Identification of cancer/testis-antigen genes by massively parallel signature sequencing

Yao-Tseng Chen*†‡, Matthew J. Scanlan†, Charis A. Venditti†, Ramon Chua†, Gregory Theiler§, Brian J. Stevenson§, Christian Iseli§, Ali O. Gure†, Tom Vasicek¶, Robert L. Strausberg‖, C. Victor Jongeneel§, Lloyd J. Old†, and Andrew J. G. Simpson†

*Weill Medical College of Cornell University, New York, NY 10021; †Ludwig Institute for Cancer Research, New York Branch at Memorial Sloan–Kettering Cancer Center, New York, NY 10021; §Office of Information Technology, Ludwig Institute for Cancer Research and Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; ¶Lynx Therapeutics, Hayward, CA 94545; and ‖The Institute for Genomic Research, Rockville, MD 20850

Contributed by Lloyd J. Old, March 30, 2005

Massively parallel signature sequencing (MPSS) generates millions of short sequence tags corresponding to transcripts from a single RNA preparation. Most MPSS tags can be unambiguously assigned to genes, thereby generating a comprehensive expression profile of the tissue of origin. From the comparison of MPSS data from 32 normal human tissues, we identified 1,056 genes that are predominantly expressed in the testis. Further evaluation by using MPSS tags from cancer cell lines and EST data from a wide variety of tumors identified 202 of these genes as candidates for encoding cancer/testis (CT) antigens. Of these genes, the expression in normal tissues was assessed by RT-PCR in a subset of 166 intron-containing genes, and those with confirmed testis-predominant expression were further evaluated for their expression in 21 cancer cell lines. Thus, 20 CT or CT-like genes were identified, with several exhibiting expression in five or more of the cancer cell lines examined. One of these genes is a member of a CT gene family that we designated as CT45. The CT45 family comprises six highly similar (>98% cDNA identity) genes that are clustered in tandem within a 125-kb region on Xq26.3. CT45 was found to be frequently expressed in both cancer cell lines and lung cancer specimens. Thus, MPSS analysis has resulted in a significant extension of our knowledge of CT antigens, leading to the discovery of a distinctive X-linked CT-antigen gene family.

germ line | human | transcript

Massively parallel signature sequencing (MPSS) generates millions of short signature sequence tags from the 3′ regions of mRNA samples, the majority of which can be unambiguously assigned to individual genes, with the number of tags present for each gene being a digital readout of corresponding mRNA abundance (1). Because individual cells contain an estimated 200,000–300,000 transcripts, sampling at this level approaches the complete cataloging of all different mRNA species expressed in a cell line or tissue. Comparison of MPSS data sets derived from different tissues provides a powerful means of defining tissue-specific gene expression.

We have exploited MPSS to seek genes encoding previously unidentified cancer/testis (CT) antigens-proteins normally expressed only in germ cells, notably in testis but also expressed in various cancer cells, often as a result of gene derepression due to hypomethylation (2). Many such proteins are immunogenic in human patients and are potential targets for cancer vaccines. To date, >40 CT antigens have been identified by either immunological screening methods or expression database analysis (3). We have used MPSS tags from 32 normal human tissues as well as cell lines that express most known CT antigens. We have discovered more than a dozen CT and CT-like genes, including a previously uncharacterized CT-antigen gene family on the X chromosome.

Materials and Methods

Tumor Tissues and Cell Lines. Specimens of tumor tissues were obtained from the Departments of Pathology at the Weill Medical College of Cornell University and Memorial Sloan–Kettering Cancer Center. Cell lines were obtained from the bank maintained at the Ludwig Institute for Cancer Research, New York Branch.

MPSS. Pooled normal human tissue RNA preparations were purchased from Clontech. In addition, mRNA was purified from two cancer cell lines, SK-MEL-37 and SK-LU-17, by using standard protocols. After DNase treatment and isolation of poly(A)$^+$ RNA, these samples were used to generate cDNA libraries according to the Megaclone protocol (4), and signature sequences adjacent to poly(A)-proximal DpnII restriction sites were obtained by serial cutting and ligation of decoding adapters (1). Each signature comprised 17 nucleotides, including the DpnII recognition sequence (GATC). Between 2,000,000 and 3,000,000 tags were sequenced from each sample, in two reading frames offset by two nucleotides. Only signatures that were seen in two independent sequencing runs and present at a minimum of five transcripts per million in at least one sample were retained for analysis.

The mapping of signatures to human transcripts was performed, as described in ref. 5, by using the National Center for Biotechnology Information (NCBI) assembly 33 of the human genome. Sequence polymorphisms present in EST sequences but not in the genomic reference sequence were taken into account for the mapping. Signatures that unambiguously matched transcribed regions were retained. Counts were pooled when multiple signatures mapped to the same gene.

***In Silico* Analysis.** To identify candidate CT genes from 1,056 MPSS-defined testis-specific genes, the expression profile of each gene in normal and tumor tissues was further evaluated by using a combination of the SAGE ANATOMIC VIEWER and its VIRTUAL NORTHERN tool (http://cgap.nci.nih.gov/SAGE/AnatomicViewer) and database searches by using BLASTN (www.ncbi.nlm.nih.gov/BLAST). The focus of the analysis was to identify UniGene clusters containing ESTs derived from testis as well as from non-germ-cell tumors and with limited expression in somatic tissues. For genes of interest, transintronic primers for RT-PCR were designed.

For some genes, e.g., CT45 (see below), the NCBI web site was used for protein-similarity searches, the identification of conserved domains, chromosomal localization, the location of DNA contigs, and transcript/protein prediction. The MyHits database (http://myhits.isb-sib.ch) was used to explore potential protein domains. Gene identifiers were retrieved from the Ensembl database (www.ensembl.org), to maintain a consistent naming convention, and short names were assigned to each previously uncharacterized gene

Abbreviations: CT, cancer/testis; MPSS, massively parallel signature sequencing.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AY743709–AY743714).

‡To whom correspondence should be addressed. E-mail: ytchen@med.cornell.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0502583102

identified in the project, using Human Gene Nomenclature Committee (HGNC)-approved symbols whenever possible.

Qualitative RT-PCR. A normalized cDNA panel was used that comprises brain, colon, heart, kidney, leukocytes, liver, lung, ovary, pancreas, placenta, prostate, skeletal muscle, small intestine, spleen, thymus, and testis [multiple tissue cDNA panels (MTC) I and II, BD Biosciences]. For evaluating the expression in tumor cell lines, RNA was prepared by the standard guanidinium thiocyanate/CsCl-gradient method. Total RNA (2 μg) was used for each 20-μl reverse-transcription reaction, and 2 μl of cDNA was used per 25-μl PCR. PCR was performed by using the Invitrogen Platinum *Taq* Supermix, with 35 cycles each consisting of 15 sec at 94°C, 1 min at 60°C, and 1 min at 72°C. PCR products were visualized on 1% agarose gels by ethidium bromide staining.

Quantitative RT-PCR. Quantitative RT-PCR was performed by using a PRISM 7000 sequence detection system (Applied Biosystems). Normal testis RNA was obtained from Ambion (Austin, TX). RNA from tumor tissue was prepared by using TRIzol reagents (Life Technologies). Two micrograms of total RNA was used per 20-μl reverse-transcription reaction, and 2 μl of cDNA was used for each 25-μl PCR. Reactions were in duplicate, and the level of expression was determined relative to the testicular preparation. A standard curve was established for each PCR plate by using testicular cDNA in 4-fold serial dilutions. Forty-five two-step cycles of amplification were undertaken, each cycle consisting of 15 sec at 95°C and 1 min at 60°C. The RNA quality of the cell lines and tissues was evaluated by amplification of β-glucuronidase and GAPDH. All specimens included in the final analysis had cycle time (Ct) values differing by fewer than four cycles, indicating similar qualities and quantities of the cDNAs used.

Results

Identification of Candidate CT Genes. MPSS data were obtained from 32 normal human tissues, including two separate preparations of testis and placenta and two CT-rich cell lines, SK-MEL-37 and SK-LC-17. Genes were considered to have testis-predominant expression when the number of corresponding MPSS tags in the testis was at least 2 times greater than the combined number of tags in all somatic tissues. A total of 1,056 such testis-predominant genes were identified, of which 39 are located on chromosome X, which is known to contain many CT-antigen genes (3). Nine these 39 genes encode known CT antigens, *NY-ESO-1*, *LAGE1*, *CT10*, and *MAGE-B1*, *-B2*, and *-B4*, *GAGE1*, *GAGE2*, and *PAGE5*, demonstrating that this approach can potentially identify new genes encoding CT antigens. Other CT-antigen genes in the 1,056 gene list included *SCP1* (chromosome 1), *CT9/BRDT* (chromosome 1), *OY-TES-1/ACRBP* (chromosome 12), *ADAM2* (chromosome 8), *ADAM21* (chromosome 14), and *TPTE* (chromosome 21).

The 1,041 genes that did not correspond to known CT genes were analyzed by using the MPSS data from SK-MEL-37 and SK-LC-17 as well as ESTs from the public database. Candidate CT genes were taken as those with ESTs or MPSS tags from cancer tissues or cell lines (excluding germ cell or testicular tumors) and where ESTs were not found in more than two normal somatic tissues, excluding fetal tissues and pooled tissues. Pooled tissues were excluded because they often include testis, and fetal tissue was excluded because its capacity to express CT antigens has yet to be determined.

Based on these criteria, 202 genes were identified, of which 36 were found to be intronless genes and were excluded from further analysis. Transintronic primers were designed for the remaining 166 genes.

mRNA Expression of CT-Candidate Genes in Normal Tissues and Cell Lines. The presence of mRNA corresponding to the 166 selected genes in normal tissue was evaluated by using the cDNA panel derived from normal tissues (see *Materials and Methods*). Successful RT-PCR amplifications were achieved for 144 of the 166 genes, of which 41 exhibited expression in the majority of tissues tested, 32 exhibited selective expression but were in three or more somatic tissues, and 71 exhibited expression only in testis, ovary, and/or placenta (41 of 71), or in these tissues and no more than two other somatic tissues (30 of 71).

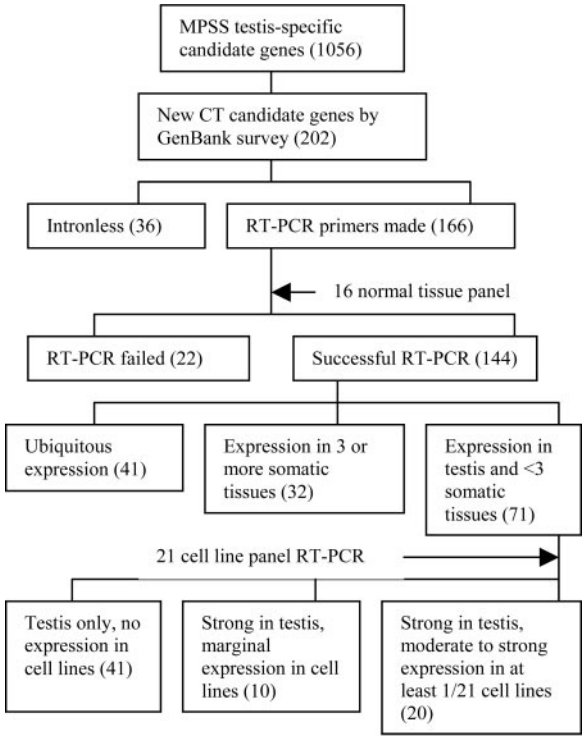

Fig. 1. Schematic summary of the genes analyzed at each step of this study. Gene numbers in each category are shown in parentheses.

The expression of the 71 genes with testis-predominant expression was evaluated by RT-PCR in 21 cancer cell lines: 7 derived from melanoma (SK-MEL-10, -24, -37, -49, -55, -80, and -128), 4 from small-cell lung cancer (NCI-H82, -H128, -H187, and -H740), 3 from non-small-cell lung cancer (SK-LC-5, -14, and -17), 3 from colon cancer (SW403, SW480, and LS174T), 1 from renal cancer (SK-RCC-1), 1 from hepatocellular carcinoma (SK-HEP-1), 1 from bladder cancer (T24), and 1 from sarcoma (SW982). Each of these cell lines expresses at least one known CT gene (data not shown).

The 71 genes fell into three groups, based on their expression in the cancer cell lines used. Forty-one genes exhibited no detectable expression in any of the cell lines, 10 exhibited only very low-level expression (relative to expression levels in testis) in one or more cell lines, and 20 exhibited moderate to strong expression in at least one cell line (Fig. 1, Table 1, and Table 2).

Quantitative RT-PCR of Selective CT Genes in Tumor Specimens. Of the 20 CT-like genes, 7 showed expression in at least five cell lines (Table 2). Two of these, ENSG00000117148 (LOC81569, UniGene Hs.2149; NM_030812) and ENSG00000140481 (Hs.383206; *FLJ32855*), exhibited strong expression in the pancreas and lung, respectively, limiting their potential utility as vaccines. These two genes may indeed encode differentiation antigens of the pancreas and lung, respectively, with concurrent testicular expression. This view is strengthened by the observation that four of five cancer cell lines expressing *FLJ32855* were small-cell lung-cancer lines and that 4 of 16 ESTs corresponding to this gene were derived from lung, the remaining being from testis, placenta, or brain. In comparison, ESTs derived from NM_030812 were found in brain and cervix in

Table 1. CT-candidate genes and their expression in normal tissues

Ensembl ID	Gene name	UniGene no.	Accession no.	Chromosome	Expression in normal tissues by RT-PCR
ENSG00000105549	*THEG*	Hs.250002	NM_016585	19	Testis only, strong expression, 2 alternative spliced forms
ENSG00000117148	*LOC81569*	Hs.2149	NM_030812	1	Strong in testis and pancreas, weak in placenta
ENSG00000187262	*MGC27005*	Hs.460933	NM_152582	X	Testis only, strong expression, 3 alternative spiced forms
ENSG00000160505	*NALP4*	Hs.351637	NM_134444	19	Strong in testis and ovary, weak in pancreas
ENSG00000133247	*COXVIB2*	Hs.329540	NM_144613	19	Strong in testis, weak in thymus, heart
NA	*LOC348120*	Hs.116287	BC047459	15	Testis only, 2 alternative spliced forms
ENSG00000140481	*FLJ32855*	Hs.383206	NM_182791	15	Strong in testis, lung, moderate in placenta, weak in ovary
ENSESTG00000023728	*LOC196993*	Hs.97823	BC048128	15	Testis only, strong expression
ENSG00000166049	*LOC139135*	Hs.160594	NM_173493	X	Testis only, strong expression
NA	*IMAGE164099*	Hs.408584	BX103208	3	Testis only, strong expression
ENSG00000104804	*TULP2*	Hs.104636	NM_003323	19	Testis only, strong expression
ENSESTG00000013526	*IMAGE1471044*	Hs.362492	AA884595	7	Testis only, strong expression
ENSESTG00000024371	*FLJ25339*	Hs.411239	BC057843	16	Testis only, strong expression, 2 alternative spliced forms
ENSG00000151962	*MGC271016*	Hs.133095	NM_144979	4	Testis only, strong expression
NA	*IMAGE4837072*	Hs.371922	BC040308	6	Testis only, strong expression
NA	*IMAGE5173800*	Hs.121221	BI818097	9	Strong in testis, weak in pancreas
ENSG00000101448	*SPINLW1*	Hs.121084	NM_181502	20	Testis only, strong expression
ENSG00000178093	*SSTK*	Hs.367871	NM_032037	19	Strong in testis, weak (+/−) in multiple tissues
ENSG00000168594	*ADAM29*	Hs.126836	NM_014269	4	Testis only, strong expression
ENSG00000173421	*LOC339834*	Hs.383008	NM_178173	3	Testis only, strong expression

NA, not available.

addition to testis, indicating that this gene is probably expressed in at least a few somatic tissues.

The other five genes ENSG00000105549 (*THEG*), ENSG00000187262 (*MGC27005*), ENSG00000160505 (*NALP4*), ENSG00000160471 (*COXVIB2*), and *LOC348120* (no Ensembl identifier; UniGene Hs.116287) are previously unidentified CT genes. Their expression was then measured in 29 lung tumors and 11 breast tumors by real-time RT-PCR (Fig. 2 and Table 3).

THEG, NALP4, COXVIB2,* and *LOC348120. *THEG* is the human ortholog of mouse *Theg* (testicular haploid expressed gene) (6). RT-PCR and DNA sequencing indicated that both known splice variants of 379

Table 2. Expression of CT-candidate genes in cell lines by qualitative RT-PCR

Cell line	*THEG*	*LOC81569*	*MGC27005/CT45*	*NALP4*	*COXVIB2*	*LOC348120*	*FLJ32855*	*LOC196993*	*LOC139135*	*IMAGE164099*	*TULP2*	*IMAGE1471044*	*FLJ25339*	*MGC271016*	*IMAGE4837072*	*IMAGE5173800*	*SPINLW1*	*SSTK*	*ADAM29*	*LOC339834*
SK-Mel-10	++	+++	++	+	+	−	−	−	++	+	+	+	−	−	+	−	−	−	−	+
SK-Mel-24	+++	+	+	+	−	+	−	−	−	+	+	+	−	+	−	+	−	−	−	−
SK-Mel-37	++	+++	+++	+++	−	+++	−	+	+++	−	+	−	−	−	++	+	+	−	−	++
SK-Mel-49	−	++	++	+++	−	++	+	−	−	−	+	−	−	−	−	+	−	+	+	−
SK-Mel-55	++	++	+++	++	−	−	+	−	−	+	+	+	+	−	++	−	−	+	−	−
SK-Mel-80	+	+	−	−	++	−	+++	+	−	−	+	+	−	−	−	+	+	−	−	−
SK-Mel-128	++	++	++	−	+++	+++	−	+	−	−	+	−	−	−	−	+	++	+	−	−
NCI-H82	+++	+++	+	+	+	−	++	+	−	+++	+	++	−	−	−	+	−	−	−	−
NCI-H128	+++	−	−	−	++	−	+++	+	−	−	+	+	++	−	−	+	−	++	−	−
NCI-H187	+++	+	−	+++	−	−	+++	−	−	−	+++	+	−	−	−	+	−	−	++	−
NCI-H740	++	−	−	−	+	−	+++	++	−	−	−	−	−	−	−	+	−	−	+	−
SK-LC-5	+++	+	+	+++	+	+++	+	++	−	++	+	−	−	−	−	+	−	−	−	−
SK-LC-14	++	++	+++	+	+	−	+	−	−	++	+	+++	−	+++	−	+	−	−	−	−
SK-LC-17	−	++	++	−	−	−	−	−	++	−	−	+	−	+++	−	+	−	−	−	−
HCT15	+++	++	++	++	+	−	−	+	−	−	++	−	++	−	−	++	−	−	−	−
LS174T	−	++	−	+++	++	+++	−	−	−	−	−	−	−	−	−	−	−	−	−	−
SW403	+	+++	−	+	+	+	−	−	−	−	+	−	−	−	−	−	−	−	−	−
SW982	++	−	+	−	−	−	−	+	−	+	−	−	−	−	−	−	−	−	−	−
SK-Hep-1	−	−	−	−	++	−	−	−	−	−	−	+	−	−	−	+	+	−	−	−
SK-RCC-1	++	+	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−
T24	++	−	+	−	−	−	−	++	−	−	−	−	−	−	−	−	−	−	−	−
Testis	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++
Key																				
++ to +++	15	11	8	7	5	5	5	3	3	3	2	2	2	2	2	1	1	1	1	1
+	2	5	5	5	7	2	4	7	0	4	12	8	1	1	1	14	3	3	2	1
−	4	5	8	9	9	14	12	11	18	14	7	11	18	18	18	10	17	17	18	19

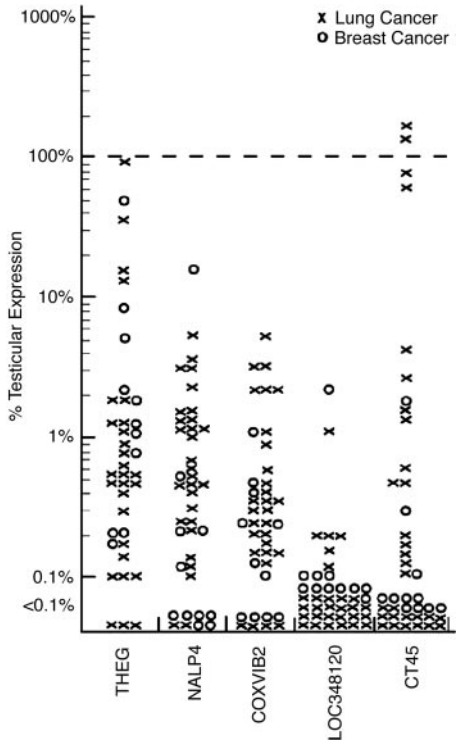

Fig. 2. Expression of five CT-antigen genes in lung and breast cancer. mRNA expression of the CT genes in non-small-cell lung cancer (X) and in breast cancer (O) was examined by real-time RT-PCR. Each symbol represents one case. The dashed line at 100% indicates the testicular level of expression with which tumor expression levels were compared.

and 344 aa are expressed in testis and in cancer. By real-time RT-PCR, expression was detected in 4 of 29 lung tumors and 1 of 11 breast tumors at >10% of the testicular level and in 9 of 29 and 7 of 11, respectively, at >1% of testicular expression.

NALP4 encodes a protein of 994 residues and contains the NTPase NACHT domain, found in apoptosis-associated proteins and in proteins involved in the transcriptional activation of the major histocompatibility genes and leucine-rich repeats probably involved in protein–protein interactions (7). Both *NALP4* and *NALP7* were identified in this study as possible CT genes. *NALP7*, however, was found to be expressed only weakly in three cell lines. In contrast, *NALP4* was expressed in seven of the lines with moderate to strong intensity. Furthermore, 11 of 29 lung tumors and 1 of 11 breast tumor specimens expressed *NALP4* at >1% of testicular expression. However, in only one breast cancer sample was expression detected at >10% of testicular expression.

COXVIB2 encodes testis-specific cytochrome *c* oxidase subunit VIb (8) and was expressed in 7 of 29 lung tumors, and 1 of 11 breast tumors expressed *COXVIB2* at >1% of the testicular level of expression, but in none was it expressed at >10% of testicular expression levels.

LOC348120 encodes a hypothetical protein of 117 aa that has no identifiable functional domains but shows significant similarity to the mouse *TLR11* (toll-like receptor 11) gene. It was expressed in only 1 of 29 lung tumors, and 1 of 11 breast tumors expressed *LOC348120* at >1% of the testicular level of expression, but none exhibited expression at >10% of testicular expression levels.

A Distinctive CT-Multigene Family on Xq26. The transcript of *MGC27005* (Hs. 460933, NM_152582) maps to chromosome Xq26.3 and was found to be expressed in 13 of 21 cell lines tested, with 8 of 13 showing moderate to strong expression. As measured by quantitative RT-PCR, the expression level in these eight cell lines

Table 3. Primer and probe sequences for quantitative RT-PCR of CT genes

Gene	Primer/ Probe	Sequence
THEG	Forward	CCAAAACCCAAGCCACATGT
	Reverse	GCACTTGTCCGACTGAGCTTT
	Probe	Fam-CAGACCATAACCGCCCTCCTTCACTTGG-Tamra
NALP4	Forward	TTGTCACCTCTCACCCATTGATT
	Reverse	CAGGATACATTCAGATACGTCAGCTT
	Probe	Fam-TGAAGTCCTTGCTGGCCTTCTAACCAACA-Tamra
COXVIB2	Forward	CCGTAACTGCTACCAGAACTTCCT
	Reverse	AGTGGTACACGCGGAAATAGTACTC
	Probe	Fam-ACTACCACCGCTGCCTCAAGACCAGG-Tamra
LOC348120	Forward	TGGATTCCAATTCATCTGACTACAG
	Reverse	CTTCCGCTTACCTCCAACTGA
	Probe	Fam-CTGCAGGTGATTCATTTGCAAGGTAAGCTG-Tamra
CT45	Forward	CTCTGCCATGTCCAAAGCAA
	Reverse	AAGTCATCAATCTGAGAATCCAATTG
	Probe	Fam-AAGCTTATGACAGGACATGCTATTCCACCCA-Tamra

Fam, 6-carboxyfluorescein; Tamra, *N,N,N′,N′*-tetramethyl-6-carboxyrhodamine.

ranged from 0.0168 to 16.2 times that in the testis. By real-time RT-PCR, 4 of 29 lung cancers (but none of the 11 breast cancers) expressed *MGC27005* at >10% of the testicular level of expression, whereas 8 of 29 and 1 of 11 of the lung and breast tumor specimens, respectively, showed *MGC27005* expression at levels >1% of the testicular expression level.

Comparison of the full-length *MGC27005* sequence (GenBank accession no. NM_152582.3) with the human genome by BLASTN identified six complete copies of extremely similar genes on chromosome X (nucleotides 133550000–133700000 on the ENSEMBL genome browser), with five having previously assigned Ensembl gene entries: ENSG00000187262, ENSG00000187264, ENSG00000187265, ENSG00000187267, and ENSG00000187245. This gene family is hereby designated CT45, following the CT nomenclature that we have proposed (3). All six CT45 gene members are products of recent gene-duplication events that differ by only 2 to 12 bp within their respective 1.0-kb transcript sequences (submitted as GenBank accession nos. AY743709 to AY743714) Thus, the CT45 transcripts detected by RT-PCR represent the accumulated expression of the CT45 gene family. Each gene spans 8–9 kb, and the genes are located in tandem within a 125-kb region (Fig. 3). The three centromeric genes are transcribed in the centromeric to telomeric direction, whereas the three telomeric genes are transcribed in the opposite direction.

An intronless copy of CT45 was identified on chromosome 5 that corresponds to the cDNA sequence of transcript variant 2 (see below), indicating that this copy on chromosome 5 is a retrogene. Although the ORF in this gene utilizes the same translational initiation site as CT45, there is a premature termination codon, resulting in a truncated 160-aa protein (versus 189-aa). This copy of CT45 on chromosome 5 is likely to be a pseudogene, which may or may not be transcribed.

In addition to these complete copies, several partial gene copies were identified within the Xq26.3 region resulting from failed duplication events, as has also been observed for other CT-gene families, such as SSX, on chromosome X (9).

Two transcript variants of CT45 were identified by aligning EST sequences against the full-length CT45 mRNA sequence (GenBank accession no. NM_152582). RT-PCR analysis and DNA sequencing confirmed both transcripts in testis and in cell lines and also identified a third transcript variant (Fig. 3). All three transcripts are derived from five exons but with exon 1 consisting entirely of a 5′ untranslated sequence varying between 85 and 256 bp. The CT45

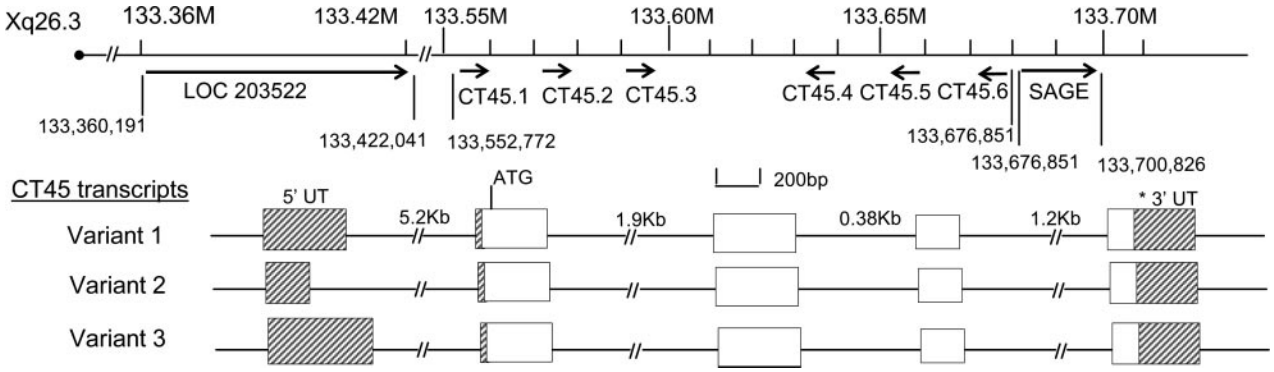

Fig. 3. CT45 gene family and transcript variants. The location of the six members of the CT45 gene family and their relation to the *LOC203522* and *SAGE* genes on the X chromosome are shown. Transcriptional orientations are indicated by arrows. The three transcriptional variants of CT45 are shown schematically, with boxes indicating exons. Untranslated regions (UT) at the 5′ and 3′ ends are shown as shaded boxes. Translational initiation sites (ATG) and termination codons (*) are indicated.

transcripts thus comprise a 5′ untranslated region ranging from 90 to 261 bp, a coding region of 570 bp, and a 3′ untranslated region of 292 bp, excluding the poly(A) tail. The CT45 protein consists of 189 aa with sequence similarity to known gene products restricted to the C-terminal 120 aa. Interestingly, the genes of two of the most similar proteins *LOC203522* (RefSeq NM_182540) and *SAGE* (RefSeq NM_018666), both map to Xq26 (see below).

CT45 Belongs to a Distinctive Protein Family. *LOC203522* is the most similar gene, with significant similarity also seen with *SAGE*, another CT gene (10), and with *DDX26* (RefSeq NM_012141; synonyms: *DICE1*, *Notch12*, *HDB*, and *DBI-1*), a DEAD-box-containing protein encoded by a gene in a region of 13q14 that has been deleted in some cancers (Fig. 4). The four proteins are of different lengths. CT45 comprises 189 aa, DDX26 comprises 887 aa, and SAGE has 904 aa. LOC203522 has several putative protein products, with a 308-aa product (GenBank accession no. AK123209) showing homology to CT45. The observed amino acid similarity among these four proteins is restricted to their C ends. Both LOC203522 and DDX26 contain an additional von Willebrand factor type A domain near their N termini that was not present in CT45.

LOC203522 is located ≈130 kb centromeric to the CT45 gene family, whereas *SAGE* is immediately (4.6 kb) telomeric to the CT45 genes. ESTs corresponding to *LOC203522* were derived from multiple somatic tissues, and RT-PCR analysis confirmed that this gene is ubiquitously expressed in normal tissues (data not shown), whereas *SAGE* and *CT45* are both CT genes.

Discussion

Of 1,056 genes initially identified with MPSS tags derived mainly from testis, a significant proportion were verified as being testis-specific by RT-PCR analysis. This finding illustrated that MPSS is a powerful tool for the identification of differentiation antigens. In this regard, MPSS should be extremely useful for identifying lineage-specific cancer-vaccine targets for tumor types for which tissue-specific autoimmunity is not a major concern, such as melanoma and ovarian or prostate cancer.

Our principle objective here was to identify CT genes of potential value as immunotherapeutic agents for use in human cancer. The first several CT antigens, including the MAGE, BAGE, and GAGE gene families, were all discovered on the basis of the autologous $CD8^+$ T cell responses they elicited in cancer patients (11). Subsequently, a further series of CT-antigen genes were identified by serological analysis of recombinant expression (SEREX) tumor cDNA libraries (12). The SEREX-defined CT antigens include the SSX family, SCP1, NY-ESO-1, CT7, CT8/HOM-TES-85, CAGE, CAGE1, and NY-SAR-35. More recently, CT antigens have been sought by identifying genes with a restricted cancer/testis mRNA-expression pattern, irrespective of their immunogenicity. This process has resulted in the identification of LAGE-1, CT9, CT10, and SAGE by representational-difference analysis (10, 13–15), and CT15, CT16, FATE, and TPTE, by EST database mining (16, 17). The present study, using MPSS to identify tissue-specific genes with therapeutic potential, is a direct extension of the concept of identifying genes encoding CT antigens by using sequence-based transcription data.

To validate the normal tissue expression, we chose to use a normalized 16 normal tissue cDNA panel (BD Biosciences–Clontech) that provided standardization across this study. However, we later found it valuable to also use a second RNA source to confirm testis restriction. For example, *THEG* showed expression, albeit at low levels, in a few somatic tissues when tested against nonnormalized cDNA synthesized from RNA of a different source (Ambion). Such discrepancies are not uncommon in studies of this kind, and expression of CT genes must ultimately be verified by protein-expression data. CT45 mRNA remained testis-restricted in both nucleic acid sources, and generation of antibody reagents against the putative protein product of this transcript is a priority.

The testis-specific genes identified in this study form three groups. The first, and largest, group consists of genes that showed expression highly restricted to testis and germ-cell tumors, with no evidence of expression in somatic tissues or in non-germ-cell cancers. This group of genes encodes true testis-differentiation antigens, some of which are known functional proteins in germ cells, often expressed from abundant mRNAs. Examples include protamine (PRM) 2, PRM1, and YBX2, which

Fig. 4. Multiple sequence alignment of the conserved region shared among SAGE, DDX26, LOC203522, and CT45. Identical sequences are shown in black, and conservative changes are shown in gray. The amino acid number in each gene that represents the starting point of the conserved segment is indicated. Although sequence similarities were also identified among SAGE, LOC203522, and CT45 at their N-terminal sequences upstream, these segments were not as conserved and were not seen in DDX26.

have 35,089, 19,397, and 5,036 corresponding MPSS tags per million, respectively (18, 19). A second group represents the true CT genes, with strong expression in a proportion of cancers. The CT45-gene family belongs to this group. The third group consists of genes that showed strong testicular expression but only marginal, low-level expression in cancer. It is clear that there is a gradient of regulation of gene expression operating in germ cells, presumably reflecting a multitude of transcriptional control mechanisms. The first group of genes is the most tightly controlled and has not yet been found to be expressed in cancers outside of germ-cell lineages. The CT genes, on the other hand, are most frequently activated in cancer, probably through hypomethylation or histone deacetylation (20, 21). However, even within this group, there is clearly a wide range of frequencies with which the genes are expressed in cancer, e.g., from >50% to <5% for the 20 CT and CT-like genes described here, in the same panel of 21 cell lines. Genes in the third group are also tightly controlled but exhibit occasional "leaky" expression in cancer. In terms of functional classification, it is debatable whether it is useful to include this third group within the CT gene category. Categorization is also complicated by the fact that some "CT genes" are expressed in selected somatic tissues. From the viewpoint of potential therapeutic utility, CT antigens that show substantial mRNA and protein expression in cancers are of most interest. Although the phenomenon of germ-line gene activation and expression in tumors is of great interest and deserves full investigation, the main focus of our efforts has been on the identification of CT antigens that are truly of immunotherapeutic potential. Of the 44 CT genes/gene families in the recently created CT database (3), we estimate that probably <12 fall into this group, most of which, intriguingly, reside on the X chromosome, including *MAGE*, *NY-ESO-1*, *SSX*, *CT7*, *CT10*, *XAGE*, *CAGE*, and *SPANX*. This group is now expanded by the discovery of *CT45*.

CT45 shares many features with other classic CT genes: (*i*) Xq localization, which is the same as *CT7* (Xq26), *SAGE* (Xq26), *CT10* (Xq27), *MAGE-A* (Xq28), *NY-ESO-1* (Xq28), and *HOM-TES*-85 (Xq24); (*ii*) multigene family, as are *MAGE*, *GAGE*, *NY-ESO-1*, and *SSX*; and (*iii*) identical or near-identical gene copies, indicating recent gene duplications, as were also described for *NY-ESO-1* (22), *SSX2*, and *SSX7* (21).

A protein similarity search using the CT45 sequence identified the two neighboring genes on Xq26.3, *SAGE* and *LOC203522*, as encoding proteins similar to CT45, suggesting that these three genes may be evolutionarily related. However, the exon–intron structures of these three genes are not conserved, and the gene and protein sizes are quite different. It would thus appear that, whereas these genes may be related, they have diverged significantly, so that their gene products are no longer functionally redundant. In this regard, the relationship between SAGE and CT45 is analogous to that between CT7 (MAGE-C1) and MAGE-A, two other X chromosomal CT-antigen genes. The CT7 protein is 1,115 aa long, with the N terminus containing 10 35-aa tandem repeats and the C-terminal sequence being nonrepetitive. It is the latter region that has similarity to the other MAGE proteins, which are typically ≈310 residues in size and lack the repetitive N-terminal sequences (23). On the other hand, SAGE is a 904-aa protein containing 13 47-aa tandem repeats and, again, with a C-terminal nonrepetitive portion that exhibits similarity to CT45, a much smaller protein.

In conclusion, we have undertaken the most comprehensive analysis to date of similarities between cancer and testis gene expression and found a considerable number of CT genes that had gone undetected by alternative technical approaches. Remarkably, among the genes identified, we have uncovered an entirely distinctive family of CT genes that has the structure and expression characteristics of the most important known CT vaccine candidates to date, *MAGE-A3*, *NY-ESO-1*, and *SSX*, and this gene family thus represents a major opportunity for study as a therapeutic target for human cancer.

This work was supported by funding from the Cancer Research Institute (to Y.-T.C., C.V.J., and A.O.G.) through the Cancer Antigen Discovery Collaborative.

1. Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., *et al.* (2000) *Nat. Biotechnol.* **18,** 630–634.
2. Scanlan, M. J., Gure, A. O., Jungbluth, A. A., Old, L. J. & Chen, Y. T. (2002) *Immunol. Rev.* **188,** 22–32.
3. Scanlan, M. J., Simpson, A. J. & Old, L. J. (2004) *Cancer Immun.* **4,** 1.
4. Brenner, S., Williams, S. R., Vermaas, E. H., Storck, T., Moon, K., McCollum, C., Mao, J. I., Luo, S., Kirchner, J. J., Eletr, S., *et al.* (2000) *Proc. Natl. Acad. Sci. USA* **97,** 1665–1670.
5. Jongeneel, C. V., Iseli, C., Stevenson, B. J., Riggins, G. J., Lal, A., Mackay, A., Harris, R. A., O'Hare, M. J., Neville, A. M., Simpson, A. J. & Strausberg, R. L. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 4702–4705.
6. Mannan, A., Lucke, K., Dixkens, C., Neesen, J., Kamper, M., Engel, W. & Burfeind, P. (2000) *Cytogenet. Cell Genet.* **91,** 171–179.
7. Tschopp, J., Martinon, F. & Burns, K. (2003) *Nat. Rev. Mol. Cell Biol.* **4,** 95–104.
8. Huttemann, M., Jaradat, S. & Grossman, L. I. (2003) *Mol. Reprod. Dev.* **66,** 8–16.
9. Gure, A. O., Tureci, O., Sahin, U., Tsang, S., Scanlan, M. J., Jager, E., Knuth, A., Pfreundschuh, M., Old, L. J. & Chen, Y. T. (1997) *Int. J. Cancer* **72,** 965–971.
10. Martelange, V., De Smet, C., De Plaen, E., Lurquin, C. & Boon, T. (2000) *Cancer Res.* **60,** 3848–3855.
11. Van Der Bruggen, P., Zhang, Y., Chaux, P., Stroobant, V., Panichelli, C., Schultz, E. S., Chapiro, J., Van Den Eynde, B. J., Brasseur, F. & Boon, T. (2002) *Immunol. Rev.* **188,** 51–64.
12. Sahin, U., Tureci, O., Schmitt, H., Cochlovius, B., Johannes, T., Schmits, R., Stenner, F., Luo, G., Schobert, I. & Pfreundschuh, M. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 11810–11813.
13. Lethe, B., Lucas, S., Michaux, L., De Smet, C., Godelaine, D., Serrano, A., De Plaen, E. & Boon, T. (1998) *Int. J. Cancer* **76,** 903–908.
14. Scanlan, M. J., Altorki, N. K., Gure, A. O., Williamson, B., Jungbluth, A., Chen, Y. T. & Old, L. J. (2000) *Cancer Lett.* **150,** 155–164.
15. Gure, A. O., Stockert, E., Arden, K. C., Boyer, A. D., Viars, C. S., Scanlan, M. J., Old, L. J. & Chen, Y. T. (2000) *Int. J. Cancer* **85,** 726–732.
16. Scanlan, M. J., Gordon, C. M., Williamson, B., Lee, S. Y., Chen, Y. T., Stockert, E., Jungbluth, A., Ritter, G., Jager, D., Jager, E., *et al.* (2002) *Int. J. Cancer* **98,** 485–492.
17. Dong, X. Y., Su, Y. R., Qian, X. P., Yang, X. A., Pang, X. W., Wu, H. Y. & Chen, W. F. (2003) *Br. J. Cancer* **89,** 291–297.
18. Steger, K., Pauls, K., Klonisch, T., Franke, F. E. & Bergmann, M. (2000) *Mol. Hum. Reprod.* **6,** 219–225.
19. Gu, W., Tekur, S., Reinbold, R., Eppig, J. J., Choi, Y. C., Zheng, J. Z., Murray, M. T. & Hecht, N. B. (1998) *Biol. Reprod.* **59,** 1266–1274.
20. De Smet, C., De Backer, O., Faraoni, I., Lurquin, C., Brasseur, F. & Boon, T. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 7149–7153.
21. Gure, A. O., Wei, I. J., Old, L. J. & Chen, Y. T. (2002) *Int. J. Cancer* **101,** 448–453.
22. Alpen, B., Gure, A. O., Scanlan, M. J., Old, L. J. & Chen, Y. T. (2002) *Gene* **297,** 141–149.
23. Chen, Y. T., Gure, A. O., Tsang, S., Stockert, E., Jager, E., Knuth, A. & Old, L. J. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 6919–6923.