D816–D819 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 5 October 2008*
doi:10.1093/nar/gkn673

CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens

Luiz Gonzaga Almeida[1], Noboru J. Sakabe[2], Alice R. deOliveira[1], Maria Cristina C. Silva[1], Alex S. Mundstein[1], Tzeela Cohen[3], Yao-Tseng Chen[3], Ramon Chua[3], Sita Gurung[3], Sacha Gnjatic[3], Achim A. Jungbluth[3], Otávia L. Caballero[3], Amos Bairoch[4], Eva Kiesler[3], Sarah L. White[3], Andrew J. G. Simpson[3], Lloyd J. Old[3], Anamaria A. Camargo[5] and Ana Tereza R. Vasconcelos[1,*]

[1]Laboratório Nacional de Computação Científica, Petrópolis, RJ, Brazil, [2]Human Genetics Department, University of Chicago, Chicago, IL, [3]Ludwig Institute for Cancer Research, New York, NY, USA, [4]Swiss Institute of Bioinformatics (SIB) and Structural Biology and Bioinformatics Department, University of Geneva, Geneva, Switzerland and [5]Ludwig Institute for Cancer Research, São Paulo, SP, Brazil

Received August 14, 2008; Revised September 21, 2008; Accepted September 22, 2008

ABSTRACT

The potency of the immune response has still to be harnessed effectively to combat human cancers. However, the discovery of T-cell targets in melanomas and other tumors has raised the possibility that cancer vaccines can be used to induce a therapeutically effective immune response against cancer. The targets, cancer-testis (CT) antigens, are immunogenic proteins preferentially expressed in normal gametogenic tissues and different histological types of tumors. Therapeutic cancer vaccines directed against CT antigens are currently in late-stage clinical trials testing whether they can delay or prevent recurrence of lung cancer and melanoma following surgical removal of primary tumors. CT antigens constitute a large, but ill-defined, family of proteins that exhibit a remarkably restricted expression. Currently, there is a considerable amount of information about these proteins, but the data are scattered through the literature and in several bioinformatic databases. The database presented here, CTdatabase (http://www.cta.lncc.br), unifies this knowledge to facilitate both the mining of the existing deluge of data, and the identification of proteins alleged to be CT antigens, but that do not have their characteristic restricted expression pattern. CTdatabase is more than a repository of CT antigen data, since all the available information was carefully curated and annotated with most data being specifically processed for CT antigens and stored locally. Starting from a compilation of known CT antigens, CTdatabase provides basic information including gene names and aliases, RefSeq accession numbers, genomic location, known splicing variants, gene duplications and additional family members. Gene expression at the mRNA level in normal and tumor tissues has been collated from publicly available data obtained by several different technologies. Manually curated data related to mRNA and protein expression, and antigen-specific immune responses in cancer patients are also available, together with links to PubMed for relevant CT antigen articles.

INTRODUCTION

More than 11 million people are diagnosed with cancer every year causing 12.5% of all deaths worldwide (World Health Organization, 2006). Novel forms of cancer treatment are desperately needed, and immunotherapy represents an approach that has yet to be fully explored. One form of immunotherapy is the therapeutic cancer vaccine that induces the immune response to recognize and destroy cancer cells. Such vaccines can be based on specific antigens, such as the CT antigens that are specifically expressed in tumors with limited expression elsewhere in the patient's tissues. Advanced clinical trials of CT antigens are underway. In 2007, GlaxoSmithKline initiated the largest-ever lung cancer trial to test the ability of a

*To whom correspondence should be addressed. Tel: +55 24 22336065; Fax: +55 24 33 6124; Email: atrv@lncc.br

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

therapeutic vaccine based on a CT antigen to delay the recurrence of resected non-small cell lung cancer (1). A second CT antigen vaccine is currently in an international phase II clinical trial for patients with resected malignant melanoma (2).

The first cancer antigen was cloned from the cells of a melanoma patient, by Thierry Boon and his colleagues (3). The antigen in question was denominated melanoma antigen-1, or MAGE-1, and subsequently renamed as MAGE-A1. An international effort to discover additional cancer antigens soon revealed an entire family expressed only in tumors and in the immunoprivileged gametogenic tissues. This group was collectively termed the CT antigens by Old and Chen (4). There are now more than 70 CT gene families, many of them promising vaccine candidates. Nevertheless, the range of discovery programs that have diversely reported these important therapeutic candidates have resulted in the terminology of a CT antigen being loosely defined and applied, adding to importance of a single carefully curated database to be able to accurately assess the relevance of individual proteins.

Due to their importance, there is a rapidly expanding body of knowledge concerning these genes widely in the literature and diverse databases. To gather and uniformly present the available information on CT antigens, we have created a user-friendly interface termed the Cancer-Testis database (CTdatabase). The database integrates heterogeneous data including basic gene, protein and expression information in normal and tumor tissues as well as immunogenicity in cancer patients. The CTdatabase contains links to external databases although a priority has been to specifically process relevant data so that it can be stored locally. The information available was expertly curated and annotated, and regular updates are planned.

CT ANTIGENS PRESENT IN THE DATABASE

A list of CT antigens was first compiled manually from the literature (see references used to compile the list at http://www.cta.lncc.br, under the link 'Gene annotations' in the main page). Computational prediction was also used. The resultant CTdatabase comprises 204 genes.

The CTdatabase has a straightforward interface where information for individual genes and their products is displayed. Each entry is listed according to the official gene symbol (or the name available at NCBI' Gene Entrez database) and information is further sorted into 'domains', displayed as 'tabs' by subject: Summary, Gene, Protein, mRNA expression, protein expression, immune response, PubMed. The domains have been populated using automatic recovery from public databases, manual annotation and novel data generated by RT-PCR.

THE 'GENE' AND 'PROTEIN' TABS

The 'Gene' tab contains general information extracted from NCBI Entrez Gene database including aliases, mRNA RefSeq accession numbers (5), gene structure, chromosomal localization, exon-intron structure, RefSeq splicing variants as well as links to the genome browsers MapViewer (6), UCSC Genome Browser (7) and Single Nucleotide Polymorphisms (6), where available.

The annotations extracted from NCBI Entrez Gene were manually curated. For example, the transcripts for SPANXE and SPANXD were found to align to a single locus. Both entries are available in the CTdatabase, but a warning is displayed for SPANXE indicating that this gene is identical to SPANXD and the user is thus directed to the SPANXD entry. Likewise, some aliases were also corrected, for example, MAGE-A4 and MAGE-A5 are different genes, but NCBI Entrez Gene reports MAGE-A4 as an alias for MAGE-A5 (as of July, 2008). The CTdatabase also annotated gene names as splicing variants when they aligned to a single genomic locus, as is the case of LAGE-1A and LAGE-1B which are variants of CTAG2, or SSX2B and SSX2A which are variants of SSX2.

Careful inspection of CT gene mRNA alignments revealed that 66 are virtually identical copies of each other, i.e. the mRNAs align with the same identity level to more than one locus (14 distinct genes, see website's link 'Gene annotations'). This information is available in the CTdatabase within the section termed 'Gene copies' under the 'Gene' tab.

In addition to nearly identical gene copies, many CT antigen genes have a common evolutionary origin. We grouped genes with >40% sequence identity as belonging to the same family [*blastp* (8), *E*-value <0.001, complexity filter off and percent identity normalized for alignment length as % identity × alignment length/length of the shorter protein]. This analysis resulted in 12 groups: CSAG, CT45, CT47, CTAG, CTAGE, GAGE, MAGE, NXF, SPANX, SSX, TSPY, XAGE1.

Using these groups as a guide, multiple alignments (non-edited) were made with Clustalx (9). Phylogenetic trees were inferred with the program MEGA 3 (10) using neighbor joining, pair wise deletion, JTT matrix and bootstrapping 100 times. The sub-families thus identified are reported in the 'Gene' tab under the section 'Phylogenetic relationships with CT genes'. Note that the family information provided by the CTdatabase is a first approach and should be used with caution.

As with the 'Gene' tab, the 'Protein' tab also contributes general information on the protein products of CTgenes, such as RefSeq accession numbers, names [from UniProt, (11)], and known protein domains. It also contains manually annotated sections on protein–protein interactions, protein localization and protein function.

THE 'mRNA Expression' Tab

CT antigens (or candidate CT antigens) were classified according to their expression patterns. Based on a collective analysis of data from CAGE, MPSS, RT-PCR and ESTs (see website for details), genes are considered to be: (a) testis-restricted, (b) testis/brain-restricted, or (c) testis-selective (Annotation field: 'Gene expression pattern').

Since the principal importance of CT antigens lies in their restricted expression in normal tissues and ample expression in cancers, a central feature of the

CTdatabase is mRNA expression data. These data are divided between 'High-throughput' (obtained using large-scale techniques), 'Tested by Ludwig Institute for Cancer Research' (RT-PCR) and 'Published literature' (manual annotation).

HIGH-THROUGHPUT DATA

Three different sources of data were utilized: Serial Analysis of Gene Expression [SAGE, (12)], Massive Parallel Signature Sequencing [MPSS, (13)] and Expressed Sequence Tags [EST, (14)].

ESTs are cDNA fragments of a hundred or more nucleotides (nt). SAGE data are comprised of 'short' or 'long' sequence tags, 10 nt and 17 nt, respectively, from the 3′-end of mRNAs. Massive Parallel Signature Sequencing tags are 13 nt tags obtained using an alternative sequencing protocol. In all cases, the number of EST, SAGE or MPSS tags reflect the number of mRNA copies in a cell; the higher the number of tags observed for a given gene, the higher the expression of the gene. The CTdatabase contains a heat map of color-coded expression levels of CT antigens. Genes not confirmed as CT antigens have their expression levels presented in the heat map, but are flagged as not being testis restricted.

SAGE AND MPSS

Both SAGE and MPSS tags are computationally predicted for each CT gene (mRNA RefSeq) using custom programs that simulate the SAGE/MPSS protocols. To do this, we located the 3′ most CATG site (the enzyme cleavage site used for SAGE) or GATC (the cleavage site used for MPSS) and extracted the putative downstream tag. To guarantee that the tag is derived from the 3′-end of a given CT antigen only mRNAs with a poly-A tail (>5 As) are used.

When a given tag is observed in more than one gene, it is not accepted as a bona-fide tag. When the tag belongs to gene copies, a warning is displayed, cautioning the user that the expression level may not be correctly reported.

The frequency of each predicted tag (expression level) in different tissues was downloaded from SAGE Genie (15) at the Ludwig Institute for Cancer Research (LICR) FTP site (ftp://ftp.licr.org/pub/databases/trome/human/) and parsed to generate a heat map. Only tissues with more than 100 000 tags are shown. Library annotation (normal/cancer, sample source, etc) were downloaded from SAGE Genie. A limited number of corrections were manually performed.

ESTs

The number of ESTs per gene contained in UniGene clusters with at least 60 000 sequences were normalized per million and also presented as a heat map. Tissue and health state annotation provided by UniGene were used to separate the cancer libraries presented. Normalized and subtracted libraries were excluded from the data to avoid sampling biases. Intronless ESTs were excluded to avoid bias from genomic DNA contaminations.

Some CT antigen genes had more than one corresponding UniGene cluster due to gene copies. For these cases, UniGene clusters were merged following manual inspection and a corresponding warning is displayed in the entry.

RT-PCR

In addition to third-party expression data, the CTdatabase provides an RT-PCR analysis of the expression levels of all CT antigen genes. A standardized analysis was undertaken in the same set of cDNA preparations from normal human tissues as well as selected human cancer cell lines. Thus the expression of 106 genes was analyzed in a panel of 22 normal tissues and 34 cancer cell lines by RT-PCR at the LICR New York Branch. Gel images are displayed and the experimental conditions, including primer sequences, PCR cycles and temperatures are provided.

LITERATURE DATA

Manually curated information retrieved from the literature is included in the CTdatabase. A list of normal tissues expressing the referred CT, as indicated by literature references, is shown. Data on expression of individual CT antigens in neoplasias were annotated and are presented according to tumor type and subtype, indicating the level of expression. A list of cell lines expressing each CT gene is also presented. For all literature information, the experimental method is provided as well as links to the PubMed references.

THE 'PROTEIN EXPRESSION' TAB

The protein expression tab includes manually reviewed information from the literature on CT protein expression in normal tissues, tumor tissues and tumor cell lines. The methodologies employed in the experiments and the PubMed references are provided. A list of antibodies raised against CT antigens as published in the literature is included.

THE 'IMMUNE RESPONSE' AND 'PubMed' TABS

The 'Immune response' tab is divided into three sections containing information manually curated from the literature respectively focused on 'Humoral immune response', 'Cellular immune response' and 'Induced immune response'. The first section contains data on spontaneous humoral immune responses to CT antigens in patients with different tumor types, including the frequency of patients with antibodies against the antigen where available. The technique used for detection of the antibodies and the PubMed article reference is shown for each entry. The 'Cellular immune response' section contains information on spontaneous cellular immune responses against the CT antigen in cancer patients. This tab also displays a table listing the peptides recognized by T cells extracted from the database at http://www.cancerimmunity.org/peptidedatabase/tumorspecific.htm. The 'Induced immune response' section lists the results, with links to PubMed, from published clinical trials in which cancer patients received CT antigen-based vaccines.

Finally, the tab 'PubMed' hosts all the references related to individual gene entries.

IMPLEMENTATION

CTdatabase runs on free software (MySQL database server, Apache WWW server with the interface written in PHP). Perl and shell scripts were written to parse downloaded data.

FUTURE DIRECTIONS

Most of the work done on CT antigens has focused on the immunologic aspects of this intriguing family of proteins. However, a central question that remains to be answered is whether CT antigen expression contributes to tumorigenesis or is a functionally irrelevant by-product of the process of cellular transformation. Initial structural and functional information on CT antigens indicate that CT antigen expression could have a fundamental role in human tumorigenesis. Knowledge on different aspects of CT antigens is continuously expanding and we plan to update the CTdatabase structure by adding new information as it appears in the literature. In addition, the high-throughput data will also be upgraded periodically.

FUNDING

Conselho Nacional de Desenvolvimento Científico e Tecnológico; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior; Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro. Funding for open access charges: Ludwig Institute for Cancer Research.

Conflict of interest statement. None declared.

REFERENCES

1. Vansteenkiste,J.F., Zielinski,M., Dahabreh,I.J., Linder,A., Lehmann,F., Gruselle,O., Therasse,P., Louahed,J. and Brichard,V.G. (2008) Association of gene expression signature and clinical efficacy of MAGE-A3 antigenspecific cancer immunotherapeutic (ASCI) as adjuvant therapy in resected stage IB/II non-small cell lung cancer (NSCLC). *J. Clin. Oncol.*, **26**, 7501.
2. Davis,I.D., Chen,W., Jackson,H., Parente,P., Shackleton,M., Hopkins,W., Chen,Q., Dimopoulos,N., Luke,T., Murphy,R. *et al.* (2004) Recombinant NY-ESO-1 protein with ISCOMATRIX adjuvant induces broad integrated antibody and CD4(+) and CD8(+) T cell responses in humans. *Proc. Natl Acad. Sci. USA*, **101**, 10697–10702.
3. van der Bruggen,P., Traversari,C., Chomez,P., Lurquin,C., De Plaen,E., Van den Eynde,B., Knuth,A. and Boon,T. (1991) A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science*, **254**, 1643–1647.
4. Old,L.J. and Chen,Y.T. (1998) New paths in human cancer serology. *J. Exp. Med.*, **187**, 1163–1167.
5. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
6. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **36**, D13–D21.
7. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2006) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
8. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–33402.
9. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
10. Kumar,S., Tamura,K. and Nei,M. (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform.*, **5**, 150–163.
11. UniProt Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
12. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
13. Brenner,S., Johnson,M., Bridgham,J., Golda,G., Lloyd,D.H., Johnson,D., Luo,S., McCurdy,S., Foy,M., Ewan,M. *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18**, 630–634.
14. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST-database for "expressed sequence tags". *Nat. Genet.*, **4**, 332–333.
15. Boon,K., Osorio,E.C., Greenhut,S.F., Schaefer,C.F., Shoemaker,J., Polyak,K., Morin,P.J., Buetow,K.H., Strausberg,R.L., De Souza,S.J. *et al.* (2002) An anatomy of normal and malignant gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 11287–11292.