Published in final edited form as:
Cancer Discov. 2012 May ; 2(5): 401–404. doi:10.1158/2159-8290.CD-12-0095.

The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data

Ethan Cerami1[1], **Jianjiong Gao**[1], **Ugur Dogrusoz**[3], **Benjamin E. Gross**[1], **Selcuk Onur Sumer**[3], **Bülent Arman Aksoy**[1,2], **Anders Jacobsen**[1], **Caitlin J. Byrne**[1], **Michael L. Heuer**, **Erik Larsson**[1], **Yevgeniy Antipin**[1], **Boris Reva**[1], **Arthur P. Goldberg**[1], **Chris Sander**[1], and **Nikolaus Schultz**[1]

[1]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York

[2]Tri-Institutional Training Program in Computational Biology and Medicine, New York, New York

[3]Computer Engineering Department, Bilkent University, Ankara, Turkey

Summary

The cBio Cancer Genomics Portal (http://cbioportal.org) is an open-access resource for interactive exploration of multidimensional cancer genomics data sets, currently providing access to data from more than 5,000 tumor samples from 20 cancer studies. The cBio Cancer Genomics Portal significantly lowers the barriers between complex genomic data and cancer researchers who want rapid, intuitive, and high-quality access to molecular profiles and clinical attributes from large-scale cancer genomics projects and empowers researchers to translate these rich data sets into biologic insights and clinical applications.

With the rapidly declining cost of next-generation sequencing, and major national and international efforts, including The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) (1), the field of cancer genomics continues to advance at an extraordinarily rapid pace. Data generated by these projects are, however, not easily or directly available to the cancer research community, hindering the translation of genomic data into new biologic insights, drugs, and clinical trials. The cBio Cancer Genomics Portal (http://cbioportal.org), developed at Memorial Sloan-Kettering Cancer Center (MSKCC), was specifically designed to address the unique data integration issues posed by large-scale cancer genomics projects and to make the raw data generated by large-scale cancer genomic projects more easily and directly available to the entire cancer research community (Fig. 1A).

Corresponding Author: Ethan Cerami, Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021. Phone: 646-888-2625; Fax: 646-888-2606; cancergenomics@cbio.mskcc.org.

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

Authors' Contributions
Conception and design: E. Cerami, U. Dogrusoz, C. Sander, N. Schultz
Development of methodology: E. Cerami, J. Gao, U. Dogrusoz, B.E. Gross, S.O. Sumer, B.A. Aksoy, A. Jacobsen, C.J. Byrne, M.L. Heuer, E. Larsson, Y. Antipin, B. Reva, A.P. Goldberg, C. Sander, N. Schultz
Writing, review, and/or revision of the manuscript: E. Cerami, J. Gao, U. Dogrusoz, B.E. Gross, B.A. Aksoy, A. Jacobsen, M.L. Heuer, E. Larsson, A.P. Goldberg, C. Sander, N. Schultz
Study supervision: E. Cerami, U. Dogrusoz, C. Sander, N. Schultz
Software development: E. Cerami, J. Gao, U. Dogrusoz, B.E. Gross, S.O. Sumer, B.A. Aksoy, A. Jacobsen, C.J. Byrne, M.L. Heuer, E. Larsson, Y. Antipin, B. Reva, A.P. Goldberg

The cBio portal currently contains 5 published data sets (2–5) and 15 provisional TCGA data sets. Provisional TCGA data sets are updated monthly, based on the latest TCGA production runs, and the portal will be continually updated as new TCGA cancer types are added. Published data sets include mutation data, but provisional data sets currently do not. As each cancer type within TCGA is finalized and somatic mutations are validated, mutation data will be released and added to the portal. In addition to mutation data, the portal includes copy number alterations, microarray-based and RNA sequencing–based mRNA expression changes, DNA methylation values, and protein and phosphoprotein levels.

Each data type is stored at the gene level and is then combined with available deidentified clinical data such as overall survival and disease-free survival intervals. The data are then organized as a function of patient and gene, and the portal's fundamental abstraction is the concept of altered genes; specifically, we classify a gene as altered in a specific patient if it is mutated, homozygously deleted, amplified, or its relative mRNA expression is less than or greater than a user-defined threshold. The notion of altered genes is a powerful simplifying concept that enables users to analyze complex data sets and to develop biologic hypotheses regarding recurrently altered gene sets and biologic pathways.

A key feature of the cBio portal is ease of use. All features of the portal are therefore available through a streamlined 4-step web interface. Specifically, users are guided to select: 1) a cancer study of interest, for example, TCGA Glioblastoma Multiforme (GBM); 2) one or more genomic profiles, for example, mutations and copy number alterations; 3) a patient case set, for example, all "complete" TCGA patients with GBM with mutation, copy number, and mRNA data; and 4) a gene set of interest: users can enter HUGO gene symbols, gene aliases, or Entrez Gene IDs and can enter arbitrary gene sets or pathways of interest. Users also have the option to automatically compute mutual exclusivity and co-occurrence between all pairs of genes. Finally, users have the option of performing cross-cancer queries, a simpler 2-step query, which requires only that users select "All Cancer Studies" and enter a gene set of interest.

For example, to visualize genomic alterations in the retinoblastoma (RB) pathway in the TCGA GBM data, one selects options 1 to 3 as described previously and in step 4 enters: *RB1*, *CDK4*, and *CDKN2A*. Based on the user input, the portal automatically generates a series of reports, each in a separate tab. The first of these reports summarizes genomic data across all patients through a concise graphical summary called an OncoPrint. In this graphical summary, individual genes are represented as rows, individual cases or patients are represented as columns, and glyphs and/or color-coding is used to compactly summarize distinct genomic alterations, including somatic mutations, copy number alterations, and mRNA expression. OncoPrints can be extremely useful for visualizing gene set and pathway alterations across a set of cases and for visually identifying trends such as trends in mutual exclusivity or co-occurrence between gene pairs within a gene set. For example, the RB OncoPrint (Fig. 1B) shows that alterations of genes within the RB pathway tend to be mutually exclusive. Statistical tests for co-occurrence/mutual exclusivity are also available in a separate tab.

Other reports, each available within a separate tab, include Network Analysis, Correlation Plots, Survival Analysis, Mutation Details, Event Map, Data Download, and Bookmark/E-mail. From these additional reports, we can, for example, observe that many *RB1* mutations may have strong functional consequences (Fig. 1C, Mutation Details), as predicted by MutationAssessor.org (6). We can further assess that *CDK4* mRNA expression is elevated in amplified cases (Fig. 1D, Plots Tab) and that cases with an RB pathway alteration have worse overall survival than cases without an RB pathway alteration (P = 0.0513, log-rank test; Fig. 1E, Survival Tab). Users can also click the Event Map or Data Download reports to copy and paste event information into an external spreadsheet application or click the Bookmark/E-mail tab to share their results with collaborators. Users can also visualize copy number details by choosing to launch a web start version of the Integrative Genomics Viewer [IGV (7)].

The network tab provides interactive analysis and visualization of networks altered in the chosen cancer study. The network consists of pathways and interactions derived from the open-source Pathway Commons Project (8). By default, the network of interest contains all neighbors of all seed genes specified by the user. If more than 50 neighbor nodes exist in the network, all genes are ranked by the frequency of genomic alteration within the specified cancer study and less frequently altered genes are automatically pruned from the network. By default, the portal also automatically overlays multidimensional genomic data onto each node, highlighting the frequency of alteration by mutation and copy number alteration (and optionally mRNA up-/downregulation). This provides an effective means of managing network complexity while automatically highlighting those genes most directly relevant to the cancer type in question. One can also download the full, nonpruned network for more complete visualization and analysis.

For example, we used the portal to identify genomic alterations in the homologous recombination (HR) DNA repair pathway in serous ovarian cancer. *BRCA1* and *BRCA2* are known to be involved in the HR pathway, but additional defects may also abrogate HR functionality and lead to potential sensitivity to PARP inhibitors (9). To identify potential HR defects in ovarian cancer, we used *BRCA1* and *BRCA2* as seed nodes for the network view and explored the resulting altered network of interest (Fig. 1F). By this means, we quickly identified alterations in *C11orf30*/EMSY (6% by amplification, 1.6% by mutation), a known interactor of BRCA2, as a possible alternate means for abrogating HR functionality (9). Users can also filter the network by alteration frequency, highlight all neighbors of a selected gene, hide specific nodes, crop to a selected set of nodes, or search the network by gene symbol. For example, we used the gene search filter to identify all altered Fanconi Anemia genes [another family of genes involved in the HR pathway (9)] and identified low frequency alterations in *FANCA* (altered in 3.5% of patients) and *FANCE* (2.8% of patients).

The portal also supports visualization of mutations in the context of protein domains from Pfam (10). For example, the most common mutations in *BRCA1* are germline frameshift mutations in codons 23 and 1756, also known as the 185delAG and 5382insC founder mutations, respectively [11 (Fig. 1G)].

Protein and phosphoprotein data integration and analysis are also available within the cBio portal. For example, large-scale proteomics data from reverse-phase protein array (12) are available for ovarian cancer, GBM, and colorectal cancer. The portal generates scatterplots of protein level versus mRNA expression for query genes if both data types are available. The portal also correlates genomic events of query genes with protein and phosphoprotein level changes. After a query from a user, all samples are separated into 2 groups: those that are altered in the query genes and those that are not. For each available protein or phosphoprotein level, a 2-sample Student *t* test for difference between the 2 groups of samples is performed and a *P* value is calculated. The user is then provided with a list of proteins or phosphoproteins that have significant changes between altered and unaltered samples. For example, using the portal, you can find that *PTEN* deletion in ovarian cancer is, as expected, tightly correlated with elevated phosphorylation of AKT (pS473 and pT308).

As an advanced feature, researchers can use the Onco Query Language to define specific types of genetic alterations for study within the cBio portal. For example, a user can specify that they only wish to see homozygous deletions and mutations, but not amplifications for

PTEN, and this setting will be reflected in the automatically generated OncoPrint and other plot and download features of the portal. The cBio portal also provides a complete web service interface and libraries for MATLAB and the R statistical package. Finally, the portal source code is freely available under the GNU Lesser GPL open-source license and hosted on Google code (http://code.google.com/p/cbio-cancer-genomics-portal/). Research groups wishing to install local instances of the portal to analyze their own data sets can do so by following the installation guide or use one of the prebuilt Amazon Machine Images but may require the assistance of system administrators.

In summary, the cBio portal facilitates access to cancer genomic data sets for the entire biomedical community. It provides a simple yet flexible interface to integrated data sets, intuitive visualization options, and a programmatic web interface, all of which can aid researchers in translating cancer genomic data into biologic insights and potential clinical applications. By integrating multiple genomic data types and lowering the barrier to access, the portal enables researchers to more easily mine genomic data, test hypotheses regarding genetic alterations in cancer, and place genomic data in the context of prior biologic knowledge. The cBio portal complements existing tools, including the TCGA and ICGC data portals (13), the IGV (7), the UCSC Cancer Genomics Browser (14), and IntOGen (15) by offering a unique focus on analyzing discrete genomic events across integrated data types, ease of use, support for exploratory data analysis, and interactive network analysis.

We anticipate several future directions for the portal. First, we intend to add many additional cancer studies, mostly from the TCGA and ICGC. Based on the current production schedules, we anticipate that the public portal will grow by at least 5 additional tumor types and more than 1,000 tumor samples by the third quarter of 2012. Second, we plan to add several new features, including complete support for miRNA expression, interactive OncoPrints, batch download of complete data sets, summary reports for cancer studies (e.g., frequently mutated genes), and further extensions to the cross-tumor query analysis.

User support for the cBio Cancer Genomics Portal is available via e-mail at: cbioportalgooglegroups.com.

Acknowledgments

We gratefully acknowledge our usability testers, Robert Sheridan (Sander Lab, MSKCC), Joyce Barlin (Levine Lab, MSKCC), and Petar Jelinic (Levine Lab, MSKCC) for providing invaluable feedback to improve the usability of the portal. We also gratefully acknowledge numerous collaborators within MSKCC and the TCGA and Stand Up To Cancer (SU2C) research networks, including Barry S. Taylor (MSKCC), Douglas Levine (MSKCC), David Solit (MSKCC), Cameron Brennan (MSKCC), Gordon Mills (MD Anderson), and Kenna Shaw (National Cancer Institute), for their generous feedback and promotion of the portal within the cancer genomics community. We also gratefully acknowledge Sinan Sonlu (Bilkent University) for implementing the custom Cytoscape Web node interface for displaying multidimensional genomic data, Gary Bader (University of Toronto) and Max Franz (University of Toronto) for their excellent documentation and technical support for Cytoscape Web, and the entire Pathway Commons team (MSKCC and University of Toronto) for developing the Pathway Commons Web Application Programming Interface (API) and making network downloads available. Finally, we gratefully acknowledge Jingchun Zhu [University of California Santa Cruz (UCSC)] and Nuria Lopez-Bigas (University Pompeu Fabra) for their feedback regarding features in the UCSC Cancer Genome Browser and IntOGen.

Grant Support

Funding for the cBio Cancer Genomics Portal is provided by the National Cancer Institute (NCI) as part of the TCGA Genome Data Analysis Center grant, NCI-U24CA143840, and NCI-R21CA135870. Funding for a separate Stand Up To Cancer (SU2C) instance of the cBio portal is provided by a Stand Up To Cancer Dream Team Translational Research Grant, a Program of the Entertainment Industry Foundation (SU2C-AACR-DT0209). Funding for network visualization and analysis within the portal is provided by the National Resource for Network Biology (NIH National Center for Research Resources grant numbers P41 RR031228 and GM103504). Funding for MutationAssessor is from the NIH NCI R01 CA132744.

REFERENCES

1. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, et al. International network of cancer genome projects. Nature. 2010; 464:993–998. [PubMed: 20393554]
2. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455:1061–1068. [PubMed: 18772890]
3. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474:609–615. [PubMed: 21720365]
4. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell. 2010; 18:11–22. [PubMed: 20579941]
5. Barretina J, Taylor BS, Banerji S, Ramos AH, Lagos-Quintana M, Decarolis PL, et al. Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. Nat Genet. 2010; 42:715–721. [PubMed: 20601955]
6. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011; 39:e118. [PubMed: 21727090]
7. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011; 29:24–26. [PubMed: 21221095]
8. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011; 39:D685–D690. [PubMed: 21071392]
9. Turner N, Tutt A, Ashworth A. Hallmarks of ‘BRCAness’ in sporadic cancers. Nat Rev Cancer. 2004; 4:814–819. [PubMed: 15510162]
10. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012; 40:D290–D301. [PubMed: 22127870]
11. Ferla R, Calo V, Cascio S, Rinaldi G, Badalamenti G, Carreca I, et al. Founder mutations in BRCA1 and BRCA2 genes. Ann Oncol. 2007; 18(suppl 6):vi93–vi98. [PubMed: 17591843]
12. Sheehan KM, Calvert VS, Kay EW, Lu Y, Fishman D, Espina V, et al. Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma. Mol Cell Proteomics. 2005; 4:346–355. [PubMed: 15671044]
13. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. Database (Oxford). 2011 bar026.
14. Sanborn JZ, Benz SC, Craft B, Szeto C, Kober KM, Meyer L, et al. The UCSC Cancer Genomics Browser: update 2011. Nucleic Acids Res. 2011; 39:D951–D959. [PubMed: 21059681]
15. Gundem G, Perez-Llamas C, Jene-Sanz A, Kedzierska A, Islam A, Deu-Pons J, et al. IntOGen: integration and data mining of multidimensional oncogenomic data. Nat Methods. 2010; 7:92–93. [PubMed: 20111033]

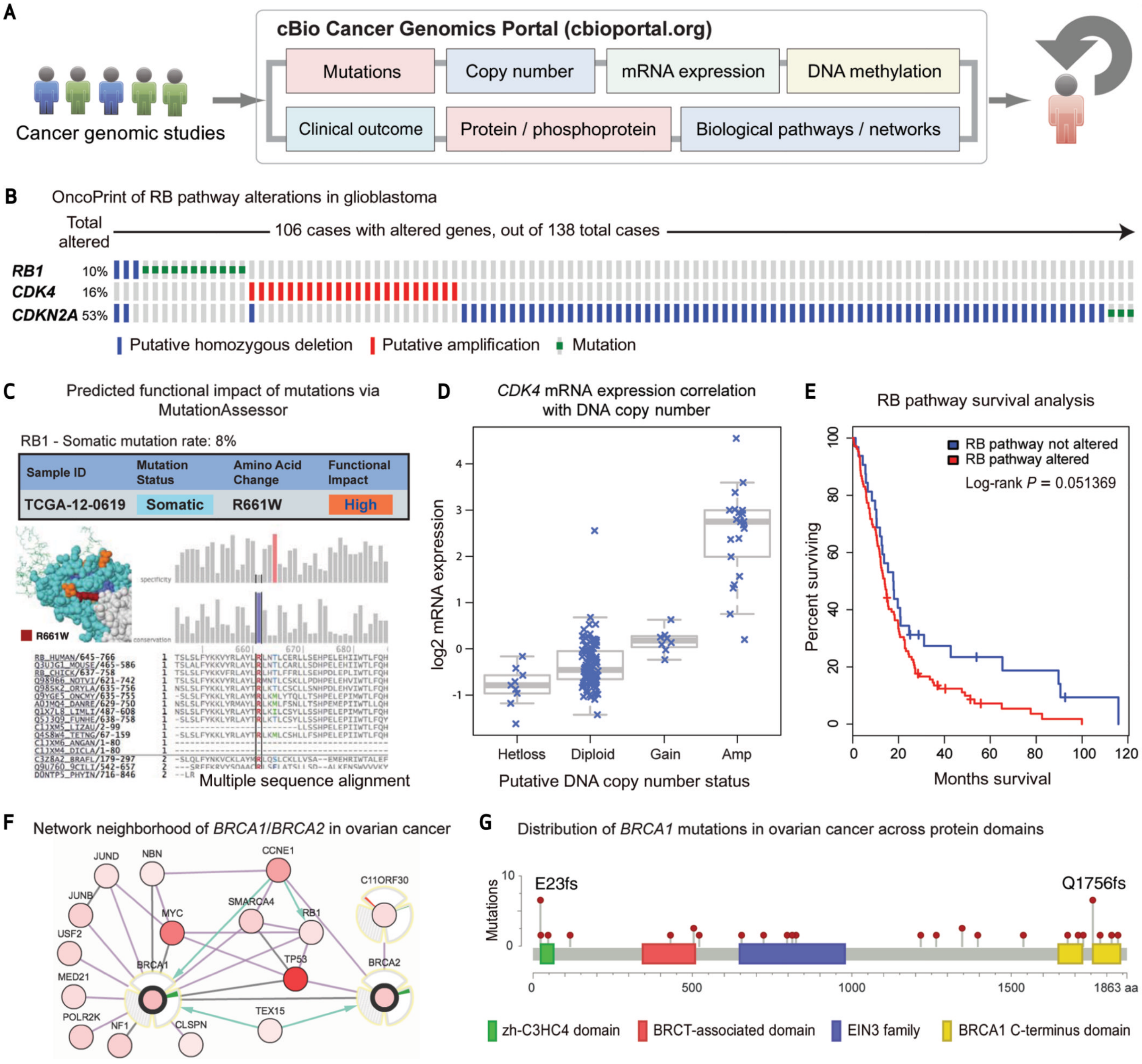

Figure 1.

The cBio Cancer Genomics Portal. **A,** the cBio Cancer Genomics Portal is an open platform for interactively exploring multidimensional cancer genomics data sets in the context of clinical data and biologic pathways. **B,** OncoPrint of RB pathway alterations in GBM. Genomic alterations of different members in the RB pathway are mutually exclusive. The OncoPrint provides an overview of genomic alterations (legend) in particular genes (rows) affecting particular individual samples (columns). **C,** mutation details for *RB1*. The predicted functional impact of the *RB1* missense mutations in GBM can be assessed through Mutation Assessor. This includes a predicted functional impact score, a multiple sequence alignment of different family members, and a 3-dimensional structure view, when available. **D,** correlation plot for CDK4. GBM samples with *CDK4* amplification have markedly increased *CDK4* mRNA expression. **E,** survival analysis. GBM cases with an RB pathway alteration have worse overall survival than cases without an RB pathway alteration. **F,** network view of the *BRCA1/BRCA2* neighborhood in serous ovarian cancer. *BRCA1* and *BRCA2* are seed genes (indicated with thick border), and all other genes are automatically identified as altered in ovarian cancer. Multidimensional genomic details are shown for *BRCA2* and *C11orf30*/EMSY. Darker red indicates increased frequency of alteration (defined by mutation, copy number amplification, or homozygous deletion) in ovarian cancer. **G,** distribution of *BRCA1* mutations in ovarian cancer across protein domains. The 2 hot spots (p.E23fs and p.Q1756fs) represent the common founder mutations 185delAG and 5382insC frequently observed in *BRCA1*.