ARTICLE

Genomics & Informatics
Vol. 10(1) 33-39, March 2012

http://dx.doi.org/10.5808/GI.2012.10.1.33
pISSN 1598-866X eISSN 2234-0742

CaGe: A Web-Based Cancer Gene Annotation System for Cancer Genomics

Young-Kyu Park[1,2], Tae-Wook Kang[1], Su-Jin Baek[1,3], Kwon-Il Kim[2], Seon-Young Kim[1,3], Doheon Lee[2]* and Yong Sung Kim[1,3]**

[1]Medical Genomics Research Center, KRIBB, Daejeon 305-806, Korea, [2]Department of Bio and Brain Engineering, KAIST, Daejeon 305-701, Korea, [3]Department of Functional Genomics, University of Science and Technology, Daejeon 305-806, Korea

Abstract

High-throughput genomic technologies (HGTs), including next-generation DNA sequencing (NGS), microarray, and serial analysis of gene expression (SAGE), have become effective experimental tools for cancer genomics to identify cancer-associated somatic genomic alterations and genes. The main hurdle in cancer genomics is to identify the real causative mutations or genes out of many candidates from an HGT-based cancer genomic analysis. One useful approach is to refer to known cancer genes and associated information. The list of known cancer genes can be used to determine candidates of cancer driver mutations, while cancer gene-related information, including gene expression, protein-protein interaction, and pathways, can be useful for scoring novel candidates. Some cancer gene or mutation databases exist for this purpose, but few specialized tools exist for an automated analysis of a long gene list from an HGT-based cancer genomic analysis. This report presents a new web-accessible bioinformatic tool, called CaGe, a cancer genome annotation system for the assessment of candidates of cancer genes from HGT-based cancer genomics. The tool provides users with information on cancer-related genes, mutations, pathways, and associated annotations through annotation and browsing functions. With this tool, researchers can classify their candidate genes from cancer genome studies into either previously reported or novel categories of cancer genes and gain insight into underlying carcinogenic mechanisms through a pathway analysis. We show the usefulness of CaGe by assessing its performance in annotating somatic mutations from a published small cell lung cancer study.

Keywords: annotation, cancer gene, high-throughput genomic technology, mutation, next-generation sequencing, pathway

Availability: CaGe is freely available at http://mgrc.kribb.re.kr/cage.

Introduction

High-throughput genomic technologies (HGTs), including next-generation DNA sequencing (NGS), microarray, and serial analysis of gene expression (SAGE), have become effective tools for cancer genomics through various applications. Especially, the applications of NGS include whole-genome, exome, and transcriptome approaches for searching for a wide range of cancer-specific genomic alterations, such as point mutations, insertions, and deletions; copy number changes; and rearrangements, leading to the development of cancer [1].

One of the hurdles in HGT-based cancer genomics is to identify causative mutations or genes out of many candidates from the analysis. A useful approach is to refer to known cancer genes and associated information [2]. The list of known cancer genes can be used to determine candidates of cancer driver mutations, while cancer gene-related information, such as gene expression, protein-protein interaction, and pathway information, can be useful for scoring novel candidates.

Some useful cancer genomic databases for this purpose exist, including Cancer Gene Census (CGC) [3] and Catalogue of Somatic Mutations in Cancer (COSMIC) [4], which have been constructed and managed by the Cancer Genome Project of the Welcome Trust Sanger Institute; and Cancer Gene Index (CGI, http://ncicb.nci.nih.gov/NCICB/projects/cgdcp), The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov), and Cancer Genome Anatomy Project (CGAP, http://cgap.nci.nih.gov), which are maintained by the National Cancer Institute of the U. S. National Institute of Health and National Human Genome Research Institute. Among them, COSMIC and CGC are the two most commonly used resources among researchers when checking reported cancer driver mutations. The COSMIC database stores information on somatic mutations and associated information extracted from the scientific literature, while the CGC da-

*Corresponding author1: E-mail: dhlee@biosoft.kaist.ac.kr
Tel +82-42-350-4316, Fax +82-42-350-8680
**Corresponding author2: E-mail: yongsung@kribb.re.kr
Tel +82-42-879-8110, Fax +82-42-879-8119
Received 26 January 2012, Revised 10 February 2012, Accepted 16 February 2012

tabase is a catalog of cancer genes with manually screened somatic mutations extracted from the literature. These cancer genes are annotated with information concerning chromosomal location, tumor types in which mutations are found, classes of mutations that contribute to oncogenesis, and other genetic properties. Also, CGI is known to be created through an automated linguistic text analysis of millions of MEDLINE abstracts, with manual validation and annotation of the extracted data by expert curators. It is a high-quality data resource consisting of genes that have been experimentally associated with human cancer diseases and/or pharmacological compounds, the evidence of these associations, and relevant annotations on the data. Thus, it can also be a valuable resource to annotate mutation data from cancer genomics studies.

Currently, few specialized tools exist for an automated analysis of a long gene list from HGT-based cancer genomics. While previously described cancer genomic information databases can be useful, they are not effective in processing results from NGS-based cancer genomic studies. Meanwhile, there are many functional annotation systems to analyze gene lists derived from high-throughput microarray experiments, including DAVID [5], GoMiner [6], GOstat [7], Onto-express [8], and gene set enrichment analysis (GSEA) [9]. DAVID is one of the most popular general purpose functional annotation systems. It provides various effective analysis tools and is also applicable in the analysis of gene lists from cancer genomics research. However, DAVID is not optimized for processing results from HGT-based cancer genomics data, especially for NGS-based cancer genomics downstream data; therefore, a more specialized cancer gene annotation system is needed.

Here, we introduce a web-accessible cancer genome annotation system, named CaGe, to provide users with information on cancer genes, mutations, pathways, and associated annotations, based on several cancer gene databases composed of reported cancer-causing genes and associated cancer pathways. For a given gene list, CaGe searches cancer gene databases with converted standard gene symbols, analyzes the biological pathways, and provides various cancer gene-related annotations through intuitive web user interfaces. It also serves additional functions, including processing various input types and formats, managing jobs for user submitted tasks, and browsing for cancer genes and pathways with various useful hyperlinks between the annotations and the external public annotation databases. We hope CaGe will be useful in identifying cancer-causing mutations and genes in HGT-based cancer genomics.

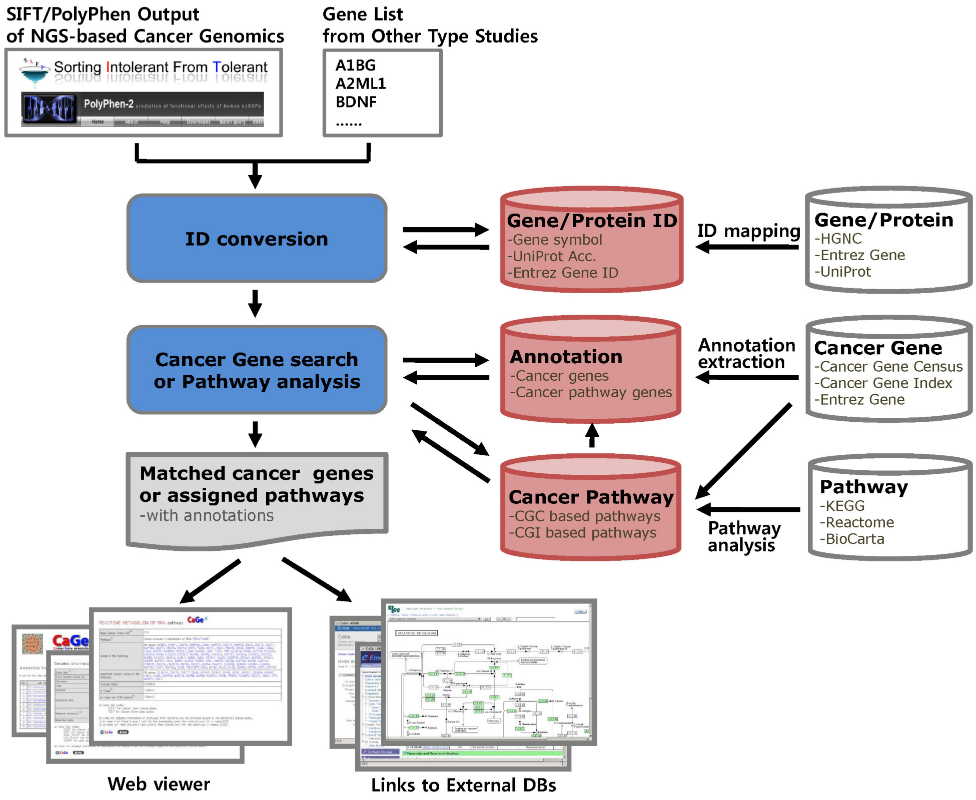


Fig. 1. Construction and main service flows of CaGe. SIFT, Sorting Intolerant From Tolerant; NGS, next-generation DNA sequencing; HGNC, HUGO Gene Nomenclature Committee; CGC, Cancer Gene Census; CGI, Cancer Gene Index.

Methods

The goal of this study was to provide researchers with a tool for classifying their candidate genes from HGT-based cancer genome studies into previously reported or novel categories of cancer genes, while providing insight into underlying carcinogenic mechanisms through a pathway analysis. To implement the cancer gene annotation function of CaGe, we constructed reported cancer gene and cancer annotation databases from public cancer genomic databases and cancer pathway-gene databases by pathway analysis with reported cancer gene sets and canonical pathways. We also constructed a gene ID database to allow various input formats for the input of gene lists and a gene functional annotation database to provide users with functional clues for the annotated candidate genes. Then, we developed a core retrieval program and web interfaces for the main functions, which include cancer gene annotation, cancer pathway annotation, cancer gene browsing, and cancer pathway browsing. The workflow for the database construction and data processing in CaGe is summarized in Fig. 1, and the cancer gene annotation page of the CaGe web interface is shown in Fig. 2.

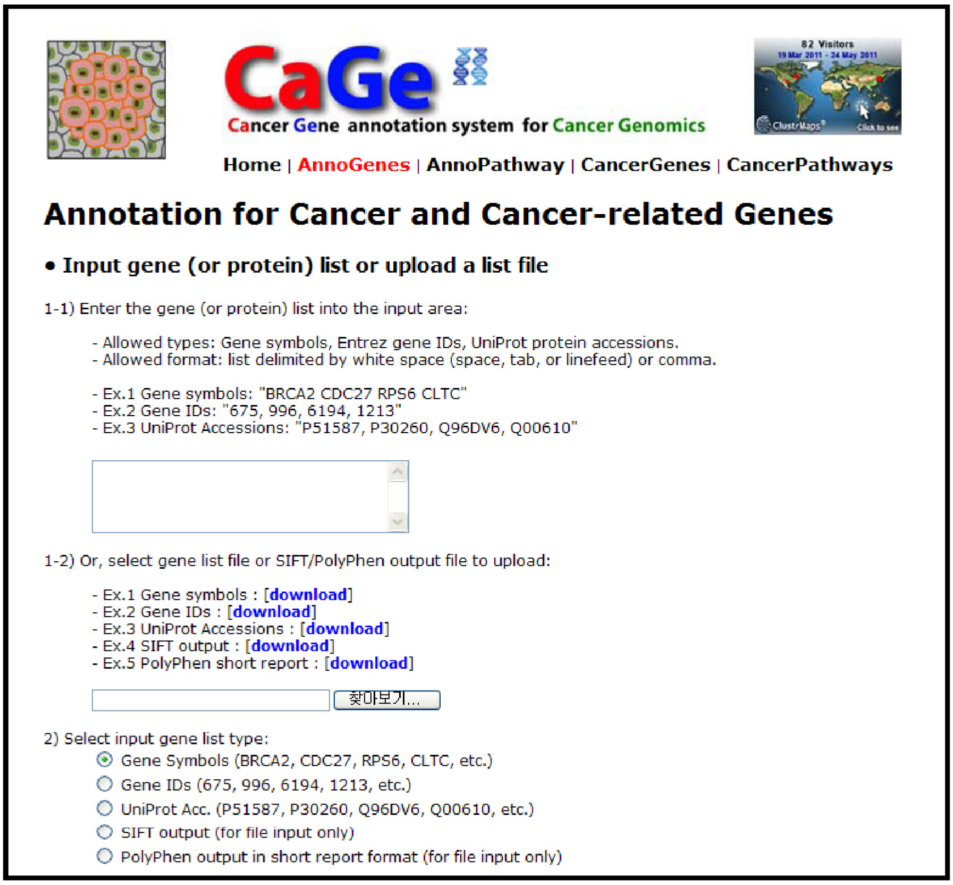
Cancer gene and annotation data

To construct the reported cancer gene database, we used gene sets from the CGC database (released on Dec, 2010) and CGI database (downloaded on Feb, 2011). The cancer pathways for the cancer pathway gene database construction were assigned based on statistical significance from one-tailed Fisher's exact test for overlapping genes between reported cancer gene sets and canonical pathways from public pathway databases, including KEGG (Release 57.0), BioCarta, and Reactome (downloaded on Feb, 2011).

We also created a gene ID database to convert various input identifiers into standard gene symbols with HUGO Gene Nomenclature Committee (HGNC) (downloaded on Feb, 2011) data for the standard gene symbols and with Entrez Gene (downloaded on Feb, 2011) and UniProt (Release 2011_03) data for the gene IDs, protein IDs, and functional annotations.

Input types

Available input types include a list of gene symbols, Entrez gene IDs, or UniProt accessions and output files


CaGe
Cancer Gene annotation system for Cancer Genomics
82 Visitors
19 Mar 2011 - 24 May 2011
ClustrMaps
Click to see
Home | AnnoGenes | AnnoPathway | CancerGenes | CancerPathways

Annotation for Cancer and Cancer-related Genes

• Input gene (or protein) list or upload a list file

1-1) Enter the gene (or protein) list into the input area:

- Allowed types: Gene symbols, Entrez gene IDs, UniProt protein accessions.
- Allowed format: list delimited by white space (space, tab, or linefeed) or comma.

- Ex.1 Gene symbols: "BRCA2 CDC27 RPS6 CLTC"
- Ex.2 Gene IDs: "675, 996, 6194, 1213"
- Ex.3 UniProt Accessions: "P51587, P30260, Q96DV6, Q00610"

1-2) Or, select gene list file or SIFT/PolyPhen output file to upload:

- Ex.1 Gene symbols : [download]
- Ex.2 Gene IDs : [download]
- Ex.3 UniProt Accessions : [download]
- Ex.4 SIFT output : [download]
- Ex.5 PolyPhen short report : [download]

2) Select input gene list type:

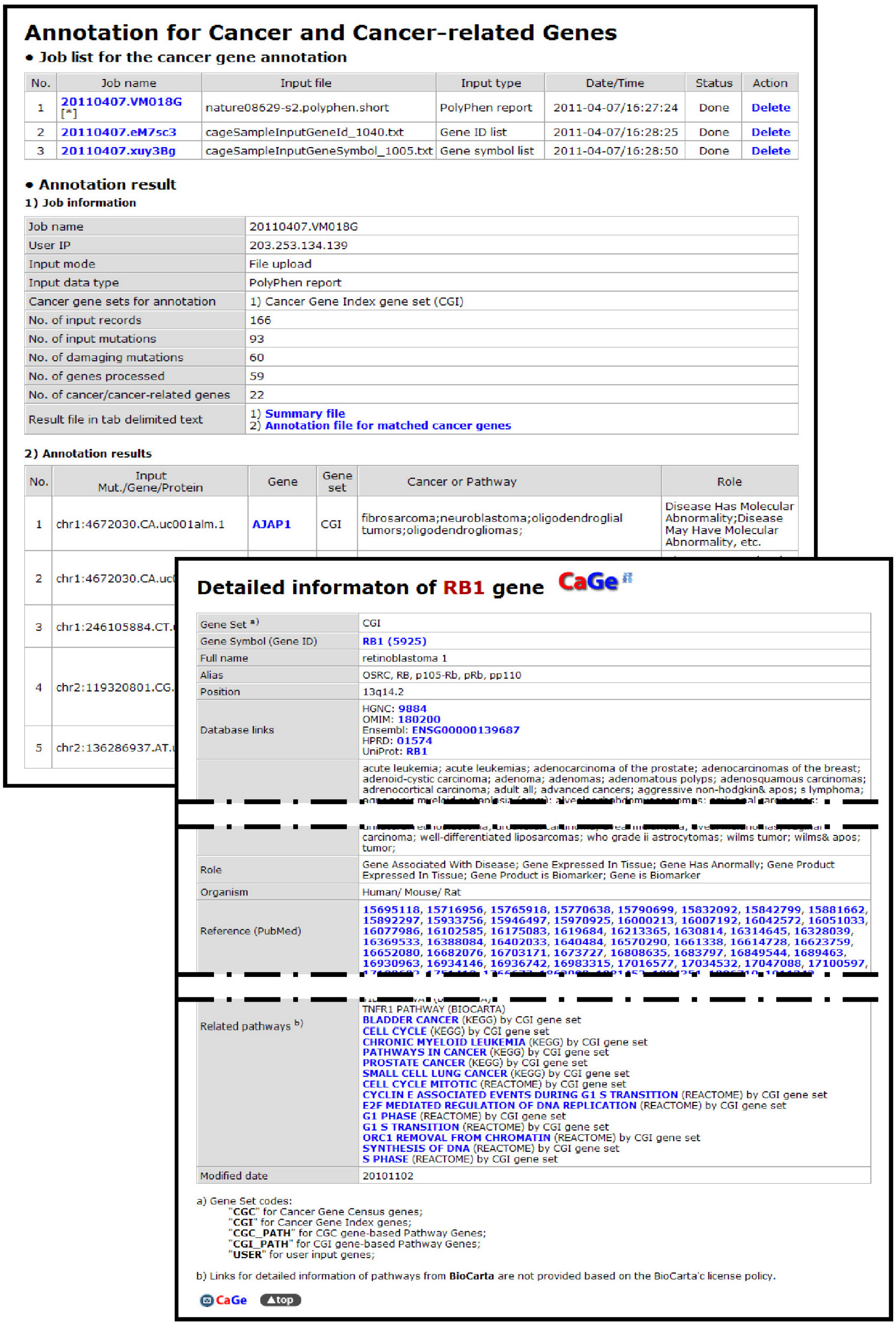
● Gene Symbols (BRCA2, CDC27, RPS6, CLTC, etc.)
○ Gene IDs (675, 996, 6194, 1213, etc.)
○ UniProt Acc. (P51587, P30260, Q96DV6, Q00610, etc.)
○ SIFT output (for file input only)
○ PolyPhen output in short report format (for file input only)

Fig. 2. Cancer gene annotation page of CaGe web interface.

from prediction tools, such as Sorting Intolerant From Tolerant (SIFT) [10] and PolyPhen [11], which evaluate functional effects of mutations on proteins. CaGe parses the output of those tools to extract the list of genes that have somatic mutations evaluated as functionally damaging, classify them into previously reported and novel candidate cancer genes, and conduct pathway analyses to give insights into underlying carcinogenic mechanisms. Thus, CaGe offers straightforward data processing from HGT-based data without additional data conversion.

Cancer gene and pathway annotation workflow

After acquiring user input, CaGe converts various input gene IDs into standard gene symbols, finds known cancer genes and pathways, links various cancer-related annotations to matched genes, and outputs them in the



Annotation for Cancer and Cancer-related Genes

- Job list for the cancer gene annotation

No.	Job name	Input file	Input type	Date/Time	Status	Action
1	20110407.VM018G [*]	nature08629-s2.polyphen.short	PolyPhen report	2011-04-07/16:27:24	Done	Delete
2	20110407.eM7sc3	cageSampleInputGeneId_1040.txt	Gene ID list	2011-04-07/16:28:25	Done	Delete
3	20110407.xuy3Bg	cageSampleInputGeneSymbol_1005.txt	Gene symbol list	2011-04-07/16:28:50	Done	Delete

- Annotation result

1) Job information

Job name	20110407.VM018G
User IP	203.253.134.139
Input mode	File upload
Input data type	PolyPhen report
Cancer gene sets for annotation	1) Cancer Gene Index gene set (CGI)
No. of input records	166
No. of input mutations	93
No. of damaging mutations	60
No. of genes processed	59
No. of cancer/cancer-related genes	22
Result file in tab delimited text	1) Summary file 2) Annotation file for matched cancer genes

2) Annotation results

No.	Input Mut./Gene/Protein	Gene	Gene set	Cancer or Pathway	Role
1	chr1:4672030.CA.uc001alm.1	AJAP1	CGI	fibrosarcoma;neuroblastoma;oligodendroglial tumors;oligodendrogliomas;	Disease Has Molecular Abnormality;Disease May Have Molecular Abnormality, etc.
2	chr1:4672030.CA.uc				
3	chr1:246105884.CT.				
4	chr2:119320801.CG.				
5	chr2:136286937.AT.				

Detailed informaton of RB1 gene CaGe

Gene Set a)	CGI
Gene Symbol (Gene ID)	RB1 (5925)
Full name	retinoblastoma 1
Alias	OSRC, RB, p105-Rb, pRb, pp110
Position	13q14.2
Database links	HGNC: 9884 OMIM: 180200 Ensembl: ENSG00000139687 HPRD: 01574 UniProt: RB1
	acute leukemia; acute leukemias; adenocarcinoma of the prostate; adenocarcinomas of the breast; adenoid-cystic carcinoma; adenoma; adenomas; adenomatous polyps; adenosquamous carcinomas; adrenocortical carcinoma; adult all; advanced cancers; aggressive non-hodgkin& apos; s lymphoma; carcinoma; well-differentiated liposarcomas; who grade ii astrocytomas; wilms tumor; wilms& apos; tumor;
Role	Gene Associated With Disease; Gene Expressed In Tissue; Gene Has Anormally; Gene Product Expressed In Tissue; Gene Product is Biomarker; Gene is Biomarker
Organism	Human/ Mouse/ Rat
Reference (PubMed)	15695118, 15716956, 15765918, 15770638, 15790699, 15832092, 15842799, 15881662, 15892297, 15933756, 15946497, 15970925, 16000213, 16007192, 16042572, 16051033, 16077986, 16102585, 16175083, 1619684, 16213365, 1630814, 16314645, 16328039, 16369533, 16388084, 16402033, 1640484, 16570290, 1661338, 16614728, 16623759, 16652080, 16682076, 16703171, 1673727, 16808635, 1683797, 16849544, 1689463, 16930963, 16934146, 16936742, 16983315, 17016577, 17034532, 17047088, 17100597,
Related pathways b)	TNFR1 PATHWAY (BIOCARTA) BLADDER CANCER (KEGG) by CGI gene set CELL CYCLE (KEGG) by CGI gene set CHRONIC MYELOID LEUKEMIA (KEGG) by CGI gene set PATHWAYS IN CANCER (KEGG) by CGI gene set PROSTATE CANCER (KEGG) by CGI gene set SMALL CELL LUNG CANCER (KEGG) by CGI gene set CELL CYCLE MITOTIC (REACTOME) by CGI gene set CYCLIN E ASSOCIATED EVENTS DURING G1 S TRANSITION (REACTOME) by CGI gene set E2F MEDIATED REGULATION OF DNA REPLICATION (REACTOME) by CGI gene set G1 PHASE (REACTOME) by CGI gene set G1 S TRANSITION (REACTOME) by CGI gene set ORC1 REMOVAL FROM CHROMATIN (REACTOME) by CGI gene set SYNTHESIS OF DNA (REACTOME) by CGI gene set S PHASE (REACTOME) by CGI gene set
Modified date	20101102

a) Gene Set codes:
"CGC" for Cancer Gene Census genes;
"CGI" for Cancer Gene Index genes;
"CGC_PATH" for CGC gene-based Pathway Genes;
"CGI_PATH" for CGI gene-based Pathway Genes;
"USER" for user input genes;

b) Links for detailed information of pathways from BioCarta are not provided based on the BioCarta'c license policy.

Fig. 3. Cancer gene annotation results page of CaGe web interface.

form of tables or text files through the web interface (Fig. 3). Another function of CaGe is to identify over-represented cancer-related or other biological pathways from the input gene list by performing one-tailed Fisher's exact test.

When processing more than one annotation job, users can manage their annotation tasks through the CaGe interface, and access to their results is maintained by the internal job database of CaGe until the jobs are deleted by the user or by a scheduled cleaning process. The IP addresses of client computers are used for secured job management without a logon process. Completed annotation jobs are listed on the job table, and the annotated results can be shown selectively by the user on the annotation result page or can be downloaded as tab-delimited text files for further analyses. The annotation on the results page has many useful links to gene and cancer-related information in the CaGe database or external public databases. In addition to the cancer gene and pathway annotation function, CaGe provides a function to browse cancer genes and pathways so that users can search cancer-related annotations without an input list. The search flows of cancer genes and pathways are connected to each other by crosslinks in the cancer gene information page or pathway information page.

Results and Discussion

Database and web interface

Based on the reported cancer-related gene sets from CGC and CGI, a cancer gene database and a cancer

Table 1. Cancer gene sets and cancer pathway gene sets in the CaGe database

No.	Cancer and pathway gene set	Gene set code	No. of genes	No. of significant pathways[a]
1	Cancer Gene Census	CGC	436	n/a
2	Cancer Gene Index	CGI	7,181	n/a
3	CGC-based Pathway Genes	CGC_PATH	5,790	143
4	CGI-based Pathway Genes	CGI_PATH	6,744	176

n/a, not available.

[a]Pathways with p-values < 0.05 from Fisher's exact test.

Table 2. Annotations for a cancer gene provided by CaGe

Name	Description	Example
Basic cancer gene annotations		
Gene set	Source database for known cancer genes or cancer related genes	CGC/CGI/CGC_PATH/CGI_PATH
Gene symbol	Standard gene symbol	*ABL/ABL2/BRCA2*
Tumor type	Type of tumor as mutation source	AML/Prostate/Breast/Ovarian
Cell type	Cell type where mutation occurs	Somatic/Germline
Mutation type	Type of reported mutation	Missense/Translocation/Deletion
Cancer syndrome	Cancer syndrome	Familial blastoma
Additional functional annotations		
Gene symbol (Gene ID)	Gene symbol (Entrez gene ID)	*BHD* (201163)
Full name	Full description of gene	Folliculin, Birt-Hogg-Dube syndrome
Alias	Other gene symbols	*BHD, DKFZp547A118, FLCL, FLJ45004, FLJ99377, MGC17998*
Position	Chromosome band where the gene having the mutation is located	17p11.2

Database links	Identifiers (IDs) with hyperlinks to the public gene and protein annotation databases	HGNC: 27310;
		OMIM: 607273;
		Ensembl: ENSG00000154803;
		HPRD: 06278;
		UniProt: *BHD*

Molecular Genetics	Type of inheritance	Recessive/Dominant
Tissue type	Tissue type of tumor	E (epithelial), M (mesenchymal)
Related pathways	Pathways where the gene is assigned	Purine metabolism (KEGG) by Cancer Gene Census (CGC) gene set

pathway gene database were constructed by the described methods. In total, 436 CGC-originated and 7181 CGI-originated cancer genes were prepared for the reported cancer gene annotation. Additionally, 5,790 CGC-based and 6,744 CGI-based cancer pathway genes were assigned for the annotation of unreported but potential candidate genes (Table 1). The gene ID database for gene symbol conversion, based on HGNC, and the functional annotation database, based on Entrez Gene and UniProt database flat files, were constructed. The types of annotations provided by CaGe are summarized in Table 2.

The main window of the CaGe web interface is shown in Fig. 2 for the cancer gene annotation process. Main menus, located in the upper part of the main window, are linked to the four main functions and home page of the system: 1) home page, 2) cancer gene annotation function, 3) cancer pathway annotation function, 4) cancer gene browsing function, and 5) cancer pathway browsing function. Detailed usages for CaGe are described in the user's guide at http://mgrc.kribb.re.kr/cage/include/cageUserGuide.pdf.

System information

CaGe was developed using PHP, R, and python languages; MySQL for database management; and Apache for the web server and is operated on a Linux platform with 8-core Intel Xeon CPUs (2.33 GHz) and 24 GB of main memory.

Performance evaluation with small cell lung cancer mutation data

To assess the capability of CaGe, we annotated genes from candidate mutations for a small cell lung cancer genome (SCLC) [12]. We had 59 genes with predicted functionally damaging mutations after applying 22,910 mutations to the PolyPhen and could annotate 22 previously reported known cancer genes successfully by applying the PolyPhen output for CaGe. Known cancer genes included *RB1*, which was mentioned as an SCLC-related gene in Pleasance's work; *DST* and *ETS2*, which were previously reported as SCLC-related genes but not mentioned in Pleasance's work; and 3 more genes (*PDGFC*, *IL16*, and *AGTR2*), which are known as lung cancer-related genes (Supplementary Table 1). The other 16 genes are known to be related to other cancer types, suggesting that they might be important in the carcinogenesis of SCLC as well. From the pathway analysis, we identified 12 known cancer genes and 11 genes in cancer-related pathways (Supplementary Table 2). Those 11 genes might also be important in the carcinogenesis of SCLC. Thus, we conclude that CaGe can annotate cancer genes effectively and suggest that CaGe will be useful in the identification of cancer-causing mutations and genes in HGT-based cancer genomics.

In this paper, we present a new cancer genomic tool, CaGe, for the assessment of candidate cancer genes having somatic mutations from HGT-based cancer genomics. CaGe provides users with information on cancer genes, mutations, pathways, and associated annotations through cancer gene annotation, cancer pathway annotation, cancer gene browsing, and cancer pathway browsing functions. It has a capacity to process SIFT or PolyPhen output files as direct input for usual NGS-based cancer genomics flows. Researchers can classify their candidate genes from cancer genome studies into previously known or novel categories of cancer genes and gain insight into the underlying carcinogenic mechanisms through a pathway analysis. We hope that CaGe will be useful for the identification of cancer-causing mutations and genes in HGT-based cancer genomics.

Supplementary materials

Supplementary data including two tables can be found with this article online at http://www.genominfo.org/html/UploadFile/article5_201203_SP.pdf.

Acknowledgments

This work was supported by grants from the KRIBB Research Initiative Program and from the Ministry of Education, Science and Technology (MEST) (2010-0012810) of the Republic of Korea. D.L. was supported by the Basic Research Laboratory grant (2009-0086964) of the Ministry of Education, Science and Technology through the National Research Foundation of Korea, and Korea Institute of Science and Technology Information (KISTI).

References

1. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* 2010;11:685-696.
2. Zang ZJ, Ong CK, Cutcutache I, Yu W, Zhang SL, Huang D, *et al*. Genetic and structural variation in the gastric cancer kinome revealed through targeted deep sequencing. *Cancer Res* 2011;71:29-39.
3. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, *et al*. A census of human cancer genes. *Nat Rev Cancer* 2004;4:177-183.
4. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY,

Beare D, *et al*. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* 2011;39:D945-D950.

5. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;4:44-57.
6. Zeeberg BR, Qin H, Narasimhan S, Sunshine M, Cao H, Kane DW, *et al*. High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). *BMC Bioinformatics* 2005;6:168.
7. Beissbarth T, Speed TP. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* 2004;20:1464-1465.
8. Khatri P, Voichita C, Kattan K, Ansari N, Khatri A, Georgescu C, *et al*. Onto-Tools: new additions and improvements in 2006. *Nucleic Acids Res* 2007;35:W206-W211.
9. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102:15545-15550.
10. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;31:3812-3814.
11. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, *et al*. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7:248-249.
12. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, *et al*. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 2010;463:184-190.