Genomics 94 (2009) 153–160

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Identification of FAM46D as a novel cancer/testis antigen using EST data and serological analysis ☆

Fabiana Bettoni [a], Fernando Camargo Filho [a], Daniela M. Grosso [a], Pedro A.F. Galante [a], Raphael B. Parmigiani [a], Murilo V. Geraldo [a], Flávio Henrique-Silva [b], Sueli M. Oba-Shinjo [c], Suely K.N. Marie [c], Fernando A. Soares [d], Helena P. Brentani [d], Andrew J.G. Simpson [e], Sandro J. de Souza [a], Anamaria A. Camargo [a,*]

[a] Laboratory of Molecular Biology and Genomics, Ludwig Institute for Cancer Research, São Paulo Branch, Hospital Alemão Oswaldo Cruz, Rua João Julião, 245, 1 andar, São Paulo, SP 01323-903, Brazil
[b] Universidade Federal de São Carlos, Rodovia Washington Luis km 235, São Carlos, SP 13565-905, Brazil
[c] Department of Neurology, School of Medicine, University of São Paulo, Av Dr Arnaldo, 455, room 4110, São Paulo, Brazil
[d] Hospital A.C. Camargo, Rua Prof. Antonio Prudente, 211, São Paulo, SP 01509-010, Brazil
[e] Ludwig Institute for Cancer Research, Memorial Sloan-Kettering Cancer Center, New York, NY 10158, USA

ARTICLE INFO

Article history:
Received 19 February 2009
Accepted 11 June 2009
Available online 18 June 2009

Keywords:
Tumor antigens
Immunotherapy
Gene expression profiling
ESTs
FAM46D

ABSTRACT

Cancer/testis Antigens (CTAs) are immunogenic proteins with a restricted expression pattern in normal tissues and aberrant expression in different types of tumors being considered promising candidates for immunotherapy. We used the alignment between EST sequences and the human genome sequence to identify novel CT genes. By examining the EST tissue composition of known CT clusters we defined parameters for the selection of 1184 EST clusters corresponding to putative CT genes. The expression pattern of 70 CT gene candidates was evaluated by RT-PCR in 21 normal tissues, 17 tumor cell lines and 160 primary tumors. We were able to identify 4 CT genes expressed in different types of tumors. The presence of antibodies against the protein encoded by 1 of these 4 CT genes (FAM46D) was exclusively detected in plasma samples from cancer patients. Due to its restricted expression pattern and immunogenicity FAM46D represents a novel target for cancer immunotherapy.

Introduction

Cancer/testis (CT) genes are expressed in the germ cells of testis and fetal ovary and in different histological types of tumors [1–4]. So far, more than 83 distinct CT gene families have been described, most of them are located on the X chromosome [1]. Most CT genes have been shown to encode proteins that are immunogenic (CT Antigens or CTA), eliciting both humoral and cellular immune responses in cancer patients [1]. Due to their restricted expression pattern and immunogenicity, CTAs are considered promising candidates for the development of therapeutic cancer vaccines. Advanced clinical trials with CTA vaccines are currently underway to access their efficacy in delaying or preventing recurrence of melanoma and lung cancer following surgical removal of primary tumors [5–8].

However, some characteristics shared by this group of proteins may represent challenges to the development of a universal cancer vaccine. Firstly, the expression frequency of a specific CT antigen is highly variable between different types of tumors. Second, the expression of a given CTA within a tumor is highly heterogeneous, with tumors showing single positive cells or small groups of positive cells to others with a homogeneous expression pattern. Finally, tumors frequently develop mechanisms to evade the host immune system and antigen loss is frequently observed in tumors.

One way to overcome tumor heterogeneity and immune escape is to design polyvalent cancer vaccines containing epitopes derived from different CTAs. The identification of a large repertoire of CTAs that could be incorporated in polyvalent cancer vaccines is thus critical for improving current immunotherapy protocols. Most of the known CTAs were identified using immunological screening methods such as T-lymphocyte epitope cloning and SEREX (Serological analysis of cDNA expression libraries) [9–14]. However, due to their characteristic expression pattern, non-immunological techniques, relying on differential mRNA expression, have also been successfully used for the characterization of novel CT genes. Methods including RDA (Representational Difference Analysis), Differential Display and cDNA micro-

☆ Sequence data from this article have been deposited with the DDBJ/EMBL/GenBank Data Libraries under Accession Nos. EF537578, EF537579, EF537580, EF537581 and EF537582.

* Corresponding author. Fax: +5511 3141.1325.
E-mail address: anamaria@ludwig.org.br (A.A. Camargo).

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.06.001

arrays have been used in the identification of CT genes expressed in testis and different tumor types, such as *LAGE-1*/*CTAG2*, *CT10*/*MAGEC2*, *CTp11*, *SPANX-C1*/*SPANXA1*, *MMA-1A*/*DSCR8* and *DBL*/*MCF2* [15–19]. These CT genes were later shown to encode immunogenic proteins recognized by the immune system of cancer patients.

Computational approaches have also been used for the identification of CT genes. In a pioneer study, Scanlan et al. [20] mined the UniGene database for gene clusters composed of expressed sequence tags (ESTs) generated from normal testis and tumor-derived cDNA libraries. A total of 1325 gene clusters were initially selected as corresponding to putative CT genes based on their EST composition. After assessing the expression pattern of 73 candidate clusters by RT-PCR, three novel CT genes (*CT15*/*ADAM2*, *CT16*/*PAGE5* and *CT17*/*LIPI*) were identified.

Using a similar strategy, Chen et al. [21] analyzed MPSS (Massively Parallel Signature Sequencing) expression data from 32 normal human tissues and identified 1056 genes that were predominantly expressed in normal testis. Further evaluation using MPSS tags from tumor cell lines and EST data from a wide variety of tumors, identified 202 of these 1056 genes as candidates for encoding novel CT genes based on their expression pattern. RT-PCR validation was carried out for a subset of 166 intron-containing candidates using cDNAs from normal testis and 21 tumor cell lines. Twenty candidates were validated as novel CT genes based on their expression pattern.

Finally, in a recent work Hofmann et al. [22] have used an *in silico* approach that combined expression information from different platforms (including ESTs, MPSS and Cap-analysis of Gene Expression) to produce a comprehensive survey of the expression profile of 153 previously described CT genes. By applying this strategy in a genome-wide survey for novel CT genes they identified over 30 CT candidate genes, of which 3 (*PEPP-2*/*RHOXF2*, *OTOA* and *AKAP4*) were confirmed to be predominantly expressed in testis and tumor cell lines by RT-PCR.

In the present work, we combined *in silico* and experimental evaluation of gene expression with serological analysis of cancer patients to identify novel CTAs. Using this approach, we were able to identify 4 CT genes frequently expressed in different types of tumors. The presence of specific antibodies recognizing the protein encoded by 1 of these 4 CT genes (FAM46D) was exclusively detected in plasma samples from cancer patients. Due to its restricted expression pattern and immunogenicity FAM46D represents a novel target for cancer immunotherapy. An additional CT gene candidate (PASD1) was described during the expression evaluation step of this work as a novel CTA frequently expressed in B-cell lymphomas [23]. Information presented in this manuscript on the high frequency of PASD1 expression in different types of solid tumors and immunogenicity in patients with cervix tumors and glioblastomas also supports PASD1 as a promising target for cancer immunotherapy.

Results

EST composition of clusters corresponding to known CTAs

ESTs can be used to identify new genes [24–28], to construct gene-based physical maps [29,30], to compare and annotate genomes of different organisms [31–33] and to study different aspects of mRNA structure such as splicing [34–39] and polyadenylation variants [40–42]. ESTs can also be used to qualitatively and quantitatively determine gene expression profiles if one considers the tissue source of the cDNA libraries and the frequency of ESTs corresponding to a given transcript in a cDNA library, respectively. In the present work, we have used the publicly available human genome sequence to cluster ESTs derived from a same gene. We were then able to select clusters composed by ESTs derived from testis and tumor cDNA libraries that would represent novel CT genes.

To address possible limitations of our strategy, we first evaluated the EST tissue composition of clusters corresponding to 20 known CTAs (*MAGEA1, MAGEA3, MAGEA10, MAGEB1, CT7*/*MAGEC1, CT10*/*MAGEC2, NY-ESO-1, SSX1, SSX2, SSX4, CT16*/*PAGE5, HOM-TES-85*/*LUZP4, BRDT*/*CT9, CTp11*/*SPANXC, OY-TES-1*/*ACRBP, CTAGE1, CT15*/*ADAM2, CSAG2, IL13RA2, HCA661*/*E2F-like*/*TFDP3*). mRNA sequences corresponding to *NY-ESO-1* and *BRDT*/*CT9* aligned in more than one location in the genome, suggesting an event of gene duplication that resulted in two identical copies [43]. Therefore the 20 CTAs were represented in our Transcriptome Database (see Materials and methods) by 23 clusters all of them containing at least one EST.

CTA clusters were generally composed by a small number of ESTs. Fifty percent of the clusters contained less than 10 ESTs, confirming the restricted expression pattern of these CTAs. Surprisingly, 85% of the CTA clusters contained ESTs derived from normal tissues and only 39.1% of these clusters contained ESTs derived from testis cDNA libraries. Indeed, only 10% of the CTA clusters were composed exclusively by ESTs derived from testis and tumor cDNA libraries as one would expect by the characteristic expression pattern of CT genes. Moreover, the majority of the 23 CTA clusters contained ESTs derived from cDNA libraries of an unknown tissue source.

Improving the transcriptome database

The presence of ESTs derived from normal tissues in CTA clusters can be explained, in part, by the existence of large gene families for most CTAs with high similarity between all members and by the fact that some members of CTA gene families have a ubiquitous expression pattern that do not correspond to that of a characteristic CT gene. ESTs corresponding to members of these gene families align to multiple regions in the genome and are represented in different clusters, therefore explaining the presence of ESTs derived from normal tissues in known CTA clusters. To overcome this problem, multiple alignments were excluded from the Transcriptome Database by giving a quality score for each alignment as described in Materials and methods and by keeping in the database only the alignments with the highest score for each sequence.

The presence of a significant number of ESTs derived from normal tissues and unknown tissue source in CTA clusters can be also explained by an erroneous description of the tissue source of cDNA libraries submitted to GenBank. To overcome this problem we have manually curated cDNA library descriptions and contacted the groups responsible for submissions when information was misleading or unavailable (Supplementary material 1). We kept in our database only EST sequences derived from cDNA libraries with a reliable and informative description.

Selection of candidate clusters corresponding to novel CT genes

The EST composition of clusters corresponding to known CT genes was also used to define parameters for computational selection of additional clusters corresponding to novel CT genes. After the improvements made in the Transcriptome Database, we observed that 20% of the CT gene clusters were composed by ESTs derived from normal testis and tumor cDNA libraries, 10% of the CT gene clusters were composed only by testis ESTs and 5% were composed only by tumor ESTs. Moreover, 65% of CT gene cluster contained a small percentage of ESTs derived from normal tissues, probably due to the spurious expression of some of the CT genes in normal tissues as has been previously reported [22].

Based on this information, we generated 3 independent lists of candidates: i) clusters composed only by ESTs derived from testis cDNA libraries, ii) clusters composed by ESTs derived from testis and tumor cDNA libraries and iii) clusters composed only by ESTs derived from cDNA libraries of tumoral origin. Clusters composed mainly by ESTs derived from testis and tumors, but containing a small percentage of ESTs from normal tissues, were not selected because, after analyzing clusters corresponding to known CT genes, we were unable to establish a reliable cut-off for the presence of these ESTs.

We have also restricted our analysis to clusters that were composed by at least one spliced EST that aligned non-contiguously to the genome sequence because EST data contain a significant fraction of DNA contaminant sequences. It should be noted, however, that this was a conservative approach that might have excluded candidate clusters composed exclusively by 3′ unspliced ESTs. Although legitimacy of 3′ ESTs can be attested by the occurrence of a bona-fide polyadenylation signal and the presence of a poly(A) tail, we decided to undertake such a conservative approach because poly(A) tails are frequently trimmed from EST sequences submitted to dbEST and the identification of polyadenylation signals is usually hampered by the low quality of the sequences near poly(A) tails and to the low signal/background ratio in the identification of reliable polyadenylation signals.

A total of 1184 candidate clusters were computationally selected of which 830 were composed only by ESTs derived from normal testis cDNA libraries, 68 were composed by ESTs derived from normal testis and tumors and 286 were composed only by ESTs derived from tumors (Supplementary material 2). We then randomly selected a subset of candidate clusters for validation of their expression pattern by RT-PCR. Cluster size requirements were not imposed as criterion for cluster selection because, as revealed by our previous analysis, clusters corresponding to known CT genes were composed by a small number of ESTs and establishment of cluster size requirements would probably eliminate good CT candidates. Clusters selected for experimental validation were manually inspected to confirm the presence of conserved donor and acceptor splicing sites (GT/AG). A total of 23 candidates from

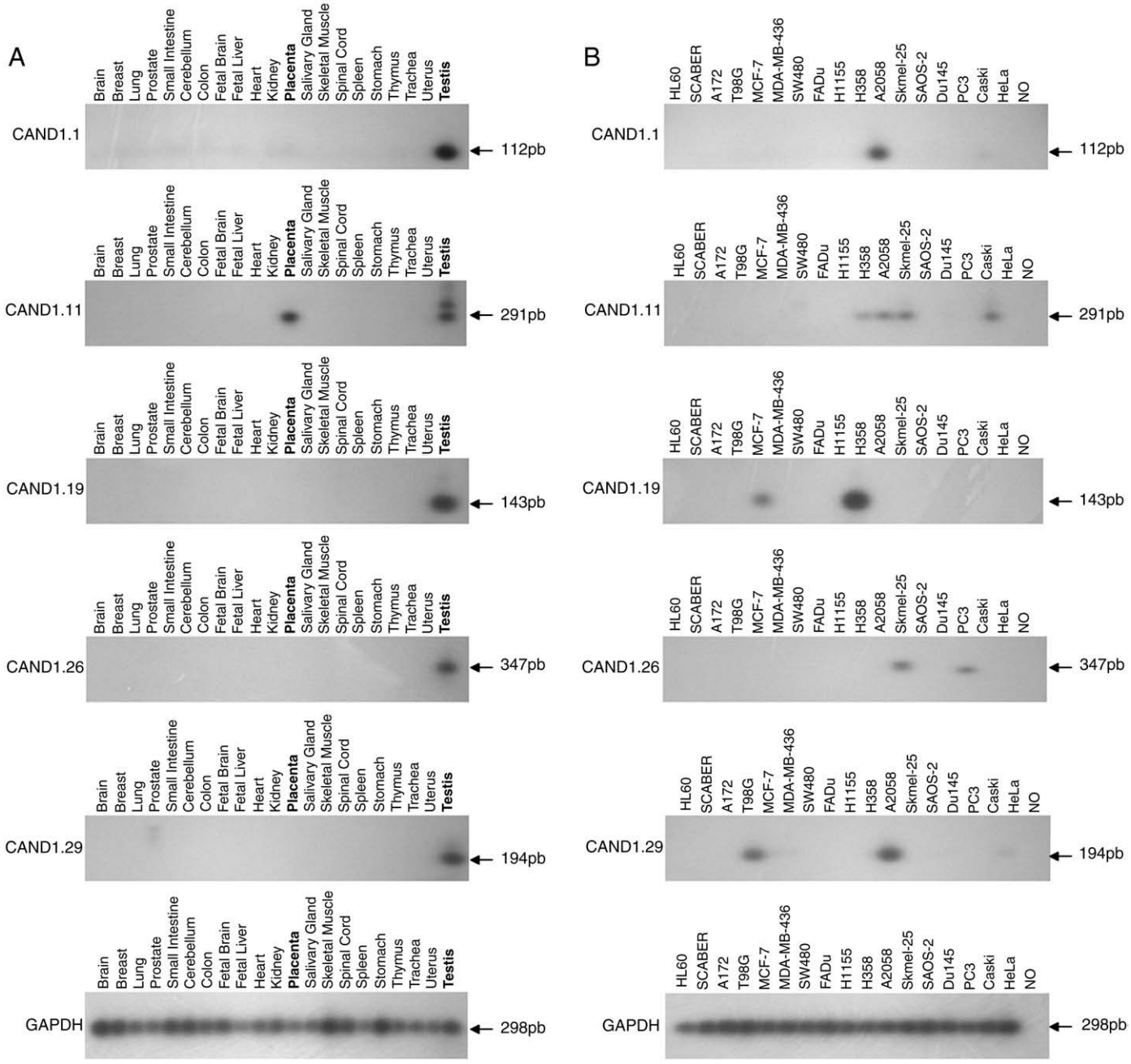

Fig. 1. mRNA expression pattern of the five candidates in normal tissues (A) and tumor cell lines (B). Southern blot of RT-PCR products amplified with each candidate specific primers. GAPDH amplification was used as positive control for cDNA synthesis. NO: cDNA-negative control.

each of the 3 lists were randomly selected for validation and we extended our analysis to 23 additional candidates from the testis only list, since all candidates validated in the initial round of RT-PCRs were derived from this particular list (Supplementary material 3).

Expression analysis of candidate clusters corresponding to novel CT genes

Expression pattern of candidate clusters was determined by RT-PCR using cDNA from 21 normal tissues, 17 tumor cell lines and 160 samples derived from 9 different types of tumors. RT-PCR conditions were standardized using testis cDNA as template. Few modifications including changes in the annealing temperature and $MgCl_2$ concentration and addition of PCR enhancers such as betaine were introduced when amplification was not achieved. Amplifications were achieved for 70 out of the 92 initially selected candidates and amplified fragments were sequenced to confirm their specificity (Supplementary material 3).

Six candidates (CAND1.1, CAND1.6, CAND1.19, CAND1.23, CAND1.26 and CAND1.29) were exclusively expressed in testis among all 21 normal tissues analyzed, five candidates (CAND1.8, CAND1.17, CAND1.25, CAND1.30 and CAND1.32) were expressed in testis and normal brain and one candidate (CAND1.11) was expressed in normal testis and placenta. The expression of six of these initial candidates (CAND1.1, CAND1.11, CAND1.17, CAND1.19, CAND1.26 and CAND1.29) with a restricted expression in normal tissues was also detected in at least one tumor cell line (Fig. 1).

During this expression evaluation step, CAND1.1 was described as a novel CTA (PASD1) frequently expressed in B-cell lymphomas and capable of eliciting humoral immune response in 40% of those patients [23]. Although we cannot claim the discovery of PASD1 as a novel CTA, we decided to evaluate its expression in different types of solid tumors

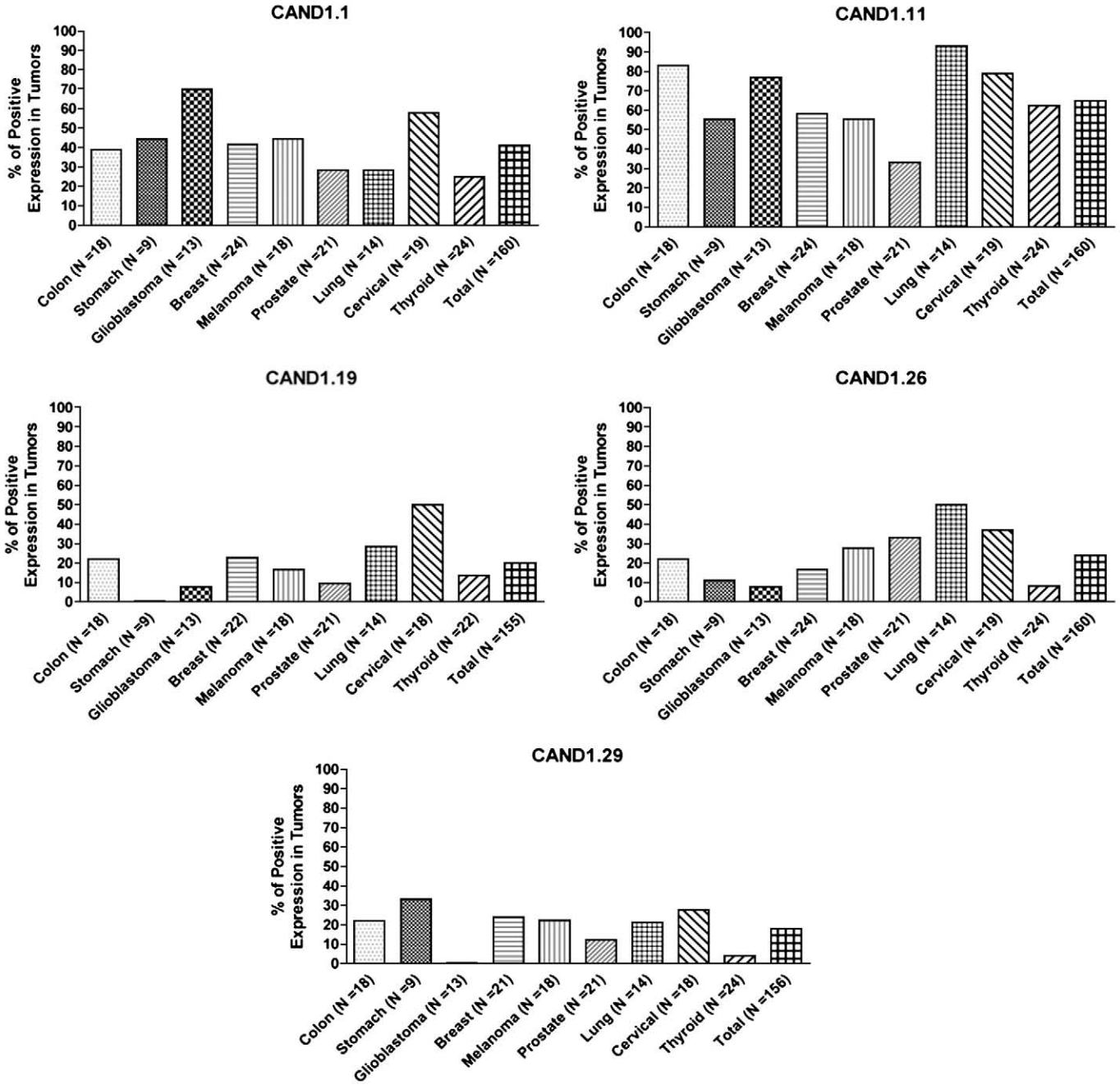

Fig. 2. Expression analysis of the 5 CT candidates in tumor samples.

to complement the information available in the literature, and to support the role of PASD1 as a therapeutic target for different tumor types other than B-cell lymphomas.

Expression analysis of PASD1 and the 4 CT candidates (CAND1.11, CAND1.19, CAND1.26 and CAND1.29) exclusively expressed in testis or testis and placenta was then carried out in 160 samples derived from 9 different histological types of tumors (Fig. 2). PASD1 was expressed in 41% (65/160) of tumor samples with the highest frequency of expression observed in glioblastomas (70%). CAND1.11 was expressed in 65% (104/160) of tumor samples with the highest expression frequency in lung tumors (93%). CAND1.19 was expressed in 20% (31/155) of tumor samples with the highest expression frequency in cervical tumors (50%). CAND1.26 was expressed in 24% (38/160) of tumor samples and was predominantly expressed in lung tumors (50%). CAND1.29 was expressed in 18% (28/156) of the tumor samples being more frequently expressed in gastric tumors (33.3%).

Candidate clusters corresponding to novel CT genes

Two (CAND1.19 and CAND1.26) out of the four CT genes are represented by a full-length mRNA sequence and code for a known human protein (Table 1). Candidate CAND1.19 corresponds to the *ASZ1* or *GASZ* gene (Germ cell-specific ankyrin, SAM and basic leucine zipper domain containing protein 1) that maps to chromosome 7q31.2 and codes for a predicted protein of 475aa containing four ankyrin repeats and a SAM (Sterile Alpha Motif) and a bZIP domains. The GASZ protein is localized in the cytoplasm of spermatocytes and oocytes at different stages of differentiation [44]. Based on its functional domains, GASZ may act as a signaling protein and/or transcriptional regulator during germ cell maturation and early embryogenesis. Candidate CAND1.26 corresponds to *FAM46D* gene (Family with sequence similarity 46, member D) mapping to chromosome Xq21.1 and encoding a 389aa protein with unknown function.

The remaining 2 CT genes (CAND1.11 and CAND1.29) were not represented by a full-length mRNA sequence at the time of this analysis and their partial 3′ sequences were extended towards their 5′ end using RACE (Rapid Amplification of cDNA Ends). Four distinct 5′ RACE fragments corresponding to CAND1.11 splicing variants were obtained (GenBank accession nos. EF537578, EF537579, EF537580, and EF537581) and a single RACE fragment (GenBank accession no. EF537582] was generated for CAND1.29. After assembling a consensus sequence for CAND1.11 and CAND1.29 using the original 3′ ESTs and the RACE fragments, no open reading frames (ORF) longer than 110aa were identified.

Serological analysis of PASD1 and CTA candidates

The presence of antibodies against the recombinant his-tagged proteins corresponding to PASD1, GASZ (CAND1.19) and FAM46D (CAND1.26) genes was then evaluated by ELISA in plasma samples from cancer patients. Samples from 19 lung cancer patients, 31 cervical cancer patients, 18 colorectal cancer patients, 25 glioblastoma patients and 50 healthy blood donors were used for the ELISA screening (Table 2). Antibodies against PASD1 and FAM46D were detected in 3.2 and 2.1% of all plasma samples from cancer patients, respectively. A humoral immune response against GASZ could not be detected among all plasma samples analyzed. Anti-PASD1 antibodies were detected in plasma samples from patients with cervical cancer (3.2%) and glioblastoma (8.0%) while anti-FAM46D antibodies were detected in samples from lung cancer (5.2%) and glioblastoma (4.0%) patients.

Table 1
Gene name and chromosome location corresponding to each of the five CT candidates.

Candidate	Reference sequence	Gene name	Chromosome region
CAND1.1	BI458651	*PASD1*	Xq28
CAND1.11	AI652043	–	11p15.4
CAND1.19	BG771896	*GASZ*	7q31.2
CAND1.26	BG722950	*FAM46D*	Xq21.1
CAND1.29	AA451827	–	Xq23

Table 2
Frequency of anti-PASD1, -GASZ and -FAM46D antibodies in plasma samples from cancer patients.

Tumor	Antibody response (%)		
	PASD1	GASZ	FAM46D
Cervix	1/31 (3.2)	0/31 (0.0)	0/31 (0.0)
Lung	0/19 (0.0)	0/19 (0.0)	1/19 (5.2)
Colon	0/18 (0.0)	0/18 (0.0)	0/18 (0.0)
Glioblastoma	2/25 (8.0)	0/25 (0.0)	1/25 (4.0)
Total	3/93 (3.2)	0/93 (0.0)	2/93 (2.1)

Discussion

CTAs are considered promising candidates for the development of therapeutic cancer vaccines. However, continued progress in the development of such vaccines will depend on the identification of a large repertoire of CTAs that could be incorporated in polyvalent cancer vaccines designed to overcome tumor heterogeneity and immune escape.

In the present work, we have used the publicly available human genome and ESTs sequences to identify novel CT genes. By examining the EST tissue composition of clusters corresponding to known CT antigens, we were able to detect and overcome initial limitations of our *in silico* approach. We believe that three major aspects of our *in silico* analysis have significantly contributed to improve our candidate selection and expression validation efficiency. The first critical aspect was the use of the human genome sequence as a scaffold for EST clustering, since it allowed us to exclude multiple alignments related to the existence of large gene families among CTAs and to correctly cluster ESTs corresponding to each family member.

Manual curation of cDNA library descriptions available in GenBank was the second important aspect of our work. Unfortunately, at the time of this analysis there was no common structured vocabulary to describe cDNA libraries and descriptions were generally poor and imprecise, leading to the misclassification of the libraries when considering the tissue source and pathologic characteristics. During the course of this work, Kelso et al. [45] developed eVOC, a controlled vocabulary for gene expression data that consists of four orthogonal controlled vocabularies describing the anatomical system, cell type, pathology and developmental stage for each cDNA library. Using eVOC the authors have curated and annotated 7016 cDNA libraries represented in dbEST. In attempt to complement and improve our own manual curation, eVOC annotation was integrated to our database, however no major improvements were observed in our candidate selection.

The analysis of EST composition of CTA clusters was the third important aspect of our analysis, since it allowed us to define broader parameters for the selection of additional EST clusters corresponding to CT gene candidates. As opposed to other *in silico* analysis, we did not limit our selection to clusters composed by ESTs derived from normal testis and tumors. Indeed, the selection of clusters composed exclusively of ESTs derived from normal testis was critical for our validation efficiency, since all CT genes we have identified in this work were represented in our Transcriptome Database by normal testis ESTs only.

Using this supervised *in silico* analysis, we were able to select a total of 1184 candidate clusters and the expression pattern of a subset of these clusters was validated by RT-PCR in normal tissues and in a large collection of tumor cell lines and tumor specimens. Our computational approach was conservative and this number is probably underestimated since we excluded clusters containing a small percentage of ESTs from normal tissues and clusters exclusively composed by bona-fide 3′ unspliced ESTs from our analysis.

Four CT genes with a restricted expression in normal tissues and broad expression in different types of tumors were identified. Southern blot analysis of RT-PCR products was used to improve the sensitivity and specificity of our analysis. Our four CT genes are frequently expressed in melanomas and lung tumors, which are classified as tumors with a high CT gene expression frequency [3,22], but they are also frequently expressed in tumors considered as moderate and low CT expressors such as breast, prostate and colon tumors, respectively.

Two of these validated CT genes (CAND1.11 and CAND1.29) were represented by partial 3′ sequences and were extended towards their 5′ end using RACE. After assembling a consensus sequence for CAND1.11 and CAND1.29 using the original 3′ ESTs and the RACE fragments, no open reading frames (ORF) longer than 110aa were identified. Although we cannot exclude the possibility that we still have partial sequences for these transcripts, it is tempting to speculate that these CT genes might belong to the growing class of non-coding regulatory RNAs. To our knowledge, these are the first descriptions of non-coding CT genes and despite the fact that they cannot be considered target for cancer immunotherapy the role of these non-coding CT genes in the tumorigenic process deserves further evaluation.

While this work was in progress, Liggins et al. [23] described PASD1 (CAND1.1) as a novel CTA eliciting humoral response in patients with diffuse large B-cell lymphomas (DLBCL). Subsequently studies have demonstrated the presence of anti-PASD1 antibodies in 35% (6/17) of sera from patients with acute myeloid leukaemia (AML) and 6% (1/16) of sera from patients with chronic myeloid leukaemia (CML) [46]. In addition, Guinn et al. [46] demonstrated that monocyte-derived dendritic cells electroporated with PASD1 mRNA could stimulate autologous T-cells to proliferate. The *PASD1* gene maps to chromosome Xq24–28. PASD1 predicted proteins have in common a PAS (Per ARNT Sim) domain, a putative leucine zipper and nuclear localization signal, suggesting that they may act as transcription factors. Although we cannot claim the discovery of PASD1 as a novel CTA, information presented in this manuscript on the high frequency of PASD1 expression in different types of solid tumors and immunogenicity in patients with cervix tumors and glioblastomas supports PASD1 as a promising target for cancer immunotherapy.

Finally, we were able to detect specific antibodies against the FAM46D protein (CAND1.26) in plasma samples from patients with lung tumors and glioblastomas. As many other CTAs, FAM46D is located on chromosome X and is a member of a large family of proteins with unknown function. Due to its restricted expression pattern and immunogenicity FAM46D represents a novel target for cancer immunotherapy.

To our knowledge, this is the first report that combined *in silico* and experimental evaluation of gene expression with serological analysis of cancer patients to identify novel CT antigens. Based on our results, FAM46D and PASD1 should be considered novel targets to the development of a cancer immunotherapy. Further experiments are ongoing to evaluate their function and immunogenicity in other tumor types.

Materials and methods

Transcriptome database

Human ESTs available at dbEST (dbEST release 070502) and mRNA sequences from UniGene release 153 were aligned to the masked human genome sequence (NCBI, build 29) by using pp-Blast [47], an implementation of MEGABLAST for a parallel cluster. The parameters used in MEGABLAST were: -f T -J F -F F -W 24. The MEGABLAST output was parsed and a MySQL database was loaded with alignments coordinates and identities. Information about the tissue origin of each expressed sequence was obtained after manual curation of cDNA library descriptions available at dbEST (Supplementary material 1) and was also uploaded in the database. Spurious and multiple alignments were excluded from the Transcriptome Database by using an additional set of alignment criteria. These included a minimum identity of 93%, coverage (percentage of sequence length aligned) greater than 45% for EST and coverage greater than 55% for mRNAs. Sequences mapping to more than one location on the genome were given a score for alignment quality. A higher score was associated with a higher identity and coverage. Only the alignments with the highest scores were kept in the database. Clustering of the expressed sequences was based on their genomic coordinates as described by Galante et al. [48]. Briefly, if two sequences shared at least partially the same gene structure (exon/intron boundaries) they were joined into the same cluster. If no exon/intron boundary was defined, a sequence had to have at least a 100-bp overlap with another sequence at the genome level to be added to the respective cluster.

Cluster selection and manual inspection

By querying the Transcriptome Database, we were able to select clusters composed by spliced ESTs derived from testis and/or tumoral cDNA libraries. Selected clusters were manually inspected to confirm the splicing structure by checking the presence of conserved donor and acceptor splicing sites (GT/AG) and to exclude those that correspond to known CT antigens. Manual inspection was carried out using the BLAT interface provided by UCSC Genome Browser (http://genome.ucsc.edu/).

Cell lines and tumor samples

Human tumor cell lines A172 and T98G (glioblastoma); FaDu (squamous cell carcinoma); SW480 (colorectal adenocarcinoma); A2058 and SKmel-28 (malignant melanoma); DU145 and PC3 (prostate carcinoma); HeLa and CasKi (cervix adenocarcinoma); MCF-7 and MDA-MB-436 (breast ductal carcinoma); HL-60 (acute promyelocytic leukaemia); H1155 and H358 (lung carcinoma); SCaBER (transitional cell carcinoma) and SAOS-2 (osteosarcoma) were obtained from American Type Culture Collection (Manassas, VA). Tumor samples were collected from patients treated at Hospital A.C. Camargo. All samples were collected after explicit informed consent and with local ethical committee approval.

RNA, cDNA synthesis and RT-PCR

Total RNA from cell lines and tumor samples was extracted by the conventional CsCl guanidine thiocyanate method [49]. Total RNA from 21 normal human tissues (testis, lung, prostate, small intestine, breast, brain, fetal brain, cerebellum, heart, uterus, placenta, colon, fetal liver, thymus, salivary gland, spinal cord, kidney, spleen, skeletal muscle, stomach and trachea) was purchased from Clontech (Palo Alto, CA). RNA samples were checked for integrity by agarose gel electrophoresis and 2 μg were used for cDNA synthesis. Reverse transcription was performed with DNA-free RNA by using SUPERSCRIPT II Reverse Transcriptase (Invitrogen) and oligo(dT) (GE Healthcare). Primers used for expression analysis (Supplementary material 2) were designed in different exons using the Primer3 program. RT-PCRs were carried out in 25 μl containing 1 μl of first-strand cDNA, 1× *Taq* DNA polymerase buffer (Invitrogen), 1 mM $MgCl_2$, 0.1 mM dNTPs (GE Healthcare), 1 U of *Taq* DNA polymerase (Invitrogen) and 0.3 μM of each primer. Amplification conditions were: initial denaturation for 4 min at 94 °C followed by 40 cycles of 45 s at 94 °C, 45 s at 60 °C and 1 min at 72 °C, with a final extension step of 6 min at 72 °C. RT-PCR was carried out in duplicates and triplicates were used to solve discrepancies between duplicates. RT-PCR products were analyzed on 8% silver-stained polyacrylamide gel and sequenced to confirm their specificity. RT-PCR products for candidates CAND1.1, CAND1.11,

CAND1.19, CAND1.26 and CAND1.29 were further analyzed on agarose gels followed by Southern blot analysis.

RACE (rapid amplification of cDNA ends)

5′-RACE was performed on normal testis poly(A)$^+$ RNA by using the Marathon cDNA Amplification Kit (Clontech, Palo Alto, CA). Amplifications reactions were performed in 25 μl by using 2.5 μl of cDNA, 0.2 mM dNTPs, 0.2 μM specific primer (RACE5′/CAND704, 5′ GATTTTCCAGACCTTGTCCAAGCTCC3′, RACE5′/CAND1001, 5′CACACA-GAATGCCGCTCGCTAGGAG3′), 0.2 μM adaptor primer, 1 U of Advantage *Taq* DNA polymerase. PCR conditions were: 5 cycles of 5 s at 94 °C and 3 min at 72 °C, 5 cycles of 5 s at 94 °C, 10 s at 70 °C and 3 min at 72 °C, and 25 cycles of 5 s at 94 °C, 10 s at 68 °C and 3 min at 72 °C. Nested PCR was carried out by using 5 μl of the first reaction product diluted 50×, specific nested primer (RACE5′/CAND704N, 5′ GTGGCCAACTGAGCTGCAGACTTCCC3′, RACE5′/CAND1001N, 5′ GATGTTTCCTCTTAGGCCTCTTCTCC3′) and nested adaptor primer. PCR conditions differ from the first reaction only by using 20 cycles. PCR products were analyzed on agarose gels, cloned and sequenced as described above.

Recombinant protein expression

CAND1.1 (773aa), CAND1.19 (475aa) and CAND1.26 longest ORFs (389aa) were amplified from normal testis cDNA using the following primers: CAND1.1F (5′- CGGAATTCATGAAGATGAGAGGGGAAAAG-3′) and CAND1.1R (5′- CCCAAGCTTTTAGCACGGCTTATTTGAGTC-3′); CAND1.19F (5′- GATATCATGGCGGCGAGCGCGCTGCGA-3′) and CAND1.19R (5′- CGGAATTCTTATTTCCTCTGGAAAGTTAG-3′); CAND1.26F (5′- CGGAATTCATGTCTGAAATCAGATTCACC-3′) and CAND1.26R (5′- CCCAAGCTTTTAACTCATACCATTTGATCC-3′). The PCR products corresponding to CAND1.1 and CAND1.26 were digested with EcoRI (New England Biolabs) and HindIII (New England Biolabs) and cloned into the expression vector pET28a (Stratagene, La Jolla, CA). The PCR product corresponding to CAND1.19 was digested with EcoRV (New England Biolabs) and EcoRI (New England Biolabs) and cloned into the expression vector pET32a (Stratagene, La Jolla, CA). After sequencing, recombinant plasmids were transformed into *Escherichia coli* BL-21 Rosetta. After induction with 0.4 mM isopropyl β-D-thiogalactoside at 37 °C for 4 h, recombinant proteins fused with a 6His-tag were purified by Ni^{2+} affinity chromatography by using the NiNTA agarose resin (Invitrogen). The purified protein was analyzed by Western blot using an anti-His-tag monoclonal antibody (GE Healthcare) to confirm the purification efficiency.

Antibody response in cancer patients

Plasmas were obtained from patients treated at Hospital A.C. Camargo. All samples were collected after explicit informed consent and with local ethical committee approval. In addition, plasma samples from 50 healthy individuals were collected from blood donors at the Hospital A.C. Camargo Blood Center. Antibodies against CAND1.1, CAND1.19 and CAND1.26 recombinant proteins were determined by direct ELISA (Enzyme-Linked ImmunoSorbent Assay). Microplates were coated with 50 μl of the recombinant protein (5 μg/ml) in phosphate buffered saline (PBS) pH 7.4 for 2 h at 37 °C. Wells were blocked with 5% (w/v) skim milk in PBS, washed 3 times with PBS-Tween (0.1%) overnight at 4 °C, and incubated with 50 μl serum diluted 1:100 in dilution buffer (5%(w/v) skim milk in PBS) for 2 h at 37 °C. After washing, 50 μl goat anti-human IgG conjugated to horseradish peroxidase (Sigma) diluted 1:5000 in PBS was used as secondary antibody for 1 h at 37 °C, followed by washing. Bound secondary antibody was visualized by the HRP enzymatic reaction using OPD tablets (Sigma). Absorbance at 492 nm was read on a SpectroCount analyzer and expressed as optical density (OD) values. All serum samples were analyzed in triplicates and corrected for background binding to an irrelevant protein from sugar cane expressed in the same expression system as CTs recombinant proteins. ELISA cut-off was established as the mean + 2SD of healthy blood donors. Experiments were carried out in duplicates and a serum sample was considered positive when reactivity was observed in two independent ELISA assays. Triplicates were used to solve discrepancies between duplicates.

Acknowledgments

We would like to thank Mônica C. Poli (Blood Center Hospital A.C. Camargo) for providing the plasma samples from the blood donors. We also thank Waleska K. Martins and Marilene H. Lopes who contributed materials essential for the study. This work was supported by the Ludwig Institute for Cancer Research, São Paulo, Brazil, by Fundação de Amparo à Pesquisa do Estado de São Paulo Grant and by Conselho Nacional de Desenvolvimento Científico e Tecnológico Grant. FB and PAFG were sponsored by fellowships from Fundação de Amparo a Pesquisa do Estado de São Paulo; RBP was sponsored by a fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior; MVG was sponsored by a fellowship from Conselho Nacional de Desenvolvimento Científico e Tecnológico.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2009.06.001.

References

[1] A.J. Simpson, O.L. Caballero, A. Jungbluth, Y.T. Chen, L.J. Old, Cancer/testis antigens, gametogenesis and cancer, Nat. Rev., Cancer 5 (2005) 615–625.
[2] M.J. Scanlan, A.O. Gure, A.A. Jungbluth, L.J. Old, Y.T. Chen, Cancer/testis antigens: an expanding family of targets for cancer immunotherapy, Immunol. Rev. 188 (2002) 22–32.
[3] M.J. Scanlan, A.J. Simpson, L.J. Old, The cancer/testis genes: review, standardization, and commentary, Cancer Immun. 4 (2004) 1.
[4] A.J. Zendman, D.J. Ruiter, G.N. Van Muijen, Cancer/testis-associated genes: identification, expression profile, and putative function, J. Cell. Physiol. 194 (2003) 272–288.
[5] D. Davis, W. Chen, H. Jackson, P. Parente, M. Shackleton, W. Hopkins, Q. Chen, N. Dimopoulos, T. Luke, R. Murphy, A.M. Scott, E. Maraskovsky, G. McArthur, D. MacGregor, S. Sturrock, T.Y. Tai, S. Green, A. Cuthbertson, D. Maher, L. Miloradovic, S.V. Mitchell, G. Ritter, A.A. Jungbluth, Y.T. Chen, S. Gnjatic, E.W. Hoffman, L.J. Old, J.S. Cebon, 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. U. S. A. 101 (2004) 10697–10702.
[6] M. Marchand, N. van Baren, P. Weynants, V. Brichard, B. Dreno, M.H. Tessier, E. Rankin, G. Parmiani, F. Arienti, Y. Humblet, A. Bourlond, R. Vanwijck, D. Lienard, M. Beauduin, P.Y. Dietrich, V. Russo, J. Kerger, G. Masucci, E. Jager, J. De Greve, J. Atzpodien, F. Brasseur, P.G. Coulie, P. van der Bruggen, T. Boon, Tumor regressions observed in patients with metastatic melanoma treated with an antigenic peptide encoded by gene MAGE-3 and presented by HLA-A1, Int. J. Cancer 80 (1999) 219–230.
[7] S. Adams, D.W. O'Neill, D. Nonaka, E. Hardin, L. Chiriboga, K. Siu, C.M. Cruz, A. Angiulli, F. Angiulli, E. Ritter, R.M. Holman, R.L. Shapiro, R.S. Berman, N. Berner, Y. Shao, O. Manches, L. Pan, R.R. Venhaus, E.W. Hoffman, A. Jungbluth, S. Gnjatic, L. Old, A.C. Pavlick, N. Bhardwaj, Immunization of malignant melanoma patients with full-length NY-ESO-1 protein using TLR7 agonist imiquimod as vaccine adjuvant, J. Immunol. 181 (2008) 776–784.
[8] P. Romero, Current state of vaccine therapies in non-small-cell lung cancer, Clin. Lung Cancer 9 Suppl. 1 (2008) S28–36.
[9] C. Traversari, P. van der Bruggen, B. Van den Eynde, P. Hainaut, C. Lemoine, N. Ohta, L. Old, T. Boon, Transfection and expression of a gene coding for a human melanoma antigen recognized by autologous cytolytic T lymphocytes, Immunogenetics 35 (1992) 145–152.
[10] P. van der Bruggen, C. Traversari, P. Chomez, C. Lurquin, E. De Plaen, B. Van den Eynde, A. Knuth, T. Boon, A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma, Science 254 (1991) 1643–1647.
[11] P. Boel, C. Wildmann, M.L. Sensi, R. Brasseur, J.C. Renauld, P. Coulie, T. Boon, P. van der Bruggen, BAGE: a new gene encoding an antigen recognized on human melanomas by cytolytic T lymphocytes, Immunity 2 (1995) 167–175.
[12] B. Van den Eynde, O. Peeters, O. De Backer, B. Gaugler, S. Lucas, T. Boon, A new family of genes coding for an antigen recognized by autologous cytolytic T lymphocytes on a human melanoma, J. Exp. Med. 182 (1995) 689–698.

[13] U. Sahin, O. Tureci, M. Pfreundschuh, Serological identification of human tumor antigens, Curr. Opin. Immunol. 9 (1997) 709–716.
[14] U. Sahin, O. Tureci, H. Schmitt, B. Cochlovius, T. Johannes, R. Schmits, F. Stenner, G. Luo, I. Schobert, M. Pfreundschuh, Human neoplasms elicit multiple specific immune responses in the autologous host, Proc. Natl. Acad. Sci. U. S. A. 92 (1995) 11810–11813.
[15] B. Lethe, S. Lucas, L. Michaux, C. De Smet, D. Godelaine, A. Serrano, E. De Plaen, T. Boon, LAGE-1, a new gene with tumor specificity, Int. J. Cancer 76 (1998) 903–908.
[16] A.O. Gure, E. Stockert, K.C. Arden, A.D. Boyer, C.S. Viars, M.J. Scanlan, L.J. Old, Y.T. Chen, CT10: a new cancer-testis (CT) antigen homologous to CT7 and the MAGE family, identified by representational-difference analysis, Int. J. Cancer 85 (2000) 726–732.
[17] A. Westbrook, A.B. Diekman, K.L. Klotz, V. Khole, C. von Kap-Herr, W.L. Golden, R.L. Eddy, T.B. Shows, M.H. Stoler, C.Y. Lee, C.J. Flickinger, J.C. Herr, Spermatid-specific expression of the novel X-linked gene product SPAN-X localized to the nucleus of human spermatozoa, Biol. Reprod. 63 (2000) 469–481.
[18] A.J. Zendman, M. Cornelissen, U.H. Weidle, D.J. Ruiter, G.N. van Muijen, CTp11, a novel member of the family of human cancer/testis antigens, Cancer Res. 59 (1999) 6223–6229.
[19] N.J. de Wit, U.H. Weidle, D.J. Ruiter, G.N. van Muijen, Expression profiling of MMA-1a and splice variant MMA-1b: new cancer/testis antigens identified in human melanoma, Int. J. Cancer 98 (2002) 547–553.
[20] M.J. Scanlan, C.M. Gordon, B. Williamson, S.Y. Lee, Y.T. Chen, E. Stockert, A. Jungbluth, G. Ritter, D. Jager, E. Jager, A. Knuth, L.J. Old, Identification of cancer/testis genes by database mining and mRNA expression analysis, Int. J. Cancer 98 (2002) 485–492.
[21] Y.T. Chen, M.J. Scanlan, C.A. Venditti, R. Chua, G. Theiler, B.J. Stevenson, C. Iseli, A.O. Gure, T. Vasicek, R.L. Strausberg, C.V. Jongeneel, L.J. Old, A.J. Simpson, Identification of cancer/testis-antigen genes by massively parallel signature sequencing, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 7940–7945.
[22] O. Hofmann, O.L. Caballero, B.J. Stevenson, Y.T. Chen, T. Cohen, R. Chua, C.A. Maher, S. Panji, U. Schaefer, A. Kruger, M. Lehvaslaiho, P. Carninci, Y. Hayashizaki, C.V. Jongeneel, A.J. Simpson, L.J. Old, W. Hide, Genome-wide analysis of cancer/testis gene expression, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 20422–20427.
[23] A.P. Liggins, B.A. Guinn, C.S. Hatton, K. Pulford, A.H. Banham, Serologic detection of diffuse large B-cell lymphoma-associated antigens, Int. J. Cancer 110 (2004) 563–569.
[24] M.D. Adams, M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelley, T.R. Utterback, J.W. Nagle, C. Fields, J.C. Venter, Sequence identification of 2,375 human brain genes, Nature 355 (1992) 632–634.
[25] M.D. Adams, A.R. Kerlavage, C. Fields, J.C. Venter, 3,400 new expressed sequence tags identify diversity of transcripts in human brain, Nat. Genet. 4 (1993) 256–267.
[26] S. Brown, J.L. Chang, W. Sadee, P.C. Babbitt, A semiautomated approach to gene discovery through expressed sequence tag data mining: discovery of new human transporter genes, AAPS PharmSci 5 (2003) E1.
[27] M.P. De Young, H. Damania, D. Scheurle, C. Zylberberg, R. Narayanan, Bioinformatics-based discovery of a novel factor with apparent specificity to colon cancer, In Vivo 16 (2002) 239–248.
[28] R.B. Parmigiani, F. Bettoni, M.D. Vibranovski, M.H. Lopes, W.K. Martins, W. Cunha, F.A. Soares, A.J. Simpson, S.J. de Souza, A.A. Camargo, Characterization of a cancer/testis (CT) antigen gene family capable of eliciting humoral response in cancer patients, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 18066–18071.
[29] B.E. Baysal, E.M. van Schothorst, J.E. Farr, M.R. James, P. Devilee, C.W. Richard 3rd, A high-resolution STS, EST, and gene-based physical map of the hereditary paraganglioma region on chromosome 11q23, Genomics 44 (1997) 214–221.
[30] T.J. Hudson, A.M. Colbert, M.P. Reeve, J.S. Bae, M.K. Lee, R.L. Nussbaum, M.L. Budarf, B.S. Emanuel, S. Foote, Isolation and regional mapping of 110 chromosome 22 STSs, Genomics 24 (1994) 588–592.
[31] Y. Lee, R. Sultana, G. Pertea, J. Cho, S. Karamycheva, J. Tsai, B. Parvizi, F. Cheung, V. Antonescu, J. White, I. Holt, F. Liang, J. Quackenbush, Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA), Genome Res. 12 (2002) 493–502.
[32] F. Mignone, A. Anselmo, G. Donvito, G.P. Maggi, G. Grillo, G. Pesole, Genome-wide identification of coding and non-coding conserved sequence tags in human and mouse genomes, BMC Genomics 9 (2008) 277.
[33] S. Tugendreich, D.E. Bassett Jr., V.A. McKusick, M.S. Boguski, P. Hieter, Genes conserved in yeast and humans, Hum. Mol. Genet. 3 (1994) 1509–1517.
[34] Q. Xu, C. Lee, Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences, Nucleic Acids Res. 31 (2003) 5635–5643.
[35] Q. Xu, B. Modrek, C. Lee, Genome-wide detection of tissue-specific alternative splicing in the human transcriptome, Nucleic Acids Res. 30 (2002) 3754–3766.
[36] P. Bonizzoni, G. Mauri, G. Pesole, E. Picardi, Y. Pirola, R. Rizzi, Detecting alternative gene structures from spliced ESTs: a computational approach, J. Comput. Biol. 16 (2009) 43–66.
[37] L. Hui, X. Zhang, X. Wu, Z. Lin, Q. Wang, Y. Li, G. Hu, Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment, Oncogene 23 (2004) 3013–3023.
[38] Z. Kan, E.C. Rouchka, W.R. Gish, D.J. States, Gene structure prediction and alternative splicing analysis using genomically aligned ESTs, Genome Res. 11 (2001) 889–900.
[39] H. Xie, W.Y. Zhu, A. Wasserman, V. Grebinskiy, A. Olson, L. Mintz, Computational analysis of alternative splicing using EST tissue information, Genomics 80 (2002) 326–330.
[40] E. Beaudoing, D. Gautheret, Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data, Genome Res. 11 (2001) 1520–1526.
[41] C. Iseli, B.J. Stevenson, S.J. de Souza, H.B. Samaia, A.A. Camargo, K.H. Buetow, R.L. Strausberg, A.J. Simpson, P. Bucher, C.V. Jongeneel, Long-range heterogeneity at the 3′ ends of human mRNAs, Genome Res. 12 (2002) 1068–1074.
[42] M. Kamasawa, J. Horiuchi, Identification and characterization of polyadenylation signal (PAS) variants in human genomic sequences based on modified EST clustering, In Silico Biol. 8 (2008) 347–361.
[43] S. Aradhya, T. Bardaro, P. Galgoczy, T. Yamagata, T. Esposito, H. Patlan, A. Ciccodicola, A. Munnich, S. Kenwrick, M. Platzer, M. D'Urso, D.L. Nelson, Multiple pathogenic and benign genomic rearrangements occur at a 35 kb duplication involving the NEMO and LAGE2 genes, Hum. Mol. Genet. 10 (2001) 2557–2567.
[44] W. Yan, A. Rajkovic, M.M. Viveiros, K.H. Burns, J.J. Eppig, M.M. Matzuk, Identification of Gasz, an evolutionarily conserved gene expressed exclusively in germ cells and encoding a protein with four ankyrin repeats, a sterile-alpha motif, and a basic leucine zipper, Mol. Endocrinol. 16 (2002) 1168–1184.
[45] J. Kelso, J. Visagie, G. Theiler, A. Christoffels, S. Bardien, D. Smedley, D. Otgaar, G. Greyling, C.V. Jongeneel, M.I. McCarthy, T. Hide, W. Hide, eVOC: a controlled vocabulary for unifying gene expression data, Genome Res. 13 (2003) 1222–1230.
[46] B.A. Guinn, E.A. Bland, U. Lodi, A.P. Liggins, K. Tobal, S. Petters, J.W. Wells, A.H. Banham, G.J. Mufti, Humoral detection of leukaemia-associated antigens in presentation acute myeloid leukaemia, Biochem. Biophys. Res. Commun. 335 (2005) 1293–1304.
[47] E.C. Osorio, J.E. de Souza, A.C. Zaiats, P.S. de Oliveira, S.J. de Souza, pp-Blast: a "pseudo-parallel" Blast, Braz. J. Med. Biol. Res. 36 (2003) 463–464.
[48] P.A. Galante, N.J. Sakabe, N. Kirschbaum-Slager, S.J. de Souza, Detection and evaluation of intron retention events in the human transcriptome, Rna 10 (2004) 757–765.
[49] J.M. Chirgwin, A.E. Przybyla, R.J. MacDonald, W.J. Rutter, Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease, Biochemistry 18 (1979) 5294–5299.