Bioinformatics, 32(19), 2016, 2891–2895
doi: 10.1093/bioinformatics/btw344
Advance Access Publication Date: 9 June 2016
Discovery Note

OXFORD

Systems biology

Evaluating the molecule-based prediction of clinical drug responses in cancer

Zijian Ding, Songpeng Zu and Jin Gu*

MOE Key Laboratory of Bioinformatics, TNLIST Bioinformatics Division & Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing 100084, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on March 1, 2016; revised on April 28, 2016; accepted on May 26, 2016

Abstract

Motivation: Molecule-based prediction of drug response is one major task of precision oncology. Recently, large-scale cancer genomic studies, such as The Cancer Genome Atlas (TCGA), provide the opportunity to evaluate the predictive utility of molecular data for clinical drug responses in multiple cancer types.

Results: Here, we first curated the drug treatment information from TCGA. Four chemotherapeutic drugs had more than 180 clinical response records. Then, we developed a computational framework to evaluate the molecule based predictions of clinical responses of the four drugs and to identify the corresponding molecular signatures. Results show that mRNA or miRNA expressions can predict drug responses significantly better than random classifiers in specific cancer types. A few signature genes are involved in drug response related pathways, such as DDB1 in DNA repair pathway and DLL4 in Notch signaling pathway. Finally, we applied the framework to predict responses across multiple cancer types and found that the prediction performances get improved for cisplatin based on miRNA expressions. Integrative analysis of clinical drug response data and molecular data offers opportunities for discovering predictive markers in cancer. This study provides a starting point to objectively evaluate the molecule-based predictions of clinical drug responses.

Contact: jgu@tsinghua.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Molecular polymorphisms and alterations play important roles in heterogeneous anti-cancer drug responses (Wang *et al.*, 2011). The relationships between molecular features and clinical drug responses lay the foundation for optimizing drug therapies based on a patient's genomic context. For example, metastatic breast tumors with HER2 overexpression are sensitive to trastuzumab (England, 2001); EGFR mutations in non-small-cell lung cancer predict response to gefitinib (Thompson *et al.*, 2004). With the fast development and reduced costs of high throughput technologies, more efforts are made to identify genomic markers that can predict drug responses (Barretina *et al.*, 2012; Garnett *et al.*, 2012). As precision medicine takes into account the genomic variability of individuals in oncology practice (Collins and Varmus, 2015; Garraway *et al.*, 2013), accurately predicting response to cancer drugs based on molecules becomes a critical issue.

Cancer cell lines have been used to find gene expression signatures which can predict *in vitro* drug sensitivities (Potti *et al.*, 2006). However, whether cell lines can capture the complex molecular alterations in real cancer patients is still in debate (Goodspeed *et al.*, 2016). A few large-scale cancer genome projects, such as The Cancer Genome Atlas (TCGA), not only provide diverse molecular data but also clinical information of cancer patients. By now, TCGA has comprehensively characterized tens of cancer types via the multi-dimensional analysis of gene mutation, copy number alteration, DNA methylation, mRNA, miRNA and protein expression (http://cancergenome.nih.gov/). TCGA has also expanded the knowledge about genomic similarities among multiple cancer lineages via

the Pan-Cancer studies (Chang *et al.*, 2013). Its molecular data promotes the development of oncology practices, such as targeted therapeutics (Rubio-perez *et al.*, 2015) and prognosis prediction (Yuan *et al.*, 2014). However, the predictive utility of the diverse molecular data for clinical drug responses has not been explored.

In this study, we aim at evaluating the utility of molecular data for predicting clinical drug responses in cancer based on TCGA data. First of all, we carefully curated the clinical data of drug treatments from TCGA, extracted 152 drugs and 2572 patients with drug response records, and then focused on four drugs (cisplatin, paclitaxel, carboplatin and fluorouracil) with relatively more clinical response records. Then we constructed a computational framework to evaluate the performances of the molecule-based prediction of clinical drug responses using copy number alterations (CNAs), DNA methylations, miRNA expressions and mRNA expressions. Compared to random classifiers, expression data shows significantly better performances on cisplatin or paclitaxel in specific cancer types. Also, we found that some signature genes are involved in important cellular processes known to mediate drug responses, such as DDB1 in DNA repair pathway and DLL4 in Notch signaling pathway. In the following multiple cancer analysis, it was found that miRNA expressions exhibit good predictive performances across cancer types and the classifier fitted to multiple cancer data can improve the predictions on certain single cancer types. Our work offers the opportunity to study diverse molecular features of clinical drug responses in primary tumors. It is a starting point to objectively evaluate the molecule-based predictions of clinical anti-cancer drug responses.

2 Materials and methods

2.1 Drug response data acquisition

We curated the records of drug treatments from TCGA clinical data. In the 'clinical_drug_cancer.txt' table, each row or entry recorded one pair of drug and patient. After deleting the pairs with missing drug response, we manually standardized the drug names according to NCI drug dictionary and DrugBank (Wishart *et al.*, 2006) (Supplementary Table S1). Then, we deleted the records of those patients who responded inconsistently to one drug during the chronology of therapy, since the different responses may be caused by progressively acquired molecular alterations (Holohan *et al.*, 2013), and established the lists of drug-patient pairs in multiple cancers (Supplementary Tables S2 & S3). Considering that most patients responded to drugs (Supplementary Fig. S1), we combined the clinical responses which used the RECIST standard (Eisenhauer *et al.*, 2009) as two types, namely responder (including complete response and partial response) and non-responder (including stable disease and progressive disease), as in previous studies (Geeleher *et al.*, 2014; Majumder *et al.*, 2015). In addition, we collected the information of neo-adjuvant therapies and the chronological orders between drug treatments and tumor resections (see more details in Supplementary Materials and Fig. S2). Based on these records, we deleted the patients who received drug treatments prior to tumor resections.

2.2 Molecular data collection

We then combined all the available molecular data with the curated clinical drug response data via TCGA patient IDs. The molecular data was downloaded from GDAC Firehose of Broad Institute, including copy number alterations (CNAs), gene mutations, DNA methylations, mRNA expressions, miRNA expressions and protein activities. More descriptions about the gene-level values were provided in Supplementary Materials and Methods.

We constructed 21 core datasets. Each dataset is for one type of molecular data, one cancer type and one drug. The datasets building and data pre-processing were described in the Supplementary Materials and Methods. These molecular data of the core datasets was also provided via link: [http://bioinfo.au.tsinghua.edu.cn/member/jgu/drug_response/].

2.3 Molecule based prediction of clinical drug responses

A computational framework was implemented to estimate the performances of individual molecular data types on predicting clinical drug responses and to identify related molecular signatures at the same time (Supplementary Fig. S3). First, the studied dataset was randomly split into training and testing subsets. Then, elastic net with bootstrapping was used to select the molecular features which can best predict the drug responses in the training dataset. The final ensemble classifier was built based on the recurrently selected features and evaluated on the testing dataset. The dataset splitting process was repeated multiple times to estimate the predictive performances. More descriptions about the framework were provided in Supplementary Materials and Methods. The practical applications of the framework were in Supplementary Table S5.

3 Results

3.1 Evaluation of the molecule-based drug responses predictions in single cancer type

Among the 21 core datasets, expression datasets frequently show high predictive performances (Fig. 1A). Two mRNA datasets of cisplatin-BLCA and cisplatin-CESC get large average AUCs as 0.843 and 0.603 (significantly larger than 0.5 with one-sided Mann Whitney U test, *P*-value 6.09e18 and 2.42e5, respectively). Meanwhile, two miRNA datasets of paclitaxel-BRCA and cisplatin-CESC get the average AUCs as 0.673 and 0.593, respectively (*P*-value 4.77e8 and 5.33 e4). Two methylation datasets of paclitaxel-BRCA and carboplatin-UCEC achieve the average AUCs as 0.611 and 0.587, respectively (*P*-value 2.12e5 and 2.20e3). And the CNA dataset of carboplatin-UCEC obtains the average AUC as 0.590 (*P*-value 8.27e5).

To evaluate the predictive performances more stringently, we performed permutation tests to assess the statistical significances (details provided in Supplementary Materials and Methods). Among the seven datasets with potential predictive power, three have *P*-values smaller than 0.001, including one miRNA dataset of paclitaxel-BRCA, and two mRNA datasets of cisplatin-BLCA and cisplatin-CESC (Fig. 1B).

3.2 Identification of molecular signatures associated with clinical drug responses

For the three datasets with significant predictive capabilities, we identified the corresponding molecular signatures and performed permutation tests to assess the statistical significance of each molecular feature in each signature (details in Supplementary Materials and Methods). Here a molecular signature is defined as a group of features that are recurrently selected by the predictive models. We also assessed the features based on survival analysis, *in vitro* cell line data, and literature mining (details in Supplementary Table S6).

Some genes in the identified signatures are involved in several important cellular processes known to mediate drug responses. For example, highly expressed DDB1 is an indicator of poor response for cisplatin-BLCA. As we know, cisplatin binds to DNA, induces

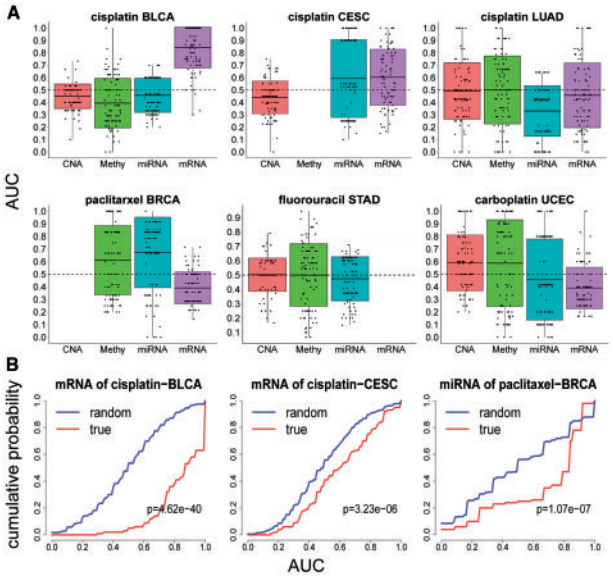

Fig. 1. Performances on predicting drug responses based on molecular data. (**A**) AUC values for four drugs in six cancer types, including Bladder Urothelial Carcinoma (BLCA), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), Lung adenocarcinoma (LUAD), Breast invasive carcinoma (BRCA), Stomach adenocarcinoma (STAD) and Uterine Corpus Endometrial Carcinoma (UCEC). When there is not enough data to build classifiers, the corresponding column is blank. The black dots represent the AUC values. In each box, the middle line represents the mean, the upper and lower lines represent the values of mean ± sd. 'CNA' is short for copy number alteration and 'Methy' for DNA methylation. (**B**) The curves of empirical cumulative distributions of AUC values for both true classifiers and random classifiers. If AUCs tend to have larger values, the curve is more warped to the bottom-right corner. *P*-values were calculated by One-sided Mann-Whitney *U* test

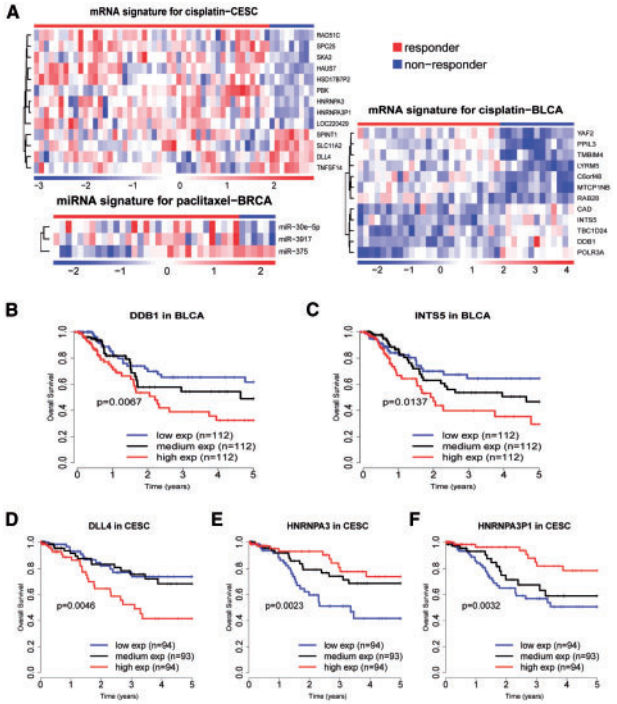

Fig. 2. Molecular signatures for the three datasets with high predictive performances. (**A**) Heat maps of the signatures. Higher expression levels are in red and lower in blue. (**B–F**) Kaplan-Meier survival plots of DDB1 and INTS5 in all patients of BLCA and DDL4, HNRNPA3 and HNRNPA3P1 in all patients of CESC from TCGA (including the patients without cisplatin response). P values were calculated by log-rank test (patients living longer than 5 years were right censored to 5 years)

DNA damages and kills cells via inducing apoptosis (Siddik, 2003). DDB1 is an important positive regulator for nucleotide excision repair (Li *et al.*, 2006) which was responsible for resistance to platinum-based agents (Galluzzi *et al.*, 2014). In TCGA data, highly expressed DDB1 is a strong indicator of poor response to cisplatin (Fig. 2A). High DDB1 expression was also observed in cisplatin resistance cancer cells (Chu and Chang, 1990). Meanwhile, survival analysis on all 336 BLCA patients found that highly expressed DDB1 was correlated with poor prognosis (Fig. 2B). INTS5, encoding a subunit of the Integrator complex, is also significantly correlated with patient survival (Fig. 2C). It's surprising that the mRNA signature as a whole can stratify patients into groups with different risks (Supplementary Fig. S4). Another signature gene, YAF2 is upregulated in the patients with good responses. YAF2 directly interacts with a transcription factor YY1, which has been reported to negatively regulate p53 transcription under genotoxic stress (Grönroos *et al.*, 2004). These results suggest that DDB1 and YAF2 may regulate two different DNA repair pathways which play complex roles in cisplatin resistance.

In the mRNA signature for cisplatin-CESC, DLL4 was identified as a poor response feature. The blockade of DLL4 can inhibit tumor growth through deregulated angiogenesis (Ridgway *et al.*, 2006) which is one cancer hallmark (Hanahan and Weinberg, 2011). In TCGA data, lowly expressed DLL4 is correlated with good response of cisplatin. What's more, lowly expressed DLL4 is correlated with good prognosis in all 281 CESC patients (Fig. 2D). Though DLL4 was not directly reported to be related with cisplatin, NOTCH1, one of DLL4's receptors, is correlated with cisplatin response (Zhou *et al.*, 2014). Besides DDL4, both HNRNPA3P3 and its pseudo gene HNRNPA3P1 are also significantly correlated with CESC patient survival (Fig. 2E, F). In the miRNA signature for paclitaxel-BRCA, miR-30e is lowly expressed in non-responders. The inhibition of miR-30e increases the self-renewal capacity and reduce apoptosis of breast tumor-initiating cells (Yu *et al.*, 2010).

3.3 Evaluation of the molecule-based drug response predictions across cancer types

Though tumors from different tissue origins have heterogeneous molecular patterns, pan-cancer similarities among genomic aberrations have been found, such as CNAs and mutations (Kandoth *et al.*, 2013; Zack *et al.*, 2013). As some drugs are widely used in multiple cancers, the construction of cross-cancer classifiers are helpful for establishing molecule-based instructions for drug usages (Chang *et al.*, 2013). Therefore, we evaluated the predictive performances on cisplatin and carboplatin across cancer types, considering as many as possible the number of available cancer types and patients (Supplementary Table S7).

In the cross-cancer analysis (training and testing on all available cancer types), the CNA and methylation datasets exhibited limited performances in predicting either cisplatin or carboplatin responses. The miRNA and mRNA expression datasets showed higher AUCs for cisplatin (Supplementary Fig. S5A), but only the miRNA expressions get significant better performance than random classifiers according to the permutation test (*P*-value 1.00e−14).

Then we evaluated the performances via testing on single cancer types based on the models trained on multiple-cancer datasets. Out of the 15 tests, eight show significantly different performances in

comparison with the single cancer type analyses (training and testing both on single cancer type, two-sided paired *t*-test *P*-value < 0.05) (Supplementary Fig. S5B and Table S8). For the miRNA expression datasets, the multiple-cancer classifiers performed much better on cisplatin-BLCA and cisplatin-LUAD datasets (*P*-value 3.13e–5 and 5.95–6, respectively. More details were presented in Supplementary Results). In short, the multiple-cancer classifier of miRNA expression significantly improves performances on predicting cisplatin responses on several single cancer types.

4 Discussion

Predicting clinical drug responses by molecular data in human cancer is one important goal of precision oncology. In this study, we carefully curated drug response records from TCGA and evaluated the performances of diverse molecular data types on predicting the clinical drug responses. In single cancer type analyses, mRNA and miRNA expressions achieve significantly better performances in specific cancer types than random classifiers. Many identified signature genes play important roles in the cellular processes known to mediate drug resistance. We also found that miRNA expressions improve performances for predicting cisplatin responses based on multiple-cancer type analysis. On the contrary, CNAs and DNA methylations show limited predictive capabilities either in single or multiple cancer type analyses.

Overall, the predictive performances of molecule-based classifiers are not so good, although some of them performed significantly better than random classifiers. The following problems may cause current limitations: first, molecular patterns are highly complex due to the tumor heterogeneity among patients (Bedard *et al.*, 2013). The available datasets cannot represent the whole picture of molecular alterations in cancer. Thus, the classifiers trained on limited data could perform poorly on predicting new patients. Second, the molecular features far outnumber the patients. Many irrelevant features may be selected by accident, although the sparse linear model was used to reduce features and avoid model over-fitting. Thirdly, the non-responders are usually much fewer. We observed that the numbers of non-responders are significantly correlated with the average AUCs (Spearman rank correlation 0.988, *P*-value 0.0016) of the mRNA datasets regardless of the drug-cancer pairs.

The predictive utility of mRNA expressions has also been reported in cancer cell lines. According to a DREAM competition, which aimed at identifying what methods and to what extent molecular data can predict *in vitro* drug sensitivities, mRNA expressions are found to provide the best predictive performance among all molecular data types (Costello *et al.*, 2014). It is also reported that protein, mRNA and miRNA abundances provide the best performance when modelling the GI50 endpoint (Cortés-Ciriano *et al.*, 2016).

The identified signature genes and miRNAs based on mRNA or miRNA expression data can provide novel insights of the mechanisms of drug resistances. DDB1 belongs to the cisplatin-resistance correlated nucleotide excision repair pathway, which is regarded as potential therapeutic targets (Pearl *et al.*, 2015). DLL4, down-regulation of which inhibited tumor growth (Ridgway *et al.*, 2006), can be targeted by Demcizumab, a drug in phase I clinical trial (Smith *et al.*, 2014). Other signature genes including INTS5, HNRNPA3 and HNRNPA3P1 also show prognostic power, but the molecular mechanisms are still unclear. miR-30c, a homolog of signature miR-30e and located in the same primary transcript, regulates paclitaxel resistance in breast cancer cells (Bockhorn *et al.*, 2013).

To our knowledge, our work is the first study to investigate the predictive utility of the molecular data for clinical drug responses from TCGA. Also, the curated pharmacogenomics data provides an important resource for future studies.

Acknowledgements

We thank Chao He for helpful discussions in designing the computational methods and Mohamed Nadhir Djekidel for careful revision of the manuscript.

Funding

This work is supported by National Basic Research Program of China [2012CB316503], National Natural Science Foundation of China [61370035 and 31361163004] and Tsinghua University Initiative Scientific Research Program.

Conflict of Interest: none declared.

References

Barretina,J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–307.

Bedard,P.L. *et al.* (2013) Tumour heterogeneity in the clinic. *Nature*, **501**, 355–364.

Bockhorn,J. *et al.* (2013) MicroRNA-30c inhibits human breast tumour chemotherapy resistance by regulating TWF1 and IL-11. *Nat. Commun.*, **4**, 1393.

Chang,K. *et al.* (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet*, **45**, 1113–1120.

Chu,G. and Chang,E. (1990) Cisplatin-resistant cells express increased levels of a factor that recognizes damaged DNA. *Proc. Natl. Acad. Sci. U. S. A.*, **87**, 3324–3327.

Collins,F.S. and Varmus,H. (2015) A new initiative on precision medicine. *N. Engl. J. Med.*, **363**, 1–3.

Cortés-Ciriano,I. *et al.* (2016) Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics*, **32**, 85–95.

Costello,J.C. *et al.* (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.*, **32**, 1–103.

Eisenhauer,E. a. *et al.* (2009) New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur. J. Cancer*, **45**, 228–247.

England,T.N. (2001) Use of chemotherapy plus a monoclonal antibody against Her2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.*, **344**, 783–792.

Galluzzi,L. *et al.* (2014) Systems biology of cisplatin resistance: past, present and future. *Cell Death Dis.*, **5**, e1257.

Garnett,M.J. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.

Garraway,L. a. *et al.* (2013) Precision oncology: an overview. *J. Clin. Oncol.*, **31**, 1803–1805.

Geeleher,P. *et al.* (2014) Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol.*, **15**, R47.

Goodspeed,A. *et al.* (2016) Tumor-derived cell lines as molecular models of cancer pharmacogenomics. *Mol. Cancer Res.*, **14**, 3–13.

Grönroos,E. *et al.* (2004) YY1 inhibits the activation of the p53 tumor suppressor in response to genotoxic stress. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 12165–12170.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

Holohan,C. *et al.* (2013) Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer*, **13**, 714–726.

Kandoth,C. *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333–339.

Li,J. *et al*. (2006) DNA damage binding protein component DDB1 participates in nucleotide excision repair through DDB2 DNA-binding and cullin 4a ubiquitin ligase activity. *Cancer Res*., **66**, 8590–8597.

Majumder,B. *et al*. (2015) Predicting clinical response to anticancer drugs using an ex vivo platform that captures tumour heterogeneity. *Nat. Commun*., **6**, 6169.

Pearl,L.H. *et al*. (2015) Therapeutic opportunities within the DNA damage response. *Nat. Rev. Cancer*, **15**, 166–180.

Potti,A. *et al*. (2006) Genomic signatures to guide the use of chemotherapeutics. *Nat. Med*., **12**, 1294–1300.

Ridgway,J. *et al*. (2006) Inhibition of Dll4 signalling inhibits tumour growth by deregulating angiogenesis. *Nature*, **444**, 1083–1087.

Rubio-perez,C. *et al*. (2015) In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities article in silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell*, **27**, 382–396.

Siddik,Z.H. (2003) Cisplatin: mode of cytotoxic action and molecular basis of resistance. *Oncogene*, **22**, 7265–7279.

Smith,D.C. *et al*. (2014) A phase I dose escalation and expansion study of the anticancer stem cell agent Demcizumab (Anti-DLL4) in patients with previously treated solid tumors. *Clin. Cancer Res*., **20**, 6295–6303.

Thompson,I.M. *et al*. (2004) Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to Gefitinib. *N. Engl. J. Med*., **350**, 2239–2246.

Wang,L. *et al*. (2011) Genomics and drug response. *N. Engl. J. Med*., **364**, 1144–1153.

Wishart,D.S. *et al*. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*., **34**, D668–D672.

Yu,F. *et al*. (2010) Mir-30 reduction maintains self-renewal and inhibits apoptosis in breast tumor-initiating cells. *Oncogene*, **29**, 4194–4204.

Yuan,Y. *et al*. (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol*., **32**, 644–652.

Zack,T.I. *et al*. (2013) Pan-cancer patterns of somatic copy number alteration. *Nat. Genet*., **45**, 1134–1140.

Zhou,Y. *et al*. (2014) MicroRNA-449a reduces cell survival and enhances cisplatin-induced cytotoxicity via downregulation of NOTCH1 in ovarian cancer cells. *Tumor Biol*., 12369–12378.