

# Deciphering Drug Action and Escape Pathways: An Example on Nasopharyngeal Carcinoma

Difeng Dong<sup>1,\*</sup>, Chun-Ying Cui<sup>2,\*</sup>, Benjamin Mow<sup>3</sup>, Limsoon Wong<sup>1</sup>

<sup>1</sup>National University of Singapore, Singapore

<sup>2</sup>Capital Medical University, China

<sup>3</sup>West Clinic Excellence Cancer Center, Singapore

\*These two authors contributed equally to this study.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

## ABSTRACT

**Motivation:** Recently, a cyclin dependent kinase (CDK) inhibitor, CYC202, is studied for its anti-tumor effect in human Nasopharyngeal Carcinoma (NPC) cells *in vitro* and *in vivo*. Both cell lines and patients in the study responded to the drug treatment differently. Our target is to understand the drug action of CYC202 in these NPC samples as well as to identify escape pathways for the drug-resistant NPC individuals. Existing computational tools focus on pathway enrichment analysis by identifying informative genes which are differently expressed between two response groups, but little information on the interplay between selected genes is provided. The identifications are too general and hardly sufficient to generate specific hypotheses for our research purpose.

**Results:** We design a drug pathway identification system, the Drug Pathway Decipherer (DPD), to identify genetic responsive pathways to drug treatment. DPD generates hypotheses of specific genetic pathways based on the knowledge of canonical biological pathways, which promises the identifications to be properly interpreted in a biological context. By applying DPD to the NPC datasets, we find the suppression of RAS-ERK cell proliferation pathway and PI3K-NF $\kappa$ B-IAP anti-apoptotic pathway correlate well with the effective CYC202 treatment in NPC cells both *in vitro* and *in vivo*. These observations are further confirmed with the associated medical assays. On the other hand, drug escape pathways may be heterogeneous for non-responders. Based on our identifications, we give suggestions to optimize the treatment of these CYC202-resistant NPC patients.

**Availability:** The Drug Pathway Decipherer is available at <http://www.comp.nus.edu.sg/~wongls/projects/drug-pathway/DPD-v1>. It is implemented in JAVA.

**Contact:** [dongdife@comp.nus.edu.sg](mailto:dongdife@comp.nus.edu.sg), [ccy@ccmu.edu.cn](mailto:ccy@ccmu.edu.cn), [bmow@westexcellence.com](mailto:bmow@westexcellence.com), and [wongls@comp.nus.edu.sg](mailto:wongls@comp.nus.edu.sg)

**Supplementary information:** Supplementary data are available at <http://www.comp.nus.edu.sg/~wongls/projects/drug-pathway>.

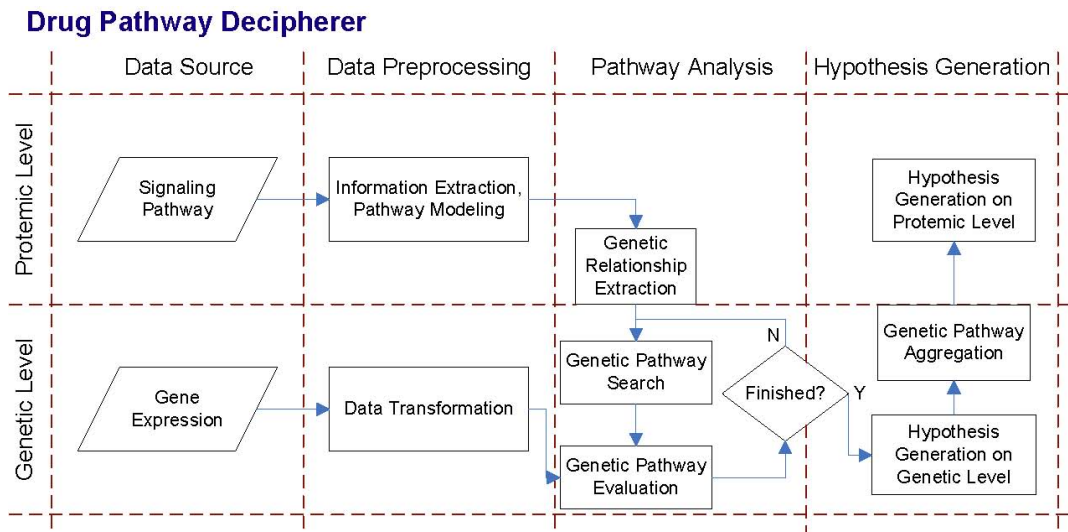
## INTRODUCTION

Biological pathway information has been incorporated into gene expression analysis to understand drug treatment response in disease populations (Soh *et al.*, 2007). Some works focus on the enrichment analysis of gene groups extracted from pathways (Zeeberg *et al.*, 2003; Doniger *et al.*, 2003; Subramanian *et al.*, 2005; Sivachenko

*et al.*, 2005, 2007). Zeeberg *et al.* (2003) and Doniger *et al.* (2003) use the hypergeometric test to determine statistically over-represented pathways in a given list of differentially expressed genes along treatment. Subramanian *et al.* (2005) propose the gene set enrichment analysis (GSEA), which uses a weighted Kolmogorov-Smirnov statistics to compare the two sets of distributions and also uses resampling to estimate false discovery rates (FDR). Sivachenko *et al.* (2005, 2007) split genes into separate regulatory groups, each sharing the same transcriptional regulators, and evaluate these gene groups in a GSEA-like manner. Other research groups identify responsive genetic networks under drug treatment (Zien *et al.*, 2000; Ideker *et al.*, 2002; Hanisch *et al.*, 2002; Guo *et al.*, 2007). Hanisch *et al.* (2002) cluster genes with a metric preferring both genetic co-expression and short distance within a network topology. Zien *et al.* (2000) exhaustively enumerate all possible gene combinations on a metabolic pathway, and select the most co-expressed gene group as the responsive pathway. Ideker *et al.* (2002) extend the method of Zien *et al.* (2000) to a protein-protein interaction network, and use an annealed random selection to generate candidate gene subnetworks for statistical evaluation. Guo *et al.* (2007) follow Ideker *et al.* (2002), but their evaluation is based on the co-expression between interacted genes rather than the significance of expression change of genes on the identified subnetworks.

However, most existing works fall short on several issues (Soh *et al.*, 2007): these works provide little information on the interplay between selected genes; the collection of pathways that can be used, evaluated and ranked against the observed expression data is limited; and the generated hypotheses are still too general to guide further research and treatment. In this paper, we present a drug pathway identification system, which we called Drug Pathway Decipherer<sup>1</sup> (DPD), to identify genetic responsive pathways of drug treatment. Different from existing works, DPD generates hypotheses of specific genetic pathways based on the knowledge of canonical biological pathways, which promises the identifications to be properly interpreted in a biological context. We apply DPD to two NPC gene expression datasets in our recent research. CYC202 (Cyclacel Ltd, Dundee, United Kingdom; Seliciclib; R-roscovitine),

<sup>1</sup> DPD is a framework for statistical evaluation of known genetic pathways against gene expression data. It consists of 4 partitions distributed on two biological levels. Figure 1 gives the diagram of its workflow.



**Fig. 1.** The workflow of DPD. (a) Data source: Structured signaling pathways and drug treatment gene expression data are taken as the input. (b) Data preprocessing: Signaling pathways are modeled as graphs, and gene expression change under drug treatment are computed. (c) Pathway analysis: Pairwise genetic relationships are extracted from the modeled signaling pathways and evaluated against the preprocessed gene expression data. (d) Hypothesis generation: Co-expressed genetic relationships are selected to be connected into complete genetic pathways, and statistical tools are performed to generate drug pathway hypotheses. Signaling pathway status are estimated based on the hypothesized genetic pathways.

a CDK inhibitor, is studied for its anti-tumor effect in human NPC cells *in vitro* and *in vivo*. 3 NPC cell lines and 13 NPC patients were treated with CYC202, and the expression of selected genes were measured during the process of the treatment. Both cell lines and patients in the study responded to the drug treatment differently. Our target is to understand the drug action of CYC202 in these NPC samples as well as to identify escape pathways for the drug-resistant NPC individuals. As a result of applying DPD to the datasets, we find that both of the identifications from the *in vitro* and *in vivo* experiments are consistent with the result of associated medical assays as well as the pathogenesis mechanism of NPC and known drug action of CYC202 from the literature. Thus, we show that DPD provides a reasonable statistical framework for genetic drug responsive pathway identification in drug treatment gene expression data. In addition, the current version of DPD allows users to construct, remove, and modify biological pathways for their own research purposes.

## SYSTEM AND METHODS

### Data source

Both NPC gene expression datasets contain 380 genes selected for apoptosis, cell proliferation, and cell cycle regulation. For the *in vitro* set, 3 cell lines, CNE1, CNE2 and HK1 were measured for their gene expression before the treatment of CYC202, and 2hs, 4hs, 6hs, 12hs and 24hs after the treatment, respectively. It was observed that CNE1 responded poorly to the treatment; CNE2 responded in a limited way; and HK1 fully responded. For the *in vivo* set, 12 NPC samples and 1 non-tumor sample were taken from NPC patients, who were traced for their response to the treatment of CYC202.

Gene expression were measured before and after the treatment. 7 patients were reported to have a molecular response to the treatment.

With respect to the selected genes in the datasets, 4 related signaling pathways are extracted from KEGG pathway database (October 17, 2007) (Kanehisa *et al.*, 2002): ERK pathway (hsa04010), JNK/p38 pathway (hsa04010), G1/S cell cycle progression (hsa04110) and Apoptosis pathway (hsa04210). The extracted pathway diagrams are available in Figure Suppl1.

### Preprocessing data source

In order to capture gene expression change in response to drug treatment, the original gene expression data are transformed into the Relative Expression<sup>2</sup> (RE). Rather than a log-ratio transformation, RE describes expression change in multiples in a linear scale, which allows the pairwise drug effect on gene expression can be measured by a linear correlation metric. Signaling pathways can be modeled as directed graphs, for a consistent denotation, we formally define: A signaling pathway  $\gamma$  is a directed graph  $(P, I)$ , with  $P$  the vertex set, representing the collection of proteins on pathway, and  $I$  the edge set, representing the collection of interactions between proteins. An interaction is a triplet  $i = \langle p_1, p_2, s \rangle$ , with  $p_1, p_2 \in P$  and  $s \in S$ , where  $S = \{\$stimulation, \$suppression\}$  is the set of terms used to denote interaction types.

### Extracting genetic relationships

Assuming  $G$  is a gene set, and  $T = \{\$positive, \$negative\}$ , is an associated terminology set used to describe relations between genes

<sup>2</sup> Given a time-course gene expression dataset  $E$ , its corresponding RE dataset is  $R$ , where  $e_{ij}$  and  $r_{ij}$  are the original expression value and RE value of gene  $i$  at time point  $j$ , respectively. If  $e_{ij} \geq e_{i0}$ , then  $r_{ij} = e_{ij}/e_{i0} - 1$ ; otherwise,  $r_{ij} = 1 - e_{i0}/e_{ij}$ .

in  $G$ , a genetic relationship (or simply a relationship) is a triplet  $q = \langle g_1, g_2, t \rangle$ , with  $g_1, g_2 \in G$  and  $t \in T$ .

The extraction of genetic relationships from a signaling pathway is a mapping from interaction to relationship. Proteins are mapped to their encoding genes, and  $\$stimulation$  and  $\$suppression$  are mapped to  $\$positive$  and  $\$negative$ , respectively. As a protein can be encoded by more than one gene, multiple genetic relationships are extracted from one interaction.

### Scoring a genetic pathway

A genetic pathway  $\vartheta$  is a string of connected genes (without loop), started with a source gene and ended with a sink gene w.r.t. the structure of a signaling pathway (see supplementary data for examples of the identified genetic pathways). It can be decomposed into a set of consecutive relationships. Thus, to score a genetic pathway, we first introduce the function to score a relationship.

Given a relationship  $q = \langle g_1, g_2, t \rangle$ , if the expression of  $g_1$  and  $g_2$  are measured at multiple time points (as our *in vitro* dataset), then the correlation of  $q$  is:

$$Corr(q) = Corr(\vec{r}_{g_1}, \vec{r}_{g_2}),$$

where  $Corr(\vec{r}_{g_1}, \vec{r}_{g_2})$  is the Pearson correlation coefficient between RE vectors  $\vec{r}_{g_1}$  and  $\vec{r}_{g_2}$ . If gene expression are only measured at two time points (as our *in vivo* dataset), then the correlation is estimated simply by comparing the post-treatment RE of the two genes:

$$Corr(q) = \frac{sgn(r_{g_1}^{post}) \times sgn(r_{g_2}^{post}) \times \min_{i=1,2} |r_{g_i}^{post}|}{\max_{j=1,2} |r_{g_j}^{post}|}.$$

$Corr(q)$  is then transformed into a  $z$ -score,  $z(q)$ , from the sample-wise correlation background.  $z(q)$  are then summed up over all  $k$  relationships in  $\vartheta$  to produce an aggregated  $z$ -score,  $z(\vartheta)$ , for the entire genetic pathway<sup>3</sup>:

$$z(\vartheta) = \frac{1}{\sqrt{k}} \sum_{q \in \vartheta} (-1)^\alpha z(q),$$

where  $\alpha = 0$  if  $q.relation = \$positive$ ;  $\alpha = 1$  if  $q.relation = \$negative$ , which suggests if two genes have a  $\$negative$  relation, their RE are expected to be negatively correlated as well.

For each pathway  $\vartheta$ , genes on pathway are permuted 10000 times to estimate the p-value of  $z(\vartheta)$ , denoted by  $score(\vartheta)$ . Intuitively, the pathway score represents the consistency between a genetic pathway and the expression change of genes on it.

### Generating hypotheses

Genetic pathways s.t. the statistical requirement of p-value<sup>4</sup> and FDR<sup>5</sup> are selected as the identified genetic pathway hypotheses.

<sup>3</sup> We follow the statistics of Ideker *et al.* (2002), which promises that if  $z(q)$  follows a standard distribution, then  $z(\vartheta)$  will also be distributed according to a normal distribution.

<sup>4</sup> Since the pathway score itself is a measurement of p-value, the statistical significance control is straight forward.

<sup>5</sup> To select the proper p-value threshold for FDR control, we first rank the scores of pathways which pass the p-value filtering. Then, we identify the maximal rank index  $j$ , s.t.  $p_j < \frac{j \cdot \alpha}{C_N \cdot N}$ , where  $p_j$  is  $j$ -th ranked p-value;  $\alpha$  is the user specified threshold;  $N$  is the total number of hypotheses; and  $C_N = \sum_{i=1}^N \frac{1}{i}$ , is the constant for dependent test (Herrington, 2002).

Since multiple genetic pathways are identified for a single signaling pathway, to evaluate the drug effect, we estimate the signaling pathway status based on the genetic identifications. The pathway score is converted into a probability metric, confidence, denoted by  $conf(\vartheta)$ , where  $conf(\vartheta) = 1 - score(\vartheta)$ . Each gene  $g$  on a genetic pathway has a relation (or indirect relation) with the downstream cellular event (denoted as virtual node in the pathway diagram, see figure Suppl1), which is called the impact of gene  $g$  on pathway  $\vartheta$ , denoted by  $impact_\vartheta(g)$ . If  $g$  is a suppressor of the downstream event, then  $impact_\vartheta(g) = -1$ , which means the downstream event has a negative correlation with the expression regulation of  $g$ ; otherwise,  $impact_\vartheta(g) = 1$ <sup>6</sup>.

Thus, for a signaling pathway  $\gamma$ , let  $\vartheta \sim \gamma$  represent the identified genetic pathway  $\vartheta$  for  $\gamma$ , and  $G_\vartheta$  represent the gene set on  $\vartheta$ . The signaling pathway status  $Z_i^\gamma$  is a weighted aggregation of RE of genes on the identified genetic pathways of  $\gamma$ , respecting to their pathway impact, at time point  $i$ , with the weight being the fraction of the confidence of a genetic pathway compared to that of whole identifications, which is in formula:

$$Z_i^\gamma = \sum_{\vartheta \sim \gamma} \sum_{g \in G_\vartheta} \left( \frac{1}{|G_\vartheta|} \times impact_\vartheta(g) \times r_{gi} \times \frac{conf(\vartheta)}{\sum_{\vartheta' \sim \gamma} conf(\vartheta')} \right).$$

Similarly, the confidence of status of  $\gamma$  is a weighted aggregation of the confidence of  $\vartheta$ , represented in formula as:

$$conf(Z^\gamma) = \sum_{\vartheta \sim \gamma} \left( conf(\vartheta) \times \frac{conf(\vartheta)}{\sum_{\vartheta' \sim \gamma} conf(\vartheta')} \right).$$

The pathway status is an aggregation of expression regulation of genes on the identified genetic drug responsive pathways of a signaling pathway. Therefore, it can be used as a benchmark to compare the regulation of signaling pathways between samples. In our study, we compare pathway status between NPC cell lines along the treatment of CYC202. To evaluate whether the identifications can well differentiate signaling pathway regulation between cell lines, we compute the maximal difference of pathway status between them<sup>7</sup>, and permute the hypothesized genes for 10000 times within the same signaling pathway to get the statistical significance of the difference.

## RESULTS AND DISCUSSION

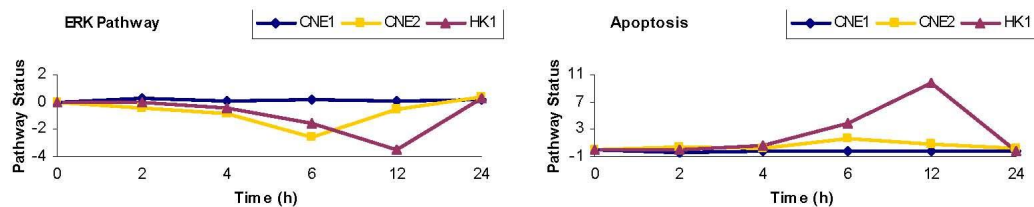
We apply DPD to our NPC gene expression datasets to identify CYC202 responsive pathways *in vitro* and *in vivo* with  $p \leq 0.05$  and  $FDR \leq 0.5$ . For cell lines, gene expression at all 6 time points are used as the input; for patients, pre- and post- treatment data are used. Table 1 shows the identified genetic pathways in 3 NPC cell lines, together with their pathway scores. We discover RAS-ERK cell proliferation pathway and PI3K-NFκB-IAP anti-apoptotic pathway in all 3 cell lines, but for the other two signaling pathways,

<sup>6</sup> Under this definition, a gene may have both positive and negative impact to the downstream cellular event, due to its multiple roles in different genetic pathways of the same signaling pathway. However, there would be no contradiction, since the measurement of gene impact is genetic pathway based.

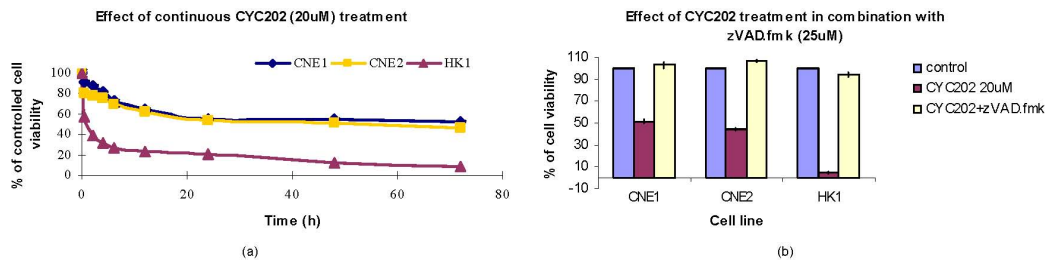
<sup>7</sup> The maximal difference of status of pathway  $\gamma$  between sample  $s_1$  and  $s_2$  is computed as  $diff_\gamma(s_1, s_2) = \max_i |Z_i^{s_1} - Z_i^{s_2}|$ .

**Table 1.** The identified genetic pathways for 3 NPC cell lines: Genes for replacement are separated by “/”; “→” and “-” represent “\$positive” and “\$negative” relation between two genes, respectively.

| Signaling Pathway | Genetic Pathway  | Score    | Confidence |
|-------------------|--|----------|------------|
| CNE1              |  |          |            |
| ERK               | Mitogen→GRB2→SOS2→HRAS→RAF1→MAP2K1→MAPK1/MAPK3→Cell Proliferation          | ≤ 0.001  | ≥ 0.999    |
| Apoptosis         | PIK3CB→PTEN→AKT2/AKT3→CHUK/IKBKB/IKBKG→NFKB2→BIRC2/BIRC5- Apoptosis        | ≤ 0.0002 | ≥ 0.9998   |
| JNK/p38           | Stress→MAP3K12→MAP2K7→MAPK9→Cell Mortality                                 | ≤ 0.04   | ≥ 0.9665   |
| G1/S              | CCND1→CDK4- RB1- E2F2/E2F3→G1/S Progression                                | ≤ 0.01   | ≥ 0.9906   |
| CNE2              |  |          |            |
| ERK               | Mitogen→GRB2→SOS1→MRAS/KRAS/NRAS/RRAS→BRAF→MAP2K1→MAPK1→Cell Proliferation | ≤ 0.03   | ≥ 0.9885   |
| Apoptosis         | PIK3CA/PIK3CB→PTEN→AKT1→IKBKB→RELA→BIRC2/BIRC5- Apoptosis                  | ≤ 0.01   | ≥ 0.9949   |
| JNK/p38           | Cytokinin→MAP4K3/TRAF2→MAP3K1→MAP2K4→MAPK8/MAPK10→Cell Mortality           | ≤ 0.04   | ≥ 0.9658   |
| HK1               |  |          |            |
| ERK               | Mitogen→GRB2→SOS1→HRAS→BRAF→MAP2K1/MAP2K2→MAPK1/MAPK3→Cell Proliferation   | ≤ 0.04   | ≥ 0.9646   |
| Apoptosis         | PIK3R1→PTEN→AKT2/AKT3→IKBKB→NFKB2/RELA→BCL2/BIRC2- Apoptosis               | ≤ 0.04   | ≥ 0.9663   |
| G1/S              | CUL1→SKP2→CDKN1A- CDK6- RB1- E2F2/E2F3→G1/S Progression                    | ≤ 0.04   | ≥ 0.9645   |



**Fig. 2.** Comparable status of ERK pathway and Apoptosis pathway of the 3 NPC cell lines along the treatment of CYC202.



**Fig. 3.** Results of the associated medical assays to measure the cell viability and apoptosis level under the treatment of CYC202 for NPC cell lines: (a) The results of trypan blue test for measuring the cell viability along the drug treatment. (b) The extent of caspase-dependent apoptosis. zVAD.fmk is a caspase activity inhibitor.

no consensus is reached (see Figure Suppl3-5 for diagrams of the identifications of cell lines highlighted on the studied signaling pathways). The identification of the anti-apoptotic pathway is interesting, since it suggests the negative control system of cell death responds more significantly to the treatment of CYC202 than the death receptor and mitochondrial regulated pro-apoptotic pathways in NPC cell lines.

To evaluate the biological significance of our identifications, we estimate the status of these signaling pathways with the identified genetic pathways along the process of treatment. Figure 2 shows the status of ERK pathway and Apoptosis pathway (the status of JNK/p38 pathway and G1/S progression are shown in Figure Suppl2). Associated significance evaluations of the

**Table 2.** The statistical significance (p-values) of the difference of signaling pathway status between cell lines.

| Comparison Group | ERK      | Apoptosis | JNK/p38 | G1/S   |
|------------------|----------|-----------|---------|--------|
| CNE1 vs. CNE2    | < 0.0001 | 0.0028    | 0.2921  | -      |
| CNE1 vs. HK1     | < 0.0001 | 0.0006    | -       | 0.4992 |
| CNE2 vs. HK1     | 0.0004   | 0.0022    | -       | -      |

difference of pathway status between cell lines are given in Table 2. The identified genetic pathways on both ERK pathway (4E-4) and Apoptosis pathway (2.8E-3) show a superior of differentiating pathway status to the other genes on the same signaling

**Table 3.** The post-treatment signaling pathway status of the NPC patients: The “response” column shows the molecular response to the treatment of CYC202.

| Patient | Response    | ERK    |       | JNK/p38 |       | G1/S   |       | Apoptosis |       |
|---------|-------------|--------|-------|---------|-------|--------|-------|-----------|-------|
|         |             | Status | Conf. | Status  | Conf. | Status | Conf. | Status    | Conf. |
| Pt5     | P(positive) | -2.25  | 0.98  | -3.08   | 0.99  | -      | -     | 1.34      | 0.99  |
| Pt8     | P           | -      | -     | -1.01   | 0.99  | -      | -     | 0.82      | 0.98  |
| Pt9     | P           | -0.97  | 0.98  | -       | -     | 0.76   | 0.95  | -         | -     |
| Pt14    | P           | -      | -     | -       | -     | -0.61  | 0.99  | -0.86     | 0.99  |
| Pt16    | P           | -0.20  | 0.99  | -0.20   | 0.95  | 0.29   | 0.99  | 1.42      | 0.97  |
| Pt17    | P           | -1.02  | 0.99  | -1.02   | 0.99  | -0.33  | 0.96  | 1.01      | 0.99  |
| Pt19    | P           | -      | -     | -0.86   | 0.98  | -      | -     | 0.91      | 0.98  |
| Pt18    | No Tumor    | -0.15  | 0.99  | -       | -     | 0.28   | 0.99  | 0.13      | 0.99  |
| Pt1     | N(egative)  | 0.21   | 0.95  | 0.52    | 0.99  | 1.06   | 0.97  | -1.00     | 0.98  |
| Pt7     | N           | -0.10  | 0.97  | -0.68   | 0.96  | 0.28   | 0.98  | 0.11      | 0.98  |
| Pt10    | N           | 1.02   | 0.99  | 1.16    | 0.99  | -      | -     | -1.57     | 0.97  |
| Pt15    | N           | -      | -     | -       | -     | -      | -     | -1.01     | 0.98  |
| Pt20    | N           | 1.30   | 0.98  | -       | -     | -0.93  | 0.96  | -1.68     | 0.99  |

pathway. ERK pathway regulates cell survival, proliferation and differentiation. Whenever this pathway is suppressed, cell viability decreases. In Figure 2, ERK pathway is significantly suppressed in the responder, HK1, but less down regulated or almost unchanged in the half-responder, CNE2, and the resister, CNE1. This observation is consistent with the known drug response of these three cell lines. We then evaluate the hypothesis of ERK pathway with the associated trypan blue test, which is used to measure the cell viability along the procedure of the drug treatment (Figure 3 (a)). The results show the coherence between the cell viability and the status of ERK pathway, which support our hypothesis. Apoptosis pathway, regulating cell death, is on the other hand induced in HK1 rather than in CNE1 and CNE2. This observation is also consistent with the known cellular drug response of the cell lines, and is further confirmed with the extra medical assays, which show that the inhibition of caspase activity prohibits apoptosis in HK1 most (Figure 3 (b)).

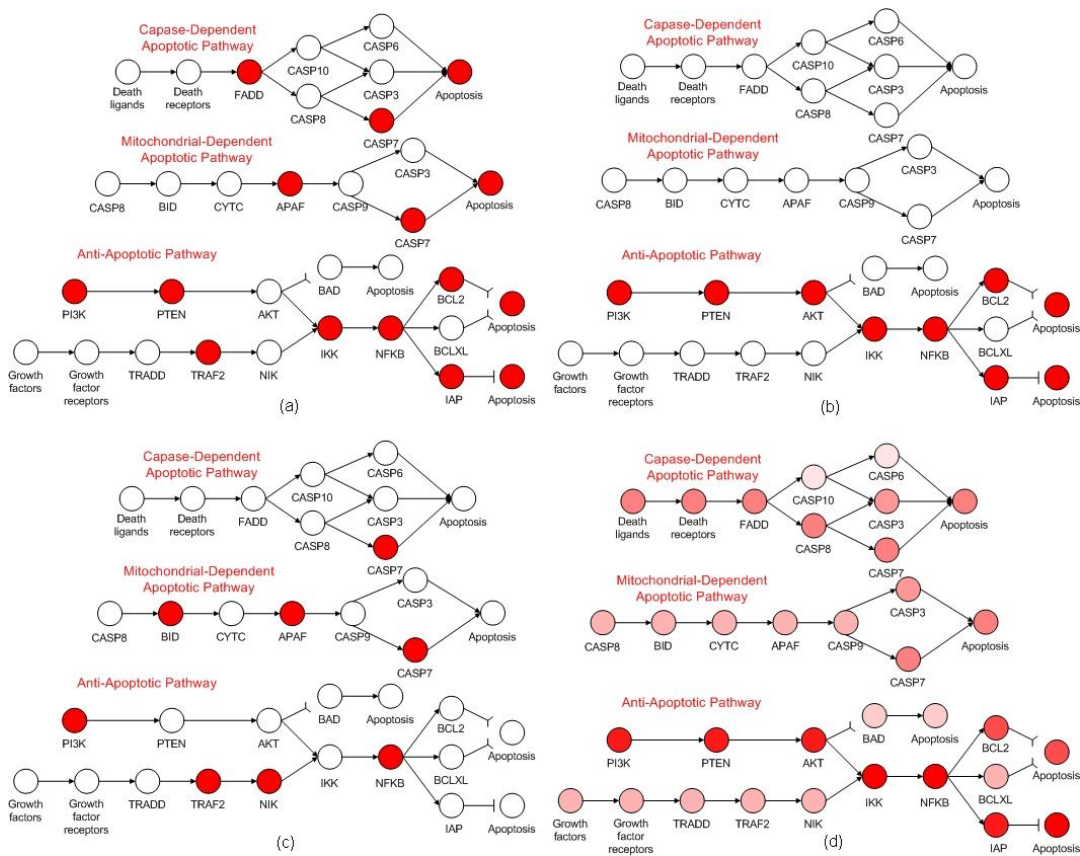
The identified genetic pathways and the post-treatment signaling pathway status of the NPC patients are shown in Figure suppl6-18 and Table 3, respectively. Pt18 is a non-tumor sample. Other patients are classified into two groups w.r.t. their molecular response to the treatment. In Table 3, the pathway status of Pt18 does not changed much under the treatment of CYC202, which can be attributed to the drug specificity to tumor cells. Thus, the pathway status of Pt18 provides a benchmark to evaluate the drug response of the other patients. An significant and interesting observation is that the post-treatment status of ERK pathway and Apoptosis pathway in two responding groups can be almost perfectly separated by that of Pt18 (Except for 14 on Apoptosis pathway). This observation suggests the suppression of ERK pathway and the induction of Apoptosis pathway is related to effective CYC202 treatment *in vivo*, which agree with the conclusion of the *in vitro* experiment.

Since DPD generate specific hypotheses on genetic pathways, we compare it with the leading edge analysis of GSEA. The same 4 signaling pathways are taken as the test gene sets for GSEA. For cell lines, CNE1 and HK1 are compared by the RE at 12hrs after the treatment. The difference of RE between two cell lines is used

as the metric for gene ranking. Gene sets are permuted 1000 times for statistical evaluation. For patients, the analysis is performed on RE between two response groups. T-statistics is used to rank the genes. The response class labels are permuted 1000 times for statistical evaluation. The statistical thresholds for p-value and FDR are 0.05 and 0.25, respectively. All other parameters for GSEA are with default values.

Only Apoptosis pathway is enriched by GSEA in both datasets with statistical significance. Figure 4 compares the results of the genes selected by the leading edge analysis of GSEA and the identified pathways of DPD on Apoptosis pathway. For cell lines, both GSEA and DPD identify the pattern of PI3K-NF $\kappa$ B-IAP anti-apoptotic pathway (Figure 4a and 4b), whose biological significance has been confirmed in the previous experiments. However, GSEA misses gene AKT on the pathway, and selects some irrelevant genes on other pathways. This is because the purpose of the leading edge analysis is to select genes expressed differently between the two response groups, but the relationships between genes of selection are ignored. When performed to the patient dataset, GSEA does not identify any strong genetic pathway pattern (Figure 4c), but for DPD, multiple pathways with different significance are identified (Figure 4d). The most significant identification is still PI3K-NF $\kappa$ B-IAP pathway, which indicates the main genetic response of patients on Apoptosis pathway is similar to that of cell lines. Another discovery is the death receptor regulated pro-apoptotic pathway. This pathway is previously undiscovered in the *in vitro* experiments, which means there exists alternative apoptosis regulation pathway for CYC202 in NPC patients rather than in cell lines. Thus, we show that, compared with the leading edge analysis of GSEA, DPD generates more biologically meaningful results, which can be used as a guide for further drug research and disease treatment. Based on our identifications, we give proposal to the CYC202-resistant NPC patients for their disease treatment (Table 4).

In addition, despite the feasibility of pathway identification for individuals, DPD is also capable of discovering consistent drug responsive pathway in a specific population. To show this point, we treat the RE of two response groups of patients as the input of DPD respectively, and compare the identifications on Apoptosis



**Fig. 4.** Contrast results of the genes identified by the leading edge analysis of GSEA and the pathways identified by DPD: (a) GSEA performed on the *in vitro* dataset. The identified genes are highlighted. (b) DPD performed on the *in vitro* dataset. Since the identifications of CNE1 and HK1 on Apoptosis pathway are the same, the pathway is highlighted in the figure. (c) GSEA performed on the *in vivo* dataset. (d) DPD performed on the *in vivo* dataset. Color density represents the frequency of a pathway identified in patients.

pathway. Results show that for responders, both PI3K-NFκB-IAP anti-apoptotic pathway and the mitochondria regulated pro-apoptotic pathway are identified, but for the non-responders, only the anti-apoptotic pathway is identified<sup>8</sup> (Figure Suppl19-20).

Epstein-Barr Virus (EBV) infection is known to play a critical role in the pathogenesis of NPC (Pathmanathan *et al.*, 1995). The dysregulation of multiple signaling pathways, including NFκB, MAPK, JAK-STAT and PI3K-AKT are induced by EBV infection (Tsao *et al.*, 2002). Particularly, it is specified that the up regulation of NFκB2 and BIRC5 (IAP) contribute in increasing resistance to apoptosis, and the role of BIRC5 in resisting apoptosis in NPC has been confirmed by RNA interference (Shi *et al.*, 2006). On the other hand, CYC202 inhibits CDK2, -7 and -9 through competitive inhibition of ATP binding (Mcclue *et al.*, 2002). CDK7 and CDK9 phosphorylate the carboxyl terminal domain of RNA

polymerase II, which initiates the gene transcription. Due to the suppression of gene transcription, the greatest effect is observed on gene products with short mRNA and protein half life, such as apoptosis regulators, including NFκB targeted genes and IAP family (Lam *et al.*, 2001). The suppression of genes involved in ERK pathway and anti-apoptotic pathway, including MAPK1, MAPK3, MCL1, BCL2, BIRC4 and BIRC5, are frequently observed associated with the treatment of CYC202 (Meijer *et al.*, 1997; Whittaker *et al.*, 2004; Alvi *et al.*, 2005; Raje *et al.*, 2005; Smith and Yue, 2006; Lacrima *et al.*, 2005). In the present study, DPD identifies the different regulation of RAS-ERK cell proliferation pathway and PI3K-NFκB-IAP anti-apoptotic pathway between two drug response groups both *in vitro* and *in vivo*, which are consistent with the known drug action of CYC202 and the pathogenesis mechanism of NPC from the literature. Thus, we conclude that these two pathways are the main drug pathways of CYC202 in human NPC cells. On the other hand, due to the diversity of individual genetic environment of patients, the identified escape pathways are heterogeneous. The dysregulation of NFκB pathway and MAPK pathway are both commonly observed in CYC202-resistant patients. More details are included in the personal treatment proposals (Table 4).

<sup>8</sup> This identification is reasonable. For GSEA, 3 genes are identified on the mitochondria regulated pro-apoptotic pathway (Figure 4(c)), which suggests these genes are differently regulated during the treatment of CYC202. The results of DPD show that the identified difference by GSEA is mainly because of the pathway regulation in the drug-responsive group.

**Table 4.** Treatment proposal for CYC202-resistant NPC patients, based on the identifications of DPD.

| Patient   | Comments  |
|-----------|---|
| Patient1  | PI3K-NFκB-IAP pathway and G1/S progression are dysregulated. JNK/p38 pathway is activated only by the cytokinin regulation. Radiotherapy is suggested to be used together with CYC202 to further activate the stress regulated JNK/p38 pathway to promote the suppression of NFκB activity and the induction of caspase activity.                               |
| Patient7  | No significant drug-resistant pathway pattern is identified for this patient. We suggest to increase the dose of CYC202 or to combine the treatment with other CDK inhibitors, such as Olomoucine and Staurosporine, to further suppress the cell cycle progression.  |
| Patient10 | The pathway regulation shows full resistance to the treatment of CYC202. Both ERK pathway and PI3K-NFκB-IAP anti-apoptotic pathway are dysregulated. It is suggested to use other therapy, such as radiotherapy, to replace the treatment of CYC202.  |
| Patient15 | This patient shows a significant resistant pattern to the drug treatment on both pro- and anti-apoptotic pathway. There is no identification for other signaling pathways. Radiotherapy is recommended to be used in stead of CYC202. An alternatively is to use drugs that regulate apoptosis via other pathways, such as p53 regulated pro-apoptotic pathway. |
| Patient20 | PI3K-BAD anti-apoptotic pathway is identified rather than the NFκB regulated one. The function of NFκB is suspected to be dysregulated. Drugs regulating apoptosis via NFκB-independent pathway are recommended.  |

## CONCLUSIONS AND FUTURE WORK

In this paper, we introduce our drug pathway identification system, DPD, to identify responsive genetic pathways under drug treatment. We apply the system to two gene expression datasets of human NPC cells treated with a CDK inhibitor, CYC202, *in vitro* and *in vivo*. The identifications suggest RAS-ERK cell proliferation pathway and PI3K-NFκB-IAP anti-apoptotic pathway are the main CYC202 regulated pathways in NPC, and for non-responders, the escape pathways are heterogeneous. In addition, DPD is compared with GSEA for the feasibility of generating biological meaningful hypotheses. It is shown that the results of DPD is more interpretable in a biological context and more useful for guiding further drug research and disease treatment. Finally, based the biologically meaningful identifications on the NPC datasets, we conclude that DPD provides a reasonable statistical framework for genetic drug responsive pathway identification in drug treatment gene expression data. However, we need to specify that due to the limitation of the NPC study, an apparent problem of our research is that only limited signaling pathways are available for evaluation in the current package of DPD, and we have not tested DPD on other datasets. To overcome this issue, we have started extracting large scale cancer related pathways from several public pathway databases, and the results of applying DPD to other datasets will be presented in our future work.

## ACKNOWLEDGEMENT

The NPC patient data are kindly provided by Dr. Boon Cher Goh, National University Hospital Singapore. This work is supported in part by a NUS research scholarship (Dong) and a MOE AcRF Tier 1 grant (Wong).

## REFERENCES

- Alvi,A. et al. (2005) A novel CDK inhibitor, CYC202 (R-roscovitine), overcomes the defect in p53-independent apoptosis in B-CLL by down-regulation of genes involved in transcription regulation and survival, *Blood*, **105**, 4484-4491.
- Doniger,S. et al. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data, *Genome Biology*, **4**, R7.
- Guo,Z. et al. (2007) Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network, *Bioinformatics*, **23**, 2121-2128.
- Hanisch,D. et al. (2002) Co-clustering of biological networks and gene expression data, *Bioinformatics*, **18**, s145-s154.
- Herrington,H. (2002) Controlling the false discovery rate in multiple hypothesis testing, <http://www.unt.edu/benchmarks/archives/2002/april02/rss.htm>.
- Ideker,T. et al. (2002) Discovering regulatory and signalling circuits in molecular interaction networks, *Bioinformatics*, **18**, s233-s240.
- Kanehisa,M. et al. (2002) The KEGG database at GenomeNet, *Nucleic Acids Research*, **30**, 42-46.
- Lacrima,K. et al. (2005) *In vitro* activity of cyclin-dependent kinase inhibitor CYC202 (Seliciclib, R-roscovitine) in mantle cell lymphomas, *Annals of Oncology*, **16**, 1169-1176.
- Lam,L. et al. (2001) Genomic-scale measurement of mRNA turnover and the mechanisms of action of the anti-cancer drug flavopiridol, *Genome Biology*, **2**(10), research004.
- Mcclue,S. et al. (2002) *In vitro* and *in vivo* antitumor properties of the cyclin dependent kinase inhibitor CYC202 (R-ROSCOVITINE), *International Journal of Cancer*, **102**, 463-468.
- Meijer,L. et al. (1997) Biochemical and cellular effects of roscovitine, a potent and selective inhibitor of the cyclin-dependent kinase cdc2, cdk2 and cdk5, *European Journal of Biochemistry*, **243**, 527-536.
- Pathmanathan,R. et al. (1995) Clonal proliferations of cells infected with Epstein-Barr virus in preinvasive lesions related to nasopharyngeal carcinoma, *The New England Journal of Medicine*, **333**, 693-698.
- Raje,N. et al. (2005) Seliciclib (CYC202 or R-roscovitine), a small-molecule cyclin-dependent kinase inhibitor, mediates activity via down-regulation of MCL1 in multiple myeloma, *Blood*, **106**, 1042-1047.
- Shi,W. et al. (2006) Multiple dysregulated pathways in nasopharyngeal carcinoma revealed by gene expression profiling, *International Journal of Cancer*, **119**, 2467-2475.
- Sivachenko,A. et al. (2005) Identifying local gene expression patterns in biomolecular networks, *Computational Systems Bioinformatics Conference (CSB)*, 180-184, Stanford University.
- Sivachenko,A. et al. (2007) Molecular networks in microarray analysis, *Journal of Bioinformatics and Computational Biology*, **5**, 429-456.
- Smith,P. and Yue.E. (2006) *Inhibitors of Cyclin-dependent Kinases as Anti-tumor Agents*, Taylor and Francis Group.
- Soh,D. et al. (2007) Enabling more sophisticated gene expression analysis for understanding diseases and optimizing treatments, *ACM SIGKDD Explorations*, **9**, 3-14.
- Subramanian,A. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Science of the United States of America*, **102**, 15545-15550.
- Tsao,S. et al. (2002) The significance of LMP1 expression in nasopharyngeal carcinoma, *Cancer Biology*, **12**, 473-487.
- Whittaker,S. et al. (2004) The cyclin-dependent kinase inhibitor CYC202 (R-Roscovitine) inhibits retinoblastoma protein phosphorylation, causes loss of cyclin D1, and activates the mitogen-activated protein kinase pathway, *Cancer Research*, **64**, 262-272.
- Zeeberg,B. et al. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data, *Genome Biology*, **4**(4), R28.
- Zien,A. et al. (2000) Analysis of gene expression data with pathway scores, *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, **8**, 407-417.