Models and Technologies

Cancer In Silico Drug Discovery: A Systems Biology Tool for Identifying Candidate Drugs to Target Specific Molecular Tumor Subtypes

F. Anthony San Lucas1,2, Jerry Fowler2, Kyle Chang1, Scott Kopetz1,3, Eduardo Vilar1,3,4, and Paul Scheet1,2

Abstract
Large-scale cancer datasets such as The Cancer Genome Atlas (TCGA) allow researchers to profile tumors based on a wide range of clinical and molecular characteristics. Subsequently, TCGA-derived gene expression profiles can be analyzed with the Connectivity Map (CMap) to find candidate drugs to target tumors with specific clinical phenotypes or molecular characteristics. This represents a powerful computational approach for candidate drug identification, but due to the complexity of TCGA and technology differences between CMap and TCGA experiments, such analyses are challenging to conduct and reproduce. We present Cancer in silico Drug Discovery (CiDD; scheet.org/software), a computational drug discovery platform that addresses these challenges. CiDD integrates data from TCGA, CMap, and Cancer Cell Line Encyclopedia (CCLE) to perform computational drug discovery experiments, generating hypotheses for the following three general problems: (i) determining whether specific clinical phenotypes or molecular characteristics are associated with unique gene expression signatures; (ii) finding candidate drugs to repress these expression signatures; and (iii) identifying cell lines that resemble the tumors being studied for subsequent in vitro experiments. The primary input to CiDD is a clinical or molecular characteristic. The output is a biologically annotated list of candidate drugs and a list of cell lines for in vitro experimentation. We applied CiDD to identify candidate drugs to treat colorectal cancers harboring mutations in BRAF. CiDD identified EGFR and proteasome inhibitors, while proposing five cell lines for in vitro testing. CiDD facilitates phenotype-driven, systematic drug discovery based on clinical and molecular data from TCGA. Mol Cancer Ther; 13(12); 3230–40. ©2014 AACR.

Introduction
Selection of targeted therapies for cancer drug development has traditionally been based on the presence or absence of specific somatic mutations and this has been shown to be an effective strategy to improve patient outcomes (1–4). However, a large number of targeted drugs and other compounds that have antitumor properties have not been linked to specific mutations, or biomarkers, that could be used to predict their selective efficacy (5). Although next-generation sequencing allows researchers to rapidly and comprehensively profile tumor mutations, the vast majority of these data have not been useful in the clinical setting since only a small number of mutations have been used to inform prognosis or guide therapeutic decisions (6–8).

Several computational approaches exist and have been implemented to predict the functional impact of mutations, and even to predict whether a specific mutation is a driver of the carcinogenesis process, based on several factors such as evolutionary conservation, predicted effects on protein structure, and observed recurrence in existing cancer datasets (9–11). However, these computational predictions provide little insight into how cellular processes are altered as a consequence of the mutations. One strategy to assess whether or not specific mutations are influential on cellular processes is to determine whether or not a mutation induces a signature of gene expression changes (12). Gene expression signatures associated with an individual mutation could then be examined to characterize its cellular impact (13) and the signature could be used as a target for candidate drug therapies (14). We have developed the Cancer in silico Drug Discovery (CiDD) platform for the purposes of characterizing tumors with specific mutations, or more generally tumors...
with specific clinicopathological or molecular characteristics, based on their putative effects on gene expression, and to identify candidate drugs to treat these tumors.

Here, we describe the general framework and integrated datasets of this novel platform. CiDD has been designed to generate hypotheses for the following three general problems: (i) to determine whether particular clinical or molecular characteristics are associated with unique gene expression signatures; (ii) to find candidate drugs to treat specific tumor subgroups based on these expression changes; and (iii) to identify cell lines that resemble the tumors being studied for subsequent in vitro experimentation. In addition, to illustrate the use of CiDD, we have applied it to a clinically relevant context in cancer drug development. We report the in silico identification of candidate drug therapies for colorectal cancers (CRC) harboring the BRAF V600E mutation. Approximately 10% of CRCs harbor the BRAF V600E mutation, which confers a poor prognosis and presents a therapeutic challenge (4, 15). We describe the analyses performed with CiDD that have identified novel targets for BRAF-mutant CRCs and drugs such as EGFR inhibitors that have already shown activity at the preclinical level in targeting this tumor subtype (4).

Materials and Methods

CiDD is a systematic drug discovery platform that integrates and analyzes large-scale cancer datasets with the primary goal of identifying candidate drugs and cell lines to be validated experimentally in vitro (see Fig. 1). The core datasets used by CiDD include The Cancer Genome Atlas (TCGA), the Connectivity Map (CMap), and the Cancer Cell Line Encyclopedia (CCLE). CiDD is purely computational and depends on publicly available clinical and experimental datasets, as well as annotation databases. CiDD is written in Python, has R package dependencies, and is command-line driven allowing it to be integrated into bioinformatics pipelines. The software and code are freely available at http://scheet.org/software.

Data assembly

Required experimental datasets for performing CiDD analyses are TCGA (16) and CMap (14). CCLE (17) is required to identify cell lines for subsequent experimentation. TCGA includes clinical, mutation, and gene expression data for thousands of samples across multiple cancer types. CiDD provides commands to download, query, and analyze these data. CMap is a collection of gene expression data for cell lines treated with small molecules paired with pattern-matching algorithms that attempt to identify biologically functional connections between drugs and gene expression profiles (14). CiDD utilizes CMap build 02, which contains more than 7,000 expression profiles representing the effects of 1,309 compounds. CCLE provides molecular profiles for 947 cancer cell lines which include DNA copy number, gene expression, and DNA mutation data (17).

The experimental data from CMap consists of rank-based gene expression values from the Affymetrix HG-U133A microarray. Thus, CMap is designed for the analysis of Affymetrix gene expression data only, which hinders using CMap with gene expression data collected from non-Affymetrix platforms. To overcome this limitation, CiDD transforms bulk-downloaded CMap data from Affymetrix probe-based rank values to Entrez gene-based ranks. Gene-based ranks are determined by taking the mean probe rank for each gene, sorting the mean rank values, and then assigning a rank for each gene based on the sorted values. This allows results from RNA sequencing and Agilent microarray technologies, such as those provided by TCGA, to be analyzed with the drug-perturbed data of CMap in a standardized way at the gene level. A similar strategy has been applied in the R package gCMap (18) that allows users to query CMap using Affymetrix probe identifiers or gene symbols. Gene expression signatures derived from both Agilent microarrays and RNA sequencing have identified validated candidate drugs when analyzed with the Affymetrix-based drug signatures of CMap (19–21) demonstrating the feasibility of a cross-platform approach.

CiDD also uses annotation datasets, which include the Molecular Signatures Database (MSigDB; ref. 13) for characterizing gene sets and drug databases including DrugBank (22), Matador (23), and KEGG Drug (24) for annotating candidate drugs. These databases provide information such as drug pharmacology, gene, and pathway targets to make the CiDD’s drug reports more informative. Public data from TCGA are automatically downloaded by CiDD, whereas data from CMap, CCLE, and MSigDB require registration at their respective websites before downloading. Upon download, CiDD automatically prepares and manages datasets for drug discovery analyses. Descriptions of these datasets are provided in Supplementary Methods.

CiDD workflow

A common workflow using CiDD is illustrated in Fig. 2. Initially, a CiDD project based on a TCGA cancer type is created and clinical, mutation, and gene expression data for TCGA samples are automatically downloaded. For an analysis, CiDD first identifies TCGA samples for use in computational experiments based on user-defined clinicopathological phenotypes or molecular characteristics, such as specific gene mutations, microsatellite instability status, tumor stage, or a variety of other patient or tumor characteristics reported through TCGA projects. On the basis of the defined phenotype, CiDD identifies two classes of samples to compare. For a mutation-based phenotype, CiDD establishes one class containing samples with a defined mutation or set of mutations and a second class containing samples that are wild-type for the genes of interest. For a clinical phenotype, the user specifies both classes explicitly, such as two classes corresponding to microsatellite unstable and stable tumors. CiDD attempts to identify a gene expression signature...
that is associated with the defined patient or tumor characteristic. If a gene expression signature exists for the phenotype of interest, that signature is characterized with MSigDB gene sets and the signature is used to identify candidate drugs through pattern-matching algorithms proposed by CMap. Subsequently, CiDD characterizes candidate drugs using databases such as DrugBank, Matador, and KEGG Drug. Finally, CiDD identifies candidate cell lines on which to test the drugs in vitro by analyzing data from CCLE. The primary results of a CiDD execution are a biologically annotated candidate drug list and candidate cell lines for subsequent drug experimentation.

Gene signature identification

TCGA provides gene expression data from Agilent microarrays, Illumina GA RNA sequencing, and Illumina HiSeq RNA sequencing. The data type to analyze can be specified as a parameter to CiDD. By default, CiDD will choose the technology that provides data for the largest number of samples with the phenotype of interest. Using the R package Limma (25) which is designed for both microarray and RNA sequencing differential expression analyses, CiDD identifies up- and downregulated genes. CiDD characterizes these results with biologic pathways by performing gene set tests using the piano Bioconductor package (26), while using gene sets defined by MSigDB.

Generation of a k-top scoring pairs classifier

For generating a classifier that is robust across gene expression technologies, CiDD takes a nonparametric approach to classification and adopts an extension of the top scoring pairs (TSP) method (27). Using the R package ktspair (28), CiDD generates a k-top scoring pairs (k-TSP) classifier for predicting the status of the phenotype of interest on independent samples. The k-TSP algorithm is described in Supplementary Methods.

Candidate drug identification

CiDD connects gene expression changes associated with the phenotype of interest with candidate drug compounds that induce a negatively correlated (or “negatively connected”) gene expression profile. CiDD compares the phenotype gene expression changes, termed a query signature, with rank-based gene expression profiles induced by CMap compounds. To compare rank-based gene expression profiles, CiDD implements nonparametric pattern-matching algorithms based on the Kolmogorov–Smirnov statistic as described by Lamb and colleagues (14). An enrichment score ranging from $-1$ to $+1$ provides a measure of the negative or positivity...
connectivity of a drug to the phenotype of interest. A permutation \( P \) value provides a measure of significance for the enrichment scores. These algorithms and the resulting metrics are described in Supplementary Methods.

**Cell line identification**

CiDD first selects CCLE cell lines based on user-specified tissue types. Then, CiDD optionally identifies cell lines that contain user-specified mutations by interrogating CCLE mutation data derived from either targeted sequencing of common cancer genes or from Oncomap 3.0, which is an SNP array that genotypes samples at known cancer-related sites. Finally, CiDD runs its k-TSP classifier on CCLE gene expression data to predict whether a cell line’s gene expression profile is representative of the phenotype being studied. Cell lines that meet these criteria are reported as candidates for use in subsequent drug experiments.

**Results**

We applied CiDD to identify candidate drugs to treat CRCs harboring \( \text{BRAF} \text{V600E} \) mutations using mutation and RNA-sequencing data from the TCGA colon and rectum projects. We also identified cell lines from CCLE that are representative of colorectal tumors with \( \text{BRAF} \) mutations, thus making them candidates for in vitro drug testing. We refer to these analyses as the TCGA-derived analyses. The detailed commands to rerun these analyses are provided in Supplementary Methods. We then compared our systematic TCGA-derived analyses generated from CiDD with analyses performed using a previously published gene expression signature for \( \text{BRAF} \text{V600E} \).
Identification of a BRAF V600E CRC gene expression signature

We used CiDD to identify 20 TCGA CRC samples with a BRAF V600E mutation and 149 BRAF wild-type samples with available Illumina GA RNA sequencing data. CiDD identified 63 upregulated and 170 downregulated genes (log fold-change ≥ 2 and Benjamini–Hochberg adjusted P value ≤ 0.05) that generated a clustering of samples representative of BRAF mutation status as shown in Fig. 3.

We identified pathways associated with the BRAF signature through CiDD using Wilcoxon-based gene set tests (26). For assessing significance of the gene set tests, CiDD performed 1,000 runs of the differential expression analyses, permuting the BRAF-mutant status of samples within each run. Fifteen KEGG gene sets were associated with the BRAF V600E status (FDR adjusted P value ≤ 0.05). To incorporate PETACC3-derived pathways as part of the pathway analysis, a list of the top 20 pathways based on an average ranking within the TCGA and PETACC3-derived pathway lists is provided in Table 1. Because raw gene expression data were not available for the PETACC3-derived signature, gene set tests were not performed. Instead, for the PETACC3-derived analysis, hypergeometric tests were applied to identify KEGG pathways enriched with genes from this signature. Twenty-seven KEGG pathways are enriched with genes from the PETACC3-derived signature (P value ≤ 0.05). The pathway ordering in Table 1 reflects the average of the P value ranks within each set (complete results are provided in Supplementary Results). These pathways are consistently related to CRC biology such as the top-ranked pathway ("CRC") and other pathways related to TGFβ signaling ("TGFβ signaling pathway"), which are well known for their role in CRC. In addition, it is known that the BRAF gene plays a role in controlling cellular proliferation and differentiation through regulation of the MAP kinase...
signaling pathway (29), and the "MAPK signaling pathway" is also represented in the top-ranked pathways.

Finally, we used CiDD to identify an 11-pair k-TSP classifier for predicting the BRAF V600E status of independent samples using the TCGA dataset. The classifier gene pairs are listed in Supplementary Table S1.

**Validation of the TCGA-derived gene-pair classifier for predicting BRAF V600E status**

To validate the TCGA-derived gene expression analyses, we compared the performance of a previously reported BRAF V600E gene expression classifier derived from the PETACC3 clinical trial (15) against the gene expression classifier that we identified from the TCGA dataset. The PETACC3-derived gene expression signature consists of 193 upregulated and 92 downregulated probes. These probes correspond to 224 unique genes. The research group also developed a 64-gene TSP classifier (these genes are defined in Supplementary Table S2) based on Affymetrix probe IDs for predicting the BRAF V600E status of CRCs. We translated these probe IDs to Entrez gene IDs so the classifier could be applied to RNA sequencing and Agilent test datasets. To assess the robustness of their gene expression results, we applied the gene-based PETACC3-derived classifier to TCGA samples that were retrieved and annotated with BRAF mutation statuses by CiDD. When applied to TCGA RNA sequencing data, the PETACC3-derived classifier resulted in 93.3% sensitivity and 83.5% specificity for detecting BRAF V600E samples.

To assess the quality of the systematic TCGA-derived classifier generated by CiDD, we compared the performance of the TCGA- and PETACC3-derived classifiers on three independent datasets (see Table 2)—two have been previously published and are available in the Gene Expression Omnibus (30, 31) and the third is the CCLE dataset. The sensitivity and specificity of both classifiers are comparable on the GSE35896 and GSE42284 datasets with the PETACC3-derived classifier exhibiting small improvements in specificity. The PETACC3-derived classifier achieved 100% sensitivity but only 30% specificity for BRAF V600E status prediction on the CCLE large intestine dataset. The TCGA-derived classifier had lower sensitivity (71%) but achieved better specificity (62%). These results suggest that the systematically obtained BRAF V600E classifier from CiDD is comparable with the published PETACC3-derived signature and that the TCGA-derived classifier may even have improved specificity for distinguishing BRAF wild-type cell lines from the BRAF-mutant cell lines.

### Table 1. The top 20-ranked pathways associated with BRAF V600E status based on systematic TCGA gene expression analyses presented with those derived from the independent PETACC3-based analyses

<table>
<thead>
<tr>
<th>Pathways</th>
<th>TCGA P</th>
<th>PETACC3 P</th>
<th>TCGA rank</th>
<th>PETACC3 rank</th>
<th>Average rank</th>
<th>Overall rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colorectal cancer</td>
<td>0.021</td>
<td>0.003</td>
<td>9</td>
<td>4</td>
<td>6.5</td>
<td>1</td>
</tr>
<tr>
<td>Bladder cancer</td>
<td>0</td>
<td>0.017</td>
<td>2</td>
<td>18</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Pathways in cancer</td>
<td>0.05</td>
<td>0.004</td>
<td>15</td>
<td>6</td>
<td>10.5</td>
<td>3</td>
</tr>
<tr>
<td>Chemokine signaling pathway</td>
<td>0.04</td>
<td>0.012</td>
<td>11</td>
<td>16</td>
<td>13.5</td>
<td>4</td>
</tr>
<tr>
<td>JAK–STAT signaling pathway</td>
<td>0.053</td>
<td>0.006</td>
<td>20</td>
<td>11</td>
<td>15.5</td>
<td>5</td>
</tr>
<tr>
<td>Axon guidance</td>
<td>0.057</td>
<td>0.003</td>
<td>26</td>
<td>5</td>
<td>15.5</td>
<td>6</td>
</tr>
<tr>
<td>FC RI signaling pathway</td>
<td>0.021</td>
<td>0.005</td>
<td>7</td>
<td>27</td>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>TGFβ signaling pathway</td>
<td>0.066</td>
<td>0.001</td>
<td>34</td>
<td>2</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>Dorsalventral axis formation</td>
<td>0.057</td>
<td>0.006</td>
<td>25</td>
<td>12</td>
<td>18.5</td>
<td>9</td>
</tr>
<tr>
<td>Peroxisome</td>
<td>0.066</td>
<td>0.006</td>
<td>33</td>
<td>10</td>
<td>21.5</td>
<td>10</td>
</tr>
<tr>
<td>MAPK signaling pathway</td>
<td>0.057</td>
<td>0.032</td>
<td>24</td>
<td>23</td>
<td>23.5</td>
<td>11</td>
</tr>
<tr>
<td>ABC transporters</td>
<td>0.068</td>
<td>0.018</td>
<td>37</td>
<td>19</td>
<td>28</td>
<td>12</td>
</tr>
<tr>
<td>ERBB signaling pathway</td>
<td>0.069</td>
<td>0.008</td>
<td>46</td>
<td>14</td>
<td>30</td>
<td>13</td>
</tr>
<tr>
<td>Fcγ R-mediated phagocytosis</td>
<td>0.062</td>
<td>0.069</td>
<td>30</td>
<td>31</td>
<td>30.5</td>
<td>14</td>
</tr>
<tr>
<td>Tryptophan metabolism</td>
<td>0.037</td>
<td>0.16</td>
<td>10</td>
<td>52</td>
<td>31</td>
<td>15</td>
</tr>
<tr>
<td>B-cell receptor signaling pathway</td>
<td>0.083</td>
<td>0</td>
<td>61</td>
<td>1</td>
<td>31</td>
<td>16</td>
</tr>
<tr>
<td>Prion diseases</td>
<td>0.04</td>
<td>0.144</td>
<td>14</td>
<td>49</td>
<td>31.5</td>
<td>17</td>
</tr>
<tr>
<td>Epithelial cell signaling in Helicobacter pylori infection</td>
<td>0.068</td>
<td>0.039</td>
<td>39</td>
<td>24</td>
<td>31.5</td>
<td>18</td>
</tr>
<tr>
<td>T-cell receptor signaling pathway</td>
<td>0.06</td>
<td>0.081</td>
<td>28</td>
<td>37</td>
<td>32.5</td>
<td>19</td>
</tr>
<tr>
<td>Neuroactive ligand receptor interaction</td>
<td>0.021</td>
<td>0.234</td>
<td>3</td>
<td>67</td>
<td>35</td>
<td>20</td>
</tr>
</tbody>
</table>

*NOTE: The table is ordered by the overall rank of each pathway, where the overall rank represents an average rank across both the TCGA- and PETACC-derived analyses. P values and ranks for pathways associated for both the TCGA- and PETACC-derived analyses are shown. These pathways are consistently related to CRC biology such as the top-ranked pathway "CRC" and the "MAPK signaling pathway," in addition to the "TGFβ signaling pathway," which is known to play a role in BRAF-mutant CRC.
**Table 2.** Performance of the TCGA- and PETACC3-derived BRAF V600E CRC classifiers when applied to independent gene expression datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TCGA-derived classifier</th>
<th>PETACC3-derived classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td>GSE35896 (n = 62)</td>
<td>4/6 (0.67)</td>
<td>39/56 (0.70)</td>
</tr>
<tr>
<td>(Affymetrix U133 Plus 2.0 Array)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSE42284 (n = 178)</td>
<td>33/36 (0.92)</td>
<td>91/142 (0.64)</td>
</tr>
<tr>
<td>(Agilent Homo sapiens 37K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DiscoverPrint_19742)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCLE LARGE_INTESTINE (n = 57)</td>
<td>5/7 (0.71)</td>
<td>31/50 (0.62)</td>
</tr>
<tr>
<td>(Affymetrix U133 Plus 2.0 Array)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: The sensitivity and specificity of both classifiers are comparable with the PETACC3-derived classifier exhibiting small improvements in specificity on the GSE35896 and GSE42284 data sets. The TCGA-derived classifier had lower sensitivity (71%) but achieved better specificity (62%) on the CCLE dataset. These results suggest that the systematically obtained BRAF V600E classifier from CiDD is comparable with the published PETACC3-derived signature and that the TCGA-derived classifier may even have improved specificity for distinguishing BRAF wild-type cell lines from the BRAF-mutant cell lines.

**Candidate drug therapies for BRAF V600E CRC**

Using both TCGA and PETACC3-derived gene expression signatures, CiDD identified candidate drugs to treat BRAF V600E CRCs. Drugs with a negative enrichment score and a permutation P value less than 0.1 using the TCGA and PETACC3-derived gene expression signatures are listed in Table 3 and Supplementary Table S3, respectively. Three compounds, gefitinib, MG-262, and trapidil, were identified in both lists. Independent research groups have recently shown that EGF inhibitors such as gefitinib and proteasome inhibitors such as MG-262 are effective drugs for treatment of colorectal tumors with BRAF mutations (4, 32). Trapidil is a novel candidate drug that inhibits phosphodiesterase and TXA2. The full candidate drug reports are provided in Supplementary Results.

**Cancer cell lines that most resemble BRAF V600E CRC**

To identify candidate cell lines for in vitro testing, CiDD analyzed data from the CCLE. From 947 cell lines in the CCLE, CiDD identified 48 large intestine samples that we consider to be representative of colorectal tumors. Then CiDD reduced this number to 7, representing those large intestine cell lines that have BRAF V600E mutations. Using the 11 gene-pair k-TSP classifier generated by CiDD, five of these cell lines were predicted to be BRAF V600E on the basis of having similar gene expression profiles to the TCGA BRAF V600E–mutated CRCs. The five identified cell lines include RKO, SNU5, CL34, COLO205, and HT29. OUMS23 and SW1417 are the two BRAF V600E large intestine cell lines that are predicted to be BRAF wild-type by the TCGA-derived classifier.

**Discussion**

As genomic technologies have ushered in the potential for targeted drug development, large-scale public genomic databases have matured in size, scope, and information content to complement this effort. It is thus advantageous, and indeed possibly necessary, to apply computational genomics to inform the drug discovery process. Although subgroup classification for prognostic assessment and therapeutic planning has been applied clinically for decades, especially among hematologic malignancies and in some solid tumors such as breast cancers, other tumor types such as CRCs appear phenotypically homogenous and are thus clinically indistinguishable. To reveal subclasses for these tumors and to generalize their genome-based classification, the use of genetic and transcriptomic analyses may prove essential. Systems biology tools such as CMap, and we believe CiDD, help fill this need of identifying candidate interventions that target specific pathways deregulated in these tumor subclasses. In this regard, CMap provided the original approach to guide drug development based on transcriptomic data. CiDD is taking this approach further by extending CMap with the clinical and molecular data of TCGA along with the high-throughput experiments of the CCLE for the purposes of systematic cancer drug discovery. Although current public resources such as that of TCGA are impressive, they are likely just a beginning. The basic logic of CiDD naturally extends to utilization of forthcoming, larger-scale databases from drug perturbation experiments and genetic and transcriptomic sequencing of tumors of a wider array of sizes and associated clinical outcomes.

We believe that CiDD is the first framework that supports systematic drug discovery based on user-specified TCGA clinical phenotypes and molecular characteristics. CiDD allows researchers to perform the following: (i) assess whether or not a mutation or clinical phenotype is associated with a gene expression signature; (ii) identify candidate drugs to target this gene expression signature; and (iii) identify cell lines for subsequent in vitro drug...
Targeted Drug Discovery with TCGA, CMap, and CCLE

experimentation. We have illustrated the power of such an approach in a meaningful application to CRCs with somatic mutations in \(BRAF\). CiDD also offers utility to researchers simply wishing to interrogate and organize TCGA data, as it can be applied to create an inventory of available TCGA data with particular clinical or genomic features, such as available datasets or patients with particular mutations, independently of its drug identification capabilities.

One of the most crucial steps in the \(BRAF\) V600E analysis was identifying a gene expression signature associated with the \(BRAF\) V600E mutation and generating a classifier for predicting mutation status. In both of these cases, we showed that the signature and classifier of the CiDD framework are comparable with those identified from the published PETACC3-derived analyses (15). Similarly to the PETACC3-derived signature and classifier, the CiDD-generated signature was composed of genes representative of known pathways associated with the \(BRAF\) V600E mutation, most notably the “MAPK signaling pathway,” and the performance of the classifier on independent datasets generated from orthogonal gene expression technologies showed robustness. The advantage of CiDD analyses is that they are systematic studies of generally available datasets. We did not have to generate any of our own experimental data, and the gene expression analyses can be relatively easily replicated and repeated for other mutation or clinical phenotypes.

Once we validated the gene expression signature, we used CiDD to identify candidate compounds for tumors harboring the well-known \(BRAF\) V600E mutation. Since the initial communication of the presence of mutations in the kinase \(BRAF\) in cancer (33), activating mutations have been described in several malignancies with different frequencies such as hairy cell leukemia (100%), melanoma (50%–60%), thyroid carcinoma (30%–50%), and CRC (10%; ref. 34). The most frequently identified mutation is a valine-to-glutamic acid substitution at codon 600 (V600E) that activates the signaling cascade downstream of MEK and ERK (33). Other mutations have been found at the same codon and are considered equivalent in terms of oncogenic activation (34). Therefore, substantial efforts were invested on developing ATP-competitive RAF inhibitors such as vemurafenib and dabrafenib to specifically target the MAPK pathway. Yet, the clinical success of \(BRAF\) inhibition has been variable and highly dependent on the tumor context. In this regard, vemurafenib has demonstrated improvement in survival in patients diagnosed with stage IV melanomas harboring the \(BRAF\) V600E mutation (35). However, this degree of clinical benefit has not been observed in the same molecular context in CRCs and papillary thyroid cancers (36). This is probably secondary to the intrinsic mechanisms of resistance to \(BRAF\) inhibition that are specific to the tumor context (34). \(BRAF\) mutations in the context of metastatic CRCs have been associated with poor prognosis and an aggressive disease course contrasting with cases in early stages. In addition, they have a characteristic clinical phenotype consistent with older age at diagnosis, female gender, right-sided location, and the presence of high levels of microsatellite instability (37, 38).

Two strategies have been suggested to overcome the primary resistance to \(BRAF\) inhibition in CRC biology. One strategy that has been supported independently by two different groups is the inhibition of the \(EGFR\) pathway by using monoclonal antibodies against \(EGFR\) (such as cetuximab) or kinase inhibitors (such as gefitinib and erlotinib) in combination with \(BRAF\) inhibitors. \(EGFR\) is activated by feedback mechanisms upon \(BRAF\) inhibition, thus reactivating ERK via \(RAS\) and \(CRAF\), therefore combinations of \(EGFR\) and \(BRAF\) inhibition will synergize in terms of activity (4, 39–41). The second strategy is based on targeting the proteasome pathway. This has demonstrated specific activity against \(BRAF\) V600E-mutant CRC cell lines and tumor xenografts. This set of experiments was

<table>
<thead>
<tr>
<th>Compound</th>
<th>Enrichment score</th>
<th>Permutation P</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gefitinib</td>
<td>−0.995</td>
<td>0.016</td>
<td>0.000</td>
</tr>
<tr>
<td>2-Deoxy-D-glucose</td>
<td>−0.977</td>
<td>0.051</td>
<td>0.022</td>
</tr>
<tr>
<td>5286656</td>
<td>−0.967</td>
<td>0.075</td>
<td>0.038</td>
</tr>
<tr>
<td>Yohimbic acid</td>
<td>−0.901</td>
<td>0.003</td>
<td>0.000</td>
</tr>
<tr>
<td>Amrinone</td>
<td>−0.884</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>Trapidila</td>
<td>−0.852</td>
<td>0.004</td>
<td>0.016</td>
</tr>
<tr>
<td>Mycophenolic acid</td>
<td>−0.735</td>
<td>0.024</td>
<td>0.048</td>
</tr>
<tr>
<td>Withaferin A</td>
<td>−0.679</td>
<td>0.026</td>
<td>0.054</td>
</tr>
<tr>
<td>MG-262a</td>
<td>−0.656</td>
<td>0.073</td>
<td>0.141</td>
</tr>
</tbody>
</table>

NOTE: Nine drugs were identified having both a negative enrichment score and a maximum permutation \(P\) value of 0.1.

*Three of these drugs were also identified using the PETACC3-derived gene expression signature.

Published OnlineFirst October 27, 2014; DOI: 10.1158/1535-7163.MCT-14-0260
performed using classical (bortezomib) and novel (carfilzomib) proteasome inhibitors and demonstrated similar activity. However, as opposed to EGFR feedback, proteasome inhibition seems to function independently of BRAF inhibition (32). CiDD has been able to identify both types of compounds (EGFR and proteasome inhibitors) as candidate drugs through an agnostic approach, thus providing a biologic validation of the value of CiDD as a screening tool to identify novel drugs to be tested and further developed in specific tumor subtypes.

CiDD also addresses the important issue of identifying appropriate publicly available cell lines as preclinical models for cancer researchers. Systematic comparisons between cancer cell lines and tumor samples from human tissues have documented substantial differences between the two, emphasizing the importance of making genomically informed choices when identifying cell lines as preclinical models of a tumor subtype (42). The CCLE provides mutation and gene expression data that allow CiDD to make these molecularly informed decisions in selecting cell lines. In our BRAF V600E analysis, CiDD identified seven large intestine cell lines harboring the BRAF V600E mutation. However, only five of the seven were predicted to be BRAF V600E based on CiDD’s gene expression classifier, suggesting heterogeneity among the BRAF V600E-mutated cell lines. CiDD prioritized those cell lines into two groups for in vitro testing, proposing that five of the seven BRAF V600E-mutated large intestine cell lines more closely resemble the TCGA CRC BRAF V600E tumors at a gene expression level. We note, however, that there may be a more ideal strategy for obtaining cell lines for in vitro testing for researchers wishing to deviate from the use of publicly available cell lines. The use of isogenic cell lines in drug experiments has been shown to be very effective, thus allowing for direct asso-
ciation of cell lines more closely resembling TCGA CRC BRAF V600E tumors at a gene expression level. We note, however, that there may be a more ideal strategy for obtaining cell lines for in vitro testing for researchers wishing to deviate from the use of publicly available cell lines.

CiDD has some limitations that could restrict its application in specific situations. Primarily, CiDD is dependent on identifying a gene expression signature representative of a phenotype of interest. In some cases, there may be no gene expression signature associated with a clinical phenotype or mutation. In other clinical contexts, such as for rare mutations or infrequent clinical phenotypes, CiDD may not have the power to identify the true underlying gene expression signature associated with the phenotype, because CiDD is limited by the number of samples available in TCGA with that specific phenotype. In these rare-phenotype analyses, CiDD may fail to identify a statistically significant gene expression signature representative of the phenotype of interest. Researchers interested in rare clinical or molecular subgroups will need to consider alternative strategies for increasing their sample sizes. These strategies may include aggregating TCGA tumor types or grouping mutations or clinical phenotypes in biologically meaningful ways, such as aggregating rare mutations at a gene or pathway level to increase the sample size. The CiDD command that generates gene expression signatures based on defined mutations provides support for aggregating mutations by listing amino acid substitutions explicitly, by specifying types of mutations (such as nonsense mutations) or by defining sets of mutations based on gene and gene set membership. In addition, the CiDD framework does not support the identification of candidate drug combinations to target tumor subtypes. CMap provides drug-perturbed data that were generated by applying compounds to cell lines one compound at a time. If future drug-perturbed datasets provide gene expression data of multiple compounds being applied to cell lines, incorporation of this data into CiDD should be relatively straightforward. As an alternative, the computational identification of multiple interacting candidate drugs based on current datasets is a potential area for future CiDD development.

Of course, these limitations apply more generally for these difficult scenarios and are not unique to CiDD. In fact, CiDD helps address these limitations by being easy to run and repeat to test multiple hypotheses quickly. Furthermore, CiDD is a framework rather than a specific method per se. As public databases evolve and expand, and as robust statistical methodologies mature for cross-platform expression-based signature identification, CiDD can be adapted to incorporate these improved components. In this sense, what we have demonstrated here is a “lower bound” of sorts, and we expect more powerful findings to emerge from such efficient systems-based computation. Finally, the field of gene expression analysis, particularly for identifying signatures of cancer sub-
types, has been criticized for failing to adhere to standards of repeatability (44). Our software facilitates repeatability and even enables replication of findings with external datasets. In all of these aspects, we expect the community of cancer genomic researchers to benefit from, and further contribute to, this framework.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors’ Contributions
Conception and design: F.A. San Lucas, J. Fowler, E. Vilar, P. Scheet
Development of methodology: F.A. San Lucas, J. Fowler, E. Vilar, P. Scheet
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): F.A. San Lucas, J. Fowler, S. Kopetz, E. Vilar, P. Scheet
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): F.A. San Lucas, J. Fowler, S. Kopetz, E. Vilar, P. Scheet
Writing, review, and/or revision of the manuscript: F.A. San Lucas, J. Fowler, E. Vilar, F. Scheet
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): F.A. San Lucas, J. Fowler, K. Chang, E. Vilar, P. Scheet
Study supervision: E. Vilar, P. Scheet

Acknowledgments
The authors thank TCGA research network for publicly sharing their data. The software and results published here are in large part based upon

Downloaded from mct.aacrjournals.org on March 30, 2017. © 2014 American Association for Cancer Research.
data generated by the TCGA project, which was established by the NCI and NHGRI. Information about TCGA and the investigators and institutions which constitute the TCGA research network can be found at http://cancergenome.nih.gov.

Grant Support
This work was supported by the Schissler Foundation (to F.A. San Lucas), the Conquer Cancer Foundation of the American Society of Clinical Oncology Young Investigator Award (to E. Vilar), the NIH grants 1R03CA176788-01A1 (to E. Vilar), U24 CA143883 (to P. Scheet), U01 GM 92666 (to P. Scheet), RO1HG05859 (to F.A. San Lucas, J. Fowler, P. Scheet), and IR01CA172670-01 (to S. Kopetz), a grant from The University of Texas MD Anderson Cancer Center Duncan Family Institute for Cancer Prevention and Risk Assessment (to E. Vilar and P. Scheet), and by The University of Texas MD Anderson Cancer Center Core Support Grant.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received March 27, 2014; revised August 22, 2014; accepted September 30, 2014; published OnlineFirst October 27, 2014.

References
Molecular Cancer Therapeutics

Cancer In Silico Drug Discovery: A Systems Biology Tool for Identifying Candidate Drugs to Target Specific Molecular Tumor Subtypes

F. Anthony San Lucas, Jerry Fowler, Kyle Chang, et al.


Updated version
Access the most recent version of this article at:
doi:10.1158/1535-7163.MCT-14-0260

Cited articles
This article cites 41 articles, 15 of which you can access for free at:
http://mct.aacrjournals.org/content/13/12/3230.full.html#ref-list-1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.