Author's response to reviews

Title: Systematically characterizing and prioritizing chemosensitivity related gene based on Gene Ontology and protein interaction network

Authors:

Xin Chen (chenxinjx@hotmail.com)
Wei Jiang (jiangwei@hrbmu.edu.cn)
Qianghu Wang (wqhmcm@yahoo.com.cn)
Teng Huang (spacewalkerht@163.com)
Peng Wang (wpggy@163.com)
Yan Li (bioinforly@yahoo.cn)
Xiaowen Chen (bioinforly@yahoo.cn)
Yingli Lv (lvyingli_1980@163.com)
Xia Li (lixia@hrbmu.edu.cn)

Version: 3 Date: 25 July 2012

Author's response to reviews: see over
Dear Editor Tim Sands,

Thanks so much for your editorial efforts regarding our manuscript and for your recommendation that we submit a revision of this manuscript. We also thank the reviewers for the constructive comments to strengthen this manuscript, which have been accommodated appropriately.

We have carefully revised the manuscript to respond to the points and concerns raised by the reviewers. In particular, we have undertaken the following major revisions over the previous version:

(1) Regarding the reviewer’s suggestion for language checking, we have carefully corrected the English errors to improve its overall clarity and readability. The revised manuscript has been internally reviewed and proofread by several professionals, and the commercial copyediting service, International Science Editing (MS ID: ISE37610), against any potential academic or grammar errors.

(2) Regarding the reviewer’s suggestion to validate the predicted drug-CRGs, we have used hypergeometric tests to evaluate the extent to which the predicted drug-CRGs appeared in the drug-CCRGs. We have added the section “Performance of the proposed method to identify drug-CRGs” in this revision.

(3) Regarding the reviewers’ suggestion to state the key results and our hypothesis in the Abstract, we have rewritten the Abstract.

(4) Regarding the reviewer’s comments on the Figure 1, we have redraw this Figure, rewritten the legend and added more detailed descriptions in the “Drug activity data and gene expression data” section.

(5) Regarding the reviewer’s comments on the legends of Figure 2 and Figure 3, we have added more detailed descriptions in their legends.

(6) Regarding the reviewer’s comment about the drug-CCRG network, we have made clear distinction of drug-curated chemosensitivity related gene (drug-CCRG) network, drug-candidate chemosensitivity related gene network and protein-protein interaction network, and added detailed description in the section “Correlation-based analysis of the drug-CCRG relationships”.
In the revised manuscript, major revisions are shown in **BLUE** color. The manuscript has been reviewed and proofread by several professionals and the commercial copyediting service to avoid any potential academic or grammar errors. The detailed responses on a point-by-point basis and revisions made are described below.

We believe that the manuscript has been greatly strengthened by the critique of the reviewer and hope that both you and the reviewers will now find the paper suitable for publication.

Thank you and the reviewers again for carefully reviewing our work.

Respectfully yours,

Xia Li, Ph.D., Dean
College of Bioinformatics Science and Technology
Harbin Medical University
194 Xuefu Road, Harbin 150081, China
Phone: 86-451-86615922
Fax: 86-451-86615922
Email: lixia@hrbmu.edu.cn
Reply to Reviewer #1 (report from 18 April, 2012)

Major revisions: Most importantly, the manuscript lacks methodological details to understand exactly how data is generated for the figures of the panels A, B, C of Figure 1 as well as Figures 2, 3, and 4. Further, the methods surrounding the GO ontology enrichment are particularly difficult to understand, provide no equations nor any references to pre-existing methods of GO prioritization.

REPLY: We thank the reviewer for the suggestion regarding methodological details in this manuscript and we have made appropriate revisions accordingly. In this study, we firstly documented 150 pairs of drug-CCRG (curated chemosensitivity related gene) from 492 published papers. Secondly, we investigated the characteristics of the drug-CCRG pairs. Previous studies which aim to identify CRG (chemosensitivity related gene) are generally based on correlation of gene expression and drug activity, and the gene whose expression is highly correlated with the drug activity is regarded as candidate CRG for the drug. However, following expression based analysis we found that the majority of drug-CCRGs exhibited a low correlation between gene expression and drug activity. Moreover, for 27/62 (44%) of drug-CCRG pairs, their correlations tended to be random by comparing \( z_i \) with \( Z_{\text{threshold}} \). When characterizing CCRGs, we found that CCRG enriched GO (Gene Ontology) terms were generally associated with chemosensitivity and exhibited higher similarity scores compared to randomly selected genes. Moreover, CCRGs played key roles in maintaining connectivity and controlling information flow in PPIN (protein-protein interaction network). Lastly, based on these characteristics we prioritized CRGs based on GO functional information and PPIN.

We have incorporated the following elaborations in this revision.

We have added “Methods” to the Abstract in the page 2.

“Methods
In this study, we proposed a method to identify CRGs based on Gene Ontology (GO) and PPIN. Firstly, we documented 150 pairs of drug-CCRG (curated chemosensitivity related gene) from 492 published papers. Secondly, we characterized CCRGs from the perspective of GO and PPIN. Thirdly, we prioritized CRGs based on CCRGs’ GO and network characteristics. Lastly, we evaluated the performance of the proposed method.”

The followings are the methodological details to understand exactly how data is generated for the figures of the panels A, B, C of Figure 1.

In Figure 1, \( D \) represents drug activity profile of the NCI 60 cell lines, each row represents a drug and each column represents a cell line, each element \( a_{ij} \) represents the drug activity (GI50) of drug \( d_j \) in cell line \( C_i, i=1,2,\ldots,4444, j=1,2,\ldots,59 \). The total
The number of drugs we analyzed in the manuscript was 4444. \( G \) represents the gene expression profiles of the NCI 60 cell lines, each row represents a gene and each column represents a cell line, each element \( e_{ij} \) represents the expression level of gene \( g_i \) in cell line \( C_j \). The total number of genes we analyzed in the manuscript was 12633.

In panel A of Figure 1, filter A is based on Gene Ontology. We characterized CCRG using GO enrichment analysis with Fisher Exact Test. We considered three aspects of GO: biological process (BP), molecular function (MF), and cellular component (CC). \( p \) represents the enrichment significance. If enriched \( p \) value is smaller than 0.01, CCRGs are significantly enriched in the GO term. Thus, we can obtain all the enriched GO terms of CCRGs, which were generally related to chemosensitivity. Moreover, we investigated whether CCRGs exhibited functional consistency. We compared the functional similarity of CCRG enriched GO terms to randomly selected genes enriched GO terms. We found that CCRG enriched GO terms exhibited higher similarity scores compared to randomly selected genes. Thus, we regarded all genes in the enriched GO terms as candidate CRGs. The similarity of CCRG enriched GO terms is shown in Table S2, which is supplemented in Additional file 4.

The first column depicts the functional aspects of the GO and the annotation depth. The second column depicts average similarity of the CCRG enriched GO terms. The third column depicts average similarity of the enriched GO terms for randomly selected genes with the same number of CCRG for 1000 times. The forth column depicts the fold change of functional similarity between CCRG and random genes. The last column depicts the \( p \) value by 1000 randomizations.

<table>
<thead>
<tr>
<th>enriched term set similarity</th>
<th>random term set similarity</th>
<th>fold of similarity (CCRG/random)</th>
<th>( p ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP-1 4.763360569</td>
<td>4.59126295</td>
<td>1.037483721</td>
<td>0.165</td>
</tr>
<tr>
<td>BP-2 7.841207898</td>
<td>6.20553737</td>
<td>1.263582416</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>BP-3 9.468104134</td>
<td>6.4027355</td>
<td>1.478759218</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>BP-4 10.1540088</td>
<td>6.56144268</td>
<td>1.547526862</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>BP-5 11.2065688</td>
<td>6.47581789</td>
<td>1.730525625</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>MF-1 3.224983443</td>
<td>4.06081228</td>
<td>0.794172009</td>
<td>0.925</td>
</tr>
<tr>
<td>MF-2 7.130407841</td>
<td>4.8006631</td>
<td>1.485296447</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>MF-3 9.482255399</td>
<td>5.6210522</td>
<td>1.686918224</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>MF-4 11.288934</td>
<td>5.31083664</td>
<td>2.125641357</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>MF-5 9.816308782</td>
<td>5.0877045</td>
<td>1.929418027</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>CC-1 3.456894625</td>
<td>2.76765466</td>
<td>1.249033947</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>CC-2 4.333211773</td>
<td>3.90003277</td>
<td>1.111070605</td>
<td>0.055</td>
</tr>
<tr>
<td>CC-3 5.666589478</td>
<td>4.37185854</td>
<td>1.29615115</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>CC-4 5.781969391</td>
<td>4.47090976</td>
<td>1.293242247</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>CC-5 5.666608925</td>
<td>4.43261552</td>
<td>1.278389452</td>
<td>&lt;0.005</td>
</tr>
</tbody>
</table>
Details for calculating the functional similarity of GO terms are described as follows. The semantic similarity measure we used was Jiang’s term similarity measures (Proc of the 10th International Conference on Research on Computational Linguistics, 1997) and best-match average (BMA, BMC Bioinformatics 2008, 9 (Suppl 5):S4). Given two terms $c_1$ and $c_2$, and their most informative common ancestor $c_A$, Jiang and Conrath's similarity measure is given by the following equation:

$$
sim_{c_1,c_2} = 1 + IC_{c_A} - \frac{IC_{c_1} + IC_{c_2}}{2}
$$

where $IC_c = -\log p(c)$, $p(c)$ is the probability of using term $c$ in the universal term set. To calculate this frequency, we firstly count the number of distinct proteins annotated to term $c$ or one of its descendant terms, and then divide the number by the total number of proteins annotated within the corresponding GO aspect.

Given two non-redundant sets of GO terms $GO(A)$ and $GO(B)$, the best-match average approach is given by the average similarity between each term in $GO(A)$ and its most similar term in $GO(B)$, averaged with its reciprocal to obtain a symmetric score:

$$sim_{BMA} = \frac{\sum_{t_1 \in GO(A)} \sum_{t_2 \in GO(B)} \frac{sim_{t_1,t_2} + sim_{t_2,t_1}}{2}}{2}$$

In panel B of Figure 1, filter B is based on PPIN. We characterized CCRG using network features of PPIN. We analyzed several network features such as degree and betweenness centrality in six PPINs. These six PPINs are BIND, BioGRID, IntAct, MINT, OPHID and HPRD. We calculated network features separately in each of the six networks, and selected the network features that were significantly different between CCRGs and random genes in all the six networks. Finally, degree and betweenness centrality were selected as network features to prioritize CRGs. The green curve represents betweenness centrality of random genes, and the vertical green line is the betweenness centrality of CCRGs. The blue curve represents degree of random genes, and the vertical blue line is the degree of CCRGs. As protein-protein interactions in HPRD are experimentally derived human protein-protein interaction data, we utilized HPRD to prioritize CRGs.

In panel C of Figure 1, filter C is based on gene expression. We characterized drug-CCRG pairs using Pearson’s correlation coefficient ($PCC$). The majority of drug-CCRGs exhibit a low correlation between gene expression and drug activity.

$$PCC_{X,Y} = \frac{\text{cov}(X,Y)}{\delta_X \delta_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\delta_X \delta_Y}$$

where $E$ is expectation, $\text{cov}$ is covariance, and $X, Y$ represent a drug and a gene,
respectively. \( \delta^2_X = E(X^2) - E^2(X) \), \( \delta^2_Y = E(Y^2) - E^2(Y) \). We ranked the absolute PCC of all drug-CCRG pairs in ascending order and set the PCC threshold as 5th percentile of all PCCs. In this way, 95% of drug-CCRGs are detected only based on PCC between drug activity and gene expression.

We have re-drawn the Figure 1 and incorporated the following elaborations in this revision.

**Figure 1. Outline of the proposed method.** Firstly, we manually curated a compendium of curated chemosensitivity related genes (CCRGs) from published papers. Then we selected genes on the Microarray that had same enriched GO categories and network characteristics with the CCRGs. These genes were considered as candidate CRGs. To get CRGs for each drug, we further filtered the initial drug-candidate CRG network based on PCC of drug-CCRGs. Filter A is based on Gene Ontology. We characterized CCRGs using GO enrichment analysis with Fisher Exact Test. We considered three aspects of GO: biological process (BP), molecular function (MF), and cellular component (CC). \( p \) represents the enrichment significance. If enriched \( p \) value is smaller than 0.01, CCRGs are significantly enriched in the GO term. Moreover, we investigated that whether CCRGs exhibited functional consistency. We compared the functional similarity of CCRG enriched GO terms to randomly selected gene enriched GO terms. We found that CCRG enriched GO terms exhibited higher similarity scores compared to randomly selected genes. Thus, we regarded all genes in the enriched GO terms as candidate CRGs. Filter B is based on protein
interaction networks. We analyzed several network features such as degree and betweenness centrality in six PPINs. Degree and betweenness centrality were selected as network features to prioritize CRGs. The green curve represents betweenness centrality of random genes, and the vertical green line is the betweenness centrality of CCRGs. The blue curve represents degree of random genes, and the vertical blue line is the degree of CCRGs. Filter C is based on gene expression. The majority of drug-CCRGs exhibit a low correlation between gene expression and drug activity. We ranked the absolute PCC of all drug-CCRG pairs in ascending order and set the PCC threshold as 5th percentile of all PCCs."

In Figure 2, we modified the original legend. The value 80.6% indicates that drug-CCRG pairs whose PCC range from -0.3 to 0.3 accounts for 80.6% of all drug-CCRG pairs; 91.9% indicates that drug-CCRG pairs whose PCC range from -0.5 to 0.5 accounts for 91.9% of all drug-CCRG pairs. To evaluate the PCC between drug activity and gene expression, only drugs with drug activity data and genes with expression data available are utilized for further analysis. The 150 drug-CCRG pairs include 64 drugs and 94 genes: 47 of 94 genes were detected for expression in NCI 60 cell lines whereas 31 of 64 drugs were detected for activity in NCI 60 cell lines; these 31 drugs and 47 genes comprised 62 drug-CCRG pairs of the original 150 drug-CCRG pairs. We then performed correlation-based analysis using these 62 drug-CCRG pairs.

We have incorporated the following elaborations in this revision.

The subsection of “Correlation-based analysis of the drug-CCRG pairs” in the page 12, 13

“A gene with expression highly correlated to drug activity is regarded as a candidate CRG for the drug. Thus, we initially investigated whether CCRGs were highly correlated with their interactive drugs. Of the 150 pairs of drug-CCRG, 62 pairs were available for correlation analysis. We evaluated the PCC between drug activity and gene expression for drug with drug activity and genes with expression available in the NCI 60 cell lines. The 150 drug-CCRG pairs included 64 drugs and 94 genes. A total of 47 of 94 genes were detected for their expression in NCI 60 cell lines and 31 of 64 drugs were detected for their activity in NCI 60 cell lines; these 31 drugs and 47 genes comprised 62 drug-CCRG pairs of the original 150 drug-CCRG pairs. We then performed correlation-based analysis on these 62 drug-CCRG pairs. In Figure 2, drug-CCRG pairs whose PCC range from -0.3 to 0.3 accounts for 80.6% of all drug-CCRG pairs while drug-CCRG pairs whose PCC range from -0.5 to 0.5 accounts for 91.9% of all drug-CCRG pairs.”

The legend of Figure 2.

“Figure 2 - The distribution of Pearson’s correlation coefficient (PCC) of drug-CCRG pairs.
The x-axis represents PCC, the y-axis represents the frequency of a certain PCC. The value 80.6% indicates that drug-CCRG pairs whose PCC range from -0.3 to 0.3 accounts for 80.6% of all drug-CCRG pairs; 91.9% indicates that drug-CCRG pairs whose PCC range from -0.5 to 0.5 accounts for 91.9% of all drug-CCRG pairs.”
In Figure 3, we did not illustrate different thresholds of Figure 3A in an explicit way. For each of the 62 drug-CCRG pairs, we compared the PCC with that of random drug-gene pairs. The statistical method we used is $z_i = |x_i - \mu|/\delta$, where $x_i$ is the PCC of drug-CCRG pair $i$, $\mu$ and $\delta$ are the mean and standard deviation of PCCs of all random drug-gene pairs. If $z_i \geq z_{\text{threshold}}$, the PCC of drug-CCRG pair $i$ is significantly different from random PCC. Figure 3A shows the number of recalled drug-CCRG pairs under different thresholds. The numbers of drug-CCRG pairs, which were identified under the corresponding $z_{\text{threshold}}$, were listed over the blue bar. The stricter the $z_{\text{threshold}}$ was, the fewer drug-CCRG pairs were identified. For example, when using 1 as the $z_{\text{threshold}}$, only 32 of 62 drug-CCRGs were identified, while using 2 as $z_{\text{threshold}}$, only 15 of 62 were identified and 6 of 62 when using 3 as $z_{\text{threshold}}$. We set $z_{\text{threshold}}$ to 0.8 in concordance with previous reports (Proc Natl Acad Sci U S A 2001, 98:10787-10792). Under this threshold, we conducted further analysis (Figure 3B, Figure 3C, and Figure 3D). In Figure 3B, “21/62” indicates that 21 of 62 drug-CCRG pairs exhibit significantly smaller PCCs than random drug-gene pairs. In this figure, we illustrate the distribution of correlations between Doxorubicin and all genes. The red line indicates the example of Doxorubicin and ABCB1 correlation. It was reported that ABCB1 overexpression predicted doxorubicin resistance (Cell Biochem Biophys 2009, 55:95-105). In Figure 3C, “14/62” indicates that 14 of 62 drug-CCRG pairs exhibit significantly larger PCCs than random drug-gene pairs. In this figure, we illustrate the distribution of correlations between NSC169517 and all genes. The red line indicates the example of NSC169517 and PRKCB correlation. It was reported that PRKCB was able to predict chemosensitivity of NSC169517 (Mol Cancer Ther 2002, 1:311-320). In Figure 3D, “27/62” indicates that 27 of 62 drug-CCRG pairs do not exhibit significantly different PCCs from random drug-gene pairs. In this figure, we illustrate the distribution of correlations between Paclitaxel (taxol) and all genes. The red line indicates the example of Paclitaxel (taxol) and GRIK1 correlation. It was reported that GRIK1 was able to predict chemosensitivity of Paclitaxel (BMC Med Genomics, 4:18).

We have incorporated the following elaborations in this revision.

The subsection of “Correlation-based analysis of the drug-CCRG relationships” in the page 13

“Figure 3A shows the number of identified drug-CCRG pairs under different thresholds. If $z_i \geq z_{\text{threshold}}$, the PCC of drug-CCRG pair $i$ is significantly different from random PCC. The numbers of drug-CCRG pairs, which were identified under the corresponding $z_{\text{threshold}}$, were listed over the blue bar. As the stricter $z_{\text{threshold}}$ was, fewer drug-CCRG pairs were identified. For example, when using 1 as the $z_{\text{threshold}}$, only 32 of 62 drug-CCRGs were identified, whereas when using 2 as the $z_{\text{threshold}}$, only 15 of 62 were identified, and when using 3 as the $z_{\text{threshold}}$ only 6 of 62 were identified. As shown in Figure 3A, we found it was not sufficient to identify drug-CCRG pairs using PCC based on random analysis.”

In Figure 4, the red line “corrOnly” represents the traditional method to identify CRGs only based on the correlation between gene expression and drug activity. The
blue line “combined filter” represents the proposed method to identify CRGs by integrating information from CCRG enriched GO terms and network features of PPIN. The ROC curve was used to evaluate the performance of both methods. For the proposed method, we rank all the genes in HPRD protein interaction network by $Q$ statistics (see details in the Methods section of the manuscript). According to $Q$ statistics and whether the genes are CCRGs, we plotted the ROC curves for our method. While for the traditional correlation method, we ranked all drug-CRG pairs using absolute $PCC$. According to $PCC$ and whether genes were CCRGs, we also plotted the ROC curves.

The area under curve (AUC) illustrated the performance of the method. The greater the area, the better the performance. From Figure 4, we concluded that our method is superior to traditional methods based only on $PCC$ of drug activity and gene expression.

We have incorporated the following elaborations in this revision.

The subsection of “Performance of the proposed method to identify drug-CRGs” in the page 16

“For the proposed method, we ranked all of the genes in predicted drug-CRGs using the $Q$ statistic (See details in Methods) in order to integrate various separate data sources. We integrated ranks of degree and betweenness centrality to determine whether CCRGs ranked at the top of the list. According to $Q$ statistics and whether genes were CCRGs, we plotted the ROC curves. For traditional correlation method, we ranked all drug-CRG pairs using absolute $PCC$ of gene expression and drug activity. According to $PCC$ and whether genes were CCRGs, we also plotted the ROC curves.”

In the manuscript, we did not clarify how we enriched the CCRGs in GO and how we utilized the enriched GO terms to prioritize CRGs. Fisher Exact test was adopted to conduct GO enrichment analysis. The details are illustrated as follows.

<table>
<thead>
<tr>
<th>In GO term</th>
<th>User Genes</th>
<th>Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>d</td>
<td></td>
</tr>
</tbody>
</table>

$$p = \frac{(a+b)(c+d)}{a} = \frac{n}{a+c}$$

where $n=(a+b+c+d)$, $a$ was the total number of user genes annotated in a GO term; $b$ was the number of genes annotated in this GO term; $c$ was the number of user genes not annotated in this GO term; $d$ was the number of background genes not annotated in this GO term. If $p \leq 0.01$, we hypothesized that the user gene lists were specifically
associated (enriched) in this GO term. We considered all three ontologies: biological process (BP), molecular function (MF) and cellular component (CC). We limited the enriched GO term to depth 5 of GO according to DAVID (Nat Protoc 2009, 4:44-57; Nucleic Acids Res 2009, 37:1-13). These enriched GO terms were used for further CRG prioritization. We assumed that all genes in these enriched GO terms were candidate CRGs. Equations and references were added according to the reviewer’s suggestion.

We have incorporated the following elaborations in this revision.

The subsection of “GO enrichment using Fisher Exact test” in the page 10, 11

“Fisher Exact test was adopted to measure the gene enrichment in annotation terms [35]. The details are illustrated as follows.

<table>
<thead>
<tr>
<th>User Genes</th>
<th>Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>In GO term</td>
<td>a</td>
</tr>
<tr>
<td>Not In GO term</td>
<td>c</td>
</tr>
</tbody>
</table>

\[
P = \frac{\binom{n}{a+c} \cdot \binom{c+d}{a}}{a! \cdot b! \cdot c! \cdot d!}
\]

where \( n = (a+b+c+d) \), \( a \) was the total number of user genes annotated in a GO term; \( b \) was the number of genes annotated in this GO term; \( c \) was the number of user genes not annotated in this GO term; \( d \) was the number of background genes not annotated in this GO term. If \( p \leq 0.01 \), we hypothesized that the user gene lists were specifically associated (enriched) in this GO term. We considered all three ontologies: biological process (BP), molecular function (MF) and cellular component (CC). We limited the enriched GO term to depth 5 of GO according to DAVID [36, 37].”

1) The manuscript would benefit to be read by a native-English speaker or a professional English language writer/Editor. It is not publishable in the current form.

REPLY: We have carefully corrected the English errors in the previous version to improve its overall clarity and readability. The revised manuscript has been internally reviewed and proofread by several professionals, and the commercial copyediting service, International Science Editing (MS ID: ISE37610), against any potential academic or grammar errors.

2) The abstract should state the key results and pvalue of the validation (predicting response to chemotherapy) in an independent dataset from the learning set or in a simulated dataset.
REPLY: We kindly accepted the reviewer’s suggestion and have validated the proposed method using hypergeometric tests.

We originally used the receiver operating characteristic (ROC) curve to compare the performance of the proposed method and traditional method based on correlation of gene expression and drug activity. For the proposed method, we evaluated the genes in predicted drug-CRGs by $Q$ statistics (See details in Methods of the main manuscript). According to $Q$ statistics and whether the genes were CCRGs, we plotted the ROC curves for our method. While for the traditional correlation method, we ranked all drug-CRGs using absolute $PCC$ of gene expression and drug activity. According to $PCC$ and whether genes were CCRGs, we also plotted the ROC curves. The area under (AUC) curve illustrates the performance of the method. The greater the area, the better the performance. The mean AUC of our method is 65.2%, whereas that of the traditional method AUC is 55.2%.

In this revision, we used hypergeometric tests to evaluate the extent to which predicted drug-CRGs appeared in the drug-CCRGs. The significance of the over-representation was calculated by the hypergeometric test:

$$P = \sum_{j \leq n} \frac{C_{n}^{j} \cdot C_{M-N}^{m-j}}{C_{M}^{m}}$$

where $M$ was the total number of all drug-candidate CRGs; $N$ was the number of predicted drug-CRGs using our method; $m$ was the number of drug-CCRGs; $n$ was the number of drug-CCRGs correctly predicted by our method. Under different degree and betweenness centrality threshold, we obtained different number of drug-CRGs. For each threshold, we evaluated the performance of our method. The results are shown in Table 3.

**Table 3. Performance of our method to predict drug-CRGs under different thresholds.**

<table>
<thead>
<tr>
<th>Threshold of degree</th>
<th>Threshold of betweenness centrality</th>
<th>Number of identified drug-CCRGs</th>
<th>enrichment significance</th>
<th>Number of identified drug-CCRGs</th>
<th>enrichment significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.01</td>
<td>6</td>
<td>1.20E-08</td>
<td>5</td>
<td>5.34E-06</td>
</tr>
<tr>
<td>0.02</td>
<td>0.02</td>
<td>8</td>
<td>0.00</td>
<td>6</td>
<td>1.05E-05</td>
</tr>
<tr>
<td>0.03</td>
<td>0.03</td>
<td>10</td>
<td>0.00</td>
<td>7</td>
<td>5.76E-06</td>
</tr>
<tr>
<td>0.04</td>
<td>0.04</td>
<td>11</td>
<td>0.00</td>
<td>7</td>
<td>2.87E-05</td>
</tr>
<tr>
<td>0.05</td>
<td>0.05</td>
<td>11</td>
<td>0.00</td>
<td>7</td>
<td>1.20E-04</td>
</tr>
<tr>
<td>0.06</td>
<td>0.06</td>
<td>11</td>
<td>0.00</td>
<td>8</td>
<td>4.56E-05</td>
</tr>
<tr>
<td>0.07</td>
<td>0.07</td>
<td>13</td>
<td>0.00</td>
<td>8</td>
<td>1.17E-04</td>
</tr>
<tr>
<td>0.08</td>
<td>0.08</td>
<td>13</td>
<td>0.00</td>
<td>8</td>
<td>2.61E-04</td>
</tr>
<tr>
<td>0.09</td>
<td>0.09</td>
<td>13</td>
<td>7.46E-08</td>
<td>8</td>
<td>5.47E-04</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Degree</th>
<th>Betweenness</th>
<th>PCC</th>
<th># Gene pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>13</td>
<td>2.01E-08</td>
</tr>
<tr>
<td>0.11</td>
<td>0.11</td>
<td>13</td>
<td>1.08E-08</td>
</tr>
<tr>
<td>0.12</td>
<td>0.12</td>
<td>14</td>
<td>6.10E-08</td>
</tr>
<tr>
<td>0.13</td>
<td>0.13</td>
<td>14</td>
<td>0.00</td>
</tr>
<tr>
<td>0.14</td>
<td>0.14</td>
<td>14</td>
<td>3.60E-07</td>
</tr>
<tr>
<td>0.15</td>
<td>0.15</td>
<td>14</td>
<td>7.63E-07</td>
</tr>
<tr>
<td>0.16</td>
<td>0.16</td>
<td>14</td>
<td>1.79E-06</td>
</tr>
<tr>
<td>0.17</td>
<td>0.17</td>
<td>14</td>
<td>2.80E-06</td>
</tr>
<tr>
<td>0.18</td>
<td>0.18</td>
<td>14</td>
<td>5.26E-06</td>
</tr>
<tr>
<td>0.19</td>
<td>0.19</td>
<td>14</td>
<td>7.77E-06</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>15</td>
<td>3.25E-06</td>
</tr>
</tbody>
</table>

*Degree threshold of 0.01 means that the reserved genes have top one percentile (0.01) degree in all genes of HPRD. Betweenness centrality threshold is set in the same way as degree threshold.

*In order to ensure the comparability of our method and the method based on correlation of gene expression and drug activity, we kept the number of predicted drug-CRGs obtained by both methods equal with each other.

We compared the performance of both methods when selecting 20 sets of thresholds. The proposed method identified a greater number of drug-CCRGS under all of the thresholds. Moreover, drug-CCRGS were much more significantly enriched in the drug-CCRGs predicted by our method.

We have incorporated the following elaborations in this revision.

The **Results** part of “Abstract” in the page 2

“Our proposed method identifies a greater number of drug-CCRGS, and drug-CCRGS that are much more significantly enriched with predicted drug-CRGs, compared to a method based on the correlation of gene expression and drug activity. The mean area under ROC curve (AUC) for our method is 65.2%, whereas that for the traditional method is 55.2%.”

The subsection of “**Performance of the proposed method to identify drug-CRGs**” in the page 15, 16

“Here, we used hypergeometric tests to evaluate the extent to which predicted drug-CRGs appeared in the drug-CCRGS. The significance of the over-representation was calculated by the hypergeometric test:

\[
P = \sum_{x \leq y} \frac{C_x^N \cdot C_{M-x}^{N-y}}{C_M^{M-N}}
\]

where \(M\) was the total number of all drug-candidate CRGs; \(N\) was the number of predicted drug-CRGs using our method; \(m\) was the number of drug-CCRGS; \(n\) was the number of drug-CCRGS correctly predicted by our method. In order to ensure the comparability of our method and the method based on gene expression, we keep number of predicted drug-CRG pairs obtained by both methods equal with each other. Using different thresholds for betweenness centrality, degree and PCC, we obtained different numbers of drug-gene pairs. In order to identify
the greatest number of drug-CCRG pairs, we set the PCC threshold to the fifth percentile (5%) of PCC for all drug-CCRG pairs. We compared the performance of both methods under 20 sets of thresholds for betweenness centrality and degree; the results are shown in Table 3.

Table 3. Performance of our method to predict drug-CRGs under different thresholds.

<table>
<thead>
<tr>
<th>Threshold of degree*</th>
<th>Threshold of betweenness centrality</th>
<th>The proposed method</th>
<th>Method based on gene expression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Number of identified drug-CCRGs</td>
<td>enrichment significance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>1.20E-08</td>
</tr>
<tr>
<td>0.01</td>
<td>0.01</td>
<td>8</td>
<td>0.00</td>
</tr>
<tr>
<td>0.03</td>
<td>0.03</td>
<td>10</td>
<td>0.00</td>
</tr>
<tr>
<td>0.04</td>
<td>0.04</td>
<td>11</td>
<td>0.00</td>
</tr>
<tr>
<td>0.05</td>
<td>0.05</td>
<td>11</td>
<td>0.00</td>
</tr>
<tr>
<td>0.06</td>
<td>0.06</td>
<td>11</td>
<td>0.00</td>
</tr>
<tr>
<td>0.07</td>
<td>0.07</td>
<td>13</td>
<td>0.00</td>
</tr>
<tr>
<td>0.08</td>
<td>0.08</td>
<td>13</td>
<td>0.00</td>
</tr>
<tr>
<td>0.09</td>
<td>0.09</td>
<td>13</td>
<td>7.46E-08</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>13</td>
<td>2.01E-08</td>
</tr>
<tr>
<td>0.11</td>
<td>0.11</td>
<td>13</td>
<td>1.08E-08</td>
</tr>
<tr>
<td>0.12</td>
<td>0.12</td>
<td>14</td>
<td>6.10E-08</td>
</tr>
<tr>
<td>0.13</td>
<td>0.13</td>
<td>14</td>
<td>0.00</td>
</tr>
<tr>
<td>0.14</td>
<td>0.14</td>
<td>14</td>
<td>3.60E-07</td>
</tr>
<tr>
<td>0.15</td>
<td>0.15</td>
<td>14</td>
<td>7.63E-07</td>
</tr>
<tr>
<td>0.16</td>
<td>0.16</td>
<td>14</td>
<td>1.79E-06</td>
</tr>
<tr>
<td>0.17</td>
<td>0.17</td>
<td>14</td>
<td>2.80E-06</td>
</tr>
<tr>
<td>0.18</td>
<td>0.18</td>
<td>14</td>
<td>5.26E-06</td>
</tr>
<tr>
<td>0.19</td>
<td>0.19</td>
<td>14</td>
<td>7.77E-06</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>15</td>
<td>3.25E-06</td>
</tr>
</tbody>
</table>

*Degree threshold of 0.01 means that the reserved genes have top one percentile (0.01) degree in all genes of HPRD. Betweenness_centrality threshold is set in the same way as degree threshold.

*In order to ensure the comparability of our method and the method based on correlation of gene expression and drug activity, we kept the number of predicted drug-CRGs obtained by both methods equal with each other.

The subsection of “Performance of the proposed method to identify drug-CRGs” in the page 17

Our findings indicated that our approach was almost exclusively superior to the traditional method based on gene expression. The mean area under ROC curve (AUC) for our method is 65.2%, whereas that for the traditional method AUC is 55.2%.

3) The conclusion is that the authors found correlation between drug activity and
genes rather than response to therapy. Yet the first sentence of the abstract is misleading states that it is imperative to predict chemotherapeutic response, which is a different concept. The difference is significant in clinical care, many “active” drugs fail to create a clinically meaningful response. The authors must either restrict the statement to in vitro cellular chemosensitivity or add an additional sentence that brings the subject of the abstract back to in vitro cellular chemosensitivity before providing results. The third sentences of the abstract are superfluous, and instead the authors could mention their hypothesis. The first sentence of the results in the abstract does not provide a clear idea that the method employs.

REPLY: Many thanks to the reviewer for this comment. We have adjusted the first and third sentences of the abstract. We restricted the statement to in vitro cellular chemosensitivity. The third sentence of the abstract was deleted, “Majority of previous methods are based on gene expression to determine CRGs”. According to the reviewer’s suggestion, we describe our hypothesis in this place. Our hypothesis is that chemosensitivity related genes (CRGs) may share functional characteristics and network features in protein interaction networks.

We have incorporated the following elaborations in this revision.

“Abstract” in the page 2

“The identification of genes that predict in vitro cellular chemosensitivity of cancer cells is of great importance. Chemosensitivity related genes (CRGs) have been widely utilized to guide clinical and cancer chemotherapy decisions. In addition, CRGs potentially share functional characteristics and network features in protein-protein interaction network (PPIN).”

4) The majority of the letters and symbols used in Figure 1 are not documented in the figure’s legend, what does C1, C2 mean?

REPLY: Thank you for your comments. In Figure 1, D represents drug activity profile of the NCI 60 cell lines, each row represents a drug and each column represents a cell line, each element aij represents the drug activity (GI50) of drug dj in cell line Cj, i=1,2,...,4444, j=1,2,...,59. The total number of drugs we analyzed in the manuscript was 4444. G represents the gene expression profiles of the NCI 60 cell lines, each row represents a gene and each column represents a cell line, each element eij represents the expression level of gene gi in cell line Cj, i=1,2,...,12633. The total number of genes we analyzed in the manuscript was 12633.

We have re-drawn the Figure 1 and added detailed description to its legend.
The legend of Figure 1.

"Figure 1. Outline of the proposed method. Firstly, we manually curated a compendium of curated chemosensitivity related genes (CCRGs) from published papers. Then we selected genes on the Microarray that had same enriched GO categories and network characteristics with the CCRGs. These genes were considered as candidate CRGs. To get CRGs for each drug, we further filtered the initial drug-candidate CRG network based on PCC of drug-CCRGs. Filter A is based on Gene Ontology. We characterized CCRG using GO enrichment analysis with Fisher Exact Test. We considered three aspects of GO: biological process (BP), molecular function (MF), and cellular component (CC). p represents the enrichment significance. If enriched p value is smaller than 0.01, CCRGs are significantly enriched in the GO term. Moreover, we investigated that whether CCRGs exhibited functional consistency. We compared the functional similarity of CCRG enriched GO terms to randomly selected gene enriched GO terms. We found that CCRG enriched GO terms exhibited higher similarity scores compared to randomly selected genes. Thus, we regarded all genes in the enriched GO terms as candidate CRGs. Filter B is based on protein interaction networks. We analyzed several network features such as degree and betweenness centrality in six PPINs. Degree and betweenness centrality were selected as network features to prioritize CRGs. The green curve represents betweenness centrality of random genes, and the vertical green line is the betweenness centrality of CCRGs. The blue curve represents degree of random genes, and the vertical blue line is the degree of CCRGs. Filter C is based on gene expression. The majority of drug-CCRGs exhibit a low correlation between gene expression and drug activity. We ranked the absolute PCC of all drug-CCRG pairs in ascending order and set the
5) Figure 2 would benefit from a better definition in the legend of the vertical spaces corresponding to the numbers 80.6% and 91.9%.

REPLY: We thank the reviewer for pointing out our incomplete definition in the legend of Figure 2. The value 80.6% indicates that drug-CCRG pairs whose PCC range from -0.3 to 0.3 accounts for 80.6% of all drug-CCRG pairs; 91.9% indicates that drug-CCRG pairs whose PCC range from -0.5 to 0.5 accounts for 91.9% of all drug-CCRG pairs.

We have incorporated the following elaborations in this revision.

The legend of Figure 2 in the page 34

“Figure 2 - The distribution of Pearson’s correlation coefficient (PCC) of drug-CCRG pairs.
The x-axis represents PCC, the y-axis represents the frequency of a certain PCC. The value 80.6% indicates that drug-CCRG pairs whose PCC range from -0.3 to 0.3 accounts for 80.6% of all drug-CCRG pairs; 91.9% indicates that drug-CCRG pairs whose PCC range from -0.5 to 0.5 accounts for 91.9% of all drug-CCRG pairs.”

6) Figure 3 also need more definitions of the different thresholds, A,B,C,D, what does 21/62 mean (it is found in the main manuscript but not the figure legend – thus the figure legend is nearly useless to interpret the results).

REPLY: Many thanks for your suggestion. To evaluate the PCC between drug activity and gene expression, we utilized only those drugs with drug activity data and genes with expression data available for further analysis. A total of 150 drug-CCRG pairs included 64 drugs and 94 genes; 47 of 94 genes were detected for their expression in the NCI 60 cell lines while 31 of 64 drugs were detected for their activity in the NCI 60 cell lines; 31 drugs and 47 genes comprised 62 drug-CCRG pairs of the original 150 drug-CCRG pairs. We then conducted correlation-based analysis on these 62 drug-CCRG pairs.

For each of the 62 drug-CCRG pairs, we compared the PCC with that of random drug-gene pairs. The statistical method we used is \( z_i = \left| \frac{x_i - \mu}{\delta} \right| \), where \( x_i \) is the PCC of drug-CCRG pair \( i \), \( \mu \) and \( \delta \) are the mean and standard deviation of PCCs of all random drug-gene pairs. If \( z_i \geq z_{\text{threshold}} \), the PCC of drug-CCRG pair \( i \) is significantly different from random PCC. Figure 3A shows the number of recalled drug-CCRG pairs under different thresholds. The numbers of drug-CCRG pairs,
which were identified under the corresponding $z_{\text{threshold}}$, were listed over the blue bar. The stricter the $z_{\text{threshold}}$ was, the fewer drug-CCRG pairs were identified. For example, when using 1 as the $z_{\text{threshold}}$, only 32 of 62 drug-CCRGs were identified, while using 2 as $z_{\text{threshold}}$, only 15 of 62 were identified and 6 of 62 when using 3 as $z_{\text{threshold}}$. We set $z_{\text{threshold}}$ to 0.8 in concordance with previous reports (Proc Natl Acad Sci U S A 2001, 98:10787-10792). Under this threshold, we conducted further analysis (Figure 3B, Figure 3C, and Figure 3D). In Figure 3B, “21/62” indicates that 21 of 62 drug-CCRG pairs exhibit significantly smaller $PCC$s than random drug-gene pairs. In this figure, we illustrate the distribution of correlations between Doxorubicin and all genes. The red line indicates the example of Doxorubicin and ABCB1 correlation. It was reported that ABCB1 overexpression predicted doxorubicin resistance (Cell Biochem Biophys 2009, 55:95-105). In Figure 3C, “14/62” indicates that 14 of 62 drug-CCRG pairs exhibit significantly larger $PCC$s than random drug-gene pairs. In this figure, we illustrate the distribution of correlations between NSC169517 and all genes. The red line indicates the example of NSC169517 and PRKCB correlation. It was reported that PRKCB was able to predict chemosensitivity of NSC169517 (Mol Cancer Ther 2002, 1:311-320). In Figure 3D, “27/62” indicates that 27 of 62 drug-CCRG pairs do not exhibit significantly different $PCC$s from random drug-gene pairs. In this figure, we illustrate the distribution of correlations between Paclitaxel (taxol) and all genes. The red line indicates the example of Paclitaxel (taxol) and GRIK1 correlation. It was reported that GRIK1 was able to predict chemosensitivity of Paclitaxel (BMC Med Genomics, 4:18).

We revised the manuscript to improve the readability and avoid confusions. We have incorporated the following elaborations in this revision.

The subsection of “Correlation-based analysis of the drug-CCRG relationships” in the page 13

“Figure 3A shows the number of identified drug-CCRG pairs under different thresholds. If $z_i \geq z_{\text{threshold}}$, the $PCC$ of drug-CCRG pair $i$ is significantly different from random $PCC$. The numbers of drug-CCRG pairs, which were identified under the corresponding $z_{\text{threshold}}$ were listed over the blue bar. The stricter the $z_{\text{threshold}}$ was, the fewer drug-CCRG pairs were identified. For example, when using 1 as the $z_{\text{threshold}}$, only 32 of 62 drug-CCRGs were identified, while using 2 as $z_{\text{threshold}}$, only 15 of 62 were identified and using 3 as the $z_{\text{threshold}}$ only 6 of 62 drug-CCRG pairs were identified. As shown in Figure 3A, we found it was not sufficient to identify drug-CCRG pairs using $PCC$ based on random analysis. We set the threshold to 0.8 in concordance with the previous reports [17]. Among the 62 drug-CCRG pairs, 21 pairs exhibit the smaller $PCC$ than random drug-gene pairs (Figure 3B), 14 pairs exhibit the larger $PCC$ than random drug-gene pairs (Figure 3C) and 27 pairs exhibit random $PCC$ (Figure 3D).”
“Figure 3 - Correlation-based analysis of the drug-CCRG relationships.
(A) For each of the 62 drug-CCRG pairs, we compared the $PCC$ of drug-CCRG with that of random drug-gene pairs. The statistical method we used is $z_i = |x_i - \mu| / \delta$, where $x_i$ is the $PCC$ of drug-CCRG pair $i$, $\mu$ and $\delta$ are the mean and standard deviation of $PCC$s of all random drug-gene pairs. If $z_i \geq z_{threshold}$, the $PCC$ of drug-CCRG pair $i$ is significantly different from random $PCC$. The numbers of drug-CCRG pairs, which were identified under the corresponding $z_{threshold}$, were listed over the blue bar. The stricter the $z_{threshold}$ was, the fewer drug-CCRG pairs were identified. For example, when using 1 as the $z_{threshold}$, only 32 of 62 drug-CCRGs were identified, while using 2 as $z_{threshold}$, only 15 of 62 were identified and 6 of 62 when using 3 as $z_{threshold}$. We set $z_{threshold}$ to 0.8 in concordance with previous reports (Proc Natl Acad Sci U S A 2001, 98:10787-10792). Under this threshold, we conducted further analysis (Figure 3B, Figure 3C, and Figure 3D). (B, C, D) Three types of $PCC$ distribution compared to random $PCC$. The x-axis shows the $PCC$ of drug-gene pair, the y-axis shows the probability density value of $PCC$. The red line represents the $PCC$ of a drug-CCRG pair, while the blue curves shows the distribution of $PCC$ of random drug-gene pairs. (B) $PCC$ of drug-CCRG is significantly smaller than $PCC$ of random drug-gene pairs. 21/62 indicates that 21 of 62 drug-CCRG pairs exhibit $PCC$s significantly smaller than random $PCC$s. We offered an example between doxorubicin and ABCB1. It was reported that ABCB1 overexpression predicts doxorubicin resistance. (C) $PCC$ of drug-CCRG is significantly larger than that of random drug-gene pairs. 14/62 indicates that 14 of 62 drug-CCRG pairs exhibit $PCC$s significantly larger than random $PCC$s. It was reported PRKCB can predict chemosensitivity of NSC169517. (D) $PCC$ of drug-CCRG is similar with that of random drug-gene pairs. 27/62 indicates that 27 of 62 drug-CCRG pairs do not exhibit $PCC$s significantly different from random $PCC$s. It was reported that GRIK1 was able to predict chemosensitivity of paclitaxel (taxol).”

7) The authors do not provide their annotated tables – only the ontologies about the tables. For reproducibility, the tables curated by the authors should be provided.

REPLY: Thank you very much for your suggestion. We didn’t clearly specify the annotated tables. It refers to Additional file 1. In the Methods section “Curating drug-CCRG pairs” of the manuscript, we changed the last sentence “All the information for drug-CCRG pairs is in Additional file 1” into “The annotated drug-CCRGs table is supplemented in Additional file 1”. Each entry in the table contained detailed information on general name of the drug, gene symbol of CCRG, the cell line where the relationship was identified, the PubMed ID of the reference that reported this relationship, and a brief description of the drug-CCRG pair. For example, the over-expression of Macrophage inhibitory cytokine-1 (MIC-1) predicts sensitivity of ribotoxic anisomycin. CCRG enriched Gene Ontology terms were supplemented in Additional file 3. All of the 204 enriched terms and their enrichment significance were listed in the table.

We have incorporated the following elaborations in this revision.
The subsection of “Curating drug-CCRG relationships” in the page 8

“For example, over-expression of Macrophage inhibitory cytokine-1 (MIC-1) predicted sensitivity of ribotoxic anisomycin. The annotated drug-CCRG table is supplemented in Additional file 1.”
Reply to Reviewer #2 (report from 28 May, 2012)

Discretionary Revisions
1) The manuscript would clearly from a careful language checking from a professional, who is native in English. At some places I have found hard to follow the story because of inappropriate usage of the language.

REPLY: We have carefully corrected the English errors in the previous version to improve its overall clarity and readability. The revised manuscript has been internally reviewed and proofread by several professionals, and the commercial copyediting service, International Science Editing (MS ID: ISE37610), against any potential academic or grammar errors.

Minor Essential Revisions:
1) In the Background section, please explain the relationship of aspargine synthetase protein to the prediction of the efficiency of different chemotharapies. It is not clear from the text, what is the connection of this enzyme to the overall topic, and why it is cited here.

REPLY: Thank you for your comment. In the Background, we did not explain the relationship between aspargine synthetase protein expression and the prediction of the efficiency of L-ASP activity. According to the reference (Mol Cancer Ther 2008, 7:3123-3128), the Pearson’s correlation coefficient between aspargine synthetase protein and L-ASP activity is -0.65 ($P=0.0014$, one-tailed). This indicates that ovarian cell lines which express low ASNS protein levels are more sensitive to L-ASP treatment. The underlying mechanism may be that ovarian cell lines produce less asparagine and are therefore more dependent on extracellular asparagine to meet metabolic demands.

We cited the reference twice in our manuscript. For the first time, we cited it to convey that aspargine synthetase protein was a predictor of L-ASP activity, at the same time, aspargine synthetase protein and mRNA expression is significantly consistent ($r=0.65$, $P=0.016$). The aspargine synthetase gene is theoretically able to predict L-ASP activity. For the second time, we want to convey that according to Pearson’s correlation coefficient between aspargine synthetase mRNA and L-ASP activity ($r=-0.21$), aspargine synthetase mRNA was not able to predict L-ASP activity. Thus, we would like to identify CRGs based on GO categories and PPIN.

We have incorporated the following elaborations in this revision.

The subsection of “Introduction” in the page 4

*Asparagine synthetase protein expression measured by immunoassay is a predictor of
L-asparaginase activity in ovarian cancer cell lines [6]. Ovarian cancer cell lines that express low ASNS protein levels are generally more sensitive to L-ASP treatment."

The subsection of “Introduction” in the page 5

“Lorenzi et al. reported that correlation coefficient of some drug-gene was not high (r=0.21). The gene would not be regarded as CRG based on correlation analysis. However, aspargine synthetase was able to predict sensitivity of L-ASP [6].”

2) Please cite the NCI-60 panel properly. Please cite e.g. Shoemaker RH 2006 Nat Rev Cancer. Oct;6(10):813-23. PMID: 16990858 in relationship with this dataset. Additionally, it is true that the NCI-60 dataset was produced in the National Cancer Institute, but you have provided the URL for such an institute with this name in Egypt, although the correct one is in Bethesda MD, USA. The correct URL (if needed) is http://genome-www.stanford.edu/nci60/index.shtml.

REPLY: Thank you very much for your carefulness and scrutiny in the reviewing process. We corrected the URL to http://genome-www.stanford.edu/nci60/index.shtml and cited Shoemaker RH 2006 Nat Rev Cancer. Oct;6(10):813-23 when referring to NCI 60 panel.

We have incorporated the following elaborations in this revision.

The subsection of “Introduction” in the page 4

“The National Cancer Institute has used a panel of 60 diverse human cancer cell lines (NCI 60 cell line) (http://genome-www.stanford.edu/nci60/index.shtml) for drug-related research [8].”

3) In the results part the authors write about recalling the drug-candidate CRGs: “majority of these drug-CCRG pairs are not big enough to be recalled”. Can you clarify this sentence?

REPLY: Thank you very much for your comment. In the Results section we did not clearly express our original meaning. The value 80.6% indicates that drug-CCRG pairs whose PCC range from -0.3 to 0.3 accounts for 80.6% of all drug-CCRG pairs; 91.9% indicates that drug-CCRG pairs whose PCC range from -0.5 to 0.5 accounts for 91.9% of all drug-CCRG pairs. Thus when we identify the drug-candidate CRGs with high PCC (PCC 3%=0.39, PCC 0.5%=0.51, both PCC thresholds are set in concordance with previous studies [39, 40]), the majority of PCCs for these drug-CCRG pairs are not large enough to pass the PCC threshold. This suggests that the majority of the drug-CCRG pairs will not be identified using traditional high PCC methods based on gene expression.

We have incorporated the following elaborations in this revision.
“In Figure 2, drug-CCRG pairs whose \( PCC \) range from -0.3 to 0.3 accounts for 80.6% of all drug-CCRG pairs while drug-CCRG pairs whose \( PCC \) range from -0.5 to 0.5 accounts for 91.9% of all drug-CCRG pairs. Thus when we identify the drug-candidate CRGs with high \( PCC \) (\( PCC0.3\% \)=0.39, \( PCC0.5\% \)=0.51, both \( PCC \) thresholds are set in concordance with previous studies [39, 40]), the \( PCC \)s of the majority of drug-CCRG pairs fall below the cut off threshold.”

4) In Methods, the authors write about how they have identified relevant papers in PubMed, they specified ‘National Cancer Institute’ as a keyword in the 'any' field of the search submission form. I find it too restrictive, and I would argue that this way the authors actually filtered out most of the relevant hits from the results. The authors should comment this, make some searches without this keyword, and either extend their base dataset or defend their practice in the light of these comparisons.

REPLY: Thank the reviewer for the critical comment with respect to the keyword search process in PubMed. Before identifying relevant papers in PubMed, we designed numerous search strategies and finally determined the searching method in the manuscript. It is an important work to construct a base dataset about human chemosensitivity related genes. If we specifically aim to construct such a base dataset, it is really too restrictive when we add ‘National Cancer Institute’ to the keyword in 'any' field of the search submission form. Because this will filter out most of the relevant hits based on other cell lines aside from the NCI 60 cell lines. As shown in Figure 1, our analysis was based on gene expression data and drug activity data of the NCI 60 cell lines. We characterized the features of drug-CCRG pairs in the NCI 60 cell lines and then prioritized chemosensitivity genes whose expressions were detected in the NCI 60 cell lines.

We used ‘NCI60’ or ‘NCI 60’ as a keyword to identify relevant papers in PubMed. This search strategy may filter out papers based on experiments implemented in only subset of NCI 60 cell lines, such as breast cancer cell line (MCF7). When we used ‘NCI60’ or ‘NCI 60’ as a keyword to search relevant papers in PubMed, we identified 215 papers. When we used ‘National Cancer Institute’ as a keyword, we identified 492 papers. From this perspective, using ‘National Cancer Institute’ as a keyword in 'any' field of the search submission form extended our dataset.

Among all of the 150 Drug-CCRG pairs, 40 pairs were identified in the “NCI 60” cell lines, while the rest of the 110 drug-CCRG pairs were all identified in subsets of NCI 60 cell lines, such as breast cancer (MCF7), colon cancer (HCT116), Non-small-cell lung carcinoma (H1650), prostate cancer (PC_3), and Leukemia (HL 60).

5) Betweenness should be correctly termed as 'betweenness centrality'. The authors should explain their selection of this particular centrality measure, as several others
have been described to be relevant for protein interaction networks (e.g. closeness centrality, eigenvector centrality, etc.).

**REPLY:** Thank you very much for the suggestion. We have revised “betweenness” as “betweenness centrality” in the revised manuscript.

Betweenness centrality is a global centrality index that quantifies to what extent a gene controls the information flow between all pairs of genes in the network. According to the reviewer’s suggestion, we calculated another two centrality measures: closeness centrality and eigenvector centrality. Both of them are not all significant features in six PPIN. In “BIND” and “MINT”, both centrality measures for CCRGs are not significantly larger than random genes. Therefore, we did not utilize them for CRG prioritizing.

### Closeness centrality of CCRG compared with random genes

<table>
<thead>
<tr>
<th></th>
<th>mean of CCRG</th>
<th>mean of random genes</th>
<th>Fold*</th>
<th>p value#</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIND</td>
<td>0.00007</td>
<td>0.00007</td>
<td>1.04719</td>
<td>0.03</td>
</tr>
<tr>
<td>IntAct</td>
<td>0.00003</td>
<td>0.00003</td>
<td>1.06977</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>MINT</td>
<td>0.00005</td>
<td>0.00005</td>
<td>1.00968</td>
<td>0.296</td>
</tr>
<tr>
<td>HPRD</td>
<td>0.00003</td>
<td>0.00003</td>
<td>1.10273</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>BioGRID</td>
<td>0.00003</td>
<td>0.00002</td>
<td>1.10634</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>OPHID</td>
<td>0.00003</td>
<td>0.00002</td>
<td>1.10566</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

*Fold is the result of mean closeness centrality of CCRG divided by that of random genes.

#1000 random gene sets was randomly selected from network, each randomization kept the number of selected genes same as that of CCRGs. Then p value was calculated over 1000 randomization.

### Eigenvector centrality of CCRG compared with random genes

<table>
<thead>
<tr>
<th></th>
<th>mean of CCRG</th>
<th>mean of random genes</th>
<th>Fold*</th>
<th>p value#</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIND</td>
<td>0.0266</td>
<td>0.0236</td>
<td>1.12719</td>
<td>0.385</td>
</tr>
<tr>
<td>IntAct</td>
<td>0.0376</td>
<td>0.0088</td>
<td>4.272727</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>MINT</td>
<td>0.0255</td>
<td>0.0176</td>
<td>1.448864</td>
<td>0.129</td>
</tr>
<tr>
<td>HPRD</td>
<td>0.0963</td>
<td>0.0234S</td>
<td>4.115385</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>BioGRID</td>
<td>0.0402</td>
<td>0.0037</td>
<td>10.86486</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>OPHID</td>
<td>0.0789</td>
<td>0.022</td>
<td>3.586364</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

*Fold is the result of mean Eigenvector centrality of CCRG divided by that of random genes.

#1000 random gene sets was randomly selected from network, each randomization kept the number of selected genes same as that of CCRGs. Then p value was calculated over 1000 randomization.

**Major Compulsory Revisions**
1) I have a concern that the authors present their approach as ‘novel’. It might be possible that this kind of prioritization is novel in relationship with chemosensitivity, but very similar approaches has been proposed in case of other phenotypes. I would suggest to take at least the following articles and explain the similarities and differences between the presented methods:


**REPLY:** Many thanks to the reviewer for the comments. We have cited the three papers, [30-32], in this revision. These studies have identified disease genes, radioresistance genes and drug target genes based on Gene Ontology and protein-interaction networks. As the reviewer mentioned, our paper focused on the prediction of drug chemosensitivity based on the characteristics of curated chemosensitivity related genes in GO and PPIN.

Ortutay and Vihinen (Nucleic Acids Res 2009, 37:622-628) identified candidate primary immunodeficiency genes by integrating Gene Ontologies and protein-interaction networks. Ortutay and Vihinen aimed to identify candidate primary immunodeficiency genes while our paper aimed to identify chemosensitivity related genes. Ortutay et al. combined the gene lists with high-network scores. They used the union of three gene lists predicted by degree, closeness centrality and vulnerability. In our method, we used the intersection of gene lists predicted by degree and betweenness centrality.

Guo et al. (Int J Oncol, 40:85-92) identified genes involved in radioresistance of nasopharyngeal carcinoma by integrating gene ontology and protein-protein interaction networks. Guo et al. identified 183 significantly differentially expressed genes (DEGs) in CNE-2R vs CNE-2 cell line. Then they employed GOEAST and STRING to identify DEG enriched GO terms and network features. Our method firstly characterized the features of CCRGs (curated chemosensitivity related gene) from Gene Ontology and protein interaction networks, and then we used CCRG enriched GO categories and network features to prioritize CRGs (chemosensitivity related gene).

Kaimal et al. proposed an integrative system biology approach (ToppGeNet) to identify and prioritize disease and drug candidate genes. ToppGeNet combined other gene information from public knowledgebases, such as Mouse Phenotype, drug, and coexpression. For our method, before using functional annotation (GO), we detected if CCRGs are much more similar in GO categories. After finding out that CCRGs were really much more similar to each other than random genes, we further utilized this function annotation for CRG prioritization.

We have incorporated the following elaborations in this revision.
“Sensitivity to a variety of compounds may be also influenced by certain aspects of Gene Ontology (GO) functionality, such as cell death, NADH dehydrogenase activity, ABC transporter, cell adhesion, G-protein coupled receptor protein signalling and macromolecule metabolism [16, 24, 26-29]. Previous studies have identified disease genes, radioresistance genes and drug target genes based on Gene Ontology and protein-interaction networks [30-32].”

2) The authors should clarify the nature and structure of the CCRG network the authors used in the further analysis. They mentioned that out of the 150 drug-gene pairs they used 62. How many gene and how many drug nodes the initial network has? How many connections? Is the graph connected, or does it consist of isolated sub-graphs?

REPLY: Maybe our unclear descriptions confused the reviewer. In the manuscript, CCRG is the abbreviation of curated chemosensitivity related gene. Drug-candidate CRG network has no abbreviation in the manuscript. The drug-candidate CRG network is a fully connected network. It is comprised of two kinds of nodes, drug nodes and gene nodes. The drug-candidate CRG network includes 4444 drugs, 12633 genes and 4444*12633 edges. Each of 4444 drugs was detected for their activity in NCI 60 panel and each of 12633 genes was detected for their expression in NCI 60 panel.

In the manuscript, we used 62 of the 150 drug-CCR pairs for correlation based analysis. The $PCC$ was evaluated between drug activity and gene expression in the NCI 60 cell lines. Therefore, only those drug with drug activity and genes with expression are available for further analysis. A total of 62 of the original 150 drug-CCR pairs were used for correlation-based analysis; based on correlation analysis, we proposed our method to prioritize CRGs.

We have incorporated the following elaborations in this revision.

The subsection of “Correlation-based analysis of the drug-CCRG relationships” in the page 12, 13

“A gene with expression highly correlated to drug activity is regarded as a candidate CRG for the drug. Thus, we initially investigated whether CCRGs were highly correlated with their interactive drugs. Of the 150 pairs of drug-CCR, 62 pairs were available for correlation analysis. We evaluated the $PCC$ between drug activity and gene expression for drug with drug activity and genes with expression available in the NCI 60 cell lines. The 150 drug-CCR pairs included 64 drugs and 94 genes. A total of 47 of 94 genes were detected for their expression in NCI 60 cell lines and 31 of 64 drugs were detected for their activity in NCI 60 cell lines; these 31 drugs and 47 genes comprised 62 drug-CCR pairs of the original 150 drug-CCR pairs. We then performed
correlation-based analysis on these 62 drug-CCRG pairs. In Figure 2, drug-CCRG pairs whose PCC range from -0.3 to 0.3 accounts for 80.6% of all drug-CCRG pairs while drug-CCRG pairs whose PCC range from -0.5 to 0.5 accounts for 91.9% of all drug-CCRG pairs.”

3) The authors should write down very clearly that they have used actually two independent graphs (networks): the drug-CCRG network and the protein-protein interaction network. The text does not make clear distinction between these, and it makes the interpretation of the whole manuscript hard.

REPLY: Many thanks to the reviewer for the comment. We actually used two independent networks, one is drug-candidate CRG network, and the other is protein-protein interaction network. CCRG in “Drug-CCRG” is the abbreviation of “curated chemosensitivity related gene”. In the manuscript, drug-CCRG specifically refers to “drug-curated chemosensitivity related gene”. Drug-candidate CRG networks have no abbreviation in the manuscript. There are two kinds of nodes, drug nodes and gene nodes. The drug-candidate CRG network included 4444 drugs, 12633 genes and 4444*12633 edges. Each of 4444 drugs was detected for activity in the NCI 60 panel and each of 12633 genes was detected for expression in the NCI 60 panel.

We have incorporated the following elaborations in this revision.

The subsection of “Drug activity data and gene expression data” in the page 9

“D represents drug activity profile of the NCI 60 cell lines, each row represents a drug and each column represents a cell line, each element a_{ij} represents the drug activity (GI50) of drug d_j in cell line C_i, i=1,2,…,4444, j=1,2,…,59. G represents the gene expression profiles of the NCI 60 cell lines, each row represents a gene and each column represents a cell line, each element e_{ij} represents the expression level of gene g_i in cell line C_j, i=1,2,…,12633. The total number of genes we analyzed in the manuscript was 12633.”

The section of “Methods” in the page 7

“ “drug-CCRG” specifically represents “drug- curated chemosensitivity related gene”.”

4) The authors should provide more information about how they used the protein interaction data from HPRD. Did they reconstructed the entire human interactome and used it for degree and network score calculation? Did they used ready or custom made software?

REPLY: Thank the reviewer for pointing out the incomplete details about HPRD. PPI (protein-protein interaction) data in Human Protein Reference Database (HPRD) (Genome Res 2003, 13:9) are experimentally derived, and have been manually
extracted from the literature by expert biologists who read, interpret and analyze the published data. We downloaded protein interaction data from HPRD on the website http://www.hprd.org/download. The number of binary non-redundant human PPIs is 36,687 in HPRD. The number of genes annotated with at least one interaction is 9,408. We did not reconstruct the entire human interactome, the HPRD network we used is the original network as we downloaded from the official website. We used six PPIN to characterize CCRGs’ network features, respectively. They are BIND, BioGRID, IntAct, MINT, OPHID and HPRD.

As protein-protein interactions in HPRD are experimentally derived human protein-protein interaction data, we further utilized HPRD to prioritize CRGs. When we calculated network scores, we utilized “matlab BGL” toolbox and R package “igraph”.

We have incorporated the following elaborations in this revision.

The subsection of “Protein-protein interaction network” in the page 11

“PPI (protein-protein interaction) data in Human Protein Reference Database (HPRD) [38] are experimentally derived and manually extracted from the literature by expert biologists who read, interpret and analyze the published data. We downloaded protein interaction data from HPRD on the website http://www.hprd.org/download. The number of binary non-redundant human PPIs is 36,687 in HPRD. The number of genes annotated with at least one interaction is 9,408. We utilized “matlab BGL” toolbox and R package “igraph” to calculate network scores.”