SUPPLEMENTARY MATERIAL

TARGETgene: A Tool for Identification of Potential Therapeutic Targets in Cancer

Chia-Chin Wu1,*, David Z. D'Argenio2, Shahab Asgharzadeh3, Timothy J. Triche3
1 Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030
2 Department of Biomedical Engineering and Biomedical Simulations Resource, University of Southern, Los Angeles, CA, 90089
3 Children’s Hospital Los Angeles and Keck School of Medicine, University of Southern California, Los Angeles, CA, 90027
*To whom correspondence should be addressed.

This supplementary document is organized as follows. Section S1 lists the data sources used for construction of the whole-genome genetic network that is used in TARGETgene. Section S2 details the network-based metrics used to identify potential therapeutic targets and driver cancer genes. Sections S3 presents some detail results of the first applications: identification of potential therapeutic targets from differentially expressed genes in several cancers. Sections S4 lists all references.

S1 CONSTRUCTION OF THE GENETIC NETWORK

Heterogeneous genomic and proteomic data (Table S1) were integrated using the RVM-based ensemble model reported in [Wu et al., 2010] in order to construct a whole-genome genetic network. The nodes in this network represent all the genes of the human genome, and the probability between any two of them indicates the strength of their functional relationship, which can reveal the tendency of genes to operate in the same or similar pathways. The constructed gene network contains critical information about gene-gene functional relationships in biological pathways that can be used to explore diverse biological questions in health and disease, including exploring gene functions, understanding complex cellular mechanisms, and identifying potential therapeutic targets.
TARGETgene uses this genetic network to map and analyze potential therapeutic target at the systems level.

S 1.1 Data Types Used for Construction of the Whole-Genome Genetic Network

Seventeen kinds of datasets (summarized in Table S1) were integrated to construct the genetic network in this work. These data sources are from the following eight categories.

**Literature**

Automatic text mining techniques are generally used to extract co-occurrence gene relations from biological literature [Li et al., 2006]. In this work, however, we used expert-curated information from the NCBI, composed of genes and their corresponding cited literatures (ftp://ftp.ncbi.nih.gov/gene/). The numbers of co-citations for each gene pair was used to define the strength of the functional relationship for a gene pair.

**Gene Ontology**

Gene Ontology characterizes biological annotations of gene products using terms from hierarchical ontologies [Ashburner et al., 2000]. Three kinds of ontologies were used representing, the molecular function of gene products, their role in multi-step biological processes, and their localization to cellular components. We determined the functional relation of a gene pair by the following steps [Rhodes et al., 2005; Qiu and Noble, 2008]:

1. Identify all GO terms shared by the two genes.
2. Count how many other genes were assigned to each of the terms shared by the two genes.
3. Identify the shared GO terms with the smallest count. (In general, the smaller the count, the greater functional relationship between two genes.)
4. A functional value of a gene pair is computed as the negative logarithm of the smallest count.
<table>
<thead>
<tr>
<th>Data Type</th>
<th># of Genes</th>
<th>Data Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature</td>
<td>26,475</td>
<td>Entrez Gene</td>
</tr>
<tr>
<td></td>
<td>16,015</td>
<td>Ashburner et al., 2000.</td>
</tr>
<tr>
<td>Protein domain</td>
<td>15,565</td>
<td>Ng et al., 2003.</td>
</tr>
<tr>
<td></td>
<td>8,787</td>
<td>Entrez Gene</td>
</tr>
<tr>
<td></td>
<td>2,166</td>
<td>Vastrik et al., 2007.</td>
</tr>
<tr>
<td>Protein-protein interaction and</td>
<td>6,982</td>
<td>Gary et al., 2003.</td>
</tr>
<tr>
<td>genetic interaction</td>
<td>9,295</td>
<td>Keshava Prasad et al., 2009.</td>
</tr>
<tr>
<td></td>
<td>6,279</td>
<td>Cline et al., 2007</td>
</tr>
<tr>
<td></td>
<td>1,959</td>
<td>Ewing et al., 2007.</td>
</tr>
<tr>
<td>Gene context</td>
<td>11,303</td>
<td>Bowers et al., 2004</td>
</tr>
<tr>
<td>Protein phosphorylation</td>
<td>5,490</td>
<td>Linding et al., 2008</td>
</tr>
<tr>
<td></td>
<td>3,205</td>
<td>Yang et al., 2008</td>
</tr>
<tr>
<td>Gene expression profile</td>
<td>19,777</td>
<td>Obayashi et al., 2008</td>
</tr>
<tr>
<td>Transcription regulation</td>
<td>937</td>
<td>Ferretti et al., 2007</td>
</tr>
</tbody>
</table>
**Protein-Protein Interactions and Genetic Interactions**

Experimental human protein-protein interactions were collected from diverse databases, including, NCBI, Reactome [Vastrik et al., 2007], BIND [Gary et al., 2003], HPRD [Keshava Prasad et al., 2009], and Cytoscape [Cline et al., 2007] (all were downloaded on December 2008). All the interactions are supported by different experiments, with most interactions in these sets derived from small-scale studies. Additional physical interactions were generated from published genome-scale screens using mass spectrometry analyses of affinity-purified protein complexes or high throughput yeast two hybrid (Y2H) assays. Since the experiments identifying the interactions can sometimes produce false-positives, we considered that number of different experiments of each gene pair as its confidence score. In addition, we also include protein-protein interactions from mass spectrometry data [Ewing et al., 2007].

**Protein Domain-Domain Interaction**

Proteins are known to interact with each other through protein domains, which represent modular protein subunits that are often repeated in various combinations throughout the genome. Thus, if two domains can physically interact, proteins containing these two domains are also likely to interact. In this work, we downloaded the predicted domain-domain interactions from the database InterDom (http://interdom.i2r.a-star.edu.sg/) [Ng et al., 2003]. These interactions were predicted based on protein structural information, and each interaction pair was assigned a confidence score. We assigned the score of each protein domain pair (inferred by InterDom) to all protein pairs containing them.

**Gene Context**

Comparative genome analyses of sequence information (Gene Context) have been successfully used to assign protein functions. The Prolinks database (http://mysql5.mbi.ucla.edu/cgi-bin/functionator/pronav) is a collection of these inference methods used to predict functional linkages between proteins [Bowers et al., 2004]. These include Gene Cluster, which uses genome proximity to predict functional linkage, Gene Neighbor, which uses both gene proximity and phylogenetic distribution to infer
linkage, Rosetta Stone, which uses a gene fusion event in other organisms to infer functional relatedness, and Phylogenetic Profile which uses the presence or absence of proteins across multiple genomes to detect functional linkages [Bowers et al., 2004]. Internal Prolinks IDs of all genes were transferred to Entrez Gene IDs. The scores of gene pairs inferred by Prolinks were assigned as the Gene Context feature.

In addition, we also generated Phylogenetic profiles from the ortholog clusters in the KEGG database [Kanehisa et al., 2010], which describes the sets of orthologous proteins in 1111 organisms. In our work, we focused only on the 188 organisms with fully sequenced genomes [Genome News Network, 2009]. The phylogenetic profile of each gene consists of a string of bits which is coded as 1 and 0 to respectively indicate the presence and absence of its orthologous protein across the 188 organisms. The functional relationship of phylogenetic profiles for any two genes was then assessed using the mutual information (MI) values [Date and Marcotte, 2003]. A gene pair whose MI value is higher was considered as more confident functional interaction.

**Protein Phosphorylation**

Regulation of proteins by phosphorylation is one of the most common ways of regulation of protein function in a pathway. Protein kinases control cellular responses by phosphorylating specific substrates in a cascade of signaling processes. The NetworKIN database (http://networkin.info) integrates consensus substrate motifs with context modeling to predict cellular kinase-substrate relationships based on the latest human phosphoproteome from the Phospho.ELM and PhosphoSite databases [Linding et al., 2007; Linding et al., 2008]. The database currently contains a predicted phosphorylation network of interactions involving 5,515 phospho-proteins and 123 human kinases. Ensemble IDs of all proteins were transferred to Entrez Gene IDs. The scores of gene pairs inferred by NetworKIN were directly assigned as the Protein Phosphorylation feature. In addition, another data source of Protein Phosphorylation, PhosphoPOINT [Yang et al., 2008], also provides 4,195 phospho-proteins, 518 serine/threonine/tyrosine kinases, and their corresponding protein interactions.
Gene Expression

Two genes in the same pathway are likely to have correlated gene expression profiles [Tavazoie et al. 1999]. Co-expression data were directly downloaded from COXPRESdb (http://coxpresdb.hgc.jp/), which was derived from publicly available GeneChip data [Obayashi et al., 2008]. It contains correlation data for 19,777 gene expression profiles in human.

Transcription Regulation (Co-Regulation)

Some genes in the same pathways are likely to be regulated by the same transcription regulators that bind to their regulatory elements. Gene co-regulation can be detected by ChIP-chip assays and may also be predicted by some computational approaches based on sequence motif information or phylogenetical conservation. In this work, the co-regulation data were downloaded directly from the PReMod database (http://genomequebec.mcgill.ca/PReMod), which describes more than 100,000 computationally predicted transcriptional regulatory modules within the human genome [Ferretti et al., 2007]. These modules represent the regulatory potential for 229 transcription factors families.

S.1.2 Construction of the Genetic Network using the RVM-based Ensemble Method

These 17 diverse data sources were all used with the previously developed Relevance Vector Machines (RVM)-based ensemble approach [Wu et al., 2010] to compute the genetic functional associations (i.e., tendency of genes to operate in the same pathways) between all gene pairs given the input data features. The RVM-based model combined two ensemble approaches, AdaBoost [Schapire and Singer, 1999] and Sub-Feature [Saar-Tsechansky and Provost, 2007], to simultaneously address the two major problems associated with constructing a genetic network: large-scale learning and massive missing data values. The Gold standard datasets for model building were generated from KEGG pathways. A complete explanation of RVM-based ensemble approach is provided in [Wu et al., 2010].
S2 NETWORK-BASED APPROACHES TO IDENTIFY IMPORTANT CANCER-RELATED GENES

Based on this constructed gene network, TARGETgene identifies potential therapeutic targets using one of two network-based metrics: 1) hub score, which uses a centrality measure to identify hub genes in a tumor-specific network, or 2) seed gene association score, which quantifies each gene's association with known cancer (disease) genes.

S 2.1 Identification using Network Centrality Metrics

In view of the complexity in cancers, potential therapeutic targets can be those genes/proteins that have a critical role in regulating multiple pathways or maintaining those malignant phenotypes. Recently, cancer-associated genes are found more likely to be signaling proteins that act as signaling hubs, actively sending or receiving signals through multiple signaling pathways [Cui et al., 2007]. In addition, under the modular structure of biological networks, intermodular hubs are found to be more associated with cancer phenotypes than intramodular hubs, since intermodular hubs interact with other intramodular hubs temporally and spatially that in turn fulfill different specific molecular functions [Taylor et al., 2009]. Therefore, potential therapeutic targets can be those hub genes in a tumor-specific network. A tumor-specific network can be generated by directly mapping the candidate gene (e.g., differentially expressed genes in a tumor) to the constructed genetic network. Two centrality measurements provided in TARGETgene can quantify the tendency of a gene to be a hub in the tumor-specific network. All candidate genes in the tumor-specific network are ranked based on their centrality measurement in the tumor-specific network. Those highly ranked hub genes can be considered as potential therapeutic targets.

Topological measures of centrality, such as total degree [Freeman, 1977], betweenness [Freeman, 1977], closeness [Freeman, 1979], and eigenvector centrality
[Newman, 2003] are typically used to determine hub genes (central nodes) in a binary network (i.e., unweighted network). However, since most gene pairs in a tumor-specific network have weighted linkages, betweenness and closeness, which are limited to calculation of the shortest path between any two gene pairs, are not used for calculating centrality in TARGETgene. Instead, the centrality metrics, weighted degree centrality and weighted eigenvector centrality [Barrat et al., 2004; Newman, 2004] are used in TARGETgene and briefly discussed below.

**Weighted degree centrality**

In a weighted network, it is intuitive to consider a definition of total degree that is based on the strength of nodes in terms of the total weight of their connections [Barrat et al., 2004; Newman, 2004].

\[
d_i = \sum_{j=1}^{n} w_{i,j}
\]

(S1)

where \( d_i \) is the centrality measurement of gene \( i \), \( w_{i,j} \) is the functional relationship between gene \( i \) and gene \( j \) in the network, and \( n \) is the number of differently expressed genes. Highly weighted nodes (larger \( d_i \)) are more central.

**Weighted Eigenvector centrality**

Eigenvector centrality is closely related to “PageRank”, a similar centrality measure used in web search engines. The eigenvector centrality \( e_i \) of a vertex in a weighted network is proportional to the weighted sum of the centralities of the vertex’s neighbors. Thus a vertex can acquire high centrality either because it is connected to a many others or because it is connected to others that themselves highly central [Newman, 2004]. We can write

\[
e_i = \lambda^{-1} \sum_{j=1}^{n} w_{i,j} e_j
\]

(S2)

where \( \lambda \) is a constant. Using matrix notation, Eq. (S2) can be written \( \lambda E = WE \), so that \( E \) is an eigenvector of the adjacency weighted matrix \( W \) of a weighted network. The eigenvector centrality of all vertexes is the eigenvector corresponding to the max eigenvalue.
S 2.2 Association with Seed Genes (Known Cancer Genes)

Genes associated with similar disease phenotypes tend to be interconnect in a biological network (i.e., participate in the same molecular pathway or the same protein complexes). Based on this concept, several network-based computational approaches [Franke et al., 2006; Köhler et al., 2008; Chen et al., 2009; Linghu et al., 2009] have been proposed to predict novel disease genes. Given a set of known genes of a disease (i.e. seed genes), functional associations (linkages) of other genes with these seed genes in biological networks can be calculated. Genes that are found to be more associated with the known disease genes are more likely involved in the disease process.

Therefore, TARGETgene also allows users to identify important cancer genes or potential therapeutic targets by associating them with user-defined seed genes (e.g., known cancer genes) in the genetic network. More specifically, the importance of each candidate gene is calculated as summation of its direct functional association with those seed genes.

\[ c_i = \sum_{j=1}^{m} w_{i,j} \]  

(S3)

where \( c_i \) is the degree of association of gene \( i \) with seed genes, \( m \) is the number of seed genes, and \( w_{i,j} \) is the functional association of gene \( i \) to seed gene \( j \) in the constructed genetic network. Genes with more associations with all the seed genes (i.e., larger \( c \) values) are likely to play more important roles in the cancer and can be potential therapeutic targets.
S3 EXAMPLE 1: IDENTIFICATION OF POTENTIAL THERAPEUTIC TARGETS FROM DIFFERENTIALLY EXPRESSED GENES

S3.1 Rank Genes Based On Their Weighted Degree Centrality in the Tumor-Specific Network

In this example, TARGETgene was applied in turn to each of three cancer types: Her2-positive breast cancer, colon cancer, and Lung Adenocarcinoma. Human Exon datasets in the Affymetrix platform for the three cancer types were collected from the National Center for Biotechnology Information Gene Expression Omnibus (GEO) [Barrett et al., 2007]. There are 10 and 20 tumor/normal paired specimens in Colon Cancer [Affymetrix sample data of exon array] and Lung Adenocarcinoma (GSE12236) [Xi et al., 2008], respectively. In addition, the case study of Breast Cancer includes 35 samples from patients with HER2 positive and three samples from normal breast tissues (GSE16534) [Lin et al., 2009]. Subsequent data analyses were done using Partek Genomic Suite 6.3 (Partek Inc.). The RMA (Robust Multichip Analysis) algorithm [Irizarry et. al., 2003] was used to do background correction, normalization and summarization. Exon-level data in each cancer type was then filtered to include only those probesets that represent 17,800 RefSeq genes and full-length GenBank mRNAs. Any effect of different microarray processing was removed using a batch removal tool of Partek Genomic Suite. ANOVA p-values and fold changes of gene expression in cancer samples against normal tissues were calculated. Finally, using a criteria of P<0.01 in the ANOVA analysis, 5203, 5,153 and 6,203 differentially expressed genes were identified in case studies of colon, breast, and lung cancer, respectively.

Differentially expressed genes in each cancer type were all ranked based on the extent of their weighted degree of centrality (Section S2.1) in a tumor-specific network, which was generated by mapping the differentially expressed genes in each cancer type to the constructed genetic network (Section S1). Figures S1.a, b, and c list the top 10 highest ranked genes for each of the three cancer types as shown in the Gene Panels of TARGETgene. The complete ranking list of genes for each of the three cancer types can
be obtained by running TARGETgene using the candidate genes list stored in the examples files and selecting the weighted degree centrality ranking option. The results show that a number of important cancer genes for each cancer type are ranked highly by TARGETgene including: AKT1 (#1), SRC (#10), ERBB2 (#25), and ESR2 (#56) in breast cancer; MYC (#174), CTNNB1 (#119), APC (#116), and DCC (#195) in colon cancer; KIT (#30), ERBB2 (#31), PPARG (#77), and PTEN (#157) in lung cancer. In addition, TARGETgene also ranks several genes highly (in the top 10%) that were recently identified as cancer-related genes in each cancer type. For example, in breast cancer we ADAM12 (rank #153) and MAP3K6 (rank #205) were recently reported to be associated with breast cancer oncogenesis [Sjoblom et al., 2006; Wood et al., 2007].

Moreover, many genes that have never been identified in each type of cancer are also ranked highly. These genes could be subject in vitro and in vivo study to evaluate their importance in each cancer type. Several of these have been identified by RNAi screens (Section S3.2.4 presents details on evaluation of predictions based on RNAi screens). For example, in colon cancer, RIPK2 and ENC1 (ectodermal-neural cortex) have a TARGETgene rank of 8 and 257, respectively. RIPK2 encodes a member of the receptor-interacting protein (RIP) family of serine/threonine protein kinases. It is also a potent activator of NF-kappaB and inducer of serine/threonine protein kinases. It is also a potent activator of NF-kappaB and inducer of apoptosis in response to various stimuli [Tao et al., 2009]. ENC1 activates p53 tumor suppressor protein and induces cell cycle arrest or apoptosis [Polyak et al., 1997]. It also has been shown to be involved in oncogenesis of brain [Seng et al., 2009] and breast cancer [Seng et al., 2007]. In breast cancer, PIK3R2 (phosphoinositide-3-kinase, regulatory subunit 2 beta) and CIT (citron) have a TARGETgene rank of 37 and 115, respectively. PIK3R2, with a 3.31 fold change in gene expression of breast cancer tissues, has been shown to be functionally involved in several cancer related pathways, such as the PI3K/Akt pathway [Radhakrishnan et al., 2008], and also associated with several other cancer types, such as ovarian cancer [Zhang et al., 2007]. CIT (citron), with a 3.06 fold change in gene expression in breast cancer tissues is a kinase that has been identified to be associated with the cell cycle [Liu et al., 2003]. In lung adenocarcinoma, MAPK13 and CBLC (Cas-Br-M (murine) ecotropic retroviral transforming sequence c) have TARGETgene ranks 19 and 173, respectively. MAPK13 is
involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development. MAPK13 has also been found to be a downstream carrier of the PKCdelta-dependent death signaling [Efimova et al., 2004]. CBLC has been reported to interact with AIP4 to cooperatively down-regulate EGFR signaling [Courbard et al., 2002]. In addition, CBLC also been shown to be a negative regulator of receptor tyrosine kinase Met signaling in B cells and to mediate ubiquitination and thus proteosomal degradation of Met, with a role in Met-mediated tumorigenesis [Taher et al., 2002]
Figure S1. Screen shots from Gene Panel for each cancer type

(a) Breast Cancer

(b) Colon Cancer

(c) Lung Cancer
S3.2 Evaluation of Predictions

TARGETgene also compares its resulting ranked genes to several benchmark gene sets, including the set of curated cancer genes, the set of genes cited in cancer literature, and the set of target genes detected by RNAi screens. Receiver Operating Characteristic (ROC) Curves are used for this evaluation.

S3.2.1 Evaluation of Predictions using Known Cancer Genes

The 1,186 curated cancer genes downloaded from the CancerGenes database [Higgins et al., 2006] are first used to evaluate if they are highly ranked by TARGETgene. These cancer genes, however, are not classified to any specific cancer type. For each cancer type, we therefore treat those genes as specific to a cancer type if they are cited by literature source related to that cancer type (Pubmed data on Dec. 2008). The curated cancer genes are considered as positive instances while other remaining genes are treated as negative instances. Figure S2.a shows TARGETgene’s prediction performance for each cancer type, evaluated using ROC curves and AUC. The high AUC values of TARGETgene’s prediction in each cancer type (all AUC > 0.85) indicate that most of known cancer genes tend to be ranked highly. (This result also reveals that the human genetic network constructed by the RVM-based model contains critical pathway information and can successfully be used to identify other important cancer genes.)

Genes that are cited by the literature of each cancer type are also used for evaluation. In this work, all Pubmed IDs of literature related to colon cancer, breast cancer, and lung adenocarcinoma were first downloaded from Pubmed on Dec. 2008. For each gene, we calculated the number of citations related to each cancer type by mapping the extracted Pubmed IDs to the gene citation information from Entrez Gene (ftp://ftp.ncbi.nih.gov/gene/), composed of genes and their corresponding cited literature. The evaluation was also based on ROC curves. Figure S2.b shows the ROC curves for the three cancer types in which genes are selected as the benchmark genes if they are cited by at least one cancer literature. The AUC values of the ROC curves for TARGETgene’s predictions are great than 0.7 for each cancer type. It is expected that the
resulting AUC’s are uniformly lower when compared to those obtained using the curated cancer genes as the benchmark, because literature citation data are noisy. The results using literature citation also depend on the number of citations (set at 1 in the results shown in Figure S2.b). In addition, as the citation cutoff number used increases (Figure S3.a-c) so do the resulting TARGETgene AUC values, indicating that genes with more citations (presumably because they are more extensively studied) also have a higher TARGETgene ranking (Figure 4.a-c). Spearman's rank correlation is also used to assess correlation between citation number and TARGETgene ranking. The resulting correlations for colon, breast and lung cancer are 0.2665, 0.3658, and 0.2927, respectively, which are all significantly higher than random expectation (P≈0.000). Recall that TARGETgene ranks many novel genes without any previous literature citations highly, which depresses the Spearman rank correlation coefficient. Nevertheless, this provides further evidence genes highly ranked by TARGETgene are also are cited more in the cancer literature.
Figure S2. ROC curve performance evaluation (true positive rate – TPR, versus false positive rate – FPR) of TARGETgene using curated cancer genes (a) and genes cited by cancer literature (one or more citations) (b).
Figure S3. ROC curve performance evaluation (true positive rate – TPR, versus false positive rate – FPR) of TARGETgene using genes cited by cancer literature with different citation number cutoff values of 1, 5 and 10.
(a). Breast Cancer

(b). Colon Cancer

(c). Lung Adenocarcinoma

Figure S4. Number of cancer literature citation of genes vs TARGETgene gene ranks (Gene Ranking Block) in the predictions of each cancer type.
**S3.2.3 Evaluation of Predictions using Gene Function Annotations**

Gorilla [Eden et al., 2009], a gene ontology enrichment analysis tool, was applied to identify enriched GO terms that appear densely at the top of TARGETgene’s ranked gene lists for each of the three cancer types. Many of identified GO process terms are known cancer-related biological processes. The examples of identified biological process terms include, regulation of cell death (GO:0010941), regulation of apoptosis (GO:0042981), regulation of cell proliferation (GO:0042127), regulation of cell migration (GO:0030334), angiogenesis (GO:0001525; GO:0060055), and regulation of cell differentiation (GO:0045595). Interestingly, several biological processes related to new hallmarks of cancers [Luo et al., 2009] are also identified. They are DNA damage (GO:0006974, GO:0042770), oxidative stress (GO:0070482), evading immune surveillance (GO:0002682; GO:0002684), metabolic stress (GO:0006796; GO:0006793), mitotic stress (GO:0007059; GO:0007346), and proteotoxic stress (GO:0009408; GO:0051603). These results indicate that genes highly ranked by TARGETgene are involved in multiple cancer-related biological processes and pathways.

Several types of molecules, such as signaling kinases, receptor tyrosine kinases, and transcription factors are often proposed as possible molecular targets in cancers [Shawver et al, 2002; Sawyers, 2004; Krause et al., 2005; Frank, 2009]. For example, protein kinases are enzymes that modify other proteins by chemically adding phosphate groups to them (protein phosphorylation). Protein phosphorylation has proven to be an important driving force in cellular signaling. Protein kinases can impact many cellular processes through their ability to control protein-protein interactions, complex formation, enzyme activity and protein degradation and translocation [Seet et al., 2006]. We find that many kinase, receptor, and transcription factor related GO function terms are enriched in highly-ranked genes in TARGETgene (Figure S5). The examples of identified molecular function terms include protein serine/threonine kinase activity (GO:0004674), protein tyrosine kinase activity (GO:0004713), kinase binding (GO:0019900), growth factor receptor binding (GO:0070851), transcription regulator activity (GO:0030528), and transcription factor binding (GO:0008134). The results indicated that many of the genes highly ranked by TARGETgene are kinase, receptor, and
transcription factor related genes.

Figure S5. Proportion of genes related to kinase, receptor, and transcription factors vs TARGETgene gene ranks (Gene Ranking Block) in the predictions of each cancer type.

**S3.2.4 Evaluation using The Results of RNAi Screens**

High-throughput RNA interference (RANi) screens are a powerful tool for genome-wide knockdown of specific gene products or perturbation of gene expression. The phenotypic results from the screen can be monitored by assaying for specific alterations in molecular and cellular endpoints, such as promoter activation, cell proliferation and viability [Iorns et al., 2007]. RNAi screens have recently been shown to be a promising tool to discover new targets for the treatment of several cancers. Therefore, data from RNAi screens can be applied to evaluate the performance of the predictions from TARGETgene. Effective targets of each cancer type detected by RNAi screens were all downloaded from GenomeRNAi, a database for cell-based RNAi phenotypes [Gilsdorf et al., 2009]. This database contains phenotypes from a number of cell-based RNA
interference screens in human cells. We selected the RNAi screens for the three cancer types whose phenotype is related to cell viability. The data sources of RNAi screens used in this work are summarized in Table S2a-c. Since lung adenocarcinoma is a type of non-small cell lung cancer (NSCLC), the result of RNAi screens in NSCLC cell lines were used for evaluation. The evaluation of prediction performance was assessed using ROC curve. In each case of cancer type, the effective targets detected by RNAi screens are treated as positive instances while others genes are treated as negative instance.

**Table S2a:** Data Sources of RNAi Screens in Breast Cancer Cell Lines

<table>
<thead>
<tr>
<th>Reference</th>
<th>Assay</th>
<th>Cell Lines</th>
<th>Number of Detected Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simpson et al., 2008</td>
<td>Cell migration and viability</td>
<td>MCF-10A</td>
<td>66</td>
</tr>
<tr>
<td>Schlabach et al., 2008</td>
<td>Cell viability</td>
<td>HCC1954</td>
<td>176</td>
</tr>
<tr>
<td>Silva et al., 2008</td>
<td>Cell viability</td>
<td>MCF-10A; MDA-MB-435</td>
<td>172</td>
</tr>
<tr>
<td>Swanton et al., 2007</td>
<td>Cell viability</td>
<td>MDA-MB-231</td>
<td>45</td>
</tr>
<tr>
<td>Turner et al., 2008</td>
<td>Cell viability</td>
<td>CAL51</td>
<td>24</td>
</tr>
<tr>
<td>Iorns et al., 2009</td>
<td>Cell viability</td>
<td>MCF-7</td>
<td>20</td>
</tr>
<tr>
<td>Brummelkamp et al., 2006</td>
<td>Cell viability</td>
<td>MCF-7</td>
<td>13</td>
</tr>
</tbody>
</table>

Total Number of detected targets 441

**Table S2b:** Data Sources of RNAi Screens in Colon Cancer Cell Lines

<table>
<thead>
<tr>
<th>Reference</th>
<th>Assay</th>
<th>Cell Lines</th>
<th>Number of detected targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schlabach et al., 2008</td>
<td>Cell viability</td>
<td>DLD1; HCT116</td>
<td>243</td>
</tr>
<tr>
<td>Moffat et al., 2006</td>
<td>Mitotic index/Cell viability</td>
<td>HT29</td>
<td>128</td>
</tr>
<tr>
<td>Swanton et al., 2007</td>
<td>Cell viability</td>
<td>HCT-116</td>
<td>45</td>
</tr>
<tr>
<td>Firestein et al., 2008</td>
<td>Wnt signaling/Cell viability</td>
<td>DLD1; HCT116</td>
<td>9</td>
</tr>
</tbody>
</table>

Total Number of detected targets 271

**Table S2c:** Data Sources of RNAi Screens in NSCLC Cell Lines

<table>
<thead>
<tr>
<th>Reference</th>
<th>Assay</th>
<th>Cell Lines</th>
<th>Number of Detected Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swanton et al., 2007</td>
<td>Cell viability</td>
<td>A549</td>
<td>45</td>
</tr>
<tr>
<td>Ji et al., 2007</td>
<td>Cell viability</td>
<td>A549</td>
<td>10</td>
</tr>
</tbody>
</table>

Total Number of detected targets 55
The result is shown in Figure S6. The high AUC in each cancer type indicates that the most effected targets identified in the genome-wide RNAi screens tend to be ranked highly by TARGETgene. Some highly ranked genes have been shown to play an important role in oncogenesis in each of the three cancer type, such as AKT1 (#1) in Breast Cancer, MET (#107) in Colon Cancer, and PTEN (#157) in NSCLC. Most interestingly, we also found that many novel targets (i.e., no citation related to the specific cancer type based on PubMed in Dec. 2008) detected by RNAi screens are also ranked highly by TARGETgene. For example, CASK (calcium/calmodulin-dependent serine protein kinase) and RUVBL1 (RuvB-like 1) are ranked 161 and 433, respectively in the prediction of breast cancer. Such results provide support from cell line models for the ability of TARGETgene to identify novel therapeutic targets in cancers. This also suggests the possibility of combination of RNAi and network-based screens for therapeutic target identification as discussed.

Figure S6 TARGETgene prediction performances (true positive rate – TPR, versus false positive rate – FPR) evaluated by the results of RNAi screens (cell viability).

S3.3 Mapping the Predictions to Drug-Target Information

Recently, information on drugs/compounds and their targets that have been approved
or are under evaluation for use in cancer treatment, have become available electronically and accessible through several public databases. This information is used by TARGETgene to report those drugs/compounds that could have action of the targets identified by TARGETgene. In this work, the information of drugs/compounds and their targets were compiled from DrugBank [Knox et al., 2011], PharmGKB [Hodge et al., 2007], and Therapeutic Target Database [Zhu et al., 2009] on Dec. 2010. The database extracted from these sources and used by TARGETgene contains nearly 4800 drug entries, including 1,350 FDA-approved small molecule drugs, 123 FDA-approved biologics (protein/peptide) drugs, 71 nutraceuticals, and 3,243 experimental drugs, as well as approximately 6,000 drug-target relationships.

After mapping the information of drugs/compounds and their targets to the ranked gene lists from TARGETgene, we found that many genes highly rank by TARGETgene are targets for some drugs that have already been in clinical trials or have been used for treatment of the three cancer types. Other identified drugs and compounds that are not used in clinical trials have also shown anti-cancer effect and could thus be considered as potential novel drug for these cancers. Table S3.a-c lists some of these drugs and compounds whose targeted genes are overexpressed and highly ranked by TARGETgene in each cancer type. In the case of breast cancer, Trastuzumab and Lapatinib have been approved for HER2 positive Breast cancer, and their main target erbB2 is very highly ranked by TARGETgene (and up-regulated). Other endocrine treatments for ER-positive breast cancer, such as Taxmoxifen, are not included because its main target ESR1 and ESR2 are not overexpressed in our analysis. Several other drugs, such as Dasatinib, UCN-01, Celecoxib, Flavopiridol, and Vorinostat, have already been in clinical trials for the treatment of breast cancer. Some of their targets are highly ranked by TARGETgene. Moreover, other drug/compounds have been shown to have anti-tumor effects and could be considered as potential novel drugs for the treatment in breast cancer, such as Alsterpaullone and Olomoucine. In addition, two naturally occurring compounds, melatonin and vitamin D (Calcidiol), are also identified by TARGETgene. Melatonin, a naturally occurring compound found in organisms, can regulate the circadian rhythms of several biological functions. Recently, a clinical trial involving a total of 643 cancer
patients using melatonin found a reduced incidence of death [Mills et al., 2005]. A study showed that women with low melatonin levels have an increased risk for breast cancer [Navara and Nelson, 2007]. Vitamin D receptors have been found in up to 80% of breast cancers, and vitamin D receptor polymorphisms have been associated with differences in survival [Buras et al., 1994; Friedrich et al., 2002; Diesing et al., 2006]. Active vitamin D compounds (Calcidiol; Calcitriol) also have been identified for their antiproliferative effects in breast cancer cells [Costa et al., 2009; Köstner et al., 2009], although the detail mechanisms are still unclear. In summary, these results provide some further evidence that genes that are highly ranked by TARGETgene be potential therapeutic targets.
<table>
<thead>
<tr>
<th>Drugs/Compounds</th>
<th>Gene and Ranking</th>
<th>Fold Changes in Cancer</th>
<th>Literatures Of Breast Cancer Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dasatinib (E)*</td>
<td>SRC(#10)</td>
<td>2.623</td>
<td>Fornier et al., 2011</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Herold et al., 2011</td>
</tr>
<tr>
<td>Celecoxib (A)*</td>
<td>PDPK1 (#14)</td>
<td>2.917</td>
<td>Fujii et al., 2008</td>
</tr>
<tr>
<td>Staurosporine (UCN-01) (E)*</td>
<td>PDPK1 (#14) MAPKAPK2 (#62) CSK (#19) GSK3B (#84)</td>
<td>2.917</td>
<td>Koh et al., 2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hawkins et al., 2005</td>
</tr>
<tr>
<td>Flavopiridol (E)*</td>
<td>CDK5 (#41) CDK2 (#108) CDK4 (#50)</td>
<td>4.640</td>
<td>Fornier et al., 2007</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Witters et al., 2004</td>
</tr>
<tr>
<td>Alsterpaullone (E)</td>
<td>CDK5 (#41) GSK3B (#84) CDC2 (#108)</td>
<td>4.640</td>
<td>Kohfeld et al., 2007</td>
</tr>
<tr>
<td>Olomoucine (E)</td>
<td>CDC2 (#108) CDK5 (#41)</td>
<td>4.382</td>
<td>Wesierska-Gadek et al., 2004</td>
</tr>
<tr>
<td>Trastuzumab (A)***</td>
<td>ERBB2 (#25)</td>
<td>46.856</td>
<td>Wardley et al., 2009.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Kaufman et al., 2009</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Esteva et al., 2009</td>
</tr>
<tr>
<td>Dexrazoxane (A)***</td>
<td>TOP2A(#302)</td>
<td>10.965</td>
<td>Gligorov and Lotz, 2008</td>
</tr>
<tr>
<td>Lithium (A)</td>
<td>GSK3B (#84)</td>
<td>2.130</td>
<td>Farina et al., 2009</td>
</tr>
<tr>
<td>Melatonin (A)</td>
<td>CALR(#651)</td>
<td>1.778</td>
<td>Navara and Nelson, 2007</td>
</tr>
<tr>
<td>Calcidiol (A)</td>
<td>VDR (#241)</td>
<td>3.358</td>
<td>Costa et al., 2009</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Köstner et al., 2009</td>
</tr>
<tr>
<td>Vorinostat (A)*</td>
<td>HDAC3 (#307) HDAC1 (#497) HDAC2 (#564)</td>
<td>2.336</td>
<td>Luu et al., 2008</td>
</tr>
<tr>
<td>Geldanamycin (17-AAG) (E)*</td>
<td>HSP90B1 (#258) HSP90AA1 (#275)</td>
<td>1.779</td>
<td>Beliakoff and Whitesell, 2004</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Perotti et al., 2008</td>
</tr>
<tr>
<td>Arsenic trioxide (A)*</td>
<td>AKT1 (#1) CCND1 (#418)</td>
<td>4.566</td>
<td>Li et al., 2004</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ye et al., 2005</td>
</tr>
</tbody>
</table>

**Note:** 1. Approved drugs are denoted as ‘A’
2. Experimental compounds are denoted as ‘E’
3. Drugs have been approved for the treatment of Breast Cancer are marked with ***
4. Drugs in clinical trials for Breast Cancer are marked with *
Table S3.b Selected Drugs Targeting High-Ranked Genes in Colon Cancer Identified by TARGETgene

<table>
<thead>
<tr>
<th>Drugs/Compounds</th>
<th>Gene &amp; Ranking</th>
<th>Fold Changes in Cancer</th>
<th>Literatures Of Colon Cancer Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Celecoxib (A)*</td>
<td>PDPK1 (#7)</td>
<td>1.256</td>
<td>Yang et al., 2010</td>
</tr>
<tr>
<td>Sorafenib (A)*</td>
<td>KDR (#25) PDGFRB (#40)</td>
<td>1.203 1.614</td>
<td>Walker et al., 2009</td>
</tr>
<tr>
<td>Sunitinib (A)*</td>
<td>KDR (#25) PDGFRB (#40)</td>
<td>1.203 1.614</td>
<td>Blesa and Pulido, 2010</td>
</tr>
<tr>
<td>Dasatinib (A)*</td>
<td>PDGFRB (#40) EPHA2 (#61)</td>
<td>1.614 1.440</td>
<td>Kopetz et al., 2009</td>
</tr>
<tr>
<td>Etoposide (A)*</td>
<td>TOP2A (#291)</td>
<td>1.748</td>
<td>Chamberlain et al., 2006 Schonn et al., 2009</td>
</tr>
<tr>
<td>Vorinostat (A)*</td>
<td>HDAC3 (#293) HDAC2 (#634)</td>
<td>1.234 1.427</td>
<td>Walker et al., 2009 Fakih et al., 2009</td>
</tr>
<tr>
<td>Bevacizumab (A)***</td>
<td>FCGR2A (#171) FCGR3A (#190) VEGFA (#872)</td>
<td>1.503 1.832 1.709</td>
<td>Giantonio et al., 2007 Hurwitz et al., 2004</td>
</tr>
<tr>
<td>Cetuximab (A)***</td>
<td>FCGR2A (#171) FCGR3A (#190)</td>
<td>1.503 1.832</td>
<td>Zhang et al., 2007</td>
</tr>
<tr>
<td>Atorvastatin (A)*</td>
<td>AHR(#540)</td>
<td>1.375</td>
<td>Yang et al., 2010 Poynter et al., 2005</td>
</tr>
<tr>
<td>Imatinib (A)*</td>
<td>PDGFRB (#40) DDR1 (#145)</td>
<td>1.614 1.308</td>
<td>Kitadai et al., 2006 Mueller et al., 2007</td>
</tr>
<tr>
<td>SU9516 (E)</td>
<td>CDK5 (#23) CDC2 (#102) CDC2 (#47)</td>
<td>1.302 1.247 1.394</td>
<td>Takagi et al., 2008 Lane et al., 2001</td>
</tr>
<tr>
<td>Flavopiridol (E)*</td>
<td>CDK5 (#23) CDK7 (#28) CDC2 (#102) CDK4 (#32) CDK6 (#45) CDK2 (#47)</td>
<td>1.302 1.171 1.247 1.666 1.515 1.394</td>
<td>Ambrosini et al., 2008 Newcomb, 2004</td>
</tr>
<tr>
<td>Staurosporine (UCN-01) (E)*</td>
<td>PDPK1 (#7) GSK3B (#70) CDK2 (#47)</td>
<td>1.256 1.208 1.394</td>
<td>Hotte1 et al., 2006</td>
</tr>
<tr>
<td>Epirubicin (A)*</td>
<td>TOP2A (#291)</td>
<td>1.748</td>
<td>Goff et al., 2008</td>
</tr>
<tr>
<td>Drugs/Compounds</td>
<td>Gene &amp; Ranking</td>
<td>Fold Changes in Cancer</td>
<td>Literatures Of Lung Cancer Treatment</td>
</tr>
<tr>
<td>----------------------</td>
<td>-------------------------------</td>
<td>------------------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>Bevacizumab (A)***</td>
<td>FCGR2A (#252)</td>
<td>1.965</td>
<td>Johnson et al., 2004</td>
</tr>
<tr>
<td></td>
<td>FCGR3A (#291)</td>
<td>2.072</td>
<td>Reck et al., 2009</td>
</tr>
<tr>
<td></td>
<td>C1QA (#747)</td>
<td>1.890</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C1QB (#1567)</td>
<td>1.825</td>
<td></td>
</tr>
<tr>
<td>Sorafenib (A)*</td>
<td>KDR (#50)</td>
<td>2.100</td>
<td>Blumenschein et al., 2009</td>
</tr>
<tr>
<td></td>
<td>KIT (#30)</td>
<td>2.440</td>
<td>Scaglioni et al., 2009</td>
</tr>
<tr>
<td></td>
<td>FLT4 (#34)</td>
<td>1.526</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PDGFRB (#69)</td>
<td>1.937</td>
<td></td>
</tr>
<tr>
<td>Sunitinib (A)*</td>
<td>FLT1 (#12)</td>
<td>1.925</td>
<td>Pal et al., 2010</td>
</tr>
<tr>
<td></td>
<td>CSF1R (#16)</td>
<td>1.560</td>
<td>Socinski et al., 2008</td>
</tr>
<tr>
<td></td>
<td>KDR (#50)</td>
<td>2.100</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KIT (#30)</td>
<td>2.440</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FLT4 (#34)</td>
<td>1.526</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PDGFRB (#69)</td>
<td>1.937</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PDGFRA (#149)</td>
<td>1.492</td>
<td></td>
</tr>
<tr>
<td>Cetuximab (E)*</td>
<td>FCGR2A (#252)</td>
<td>1.965</td>
<td>Hanna et al., 2006</td>
</tr>
<tr>
<td></td>
<td>FCGR3A (#291)</td>
<td>2.072</td>
<td>Horn and Sandler, 2009</td>
</tr>
<tr>
<td></td>
<td>C1QA (#747)</td>
<td>1.890</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C1QB (#1567)</td>
<td>1.825</td>
<td></td>
</tr>
<tr>
<td>Trastuzumab (A)*</td>
<td>FCGR2A (#252)</td>
<td>1.965</td>
<td>Ferrone and Motl, 2003</td>
</tr>
<tr>
<td></td>
<td>FCGR3A (#291)</td>
<td>2.072</td>
<td>Cappuzzo et al., 2006</td>
</tr>
<tr>
<td></td>
<td>C1QA (#747)</td>
<td>1.890</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C1QB (#1567)</td>
<td>1.825</td>
<td></td>
</tr>
<tr>
<td>Dasatinib (A)*</td>
<td>FYN (#5)</td>
<td>1.691</td>
<td>Haura et al., 2010</td>
</tr>
<tr>
<td></td>
<td>ABL1 (#1)</td>
<td>1.524</td>
<td>Li et al., 2010</td>
</tr>
<tr>
<td></td>
<td>KIT (#30)</td>
<td>2.440</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PDGFRB (#69)</td>
<td>1.937</td>
<td></td>
</tr>
<tr>
<td></td>
<td>STAT5B (#212)</td>
<td>1.581</td>
<td></td>
</tr>
<tr>
<td>Sulindac (A)*</td>
<td>MAPK3 (#55)</td>
<td>1.618</td>
<td>Attia et al., 2008</td>
</tr>
<tr>
<td></td>
<td>PTGS1 (#2914)</td>
<td>1.841</td>
<td>Jin et al., 2008</td>
</tr>
<tr>
<td></td>
<td>PTGS2 (#1213)</td>
<td>2.903</td>
<td></td>
</tr>
<tr>
<td>Staurosporine (UCN-01) (E)</td>
<td>PRKCQ (#2)</td>
<td>1.409</td>
<td>Edelman et al., 2007</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Wang et al., 2009</td>
</tr>
</tbody>
</table>


Diesing, D.et al. (2006) Vitamin D–metabolism in the human breast cancer cell line
Ding, L. et al. (2008) Somatic mutations affect key pathways in lung adenocarcinoma,
Nature 455 (7216):1069-75.
Edelman, M. J. et al. (2007) Phase I and pharmacokinetic study of 7-
hydroxystaurosporine and carboplatin in advanced solid tumors. Clin Cancer Res.
13(9): 2667-74.
Eden, E. et al. (2009) GOrilla: A Tool for discovery and visualization of enriched GO
Efimova, T. et al. (2004) Protein kinase Cdelta regulates keratinocyte death and survival
by regulating activity and subcellular localization of a p38delta-extracellular signal-
Esteva, F. J. et al. (2009) Molecular predictors of response to trastuzumab and lapatinib in
mass spectrometry. Mol Syst Biol. 3: 89.
Fakih, M. G. et al. (2009) A phase I, pharmacokinetic and pharmacodynamic study on
vorinostat in combination with 5-fluorouracil, leucovorin, and oxaliplatin in patients
Ferretti, V. et al. (2007) PReMod: a database of genome-wide mammalian cis-regulatory
Firestein, R. et al. (2008) CDK8 is a colorectal cancer oncogene that regulates beta-
Fornier, M. N. et al. (2007) Phase I dose-finding study of weekly docetaxel followed by
Fornier, M.N. et al. (2011) A phase I study of dasatinib and weekly paclitaxel for
Frampton, J. E. (2009) Lapatinib: a review of its use in the treatment of HER2-
overexpressing, trastuzumab-refractory, advanced or metastatic breast cancer. Drugs.
application for prioritizing positional candidate genes. Am J Hum Genet. 78(6):1011-
25.
Freeman, L. C. (1977) A set of measures of centrality based on betweenness. Sociometry
40: 35–41.
Networks 1: 215-239.


Sjöblom, T et al. (2006) The consensus coding sequences of human breast and colorectal