D968–D973 *Nucleic Acids Research, 2015, Vol. 43, Database issue* *Published online 11 November 2014*
doi: 10.1093/nar/gku1140

Cancer3D: understanding cancer mutations through protein structures

Eduard Porta-Pardo†, Thomas Hrabe† and Adam Godzik*

Bioinformatics and Systems Biology Program, Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA

Received August 14, 2014; Revised October 20, 2014; Accepted October 27, 2014

ABSTRACT

The new era of cancer genomics is providing us with extensive knowledge of mutations and other alterations in cancer. The Cancer3D database at http://www.cancer3d.org gives an open and user-friendly way to analyze cancer missense mutations in the context of structures of proteins in which they are found. The database also helps users analyze the distribution patterns of the mutations as well as their relationship to changes in drug activity through two algorithms: e-Driver and e-Drug. These algorithms use knowledge of modular structure of genes and proteins to separately study each region. This approach allows users to find novel candidate driver regions or drug biomarkers that cannot be found when similar analyses are done on the whole-gene level. The Cancer3D database provides access to the results of such analyses based on data from The Cancer Genome Atlas (TCGA) and the Cancer Cell Line Encyclopedia (CCLE). In addition, it displays mutations from over 14 700 proteins mapped to more than 24 300 structures from PDB. This helps users visualize the distribution of mutations and identify novel three-dimensional patterns in their distribution.

INTRODUCTION

Some of the most practical issues to explore in cancer research are the detection of cancer drivers and the identification of biomarkers that predict a patient's response to a drug (1–4). Recent publication of several large-scale cancer datasets, such as The Cancer Genome Atlas (TCGA) (5) or the Cancer Cell Line Encyclopedia (CCLE) (6), provides means to explore such questions from a genome-wide perspective in hundreds or thousands of samples. Various tools and databases have been developed to provide access and allow analysis of these data, such as UCSC's Cancer Genomics Browser (7), canSAR (8), cBioPortal (9) or COSMIC (10). Most of these databases, however, do not integrate genomic data such as missense mutations with protein-structure information. Even if they do, as in the case of canSAR, or cBioPortal they invariably focus either on whole-proteins or individual mutations. None of the existing databases use information regarding the different protein functional regions (PFRs) in a protein in their analyses. Integrating PFR annotations is important because gene-centric methods can confuse or dilute signals where only mutations at some specific positions within a protein are relevant (11–14), whereas analyses focusing on individual mutations usually lack statistical power. Last but not least, many databases do not allow the exploration of cancer drivers and biomarkers at the same time.

Cancer3D integrates data from TCGA and CCLE and allows users to explore the biomarker and driver problems at the same time through two novel algorithms: e-Driver (13) and e-Drug (15). These algorithms are unique in using information about the modular structure of a protein to predict novel cancer drivers or drug biomarkers, respectively. Statistics are calculated separately for each region in each protein, including known PFAM domains (16), predicted intrinsically disordered regions and over 1300 potential novel domains in the human proteome detected by AIDA (17). Another important feature of Cancer3D is that it maps somatic missense mutations from over 18 000 human proteins to, wherever available, experimental or predicted protein three-dimensional structures. The Cancer3D database not only displays the mutated positions of a protein in their corresponding structures, but also conveys important information such as drug activity or mutation frequency by color-coding. This helps users to quickly identify three-dimensional patterns in data such as clustering of mutation hotspots around particular regions (Figure 1).

MATERIALS AND METHODS

Data sources

Mutation data used in Cancer3D come from the CCLE and TCGA pancancer analysis projects. In both cases, we used the Variant Effect Predictor Tool to map mutations from genomic coordinates to all ENSEMBL protein isoforms (18).

*To whom correspondence should be addressed. Tel: +1 858 646 3168; Fax: + 1 858 795 5249; Email: adam@godziklab.org
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors

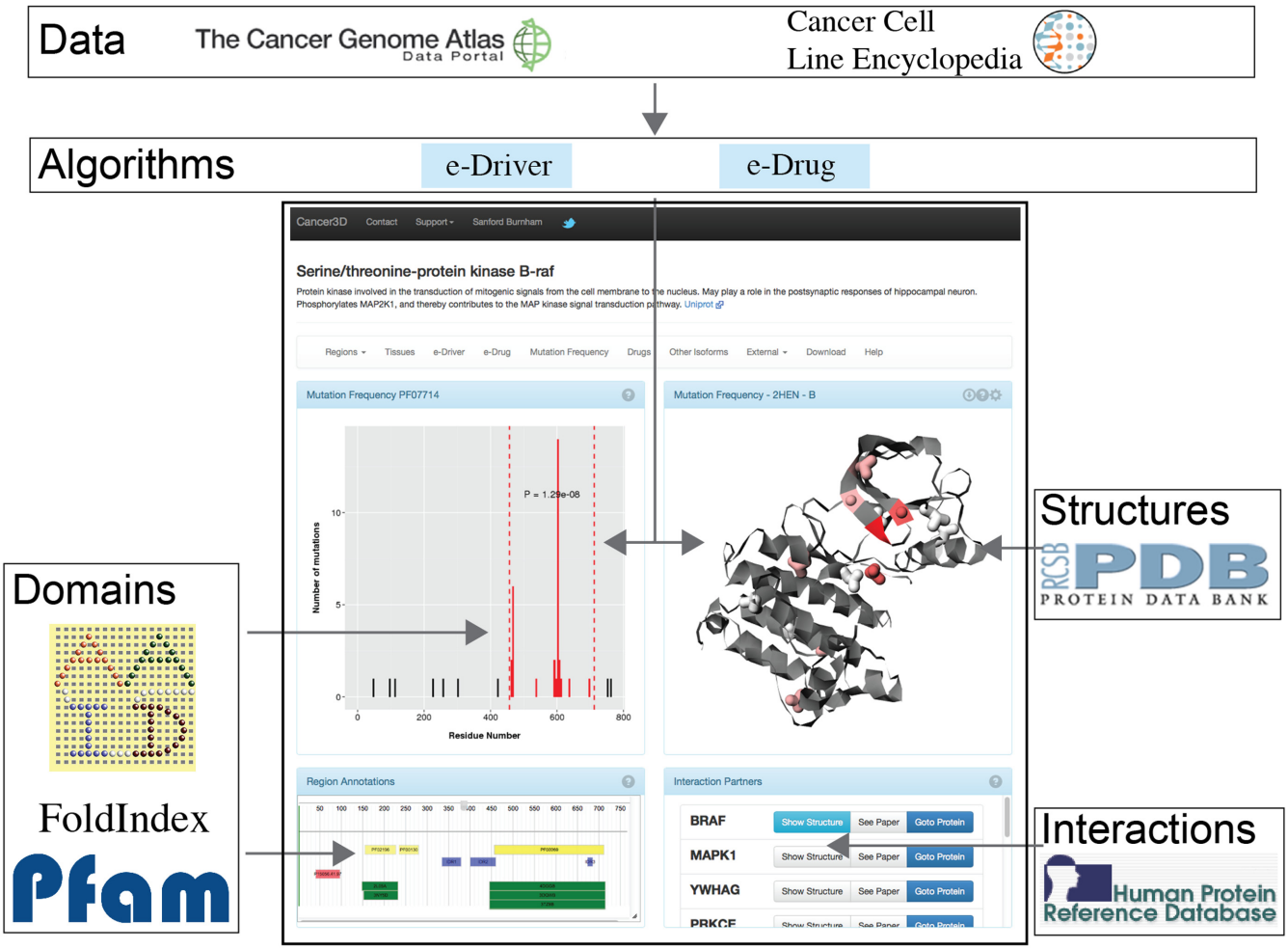

Figure 1. Database sources, content and main view. The database allows users to simultaneously access two types of cancer data: mutation frequency (from TCGA) and pharmacogenomic profiles (from CCLE). When a user queries the database with a protein name, Cancer3D retrieves these data and analyzes them using e-Driver and e-Drug, respectively. The user can also view where the mutations are located in different structures from PDB and navigate through the different protein regions and structures using the protein viewer. Finally, Cancer3D also provides information on which proteins are interacting with the query according to Human Protein Reference Database (HPRD), allowing users to either go to references describing the interaction or to query Cancer3D with those proteins.

We have a total of 275 648 mutations mapped onto 103 918 protein isoforms that belong to 23 226 unique genes.

Protein domains were assigned using Pfam HMM models as retrieved from ENSEMBL through its application programming interface (API), whereas intrinsically disordered regions were predicted using Foldindex (19) (we used regions with a score below – 0.1). The coordinates of novel protein domains were identified by the AIDA server (17) and curated manually. Overall results for 272 188 individual regions in the human proteome are currently stored in Cancer3D: 156 147 Pfam domain annotations, 103 986 intrinsically disordered regions and 12 055 novel domains. Note that these numbers include all annotations for all isoforms of a gene. For example, since Epidermal Growth Factor Receptor (EGFR) has three different isoforms with an instance of PF00069 (protein kinase domain), each of them will be counted separately. We have also included a list of protein–protein interactions from HPRD (20) in order to allow users to explore not only their favorite protein, but also the protein's interaction partners.

Drug activity data for the 24 anticancer drugs were downloaded from the CCLE and calculated using eight-point dose–response curves in 479 different cell lines. These curves are adjusted to a logistical-sigmoidal function and described by four different variables: (i) the maximal effect level (A_{max}), (ii) the drug concentration at half-maximal activity of the compound (EC_{50}), (iii) the concentration at which the drug response reached an absolute inhibition of 50% (IC_{50}) (iv) and the activity area that is the area above the dose–response curve. In order to simplify the analysis in Cancer3D, we used values regarding the activity area only. According to the CCLE, the activity area captures simultaneously both variables of drug activity: its efficacy and its potency.

Algorithms

We have previously developed two algorithms called e-Driver (13) and e-Drug (15) that allow the identification of candidate cancer driver genes and drug biomarkers, respectively. The main novel feature of both algorithms is that instead of analyzing data in terms of genes, they focus on regions corresponding to Pfam domains or intrinsically disordered regions. Similar approaches have been previously used in the context of mutations associated with Mendelian disorders (11,12). Users can find detailed explanations of the algorithms on Cancer3D's help page and in the individual manuscripts.

Structure mapping

We used BLAST (21) to match sequences of three-dimensional structures to all our domains. The full PDB (March 2014 including non-human proteins) was queried for each protein sequence in Cancer3D. PDB structures were assigned to each protein that had an *e*-value below 1e-6. The BLAST output was used to map mutated positions onto the structures.

USING THE DATABASE

In the following paragraphs, we will present a use-case scenario with all input and views based on the BRAF protein. This is a well-known oncogene for which a drug targeting the V600E mutation has been recently approved for metastatic melanoma treatment in the USA and in Europe in 2011 and 2012, respectively (22). The drug, which goes under the commercial name of Vemurafenib, is part of the CCLE under the synonym PLX4720.

Input and isoform selection

The start page contains one input field to search for the user's protein of interest. The user can input a gene or a protein name, a Uniprot ID, or an ENSEMBL Gene/Protein ID. If the name is recognized by the database, a list of protein isoforms identified by their ENSEMBL protein ID will be listed. Information displayed also includes the respective sequence length and the number of mutations in both TCGA and CCLE. Isoforms are sorted by their individual sequence lengths, and the longest isoform is highlighted for quick selection. The user also has an option to select either e-Driver or e-Drug results for the next page.

In our use-case scenario, the user would specify BRAF in the input menu and select the first isoform (ENSEMBL protein id: ENSP00000288602) and e-Driver. For the user interested in just browsing and getting familiar with the database, we provide a link to a 'protein of the month' that leads to a preselected protein.

After the initial selections, the user is forwarded to the main page (Figure 2). The main page can essentially have two views: (i) the e-Driver or (ii) the e-Drug view. Switching between these views is possible by selecting a region or region–drug combination in the respective menus. The 'Mutation Frequency' or 'Drug' buttons in the menu bar allow the user to immediately switch between the e-Driver and e-Drug views for the currently selected region. The user can access view-specific help texts for the current view state by clicking on one of the question-mark symbols.

Main e-Driver view

After the user has chosen an isoform, the main analysis page will be displayed for the combination of the selected isoform and view. For the e-Driver view, we display properties of the region with the lowest *P*-value calculated by the e-Driver algorithm. The upper left view initially contains the '*Mutation Frequency*': a histogram plot depicting the mutation frequency for each residue where the current domain is highlighted in red (Figure 2). The upper right view displays a placeholder image for the three-dimensional model of the region (experimental or predicted). In order to view the interactive three-dimensional structure, the user has to click on the placeholder image. The structure displayed will have each residue colored according to its mutation frequency. Residues with higher mutation rates are colored in more-intense red tones, while positions with lower mutation rates are colored in less-intense red tones.

Viewing other regions is possible by selecting them either from the region viewer, where they are listed according to their occurrence in sequence, or from the e-Driver menu, where they are sorted according to their *P*-value.

For our BRAF use-case scenario, the user would see the '*Mutation Frequency*' displayed for the PF00069 domain and chain B of the PDB structure 2HEN with the mutation sites colored according to the mutation frequency. The algorithm identifies BRAF's kinase domain as a potential cancer driver domain because it is highly enriched in missense mutations in the TCGA dataset ($P < 1e{-}6$), particularly in the V600 position (shown in red in the structure).

Main e-Drug view

If the user enables the e-Drug option on the start page, the main page will load into the e-Drug view. The graph in the left view will display the '*Drug Activity Area*' for the region–drug combination with the lowest *P*-value. The region-specific drug activity region is highlighted in red. The structure on the right visualizes the spatial drug specificity for each residue, but must be enabled by clicking on the placeholder image first. All other region–drug combinations are listed in the e-Drug menu, sorted by their *P*-values. Selection of either one of these combinations will display the '*PFR-Drug Scatterplot*' in the top left view and the '*Drug Boxplot*' for the selected region and drug. Here, activity of the current drug is compared in three different groups of cell lines: cell lines with mutations in the selected region, cell lines with mutations in other regions of the protein and cell lines with no somatic missense mutations in any region of the current protein (shown as WT in the boxplot).

In order to investigate the region–drug combination with the lowest *P*-value in our BRAF use-case scenario, the user would select the PF07714 (tyrosine kinase) domain—PLX4720 (Vemurafenib) drug combination in the e-Drug menu. Selecting this combination would display the respective '*PFR-Drug Scatterplot*' and '*Drug Boxplot*' in the top views. Clicking the assigned PDB structure displays the latter with all mutation sites colored according to the drug

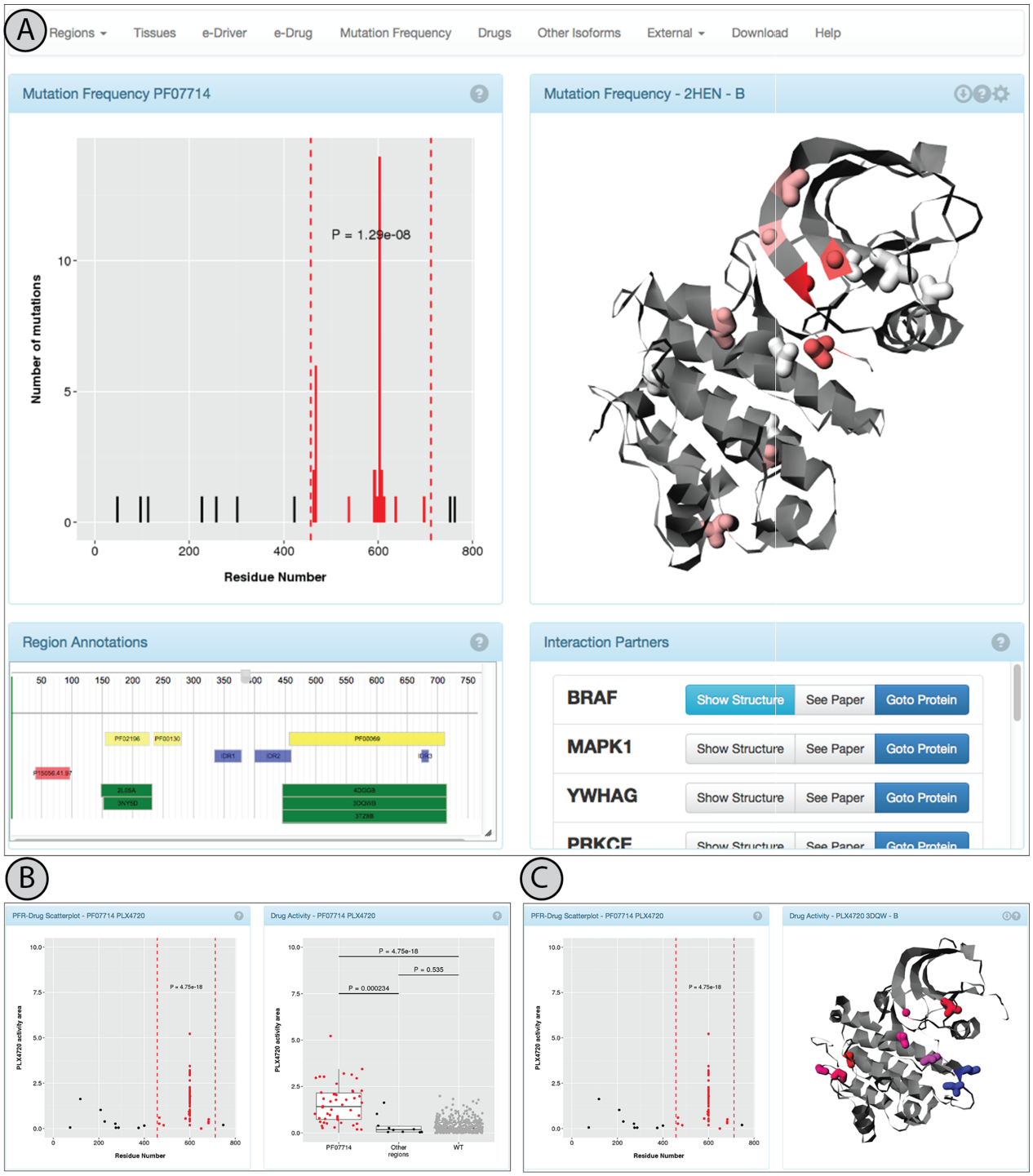

Figure 2. Different states of the main view. (**A**) Consistent with our use-case scenario of BRAF (ENSEMBL protein id: ENSP00000288602), the user would see this initial view. The upper left box contains the '*Mutation Frequency*' plot for the whole protein as detected by the e-Driver algorithm. The red region in the plot highlights the region with the lowest *P*-value (PF00069). The three-dimensional structure in the right box is the best homolog for the current domain. In this structure, mutated residues are highlighted according to the observed mutation frequency where white is the lowest and red is the highest observed mutation frequency. The '*Region Annotations*' box allows the user to select alternative regions: either PFAM domains (yellow), intrinsically disordered regions (blue) or newly annotated domains (red). The green boxes mark regions for which structures have been found. The interaction box lists all interacting proteins for the currently selected one and allows the user to either view the paper in which this particular interaction was described or investigate a particular interaction partner with Cancer3D. The first menu element, 'Regions', in the menu bar allows the user to select regions (similar to '*Region Annotations*'). e-Driver allows users to select regions sorted by their *P*-values. The e-Drug menu element allows the user to browse through region–drug combinations sorted according to their *P*-values detected by the e-Drug algorithm. (**B**) By clicking on one of the entries in the e-Drug menu, the two upper boxes will display the 'PFR-Drug Scatterplot' and the 'Drug boxplot' for the particular region–drug combination. Notably, the structure view disappears but can be reactivated by selecting the corresponding PDB domain in the '*Region Annotations*' view. (**C**) Now, all mutated residues are highlighted based on their drug activity, where red residues have low activity and blue residues have high activity.

activity. Mutations found in cell lines with lower drug activity are colored in red, whereas those found in cell lines with higher drug activity are shown in blue. When a mutation is found in multiple cell lines, the coloring reflects the average activity in such cells. In the case of the BRAF kinase—PLX4720 (Vemurafenib) pair from our use case (Figure 2)—one can observe in the Drug Scatterplot that the cell lines with mutations in BRAF that mostly respond to this drug are those with mutations in the V600 position. The mutation can also be quickly identified in the structure, as the V600 position appears clearly shown in blue.

Region viewer

We use a custom region viewer plugin to visualize the position of PFAM domains, intrinsically disordered regions and new hypothetical domains in the protein. These three types are displayed in yellow, blue and red boxes, respectively. Green boxes indicate regions matching or homologous to available protein structures. The user can click and select any region of interest, leading to the specific analysis (either for e-Driver or for e-Drug, depending on the current view) for that particular region.

Interaction viewer

In the bottom right area, we list all interaction partners for the protein under scrutiny. For each interaction, we provide the user with a link to the respective publication where the interaction between both proteins has been described. A link to all isoforms in the database for the interacting partner allows the user to continue investigating e-Driver and e-Drug results for their respective interaction partners. Finally, the 'Show structure' button allows the user to view a three-dimensional model of the interaction in the upper right window. The button is activated and highlighted in blue whenever a matching complex for the interacting proteins was found. Gray 'Show Structure' buttons indicate that no structures for the particular interaction were found.

Tissue selection

Cancer3D returns, by default, the results obtained using all the available data. While it increases the statistical power of the analysis, it hides tissue-specific results. For example, we have previously shown that the mutation patterns in EGFR are drastically different in the glioblastoma and lung adenocarcinoma datasets from TCGA (13). The user can focus on a specific cancer type by clicking on the 'Tissue' button and selecting the desired option.

Precomputed results

The 'Precomputed' menu provides direct links to results discussed in our previous publications where we first introduced the algorithms. These links aim at helping the user to find biologically relevant results identified by our algorithms.

Other pages

In order to help making the most of Cancer3D, we provide a series of pages with tutorials and a guideline explaining issues regarding *P*-values, multiple testing corrections and specific details of the algorithms and datasets that are part of Cancer3D. The user finds this page in the 'Support' menu by selecting then the '*P*-Value explained' entry.

CONCLUSIONS AND FUTURE DEVELOPMENTS

We developed Cancer3D to present and distribute results of our novel algorithms e-Driver and e-Drug in a user-friendly manner. These algorithms exploit the knowledge of the inner structure of proteins to detect novel drivers and drug biomarkers, respectively. The database also provides a means to explore somatic missense mutations from TCGA and CCLE mapped onto over 24 300 structures, as well as 1300 potential novel protein domains identified by the AIDA server.

Our ongoing research in this field will expand the content of this database, and we will continuously extend the server with results stemming from our future algorithms. In the short term, we plan to integrate other tools developed by our group into the database to expand the result spectrum on the web pages. For example, we are working on an interface to AIDA to allow users to generate models of their proteins directly from Cancer3D. This feature would enable three-dimensional visualization of mutations for structurally unknown domains. We will provide an API to allow developers to access our database programmatically. In order to extend our structural mappings, we plan to include mappings from FFAS (23).

Finally, we intend to expand the database by including other types of mutations relevant to cancer, such as copy-number variations (24) or synonymous mutations (25), as well as data from other cancer genomics projects, such as the International Cancer Genome Consortium (26) or the Genomics of Drug Sensitivity in Cancer (27).

ACKNOWLEDGEMENTS

The authors would like to thank Mayya Sedova for support integrating the Domain viewer and Cindy Cook and Andrew LeBlanc for help with editing.

FUNDING

Human Frontiers Science Program [RGP0027/2011]; National Institutes of Health [R01 GM101457] and Sanford-Burnham Medical Research Institute (SBMRI). Funding for open access charge: Human Frontiers Science Program [RGP0027/2011]; National Institutes of Health [R01 GM101457] and Sanford-Burnham Medical Research Institute (SBMRI).
Conflict of interest statement. None declared.

REFERENCES

1. Kandoth,C., McLellan,M.D., Vandin,F., Ye,K., Niu,B., Lu,C., Xie,M., Zhang,Q., McMichael,J.F., Wyczalkowski,M.A. *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333–339.

2. Kelloff,G.J. and Sigman,C.C. (2012) Cancer biomarkers: selecting the right drug for the right patient. *Nat. Rev. Drug Discov.*, **11**, 201–214.
3. Yuan,Y., Van Allen,E.M., Omberg,L., Wagle,N., Amin-Mansour,A., Sokolov,A., Byers,L.A., Xu,Y., Hess,K.R., Diao,L. *et al.* (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol.*, **32**, 644–652.
4. Valencia,A. and Hidalgo,M. (2012) Getting personalized cancer genome analysis into the clinic: the challenges in bioinformatics. *Genome Med.*, **4**, 61.
5. Cancer Genome Atlas Research, N., Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
6. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehar,J., Kryukov,G.V., Sonkin,D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
7. Cline,M.S., Craft,B., Swatloski,T., Goldman,M., Ma,S., Haussler,D. and Zhu,J. (2013) Exploring TCGA Pan-Cancer data at the UCSC Cancer Genomics Browser. *Sci. Rep.*, **3**, 2652.
8. Bulusu,K.C., Tym,J.E., Coker,E.A., Schierz,A.C. and Al-Lazikani,B. (2014) canSAR: updated cancer research and drug discovery knowledgebase. *Nucleic Acids Res.*, **42**, D1040–D1047.
9. Cerami,E., Gao,J., Dogrusoz,U., Gross,B.E., Sumer,S.O., Aksoy,B.A., Jacobsen,A., Byrne,C.J., Heuer,M.L., Larsson,E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, **2**, 401–404.
10. 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.
11. Zhong,Q., Simonis,N., Li,Q.R., Charloteaux,B., Heuze,F., Klitgord,N., Tam,S., Yu,H., Venkatesan,K., Mou,D. *et al.* (2009) Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.*, **5**, 321.
12. Wang,X., Wei,X., Thijssen,B., Das,J., Lipkin,S.M. and Yu,H. (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.*, **30**, 159–164.
13. Porta-Pardo,E. and Godzik,A. (2014) e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics*, **30**, 3109–3114.
14. Nehrt,N.L., Peterson,T.A., Park,D. and Kann,M.G. (2012) Domain landscapes of somatic mutations in cancer. *BMC Genomics*, **13**(Suppl. 4), S9.
15. Porta-Pardo,E. and Godzik,A. (2014) Analysis of individual protein regions provides novel insights on cancer pharmacogenomics. *PLoS Comp. Biol.*, in press.
16. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
17. Xu,D., Jaroszewski,L., Li,Z. and Godzik,A. (2014) AIDA: ab initio domain assembly server. *Nucleic Acids Res.*, **42**, W308–W313.
18. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
19. Prilusky,J., Felder,C.E., Zeev-Ben-Mordehai,T., Rydberg,E.H., Man,O., Beckmann,J.S., Silman,I. and Sussman,J.L. (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**, 3435–3438.
20. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
21. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
22. Bollag,G., Tsai,J., Zhang,J., Zhang,C., Ibrahim,P., Nolop,K. and Hirth,P. (2012) Vemurafenib: the first drug approved for BRAF-mutant cancer. *Nat. Rev. Drug Discov.*, **11**, 873–886.
23. Jaroszewski,L., Li,Z., Cai,X.H., Weber,C. and Godzik,A. (2011) FFAS server: novel features and applications. *Nucleic Acids Res.*, **39**, W38–W44.
24. Ciriello,G., Miller,M.L., Aksoy,B.A., Senbabaoglu,Y., Schultz,N. and Sander,C. (2013) Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, **45**, 1127–1133.
25. Supek,F., Minana,B., Valcarcel,J., Gabaldon,T. and Lehner,B. (2014) Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, **156**, 1324–1335.
26. International Cancer Genome, C., Hudson,T.J., Anderson,W., Artez,A., Barker,A.D., Bell,C., Bernabe,R.R., Bhan,M.K., Calvo,F., Eerola,I. *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.
27. Yang,W., Soares,J., Greninger,P., Edelman,E.J., Lightfoot,H., Forbes,S., Bindal,N., Beare,D., Smith,J.A., Thompson,I.R. *et al.* (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, **41**, D955–D961.