line data and in other tumor types are therefore necessary, prior to clinical translation.”

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.
We have changed the title accordingly to “Bioinformatic analyses identifies novel protein-coding pharmacogenomic markers associated with paclitaxel sensitivity in NCI60 cancer cell lines”. In addition, we have made the appropriate corrections in text. In some cases we have left the use of the phrase “paclitaxel response” when we are speaking in reference to the future goal of the study (i.e., in the first part of the abstract).

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.

As suggested by the reviewer, this decision was made after seeing the cut off for FDR<0.1 (providing 68 SNPs) to limit the number of SNPs and therefore increase the stringency of SNPs that were detected as being significantly associated with paclitaxel response. This helped to identify the top choices of SNPs to investigate further for this paper and study.

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.

The sentence in the abstract has been revised as “The identified haplotypes highlight the role of utilizing SNP combinations within genomic loci of interest to improve the risk determination associated with drug response.”

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

This has been revised in text.

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.

This has been corrected in text as “microtubule-targeting drugs”.

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.
This has been corrected in the text as “Since the haplotype association analyses were also exploratory in nature, we chose to not perform corrections for multiple comparisons.”

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.”

Figure 1 has been updated and corrected. The relevant section on page 14 has been revised as “The GI50 response data was extrapolated from a range of doses which started at a $10^{-6}$ M dose and this was used as the basis for our case-control analysis”.

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.

We have accounted for this valid point in the text by making the numerical changes as well as stating in the methods section as “A total of 62 cell lines had drug response data available from this drug screening protocol.” and we did originally state that those cell lines without genotype data were removed from the before the GWAS leaving 58 cell lines (Page 7). We have also applied a similar note to the captions (see below) to make it clear to the reader.

In Figure 1, as “The 62 cell lines which had drug response data available from DTP were used in this categorization”.

In Table 1, as “This GWAS was performed using only 58 of the 62 cell lines which had drug response data available since the remaining 4 cell lines lacked available genotype data”.

In Table 2, as “A total of 20 probes were available to help analyze the expression pattern of the 8 genes listed above. As well, only 54 of the cell lines (from the 58 used in the GWAS) were used in this analysis as the remainder lacked gene expression data.

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).
We are grateful to the reviewer for pointing this out. However, our focus for this paper was on the protein-coding SNPs as these are the ones that are more likely to have functional consequences. We will focus on non-protein coding SNPs in a future publication and project. Furthermore, in correcting the title, we have included in it the phrase “protein-coding” to be clear about our current scope. However, we included the following statement in the discussion to acknowledge this point: “Furthermore, one of our identified SNPs at 9p23 (marker number: 2836880) was found as part of the identified \textit{PTPRD} haplotype. This suggests that by performing haplotype analyses, SNPs identified in the GWAS that were originally unclear on their role in paclitaxel sensitivity, may be found related to part of the gene through haplotypes.”

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.

Many thanks. We have included the following text in the end of the section “Although FastSNP did accurately predict the potential effect for the 4 SNPs serving as intronic enhancers along with the intronic SNPs in \textit{ROBO1} and \textit{LPHN2} as not having any effects, these predicted changes are only suggestive at this stage and will need further validation using, for example, \textit{in vitro} cell based functional assay systems.”

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.

On page 22 (this section was formerly on page 20), we have now included several sentences to this effect. We mentioned the differences in methods that were used to identify markers of sensitivity as a potential reason for why our results differed and along with potential a mechanism on why this may have been the case.

Discretionary Revisions:
\textit{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.}

Agreed. We have applied this to programs which do not have a reference associated with them: Affymetrix, IPA and SAS.

\textit{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.”}

This has been corrected in text as “The 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).

This has been corrected in text and the change has been tracked.

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…”

We have corrected for this in the manuscript and used a similar phrasing technique as “One approach to improve drug response is by increasing dosages beyond the typical dose to increase the efficacy, but this is not an available approach due to the existence of dose limiting toxicities”.

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.

This has been corrected in text and the change has been tracked.

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.

Originally, a total of 124,000 SNPs were meant to be genotyped on the array. However, only 118,409 SNPs were actually genotyped in the dataset we obtained. As well, we believe that when calculating the number of SNPs which failed each criteria, some SNPs were counted twice by PLINK (i.e, failed both minor allele frequency test and genotyping criteria and counted once in each group). This has been noted and corrected in the text as “A total of over 124,000 SNPs were genotyped on this array, but the data for only 118,409 SNPs was available [12].”. As well, we have incorporated your suggestion about large numbers.

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.

Agreed. This has been corrected in the table and the change has been tracked.
8) Pg 11 typo: “were predicted to having no exonic” I think it should be “were predicted to not have”
This has been corrected in text and the change has been tracked.

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.

   We agree that this will help make things clearer for readers and we have included this in our manuscript as “These investigations helped us classify the identified genes into two groups when determining the cellular response to paclitaxel: those such as PTPRD and BTBD12, interacting with the p53 and β-catenin axis, and those including ROBO1, CFTR, ZNF607, GRIK1, LPHN2, DCT, SGCD, SNTG1 that interact with cellular microtubules”.

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.
   This has been included in the table 2 caption as “A total of 20 probes were available to help analyze the expression pattern of the 8 genes listed above”.

11) Pg 13 typo: I think it should be “transcriptomic” instead of “transcriptonomic”
   This has been corrected in text and the change has been tracked.

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.

   We were trying to state that this rationale – about how SNPs can cause changes at a cellular level in ways other than changing gene expression, could be similarly applied to the case of SNPs identified in other pharmacogenomic studies. We have removed this statement since we feel that the reader can extrapolate this from our reasoning.

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”

   Both of these suggestions have been corrected in text and the changes have been tracked.

14) In vivo and in vitro should probably be italicized throughout the manuscript
   This has been corrected in text and the change has been tracked.
15) Pg 18 typo: “GRIK1 is a kainite receptor used interneuronal communication” should probably have “for” between “used” and “interneuronal”

This has been corrected in text and the change has been tracked.

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

This has been corrected in text and the change has been tracked.

On behalf of the authors of this manuscript, we are grateful for the editorial assistance provided by yourself and your administrative staff in handling our manuscript. We would also like to thank the reviewers for their valuable comments that have helped to improve the clarity and quality of our manuscript and project. We hope that this manuscript is acceptable in its present form.

Best Wishes,

Lawson Eng, Hon. B. Sc
University of Toronto – Doctor of Medicine Candidate, Class of 2014
Princess Margaret Hospital – Liu Lab – Comprehensive Research Experience for Medical Student Program Student
Mount Sinai Hospital – Samuel Lunenfeld Research Institute – Research Student