D960–D967 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 15 October 2010*
doi:10.1093/nar/gkq910

CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge

Jessica Ahmed, Thomas Meinel, Mathias Dunkel, Manuela S. Murgueitio, Robert Adams, Corinna Blasse, Andreas Eckert, Saskia Preissner and Robert Preissner*

Charité – University Medicine Berlin, Institute for Physiology, Structural Bioinformatics Group, Thielallee 71, 14195 Berlin, Germany

Received August 12, 2010; Revised September 21, 2010; Accepted September 23, 2010

ABSTRACT

During the development of methods for cancer diagnosis and treatment, a vast amount of information is generated. Novel cancer target proteins have been identified and many compounds that activate or inhibit cancer-relevant target genes have been developed. This knowledge is based on an immense number of experimentally validated compound–target interactions in the literature, and excerpts from literature text mining are spread over numerous data sources. Our own analysis shows that the overlap between important existing repositories such as Comparative Toxicogenomics Database (CTD), Therapeutic Target Database (TTD), Pharmacogenomics Knowledge Base (PharmGKB) and DrugBank as well as between our own literature mining for cancer-annotated entries is surprisingly small. In order to provide an easy overview of interaction data, it is essential to integrate this information into a single, comprehensive data repository. Here, we present CancerResource, a database that integrates cancer-relevant relationships of compounds and targets from (i) our own literature mining and (ii) external resources complemented with (iii) essential experimental and supporting information on genes and cellular effects. In order to facilitate an overview of existing and supporting information, a series of novel information connections have been established. CancerResource addresses the spectrum of research on compound–target interactions in natural sciences as well as in individualized medicine; CancerResource is available at: http://bioinformatics.charite.de/cancerresource/.

INTRODUCTION

Drug–protein interactions, or more generally, compound–target interactions, are becoming increasingly available for several layers of information according to the different interests in biological, physical or pharmacological research. Consequently, a broad set of data resources have been established and it is therefore not easy for biological, chemical or pharmaceutical scientists to deal with the often widespread and vast amounts of data. However, it is straightforward to use the capability of the Internet (1)—this includes up-to-date techniques like Web Services (2) to access existing repositories—for discovering compound–target interactions or determining the druggability of genes.

CancerResource addresses the complexity of cancer by covering not only a large but specific set of compound–target interactions, experimental data and supporting information but also by allowing individual data to be processed for advanced analyses. This article describes the database content and access to stored data together with the usage of provided tools and tool combinations toward workflows.

CANCER LITERATURE AND TEXT MINING

In the past three decades, huge effort was spent on research into cancer by an overwhelming number of

*To whom correspondence should be addressed. Tel: +49 30 450540755; Fax: +49 30 450 540 955; Email: robert.preissner@charite.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

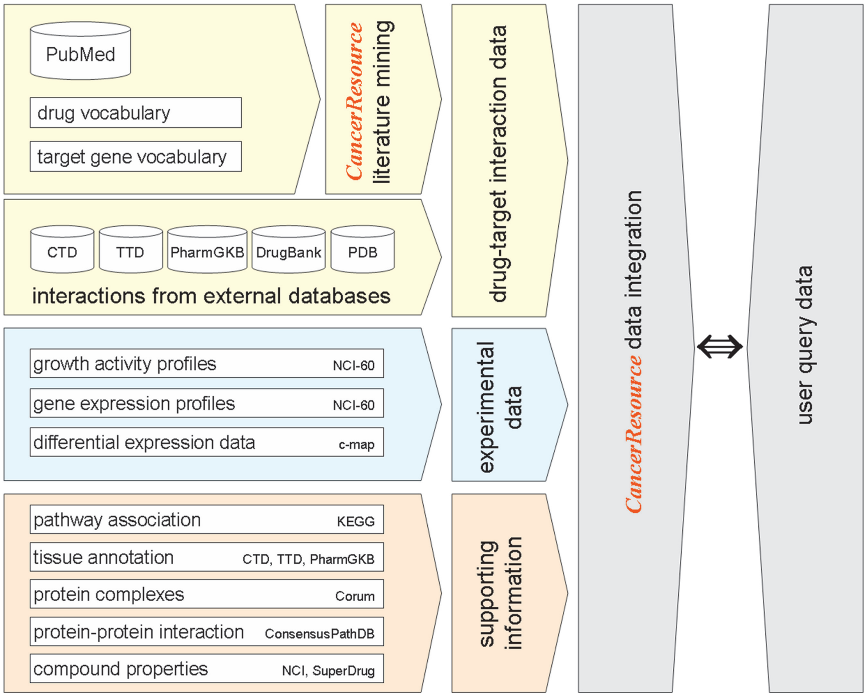
PubMed
drug vocabulary
target gene vocabulary
CancerResource literature mining
CTD
TTD
PharmGKB
DrugBank
PDB
interactions from external databases
drug-target interaction data
growth activity profiles NCI-60
gene expression profiles NCI-60
differential expression data c-map
experimental data
pathway association KEGG
tissue annotation CTD, TTD, PharmGKB
protein complexes Corum
protein-protein interaction ConsensusPathDB
compound properties NCI, SuperDrug
supporting information
CancerResource data integration
⇔
user query data

Figure 1. Data integration in CancerResource with three levels of data information: compound–target interaction data- PubMed (14), CTD (8), TTD (10), PharmGKB (9), DrugBank (11), PDB (18) -, experimental data - DTP at NCI (12), Connectivity Map (31) - and supporting information. - KEGG (29), CORUM (24), ConsensusPathDB (23), SuperDrug (15). CancerResource can be explored with queries or user-defined external data.

single studies. Literature on genetic disorders in cancer is extracted and made available in the Online Mendelian Inheritance in Man (OMIM) (3) database. Multiple data collections have arisen from the available repertoire of knowledge on cancer by text mining. They are often specialized like the Catalogue Of Somatic Mutations in Cancer (COSMIC) (4), a web resource on mutations in cancer genes that are detected in somatic tissues but also in cultured tissue samples. Cancer-relevant genes have been intensively studied and a fundamental model of cancer was established by Hanahan and Weinberg (5). On the other hand, the druggable genome (6) is, independent from the kind of disease, a set of proteins that are regarded as possible drug target candidates. Genome-scale targeting was identified originally by literature mining but has been successively developed by adding other information resources (7). The overlap between both perspectives, cancer-specific genes and druggable genome entities, forms the theoretical background of the CancerResource approach to the target genes; practically, target genes are derived from literature mining approaches.

Existing repositories like the Comparative Toxicogenomics Database (CTD) (8), the Pharmacogenomics Knowledge Base (PharmGKB) (9), the Therapeutic Target Database (TTD) (10) and the DrugBank (11) provide rich information on interactions of drugs (or drug-like compounds) with target genes or proteins. After inspecting cancer-relevant compound–target interactions, we found, surprisingly, that the data sets of these resources are more or less disjunct (Table 1) even when the results of the CancerResource literature mining are considered. This analysis indicated that there is need for integrating compound–target interactions from external data sets into one source and hence stimulated the creation of the CancerResource.

Table 1. Numbers of known interactions in external databases and from the CancerResource literature text mining

Resource	References	Numbers of compound–target relationships		Uniqueness
		Redundant	Unique	Degree (%)
External databases				
CTD	(8)	3875	3748	96
PharmGKB	(9)	1307	1158	88
TTD	(10)	282	163	58
DrugBank	(11)	4949	4763	96
CancerResource				
Literature mining	(this article)	1122	992	88
CancerResource				
Full data integration	(this article)	11 585	10 824	93

Data from CTD, PharmGKB and TTD are filtered according to cancer-related disease annotations, data for DrugBank are unfiltered. Relationships unique to each approach include the CancerResource literature mining result. The full integration result is presented additionally. The degree of uniqueness reveals that the data sets are more or less disjunct.

Cancer is often studied using somatic tissues, which are cultured for research as tissue samples of various cancer types and established as human standard cell lines. This inhomogeneous spectrum of cancers is well characterized and analyzed in large experimental studies investigating gene expression or cell growth activity under the influence of chemicals (12). This compound set of the National

Cancer Institute (NCI) is a rich resource for knowledge and research on gene dysfunctions in cancer. A data integration tool like CancerResource demands extended functionality. It is obvious that, similar to other compound–target interaction resources presented in a toxicological perspective (13), additional data such as experimental results and further supporting information enhance the knowledge of interactions together with features like: relationship of genes in pathways, druggability of the genes in the interactome, capability for user-defined data analyses and data mining and curation.

DATA INTEGRATION PROCEDURES AND METHODS

Compound–target gene interactions: CancerResource text mining

Compound–target relationships were automatically detected by own literature text mining over 19 million PubMed (14) abstracts using our vocabularies for drugs and targets. The drug vocabulary was generated from compounds having a cancer-related classification with respect to the Anatomical Therapeutic Chemical (ATC) Classification system via SuperDrug (15) or if the compound and its synonymous name are in the NCI compound set. The cancer relationship of a gene was determined from annotations in cancer-related pathways (see sub-section 'KEGG pathways') and the Gene Ontology (GO) (16). Abstracts, titles and Medical Subject Headings (MeSH) terms were converted into a text index using the LingPipe (http://alias-i.com/lingpipe/index.html) and the Lucene software packages (17). Both vocabularies were searched against each indexed abstract and the result was scored by an own rule-based validation algorithm. After this automatic procedure and a subsequent ranking revealing about 8000 publications, a manual revision of the hits followed resulting in about 900 highly significant publications of direct interactions.

Compound–target gene interactions: more data

Important interaction resources are integrated in CancerResource: CTD, TTD and PharmGKB. Sub-sets of cancer-specific interactions are filtered out according to the cancer vocabulary that is inherent in the three resources. The cancer-specific vocabulary is searchable and consists of more than 400 redundant cancer expressions. These are grouped into about 30 (mostly tissue-related) categories. To explore the impact of a particular drug on genes that are not just connected with cancer we integrated cancer-unspecific information on interactions provided by DrugBank (11). For ligands that are entries in the NCI compound set ligand–protein interactions from the Protein Data Bank (PDB) (18) were integrated into CancerResource.

PubMed references are extracted for identified compound–target relations to be cited in the web interface (if available; otherwise the relation is referenced by linking to the data resource by the resource's identifier).

(Drug-like) compounds and target genes

Core information of compounds and drugs was collected from different databases like the Developmental Therapeutics Program (DTP) at the NCI (12), PubChem (19), SuperTarget (20) and SuperDrug (15). CancerResource contains more than 40 000 cancer-relevant compounds.

The current set of target genes or proteins with cancer relevance are confirmed by the own text mining and complemented by genes extracted from existing interaction databases. The drug association is generally given (and searchable) at gene level and, if available, additionally at protein level. Core information on proteins and genes is based on UniProt (21) and Ensembl (22). Supporting information on cancer-relevant genes or proteins are provided by ten of thousands protein–protein interactions from ConsensusPathDB (23), affiliation of proteins to more than a thousand protein complexes from the Comprehensive Resource of Mammalian protein complexes (CORUM) (24); hundreds of gene mutations in NCI-60 tissue samples from COSMIC (4) and information by Web Service requests or virtual data links to iHOP, Reactome, Pfam and SYSTERS (25–28), see also Supplementary Table S1.

KEGG pathways

To put compound–target relations into a cellular context, we analyzed KEGG (signaling) pathways (29) according to their relevance in cancer emergence and cancer development. Forty four KEGG pathways were integrated into the CancerResource environment. This set comprises cancer-specific pathways, pathways related to cell-cycle regulation, replication, immune response and drug metabolism. Pathway maps are dynamically retrieved via Web Service from KEGG facultative with highlighted expression data if gene expression is computed online before. KEGG genes were excerpted from the set of analyzed pathways and used in the gene vocabulary for the text mining.

Cancer cell lines

Sixty human cancer cell lines of the NCI (NCI-60 set) were selected with respect to the availability of expression data as well as data of changes in biological activity by compound treatment. (Human cancer cell lines and cancer types are described in the Supplementary Data.)

Biological activity profiles: cellular fingerprints

Biological activity profiles indicate the influence of compounds on the growth rates of human cancer cell lines, wherein a GI-50 value indicates the compound concentration that induces 50% growth inhibition after treatment. More than 40 000 biological activity profiles are obtained for each compound. All activity profiles are translated into cellular fingerprints which allow the fast computation (30) of profile differences.

Gene expression data

Expression data of NCI-60 cancer cell lines were retrieved from DTP at the NCI and re-calculated to be comparable to external data sets in three steps (see Supplementary Data). Over the whole microarray data set (Affymetrix U133A chips), we introduced (i) the median normalization on Affymetrix probe set expression and (ii) compared normalized expression values of each probe set across NCI-60 tissue samples by introducing the relative abundance over all 60 cancer cell lines. Expression intensities of probe sets are ignored if they are associated with multiple genes. For each gene that is, according to Ensembl, associated with multiple probe sets (iii) the average of respective expression intensities is calculated.

Differential expression of genes after treatment

The Connectivity Map (31) provides differential expression data for five human cancer cell lines from the NCI-60 set before and after treatment with more than 200 compounds. Data correspond to the Gene Expression Omnibus (GEO) (32) data set GSE5258 and ratios are retrieved by Web Service from the GenomeMatrix repository (33), see also a detailed description in (34).

RESULTS

Currently, CancerResource comprises more than 10 800 non-redundant compound–target relations. More than 6000 (56%) are associated with cancer and over 4700 relations from DrugBank that do not have a disease specification. However, integration of DrugBank data enables high-quality searches for alternative targeting, which is, in the context of pharmacogenomic research, also known as drug repositioning. The CancerResource literature text mining revealed 992 new compound–target interactions (Table 1), which are ~16% of the cancer-related drug–target interactions or ~10% of all unique interactions in CancerResource. This ostensibly low number is owing to mining abstract texts only. Even after integration of our text mining results, the degree of uniqueness for the CancerResource is still ~90%, which indicates that all four text mining strategies with focus on cancer are obviously different to each other. In the whole CancerResource interaction data set, 2392 cancer-related target genes from CancerResource text mining, CTD, PharmGKB and TTD and additionally 995 genes from DrugBank cover 30% of the druggable genome (7); additionally 728 cancer-related genes not present in the druggable genome are found having compound–target interactions. (More issues and numbers on integrated data can be inspected in Supplementary Table S2.)

The integration of the set of more than 40 000 NCI compounds, dedicated as experimental drugs, extends the set of Food and Drug Administration (FDA) approved drugs by a factor of 100. It enriches CancerResource as an information resource for better understanding cancer and its treatment with drugs with a huge experimental background.

MUTUAL ACCESS TO COMPOUNDS AND GENES

CancerResource provides the referencing to interaction literature by links to citations in PubMed. In the web interface, such relations can be accessed by a drug, a target or a cancer feature; each of the three subjects can be used to query the web tool. Both molecular instances, target genes and drugs, can be mutually accessed (see Figure 2a). Respective web pages are organized into three parts that describe in detail (i) the relevance of a drug or a gene to cancer, (ii) compound–target interactions and (iii) supporting information.

At several sites in the web tool, interaction matrices of compounds and target genes provide information on single drugs targeting multiple genes (ambiguity) as well as multiple drugs targeting a single gene (redundancy). Such information on alternative targeting, which is helpful for the potential repositioning of compounds, is amplified through the integration of non-specific interaction information by DrugBank entries.

ACCESS TO EXPERIMENTAL DATA

Compound–target interaction information is more valuable by integrating experimental and supporting information. Therefore, CancerResource provides experimental data in addition to the information on interactions. Thereby, data stored in CancerResource can be compared with the user's own data. Several ways for accessing the web tool are presented in this section.

Similarity of compounds by structure or biological activity

The influence of a compound on the growth of cancer cell lines is a frequently used approach for the characterization and development of drugs. The biological activity of two compounds across the 60 NCI cell lines can be compared by a similarity measure, the Tanimoto coefficient of cellular fingerprints (30); this comparison by biological characteristics of a compound is a strong feature of the CancerResource web tool that complements the comparison of compounds by 2D structures. Here, the Tanimoto coefficient of structural fingerprints (35) enables the comparison of 2D similarities independently from the biological activity of a compound. CancerResource suggests thereby substitutability, alternative compound applications and support thereby drug research and drug treatment. Similar compounds are searchable by a given activity profile or the profile of a particular compound (query options are given in Figure 2b). Moreover, activity profiles can be found for a given compound structure.

Most active drugs against a cancer cell line

Alternatively to the compound characteristic defined by all cancer cell lines CancerResource enables the searching for compounds that are most biologically active against a single cell line (second part in Figure 3). In clinical medicine, one of the most successful approach to treat cancer is the growth inhibition of the cancer tissue. Therefore, CancerResource implies a module to find

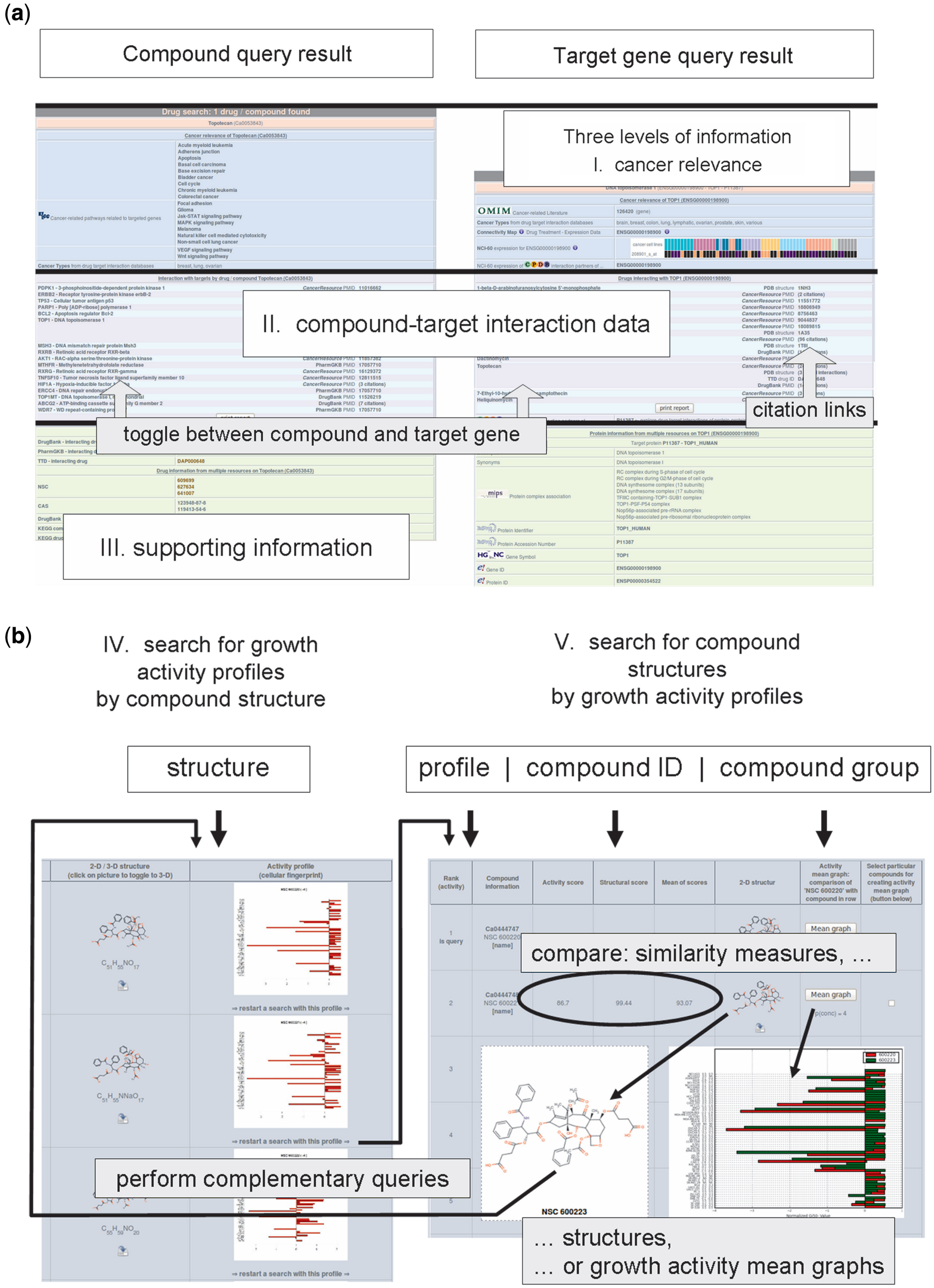
(a)
Compound query result
Target gene query result
Drug search: 1 drug / compound found
Three levels of information
I. cancer relevance
II. compound-target interaction data
citation links
toggle between compound and target gene
III. supporting information
(b)
IV. search for growth activity profiles by compound structure
V. search for compound structures by growth activity profiles
structure
profile | compound ID | compound group
compare: similarity measures, …
perform complementary queries
… structures,
… or growth activity mean graphs

Figure 2. Knowledge retrieval in CancerResource: (**a**) Access to detailed compound/drug and target gene information in CancerResource. A similar layout for both information layers, compounds (left) and target genes (right), comprises three information sections: (i) cancer relevance of a target gene or a compound with KEGG cancer pathways, involved somatic cancer types, information on expression across cancer cell lines; (ii) information on interactions with a toggle option between compound and target gene, source of information and link to the original literature source; (iii) specific compound or gene information. (**b**) Access to complementary information on growth activity across NCI-60 human cancer cell lines and structures of acting compounds. A search by compound structures (iv) reveals similar structures and associated growth activity profiles. The search by activity profiles (v) enables the user to compare structure formulas, activity profiles (pairwise mean graphs) and similarity measures for both growth activity and structures. Complementary queries can be performed by structures after downloading or by implemented links for a profile.

most effective compounds (that are inducing highest inhibition) against a single cell line.

Gene expression data of NCI-60 cancer cell lines

In CancerResource, gene expression data are available for about 4000 genes (see 'Results' section) and 60 NCI-60 cancer cell lines in both dimensions: genes are described and can be compared by expression profiles, the arrays of expression values across cell lines; NCI-60 cell lines are described and can be compared by a profile across genes. Relative abundances (data are calculated online if external data are uploaded) are displayed in the web tool by an array of colored boxes, each corresponding to a single gene. The blue/black/yellow color scheme is used for lower/non-significant/higher expression relatively to the average across all cell lines.

Several entry points for expression data with respect to genes are enabled: genes are searchable by the affiliation of genes to KEGG pathways, affiliation to protein–protein interaction data and for genes with low or high relative abundance in a couple of cancer cell lines. Resulting expression profiles over all 60 cell lines are characteristics for genes. They can be ranked by similarity (Pearson's correlation) if a gene is selected as center; protein–protein interactions and expression profile similarity are combined features here.

Furthermore, the NCI-60 cell line closest to a user-defined expression set (chip experiment; expression data sets are compared by Pearson correlation) can be searched both for genes (or probe sets) of a whole microarray or for selected probe sets or genes.

Differential gene expression

CancerResource allows the genome-wide online validation of two microarray chip experiments by computation of differential expression via ratios. Either external data are compared to a NCI-60 cell line or two external data sets can be compared to each other. Normalization for a subset of genes is regarded as a positive selection feature. It is enabled in CancerResource, which hence supports tumor/normal tissue comparisons or drug-treatment/control experiments. Alternatively, pre-calculated ratios associated with Ensembl gene IDs can be uploaded to enable the import of results from other experiment types (e.g. data collected using other micro-array platforms, next generation sequencing, protein chips, etc).

Ratios for differential expression are displayed in the web tool by the green/black/red color scheme (down/non-significant/up). Arrays of colored boxes are arranged according to the affiliation of respective genes to chromosomes or KEGG pathways. For the latter, differentially expressed genes are analyzed in order to estimate the over-representation in a pathway. This is calculated by a *P*-value using the hypergeometric function and distribution, see details in (36).

Connectivity map

The Connectivity Map (31) was intended to aid the discovery of functional connections among diseases, genetic perturbation and drug action. The influence of more than 200 compounds on differential gene expression was determined for the whole genome of five cancer cell lines. Two query options in CancerResource provide access to expression profiles (i) for the influences of the set of compounds in the five cell lines on a single gene and (ii) for the influence of a single compound on all genes in a single cell line. The visualization, again by arrays of colored boxes, is restricted to target genes that possess interactions integrated in CancerResource.

Direct and indirect knowledge on compound–target interaction

In the Connectivity Map data set, the influence of a row of compounds on genes is experimentally studied by differential expression, which is indirect knowledge about gene targeting (but no about cause-and-effect relationships). The simultaneous comparison (Supplementary Figure S1) with compound–target interactions from the literature mining ('direct knowledge') facilitates considerations about druggability and targeting of genes.

PROPOSED WORKFLOWS

CancerResource facilitates complex searches by the implementation of several ways of accessing the data. Two workflows are demonstrating suggested research use cases.

Finding alternative, most effective drugs for a (somatic) tissue similar to a cancer cell line

An external tissue sample can be identified as most similar to a single NCI-60 cell line by expression profiles across genes or probe sets. Figure 3 explains how the most similar ('best') NCI-60 cell line can be determined with differentially expressed genes by calculation of Pearson correlations between the upload data and all 60 tissues samples (see above). In the next step, the most effective drugs will be determined for this cell line (which is basing on the growth inhibition of a compound is measured for all NCI-60 cancer cell lines and is described above). Finally, for the identified compounds the tool displays the genes they target including the alternative targeting.

Finding alternative compound–target gene interactions for differentially expressed genes via pathway information

KEGG (signaling) pathways elucidate the context of genes according to functionality. To visualize the differential regulation of genes in a pathway, colored pathway maps are dynamically generated in CancerResource. The workflow starts with the loading of expression data (Supplementary Figure S2), which is possible in multiple forms. The data are re-calculated and displayed as an array of colored boxes for each KEGG pathway; overrepresentation analyses are available for each pathway and for both up- and down-regulated genes; the pathway map generation can be started from here to display integrated expression of genes, either for a single gene or for all genes in the pathway. Finally, drug information is available (via the pathway map) and,

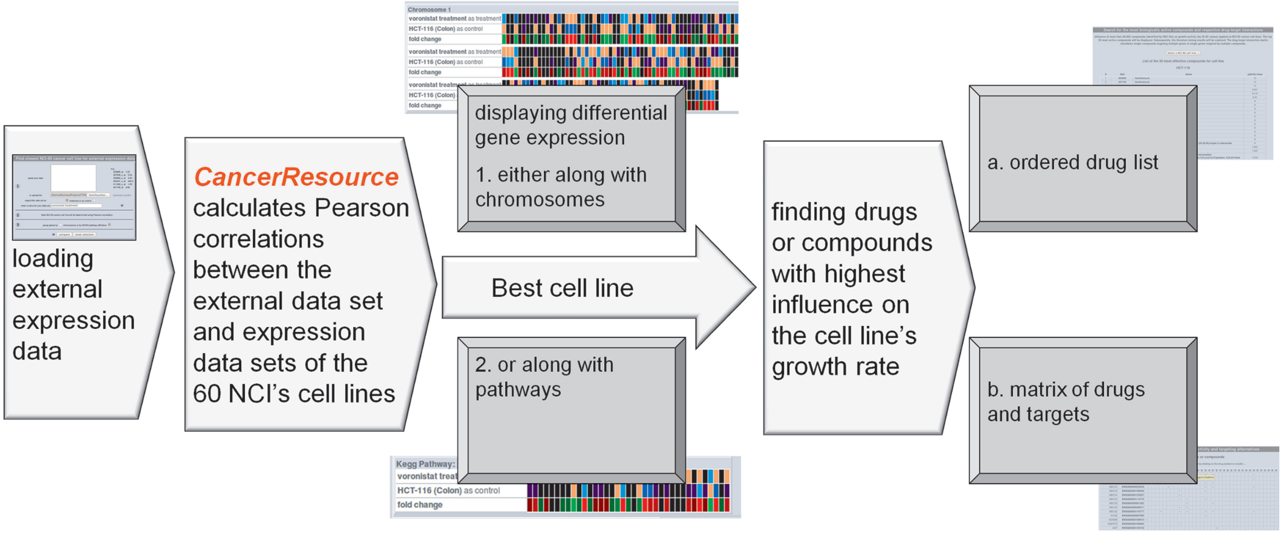
loading external expression data
CancerResource calculates Pearson correlations between the external data set and expression data sets of the 60 NCI's cell lines
displaying differential gene expression
1. either along with chromosomes
Best cell line
2. or along with pathways
finding drugs or compounds with highest influence on the cell line's growth rate
a. ordered drug list
b. matrix of drugs and targets
Chromosome 1
voronistat treatment as treatment
HCT-116 (Colon) as control
fold change
Kegg Pathway:
fold change

Figure 3. Finding the most similar cell line from the NCI-60 set and, subsequently, compounds or drugs having the highest influence on that cell line (this workflow includes two user interactions).

subsequently, the compound–target matrix for alternative targeting. The integration of dynamically assigned pathway maps makes CancerResource into a systems biology approach.

CONCLUSIONS AND FUTURE DIRECTIONS

The feedback by many scientists shows that there is a need for specialized resources that not only cover a specific set of interaction data but also deliver tools that are specialized for the further analysis of the respective data set. CancerResource tries to cover both levels of scientific work, the support of scientists who try to develop novel drugs and the medic who is reliant on advice for the development of individualized therapy approaches.

Cancers, even of the same tissue type, are extremely divergent in terms of gene alterations. Individual therapy will be made possible by understanding single nucleotide polymorphisms (SNPs), complete or partial gene deletions, copy number variations, gene aberrations or gene fusions. All of those issues may cause substantial dysfunctions of defected genes that have influence on gene regulation in the whole cell of an individual. Additionally to those integration issues, new data integration concepts will be required or are planned to be integrated in CancerResource for coping with personalized therapies. The literature mining will be extended to full text mining, manual upload of single relationships and enhanced specificity in cancer annotations. Expression data will be comparable for platforms other than Affymetrix U133A. Large studies performed on the basis of new techniques (Next Generation Sequencing; e.g. Genetics of 1000 Tumors) are highly interesting objectives to be made available in CancerResource. Updating of data is projected to occur once a year.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors like to thank Björn Grüning for the development of the initial database and Ilona Gawronska and Ulrike Schmidt for manual validation of cancer-relevant drug–target relations.

FUNDING

The International Research Training Group IRTG (Genomics and Systems Biology of Molecular Networks) Deutsche Krebshilfe; Deutsche Forschungsgemeinschaft DFG; Federal Ministry of Education and Research BMBF (MedSys Project); European Union EU (SynSys Project). Funding for open access charge: DFG, BMBF, EU.

Conflict of interest statement. None declared.

REFERENCES

1. Wishart,D.S. (2007) Discovering drug targets through the web. *Comp. Biochem. Physiol. Part D Genomics Proteomics*, **2**, 9–17.
2. Meinel,T. and Herwig,R. (2010) SOAP/WSDL-based Web Services for Biomedicine. In Lazakidou,A. (ed.), *Web-Based Applications in Healthcare and Biomedicine*. Springer, New York, USA, pp. 101–116.
3. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–796.
4. Forbes,S.A., Tang,G., Bindal,N., Bamford,S., Dawson,E., Cole,C., Kok,C.Y., Jia,M., Ewing,R., Menzies,A. *et al.* (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res.*, **38**, D652–D657.
5. Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.

6. Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.*, **1**, 727–730.
7. Sophic Alliance. (2010) *The Integrated Druggable Genome Database*. Sophic Systems Alliance Inc, Rockville, MD, USA.
8. Davis,A.P., Murphy,C.G., Saraceni-Richards,C.A., Rosenstein,M.C., Wiegers,T.C. and Mattingly,C.J. (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res.*, **37**, D786–D792.
9. Hernandez-Boussard,T., Whirl-Carrillo,M., Hebert,J.M., Gong,L., Owen,R., Gong,M., Gor,W., Liu,F., Truong,C., Whaley,R. *et al.* (2008) The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res.*, **36**, D913–D918.
10. Zhu,F., Han,B., Kumar,P., Liu,X., Ma,X., Wei,X., Huang,L., Guo,Y., Han,L., Zheng,C. *et al.* (2010) Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.*, **38**, D787–D791.
11. Wishart,D.S., Knox,C., Guo,A.C., Cheng,D., Shrivastava,S., Tzur,D., Gautam,B. and Hassanali,M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.
12. Shoemaker,R.H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, **6**, 813–823.
13. Mattingly,C.J. (2009) Chemical databases for environmental health and clinical research. *Toxicol. Lett.*, **186**, 62–65.
14. McEntyre,J. and Lipman,D. (2001) PubMed: bridging the information gap. *Can. Med. Assoc. J.*, **164**, 1317–1319.
15. Goede,A., Dunkel,M., Mester,N., Frommel,C. and Preissner,R. (2005) SuperDrug: a conformational drug database. *Bioinformatics*, **21**, 1751–1753.
16. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
17. McCandless,M., Hatcher,E. and Gospodnetić,O. (2010) *Lucene in Action*, 2nd edn. Manning Publications Co, Greenwich, CT, USA.
18. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
19. Wang,Y., Bolton,E., Dracheva,S., Karapetyan,K., Shoemaker,B.A., Suzek,T.O., Wang,J., Xiao,J., Zhang,J. and Bryant,S.H. (2010) An overview of the PubChem BioAssay resource. *Nucleic Acids Res.*, **38**, D255–D266.
20. Gunther,S., Kuhn,M., Dunkel,M., Campillos,M., Senger,C., Petsalaki,E., Ahmed,J., Urdiales,E.G., Gewiess,A., Jensen,L.J. *et al.* (2008) SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.*, **36**, D919–D922.
21. UniProtConsortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
22. Flicek,P., Aken,B.L., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S. *et al.* (2010) Ensembl's 10th year. *Nucleic Acids Res.*, **38**, D557–D562.
23. Kamburov,A., Wierling,C., Lehrach,H. and Herwig,R. (2009) ConsensusPathDB: a database for integrating human functional interaction networks. *Nucleic Acids Res.*, **37**, D623–D628.
24. Ruepp,A., Brauner,B., Dunger-Kaltenbach,I., Frishman,G., Montrone,C., Stransky,M., Waegele,B., Schmidt,T., Doudieu,O.N., Stumpflen,V. *et al.* (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.*, **36**, D646–D650.
25. Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nat. Genet.*, **36**, 664.
26. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
27. Meinel,T., Krause,A., Luz,H., Vingron,M. and Staub,E. (2005) The SYSTERS Protein Family Database in 2005. *Nucleic Acids Res.*, **33**, D226–D229.
28. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
29. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
30. Fullbeck,M., Dunkel,M., Hossbach,J., Daniel,P.T. and Preissner,R. (2009) Cellular fingerprints: a novel approach using large-scale cancer cell line data for the identification of potential anticancer agents. *Chem. Biol. Drug Des.*, **74**, 439–448.
31. Lamb,J., Crawford,E.D., Peck,D., Modell,J.W., Blat,I.C., Wrobel,M.J., Lerner,J., Brunet,J.P., Subramanian,A., Ross,K.N. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.
32. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
33. Hewelt,A., Ben Kahla,A., Hennig,S., Nagel,A., Himmelbauer,H., Zehetner,G., Haas,S., Vingron,M., Yaspo,M.L. and Lehrach,H. (2002) *Human Genome Meeting HGM2002*. Shanghai, China, Abstract 23.
34. Meinel,T., Mueller,M.S., Ahmed,J., Yildirimman,R., Dunkel,M., Herwig,R. and Preissner,R. (2010) SOAP/WSDL-based web services for biomedicine: demonstrating the technique with the CancerResource. In Bamidis,P.D. and Pallikarakis,N. (eds), *XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010*. Springer, Berlin Heidelberg, Chalkidiki, Greece, pp. 835–838.
35. Thimm,M., Goede,A., Hougardy,S. and Preissner,R. (2004) Comparison of 2D similarity and 3D superposition. Application to searching a conformational drug database. *J. Chem. Inf. Comput. Sci.*, **44**, 1816–1822.
36. Masseroli,M. and Tagliasacchi,M. (2010) Web resources for gene list analysis in biomedicine. In Lazakidou,A. (ed.), *Web-Based Applications in Healthcare and Biomedicine*. Springer, New York, USA, pp. 117–141.