BIOINFORMATICS DISCOVERY NOTE

Vol. 30 no. 6 2014, pages 748–752
doi:10.1093/bioinformatics/btt606

Data and text mining

Advance Access publication October 24, 2013

Inferring therapeutic targets from heterogeneous data: HKDC1 is a novel potential therapeutic target for cancer

Gong-Hua Li[1] and Jing-Fei Huang[1,2,*]

[1]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences and [2]Kunming Institute of Zoology, Chinese University of Hongkong Joint Research Center for Bio-resources and Human Disease Mechanisms, Kunming, Yunnan 650223, China

Associate Editor: Gunnar Ratsch

ABSTRACT

Motivation: The discovery of therapeutic targets is important for cancer treatment. Although dozens of targets have been used in cancer therapies, cancer remains a serious disease with a high mortality rate. Owing to the expansion of cancer-related data, we now have the opportunity to infer therapeutic targets using computational biology methods.

Results: Here, we describe a method, termed anticancer activity enrichment analysis, used to determine genes that could be used as therapeutic targets. The results show that these genes have high likelihoods of being developed into clinical targets (>60%). Combined with gene expression data, we predicted 50 candidate targets for lung cancer, of which 19 of the top 20 genes are targeted by approved drugs or drugs used in clinical trials. A hexokinase family member, hexokinase domain-containing protein 1 (HKDC1), is the only one of the top 20 genes that has not been targeted by either an approved drug or one being used in clinical trials. Further investigations indicate that HKDC1 is a novel potential therapeutic target for lung cancer.

Conclusion: We developed a protocol to identify potential therapeutic targets from heterogeneous data. We suggest that HKDC1 is a novel potential therapeutic target for lung cancer.

Contact: huangjf@mail.kiz.ac.cn

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

Received on January 25, 2013; revised on August 1, 2013; accepted on October 19, 2013

1 INTRODUCTION

In the past decade, targeted cancer therapies have improved cancer treatment (Aggarwal, 2010; Sawyers, 2004). Dozens of molecular targets have been used in cancer treatment, including EGFR (Ciardiello and Tortora, 2008) and VEGFR (Tugues *et al.*, 2011). Although multiple drugs that selectively inhibit these targets have been developed and are used in cancer treatment, cancer is still a highly challenging disease (Siegel *et al.*, 2012). Therefore, discovering novel therapeutic targets is still an important and challenging task for cancer treatment.

Several hallmarks of cancer have been defined and have helped guide cancer therapy (Hainaut and Plymoth, 2013; Hanahan and Weinberg, 2011). Of these hallmarks, uncontrolled growth is the most critical for cancer propagation. Therefore, a cancer therapeutic target should have the following two characteristics: (i) the target should be essential for the growth of cancer cells (Ngo *et al.*, 2006; Sethi *et al.*, 2012; Tiedemann *et al.*, 2012), and inhibition of the target should directly or indirectly suppress cancer cell growth; and (ii) disturbing the target should have minimal side effects in normal cells. It would be 'perfect' if the target was not expressed in normal cells but highly expressed in cancer cells.

In this study, we used a computational biology method to infer potential therapeutic targets from heterogeneous data. We first used our previously published method CDRUG (Li and Huang, 2012) to predict anticancer ligands in the CHEMBL database (Gaulton *et al.*, 2012). CDRUG is a web server (or method) used to predict whether a chemical compound has anticancer activity (Li and Huang, 2012). CHEMBL is a manually curated chemical database of bioactive molecules that includes >9000 genes and >1 million compounds (Gaulton *et al.*, 2012). We then performed a hypergeometric test, termed anticancer activity enrichment analysis (ACEA), to determine genes with significant enrichment of anticancer ligands. Further investigation revealed that these anticancer ligand-enriched genes have high potentials to become clinical targets.

After overlapping these genes with expression data from lung cancer tissues, we inferred 50 candidate therapeutic targets for lung cancer. We further propose that HKDC1, one of these 50 genes, is a novel potential therapeutic target for lung cancer.

2 METHODS

To infer novel therapeutic targets, we first collected a large assortment of datasets. These datasets included ligand–protein interaction data (CHEMBL version 13) (Gaulton *et al.*, 2012), the NCI-60 GI50 data (Shoemaker, 2006), microarray-based NCI-60 cell line expression data (Reinhold *et al.*, 2012), RNA-seq-based expression data (Krupp *et al.*, 2012), RNA-seq-based expression data from lung cancer and adjacent normal lung cells (Seo *et al.*, 2012) and all known anticancer drugs, including approved drugs and those still in clinical trial, from the Thomson Reuters Integrity^{SM} database. Detailed information concerning these datasets can be found in Supplementary Table S1.

Next, we filtered the CHEMBL dataset based on a half maximal inhibitory concentration (IC50) of $<10\,\mu M$ or a $Ki < 10\,\mu M$ and obtained 206 173 ligands that belong to 1776 human genes. The anticancer activities of all the ligands were predicted using CDRUG (Li and Huang, 2012). CDRUG is based on chemical fingerprint similarity and uses a confidence level (*P*-value) to predict whether a compound has anticancer activity. Thus, we predicted 4018 anticancer ligands using the default

*To whom correspondence should be addressed.

($P<0.05$) cutoff and 1071 anticancer ligands using the strict ($P<0.01$) CDRUG cutoff.

Then, a novel method, termed ACEA, was developed to measure whether a gene is essential for the growth of cancer cells. ACEA is based on the results of the CDRUG analysis (Li and Huang, 2012) and uses a hypergeometric distribution to perform enrichment analysis. The *P*-value of each gene can be calculated using the following equation:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \quad (1)$$

Here, N and n are the total number of ligands and the total number of anticancer ligands in the filtered CHEMBL datasets, respectively; m and k represent the number of ligands and the number of anticancer ligands in a gene, respectively. Both n and k are calculated using CDRUG.

Two runs of ACEA were performed to obtain a list of anticancer ligand-enriched genes. The first and second ACEA runs were performed using the default ($P<0.05$) and the strict ($P<0.01$) CDRUG cutoffs, respectively. For example, in the second ACEA run, N equaled 206 173, and n equaled 1071. When we used ACEA to examine EGFR, we obtained 3688 ligands that interacted with EGFR, and 114 of these 3688 ligands were predicted to have anticancer activity. Therefore, m and k were 3688 and 114, respectively. The *P*-value for EGFR was then calculated to be 1.3×10^{-51}.

Because multiple tests (1776 genes) were performed, the Bonferroni correction method was used to adjust the *P*-value determined by ACEA:

$$p\_adj = p \times N_g \quad (2)$$

Here, p is the *P*-value of ACEA, p_adj is the adjusted *P*-value of ACEA and N_g is the number of genes in the filtered CHEMBL datasets. In the EGFR study, N_g was 1776. Therefore, the second ACEA run had a p_adj value of 2.2×10^{-48}. Only genes with $p\_adj<0.05$ in both the first and second runs were retained. Using this process, we predicted 102 anticancer ligand-enriched genes.

Next, to validate the predicted targets, we separated the known cancer drug targets within the 1776 CHEMBL genes. We also obtained information on all developed anticancer drugs from the Thomson Reuters Integrity^{SM} database; this included 743 anticancer compounds, of which 274 have been approved for treatment. The remaining 469 compounds are currently undergoing clinical trials (Supplementary Table S2). Then, an all-against-all ligand similarity search was performed to map these drugs to the CHEMBL datasets (Supplementary Table S3). Thus, we obtained 239 approved cancer drug targets and 425 targets of trial drugs from the 1776 CHEMBL genes (Supplementary Table S3). These targets were then used to annotate the predicted targets.

Finally, to infer potential therapeutic targets for the treatment of lung cancer, the expression profiles of 102 genes in lung cancer cells were gathered from the collected RNA-Seq data (GSE40419, including 87 lung cancers and 77 adjacent normal tissues) (Seo *et al.*, 2012). Of the 102 genes, 50 genes were significantly overexpressed in lung cancer. In this process, genes with an RPKM (reads per kilo bases per million reads) of <3.0 were considered silent or expressed at low levels (Mortazavi *et al.*, 2008). The functional enrichment analysis of these 50 genes was performed using the DAVID server (Huang *et al.*, 2009).

3 RESULTS

3.1 Inferring potential therapeutic targets using ACEA

To infer potential therapeutic targets, we first filtered the CHEMBL datasets and obtained 1776 genes that contain 206 173 ligands. Of these 1776 genes, ~13% (239 of 1776) are approved cancer drug targets, and 24% (425 of 1776) are targeted by drugs currently in clinical trials (Supplementary Table S3). Then, we predicted the anticancer activity of the 206 173 ligands using our previously published method (CDRUG), which is based on chemical fingerprint similarity (Li and Huang, 2012). Finally, the anticancer ligand enrichment method, ACEA, was developed to determine genes enriched for anticancer ligands.

After two runs of ACEA, we obtained 102 anticancer ligand-enriched genes (Supplementary Table S4). Approximately 40% of these genes (43 of 102) are approved therapeutic targets, whereas 60% of these genes (67 of 102) are being targeted in clinical trials (Fig. 1). The percentage of known cancer targets (approved or clinically used targets) within these 102 genes is approximately three times the percentage of therapeutic targets found in the total list (1776 CHEMBL genes). In other words, the known therapeutic targets are significantly enriched in the list of 102 potential targets ($P<10^{-12}$, hypergeometric test) (Fig. 1). These results indicated that the anticancer ligand-enriched genes have high likelihoods of being developed into clinical targets (>60%). Thus, these 102 genes represent potential therapeutic targets that should be validated further using other methods.

3.2 Inferring lung cancer therapeutic targets

Different types of cancer usually exhibit different characteristics, including gene expression profiles; therefore, the predicted 102 potential targets should be further filtered for a given cancer type. Using RNA-seq expression data from lung cancer tissues, which includes 87 lung cancers and 77 adjacent normal tissues, we extracted 50 targets that are overexpressed in lung carcinomas (Supplementary Table S5). These targets are involved in cell growth or survival-related biological processes (adjusted $P<0.01$), including histone deacetylation, oxidation–reduction, cell division and the electron transport chain. Approximately 40 (19/50) or 60% (30/50) of these genes are targeted by approved drugs or drugs that are currently being used in clinical trials, respectively.

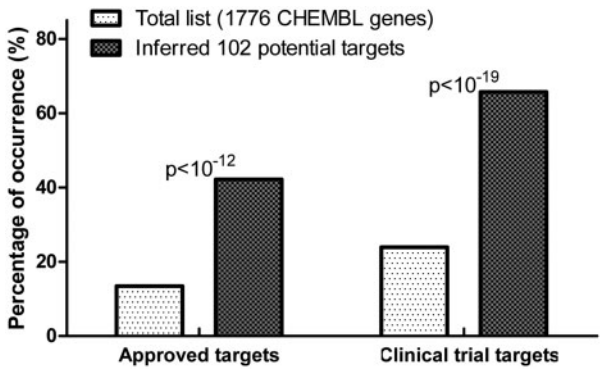

Fig. 1. Comparison of the percentages of known therapeutic targets between the total list (1776 CHEMBL genes) and the 102 potential targets determined by ACEA. The percentages of approved targets in the total list and in the ACEA list are 13.5 and 42.2%, respectively. The percentages of the clinical trial targets in the total list and in the ACEA list are 23.9 and 65.7%, respectively. Known therapeutic targets, including both approved and clinically used targets, were significantly enriched in the ACEA target list ($P<10$–12, hypergeometric test)

Table 1. Top 20 predicted lung cancer therapeutic targets

Rank	Gene	Approved target?	Clinical trial target?	Expressed in normal lung tissues?	Significance of over-expression in lung cancer tissues	ACEA adjusted *P*-value
1	DHFR	√	√	√	***	7.92E-125
2	TOP1	√	√	√	***	2.61E-123
3	**TUBB3**		√	×	***	**1.72E-94**
4	HSP90AB1	√	√	√	***	2.48E-81
5	PPIA	√	√	√	***	9.93E-74
6	HDAC1	√	√	√	***	9.89E-71
7	HSP90AA1	√	√	√	**	3.69E-65
8	HDAC2	√	√	√	***	2.72E-64
9	HDAC3	√	√	√	***	4.51E-63
10	HDAC6	√	√	√	***	3.53E-61
11	HDAC8	√	√	√	***	4.30E-60
12	EGFR	√	√	√	**	2.24E-48
13	PPIB	√	√	√	***	1.19E-39
14	HDAC10	√	√	√	**	5.25E-34
15	**CDK1**	√	√	×	***	**6.11E-21**
16	**HKDC1**			×	***	**1.11E-20**
17	ERBB2	√	√	√	***	9.21E-19
18	**CCNB2**		√	×	***	**2.41E-16**
19	**CCNB1**		√	×	***	**2.54E-16**
20	TUBA4A		√	√	***	2.02E-15

Note: Significances of overexpression of the lung cancer targets are cataloged as **P*-value of 0.05–0.01, ***P*-value of 0.01–10^{-6} or ****$P < 10^{-6}$, respectively. The targets that are not expressed or only expressed at low levels in normal lung cells (RPKM < 3) are shown in bold (underlined). Note: only the adjusted *P*-values of the second ACEA run are shown in this table.

In these 50 candidate lung cancer therapeutic targets, we observed that 19 of the top 20 targets were already targeted by drugs that have been approved or are in clinical trials (Table 1). These results indicate that ACEA can precisely predict lung cancer targets. Notably, HKDC1 is the only one of the top 20 targets that has not been targeted in clinical trials (Table 1). We also observed eight targets (TUBB3, CDK1, HKDC1, CCNB2, CCNB1, KIF11, TOP2A and HSD17B1) that are not expressed or only expressed at low levels in normal lung cells (RPKM < 3.0) (Supplementary Table S5). Except for HKDC1, all of these genes are targeted by approved drugs or drugs currently in clinical trials (Supplementary Table S5). These results suggest that HKDC1 is a novel potential therapeutic target for lung cancer.

3.3 HKDC1 may be a novel therapeutic target for lung cancer

To further validate whether HKDC1 could be used to target lung cancer, we investigated the heterogeneous data related to HKDC1. Twelve inhibitors of HKDC1 were screened in the NCI-60 DTP project (Shoemaker, 2006) (Fig. 2A), and 8 of the 12 inhibitors have GI50s (50% growth inhibition) of <10 μM (Fig. 2B). All of these inhibitors had GI50s of <30 μM (Fig. 2B). In addition, HKDC1 is expressed in nearly all the NCI-60 cell lines (Fig. 2C), assuming that genes with a GCRMA value >3.32 are considered to be expressed (Siddiqui *et al.*, 2006). These observations suggest that inhibition of HKDC1 could result in the suppression of cancer cell growth.

As mentioned earlier, a 'perfect' cancer therapeutic target should not only be essential to cancer growth but also should not be expressed in normal cells. Thus, we investigated the expression profiles of HKDC1 in normal tissues using data from RNA-Seq Atlas (Krupp *et al.*, 2012). The results show that HKDC1 is either not expressed or expressed at low levels in normal tissues, except for the kidney (RPKM = 4.82) (Fig. 2D). As shown in Figure 2E, HKDC1 is not expressed or is expressed at low levels in all of the normal lung tissue samples (77 samples). In contrast, HKDC1 is expressed or highly expressed (RPKM > 3.0) in 60% (52/87) of the lung carcinoma tissues (Fig. 2E). These results indicate that HKDC1 is a potential therapeutic target for lung cancer and could be applied to ~60% of lung cancer patients.

4 DISCUSSION

We have developed a computational biology method, ACEA, to determine potential therapeutic target genes. Further analysis shows that these genes have high likelihoods of being developed into clinical targets. Combined with gene expression data, we predicted 50 candidate targets for the treatment of lung cancer. Further validation investigations suggest that HKDC1 is a novel therapeutic target for lung cancer.

HKDC1 encodes the fifth mammalian hexokinase, which phosphorylates hexoses (Wilson, 2003). Because phosphorylation

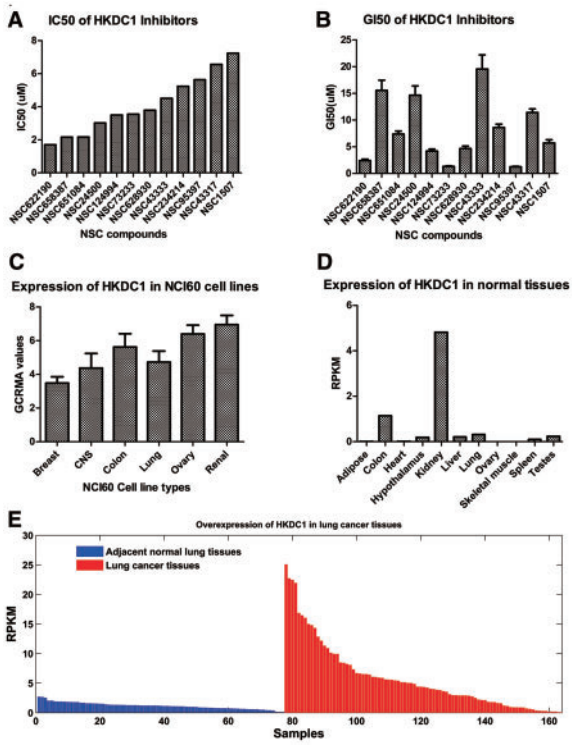

Fig. 2. HKDC1 is a novel potential therapeutic target for lung cancer. (**A**) IC50 of HKDC1 inhibitors screened in the NCI-60 DTP project. (**B**) GI50 of the HKDC1 inhibitors. (**C**) Microarray-based expression data of HKDC1 in the NCI60 cell lines. (**D**) RNA-seq-based expression data of HKDC1 in normal tissues. (**E**) Waterfall plot of RNA-seq-based differential expression between adjacent normal lung tissues and lung cancer tissues

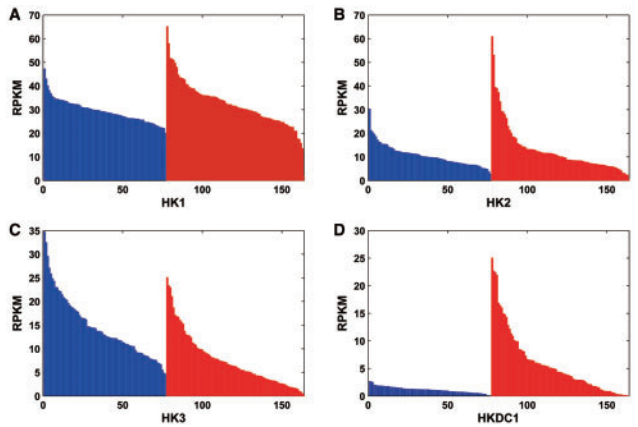

Fig. 3. Waterfall plots of the expression profiles of different hexokinases between normal lung tissues and lung cancer tissues. The 77 normal lung tissues (range 1–77) are colored blue, whereas the 87 lung cancer tissues (range 78–164) are colored red. Note: the hexokinases include HK1, HK2, HK3, GCK (HK4) and HKDC1. The expression profile of HK4 was not shown because HK4 is not expressed in either normal lung tissues (average RPKM <0.2) or in lung cancer tissues (average RPKM <0.2)

is the first step in glucose metabolism, hexokinases may play an important role in regulating energy metabolism and thus regulating cell growth. Hexokinase type II (HK2) is a well-studied therapeutic target (Rempel *et al.*, 1996; Tennant *et al.*, 2010; Wolf *et al.*, 2011). Figure 3B shows that HK2 is overexpressed in lung cancer tissues (fold change <1, $P=0.03$). However, only HKDC1 showed a dramatic overexpression in cancer tissues when compared with other hexokinases (fold change >2, $P<10^{-10}$) (Fig. 3). This result suggested that HKDC1 might play an important role in cancer growth that is different from other hexokinases. Further experimental elucidation of the exact role of HKDC1 in cancer growth will be important to its development as a therapeutic target.

5 CONCLUSION

We developed the novel computational biology method ACEA to determine which genes are significantly enriched for anticancer ligands. We inferred 50 candidate targets for the treatment of lung cancer, and we suggest that HKDC1 is a novel potential therapeutic target for lung cancer.

Funding: This work was supported by the National Basic Research Program of China (grant number 2013CB835100), and the National Natural Science Foundation of China (grant number 31123005 to J.F.H.).

Conflicts of Interest: none declared.

REFERENCES

Aggarwal,S. (2010) Targeted cancer therapies. *Nat. Rev. Drug Discov.*, **9**, 427–428.
Ciardiello,F. and Tortora,G. (2008) Drug therapy: EGFR antagonists in cancer treatment. *N. Eng. J. Med.*, **358**, 1160–1174.
Gaulton,A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.
Hainaut,P. and Plymoth,A. (2013) Targeting the hallmarks of cancer: towards a rational approach to next-generation cancer therapy., *Cur. Opin. Oncol.*, **25**, 50–51.
Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.
Huang,D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
Krupp,M. *et al.* (2012) RNA-Seq Atlas-a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics*, **28**, 1184–1185.
Li,G.H. and Huang,J.F. (2012) CDRUG: a web server for predicting anticancer activity of chemical compounds. *Bioinformatics*, **28**, 3334–3335.
Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
Ngo,V.N. *et al.* (2006) A loss-of-function RNA interference screen for molecular targets in cancer. *Nature*, **441**, 106–110.
Reinhold,W.C. *et al.* (2012) CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res.*, **72**, 3499–3511.
Rempel,A. *et al.* (1996) Glucose catabolism in cancer cells: amplification of the gene encoding type II hexokinase. *Cancer Res.*, **56**, 2468–2471.
Sawyers,C. (2004) Targeted cancer therapy. *Nature*, **432**, 294–297.
Seo,J.S. *et al.* (2012) The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res.*, **22**, 2109–2119.
Sethi,G. *et al.* (2012) An RNA interference lethality screen of the human druggable genome to identify molecular vulnerabilities in epithelial ovarian cancer. *PLoS One*, **7**, e47086B.
Shoemaker,R.H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, **6**, 813–823.

Siddiqui,A.S. *et al.* (2006) Sequence biases in large scale gene expression profiling data. *Nucleic Acids Res.*, **34**, e83.

Siegel,R. *et al.* (2012) Cancer statistics, 2012. *CA Cancer J. Clin.*, **62**, 10–29.

Tennant,D.A. *et al.* (2010) Targeting metabolic transformation for cancer therapy. *Nat. Rev. Cancer*, **10**, 267–277.

Tiedemann,R.E. *et al.* (2012) Identification of molecular vulnerabilities in human multiple myeloma cells by RNA interference lethality screening of the druggable genome. *Cancer Res.*, **72**, 757–768.

Tugues,S. *et al.* (2011) Vascular endothelial growth factors and receptors: anti-angiogenic therapy in the treatment of cancer. *Mol. Aspects Med.*, **32**, 88–111.

Wilson,J.E. (2003) Isozymes of mammalian hexokinase: structure, subcellular localization and metabolic function. *J. Exp. Biol.*, **206**, 2049–2057.

Wolf,A. *et al.* (2011) Hexokinase 2 is a key mediator of aerobic glycolysis and promotes tumor growth in human glioblastoma multiforme. *J. Exp. Med.*, **208**, 313–326.