Large-scale Transcriptome Analyses Reveal New Genetic Marker Candidates of Head, Neck, and Thyroid Cancer

Eduardo M. Reis,[1,2] Elida P.B. Ojopi,[3] Fernando L. Alberto,[7] Paula Rahal,[9] Fernando Tsukumo,[8] Ulises M. Mancini,[9] Gustavo S. Guimarães,[5] Gloria M.A. Thompson,[11] Cleber Camacho,[5] Elisabete Miracca,[4] André L. Carvalho,[6] Abimael A. Machado,[2] Apuã C.M. Paquola,[2] Janete M. Cerutti,[5] Aline M. da Silva,[1] Gonçalo G. Pereira,[8] Sandro R. Valentini,[11] Maria A. Nagai,[4] Luiz Paulo Kowalski,[6] Sergio Verjovski-Almeida,[1] Eloiza H. Tajara,[10] Emmanuel Dias-Neto,[3] and Head and Neck Annotation Consortium

[1]Departamento de Bioquímica, [2]Laboratório de Bioinformática, Instituto de Química, [3]Laboratório de Neurociências (LIM-27), Instituto e Departamento de Psiquiatria, and [4]Disciplina de Oncologia, Departamento de Radiologia, Faculdade de Medicina, Universidade de São Paulo; [5]Laboratório de Endocrinologia Molecular, Departamentos de Medicina e Morfologia, Universidade Federal de São Paulo; [6]Departamento de Cirurgia de Cabeça e Pescoço e Otorrinolaringologia, Hospital do Câncer A.C. Camargo, São Paulo, SP, Brazil; Laboratórios de [7]Biologia Molecular e Genômica, Hemocentro and [8]Genômica e Expressão, Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, SP, Brazil; [9]Departamento de Biologia, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista; [10]Departamento de Biologia Molecular, Faculdade de Medicina de São José do Rio Preto, São José do Rio Preto, SP, Brazil; and [11]Departamento de Ciências Biológicas, Escola de Farmácia, Universidade Estadual Paulista, Araraquara, SP, Brazil

Abstract

A detailed genome mapping analysis of 213,636 expressed sequence tags (EST) derived from nontumor and tumor tissues of the oral cavity, larynx, pharynx, and thyroid was done. Transcripts matching known human genes were identified; potential new splice variants were flagged and subjected to manual curation, pointing to 788 putatively new alternative splicing isoforms, the majority (75%) being insertion events. A subset of 34 new splicing isoforms (5% of 788 events) was selected and 23 (68%) were confirmed by reverse transcription–PCR and DNA sequencing. Putative new genes were revealed, including six transcripts mapped to well-studied chromosomes such as 22, as well as transcripts that mapped to 253 intergenic regions. In addition, 2,251 noncoding intronic RNAs, eventually involved in transcriptional regulation, were found. A set of 250 candidate markers for loss of heterozygosis or gene amplification was selected by identifying transcripts that mapped to genomic regions previously known to be frequently amplified or deleted in head, neck, and thyroid tumors. Three of these markers were evaluated by quantitative reverse transcription–PCR in an independent set of individual samples. Along with detailed clinical data about tumor origin, the information reported here is now publicly available on a dedicated Web site as a resource for further biological investigation. This first *in silico* reconstruction of the head, neck, and thyroid transcriptomes points to a wealth of new candidate markers that can be used for future studies on the molecular basis of these tumors. Similar analysis is warranted for a number of other tumors for which large EST data sets are available. (Cancer Res 2005; 65(5): 1693-9)

Introduction

Head and neck squamous cell carcinomas (HNSCC) are common and aggressive neoplasms, with an estimated annual incidence of 615,900 new cases worldwide (1) and an increasing incidence rate in many areas of the world (2–4). Thyroid cancer is the most common neoplasia from the endocrine system with an estimated annual incidence of 122,800 worldwide. Most of the incidence variations are probably due to ethnic or environmental factors such as radiation, dietary habits, or other external factors.

Cancer phenotype is defined by the accumulation of mutations and epigenetic changes that may alter protein function and/or cause alterations in transcriptional patterns. These events usually confer competitive advantages to a cell, ultimately leading to the malignant phenotype. Transcriptome wide analysis is an important tool to reveal some of these molecular mechanisms that underlie human malignancies (5, 6).

To generate a database of transcriptional profiles derived from tumor and nontumor tissues, a high-throughput cDNA-sequencing project, based on normalized mini-cDNA libraries generated with the open reading frame expressed sequence tags (ORESTES) methodology, was conducted by our group, yielding a set of ~1.2 million expressed sequence tags (EST) from diverse human tissues, from which a filtered set was deposited in public databases (6, 7). ORESTES sequences are predominantly derived from the central coding region of genes, thus favoring the identification of gene function based on protein similarity. In addition, this methodology partially normalizes the population of tagged genes, enabling the study of rare transcripts (8). The present study reports on a detailed informatics analysis of 134,495 ORESTES from head and neck as well as 79,141 ORESTES from thyroid tissues followed by experimental validation of new transcripts and new splicing isoforms and by an evaluation of putative markers for these malignancies.

Materials and Methods

Patients and Biological Samples. Subjects were selected at random from surgical patients treated at the Head and Neck Surgery and Otorhinolaryngology Department of the Hospital do Câncer A.C. Camargo in São Paulo, Brazil. All patients were treated and followed up according to prevailing protocols. Eligible patients were asked to sign an informed

Note: E.M. Reis and E.P.B. Ojopi contributed equally to this work.

Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

A complete list of authors is presented in the Appendix.

Requests for reprints: Emmanuel Dias-Neto, Laboratory of Neurosciences (LIM-27), Instituto de Psiquiatria, Faculdade de Medicina, Universidade de São Paulo, R. Dr. Ovidio de Campos, s/n Consolação 05403-010, São Paulo, SP, Brazil. Phone: 55-11-3069-7267; Fax: 55-11-3069-8010; E-mail: emmanuel@usp.br.

consent previously approved by the hospital's institutional review board. The criteria for eligibility were determined to insure a good representation of the transcriptional diversity of the tissues studied. On this basis, selected patients had to be at least 18 years old and have a histologically confirmed diagnosis of a non–previously treated (*a*) oral, pharyngeal, hypopharyngeal, or laryngeal squamous cell carcinoma (31 samples) or (*b*) thyroid carcinoma, adenoma, or goiter (45 samples). Detailed composition of samples, in terms of tissue origin and pathologic state, is given in Supplementary Methods. Tumor and adjacent nontumor tissues were collected, snap-frozen, and stored; tumor margins that were considered disease free after histologic examination were used here as nontumor samples. Before RNA extraction, all pathologic diagnoses were reconfirmed and samples were macrodissected, reducing nontumor tissues to <20%.

RNA Extraction and ORESTES. PolyA$^+$ RNA was isolated, treated with DNase and purified as described before, (8) and used for constructing 1,576 ORESTES cDNA minilibraries (8) that were cloned and sequenced (7). The detailed composition of each of these cDNA minilibraries, in terms of tissue origin and pathologic state, is given in Supplementary Methods and in Supplementary Table 1. A total of 1,003 cDNA minilibraries were sequenced, generating 18,620 sequences from nontumor tissues, 190,253 from tumor tissues, and 4,763 from thyroid goiter samples; 134,495 sequences were derived from head and neck and 79,141 from thyroid tissues. mRNA derived from a second set of tumors and nontumor samples as well as from four human cell lines (FaDu, Hep2, HeLa, and Siha) was extracted as described above and used for the experimental validation of new splicing isoforms, new genes, and potential tumor markers predicted by bioinformatics analyses.

cDNA Sequence Analyses. A total of 213,636 ESTs generated from head and neck and thyroid tissues were analyzed along with local copies of the human genome sequence draft (University of California at Santa Cruz, June 2002) and of the RefSeq and mRNA public data sets (www.ncbi.nlm.nih.gov). The complete set of 1,187,342 ESTs produced in the Human Cancer Genome Project (HCGP; ref. 7) from more than 20 tissues was filtered to reduce the presence of low-quality sequences or contaminants, and assembled in 956,456 reads, which included 173,640 sequences derived from head, neck, and thyroid tissues. The data set was clustered using BLAST (http://www.ncbi.nlm.nih.gov/) and assembled with Phrap (http://www.phrap.org). To further reduce the residual redundancy and to eliminate misassembled contigs or spurious sequences that passed the filtering step, all contigs and singlet sequences were aligned to the draft human genome sequence using BLAT (9). All mapped sequences that exhibited genomic overlap were merged into a unique genomic cluster. Only sequences that aligned with >90% identity to the draft sequence through at least 50% of their length were considered for further analyses. A subset of 119,552 ESTs that were used here and were not analyzed in previous publications (6–8) were now deposited in dbEST GenBank under accession numbers CV309219 to CV428770.

Identification of Tissue-Specific Transcriptional Markers. Genes that were preferentially expressed in certain head and neck sites (oral cavity, pharynx, and larynx) or thyroid were identified using a Bayesian model, as described in detail in Supplementary Methods. Using this approach, EST count thresholds were established to define 95% confidence intervals for the absence or presence of expression of a certain gene in each anatomic site.

Analysis of New Splicing Isoforms and New Human Transcripts. Four trained investigators, using the same criteria, visually inspected the genomic mapping of all ESTs with putative new splicing isoforms. For all new splicings approved by visual inspection, additional data were collected regarding splicing classification (10), size of the event in terms of nucleotides involved, donor/acceptor sites, and confirmation of the event by other cDNA sequences. Reverse transcription–PCR (RT-PCR) and cDNA sequencing were done to confirm a subset of these new gene isoforms. To reduce individual variations, isoform validations were done using human cell lines as well as pooled samples from different patients and topological locations.

Candidates for new human transcripts were identified by selecting clusters or single ESTs with no similarity to known human genes, as determined by BLASTN queries against a complete set of 162,104 human mRNA full-length sequences deposited in GenBank (cutoff *E* value = 10^{-15}). In addition, clusters and singlets were identified that mapped to a 2.37-Mb region of chromosome 22 (22q11.2), flanking marker D22S421, previously shown to be deleted in larynx tumors (11). These were selected for manual annotation, which included sequence alignment with *in silico* gene predictions (GeneScan) and comparison with mouse full-length transcripts and ESTs, using the UCSC Genome Browser (http://genome.ucsc.edu/). A set of putative new transcripts was validated by RT-PCR, using pooled mRNAs from nontumor or tumor larynx and further evaluated in 45 different mRNA samples derived from larynx (14 nontumor and 13 tumor), tongue (9 nontumor and 3 tumor), and tonsil (4 nontumor and 2 tumor).

Evaluation of Differential Gene Expression. Experimental evaluation of differential expression was done for a subset of ORESTES contigs mapped to genomic regions that, according to the literature data, exhibit recurrent loss of heterozygosity or amplifications in the selected tumors. Three genes were selected for experimental validation of their differential expression by quantitative real-time PCR analysis in a GenAmp 5700 (Applied Biosystems, Foster City, CA). cDNA was produced from individual tumor and paired normal adjacent tissue samples from the larynx and oral cavity. To reduce individual sample variations possibly arising from contamination of normal tissue with adjacent tumor, the average expression level of each gene in all normal samples was used as a reference for each tissue examined. For each patient, fold change of expression was calculated as the ratio between the individual tumor sample and the average level of normal tissue samples. Each cDNA sample was analyzed in duplicate reactions using the SybrGreen PCR Core Reagent (Applied Biosystems), and normalization was done according to the manufacturer's instructions using *ACTB* or *GAPDH* as the reference gene.

Web Site. A dedicated Web site was prepared for data analysis, enabling visual inspection of the assemblies and of their genome map coordinates as well as searches using keywords, Gene Ontology classification, or chromosomal location of transcripts detected in head, neck, or thyroid tumor types under study. The site is publicly available at http://verjo19.iq.usp.br/java/jsp/head_neck/.

Results and Discussion

Assembly and Mapping the Head, Neck, and Thyroid EST Dataset. A set of 213,636 EST sequences derived from head, neck and thyroid tumors was analyzed here, being part of the over 1.2 million ESTs that were generated by the HCGP (7). After filtering for contaminants, the remaining 173,640 ESTs (81%) were assembled together with all HCGP sequences as described under Materials and Methods. This assembly resulted in 32,576 contigs and 24,434 singlets that were aligned to the draft sequence of the human genome. All mapped sequences that exhibited genomic overlap were included in the same genomic cluster, thus reducing the data set to a nonredundant collection of 48,001 sequences (contigs plus singlets) grouped into 20,348 genomic clusters (Table 1). This corresponds to 38% of the total genomic groups formed by all ORESTES (20,348/53,105), and includes a number of clusters containing reads exclusively derived from head and neck (2,712) or thyroid (1,331) tissues.

Patterns of Gene Expression and Tissue Markers. A total of 10,228 distinct RefSeq genes were sampled by at least one EST derived from head and neck or thyroid, as determined by merging all nonoverlapping EST clusters that mapped within the same RefSeq gene sequence. Different amounts of ESTs were generated from each head and neck or thyroid tissue using different sets of primers, which preclude a direct assessment of genes differentially expressed at a distinct topology. Aiming to normalize these experimental factors, we applied a Bayesian statistical model (12) to select only EST counts that were

Table 1. Genomic mapping of all ORESTES sequences and of head and neck or thyroid assembled sequences

	Assembled contigs + singlets	Percent of total	Genomic clusters	Percent of total	Genomic clusters colinear with RefSeqs*	Percent of total in each category
Containing all ORESTES†	173,329	100	53,105	100	33,022	62
Not containing head and neck or thyroid sequences	125,328	72	32,742	62	19,003	58
Containing head and neck or thyroid sequences	48,001	28	20,348	38	14,019	69
Containing only head and neck or thyroid sequences	23,474	14	4,186	8	2,162	52
Containing only head and neck sequences (normal or tumor)	14,815	9	2,712	5	1,396	4
Containing only HNSCC sequences	14,305	8	2,621	5	1,338	4
Containing only normal head and neck sequences	473	0.3	66	0.1	38	0.1
Containing only thyroid sequences	8,316	5	1,331	3	651	2
Containing only thyroid tumor sequences	6,591	4	1,099	2	529	2
Containing only normal thyroid sequences	1,651	1	208	0.4	106	0.3

*A total of 20,602 RefSeqs mapped to the draft sequence of the human genome with 50% of their length with identity ≥90%.
†Only sequences aligned to the draft sequence of the human genome through at least 50% of their length with identity ≥90%.

detected within a confidence interval of 95% (see Materials and Methods). When this statistical model was applied, ESTs representing exons of 238 RefSeq genes were selected. The analysis suggested larynx/oral cavity as the most similar pair of tissues in terms of gene expression, followed by oral cavity/pharynx and larynx/pharynx (Fig. 1*A*). The same result was obtained when EST clusters that mapped to genomic regions with no RefSeq genes were included in the analysis (not shown). In addition, the trio larynx/oral cavity/pharynx contains more genes in common than trios that include thyroid (Fig. 1*B*). The 238 RefSeq genes that were identified as being differentially expressed in different topologies may be used as putative markers for identifying the tissue origin of circulating tumor cells, as well as for distinguishing primary tumor from metastatic lesions, a subject that deserves further evaluation.

This analysis revealed genes preferentially expressed in each head and neck site or in thyroid with statistical criteria, being a source of tissue-specific candidate markers for HNSCC or thyroid tumors that may be selected for further validation (list of genes is available as Supplementary Table 2).

New Genes and Noncoding RNAs. Out of the total of 20,348 genomic clusters containing sequences from head and neck or thyroid, a significant fraction (25%, i.e., 5,038 clusters) represents possible new human transcripts because they have no significant similarities to previously known human genes (see Materials and Methods). Among these, a small fraction has evidence of splicing (16%, i.e., 835) and only 507 sequences (10%) have a normalized ESTscan score (13) >1, which is indicative of high protein coding potential. Nevertheless, none of them contain known protein motifs when compared with the PFAM data set. Among these new coding transcript fragments, 254 were found to map to intronic regions of RefSeq genes and may represent new coding exons (see Supplementary Table 3). The remaining 253 map to intergenic regions and may represent fragments of new human genes (see Supplementary Table 4).

The remaining 4,531 clusters that represent new transcripts have a normalized ESTscan score lower than 1; that is, they have a low coding potential. Remarkably, a large fraction of these (50%, i.e., 2,251 clusters) map to intronic regions of RefSeq genes (Supplementary Table 5). Whereas some of these may reflect a certain degree of genomic DNA contamination that might have persisted even after DNase treatment and polyA$^+$ mRNA selection, it is likely that most of these ESTs represent bona fide mRNA transcripts. In fact, recent data showed that similar fractions of intronic and exonic transcripts are expressed in 47 samples of tumor and normal prostate, as detected by cDNA microarrays constructed with ORESTES clones (14).

Recent work describing the transcriptional output of the human genome points to the existence of a significant number of noncoding RNA transcripts (15, 16) and that 10% to 20% of human transcripts might form sense-antisense pairs (17, 18). Furthermore, comparable fractions of transcriptional activity were detected within exons or introns of annotated genes, and nearly half of these intronic transcripts were expressed antisense to their respective well-characterized introns (16, 19). Although the role of these antisense intronic noncoding messages in RNA regulation is not yet fully understood (20), it has been recently shown that intronic noncoding transcripts expressed antisense to their respective introns have their expression level significantly correlated to the degree of tumor differentiation in prostate cancer (14). Identification in the present work of a large set of intronic noncoding transcripts expressed in head-neck and thyroid suggests that noncoding intronic transcripts may also play an important role in HNSCC and thyroid tumors.

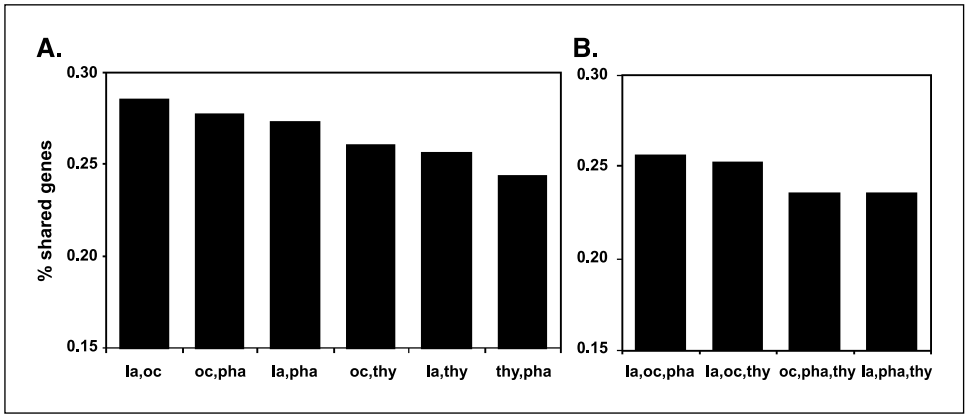

Figure 1. Transcriptome sharing between HNSCC and thyroid. 238 RefSeq genes detected with 95% confidence in the EST data set were scored (see Materials and Methods). Fractions of genes shared by each pair (*A*) or trio (*B*) of topologies are shown. *la,* larynx; *oc,* oral cavity; *pha,* pharynx; *thy,* thyroid.

Significantly, the majority of the new transcripts containing sequences from head and neck or thyroid (2,832 of 5,038, i.e., 56%) map to genomic regions previously associated to HNSCC or thyroid tumors (21). We found 12 putative new human transcripts mapping to a 2.37-Mb sequence from human chromosome 22q11-12, a region selected for its involvement with larynx tumors (11). These include three EST clusters containing head, neck, and thyroid sequences. All three had their expression confirmed in larynx by RT-PCR and were further verified by semiquantitative RT-PCR in 45 different mRNA samples obtained from pairs of nontumor/tumor tissues (see Materials and Methods). Positive amplification was obtained in 42 of 45 (LOC91353), 29 of 45 (LOC164615), and 21 of 45 (LOC91355) of the evaluated samples. Although no evidence for differential expression between nontumor and tumor samples was observed for this small set of genes, these results confirm the potential of this approach to identify new human genes.

New Splicing Isoforms. Genomic alignments of all EST clusters matching known human genes were visually inspected to evaluate the presence of putative new splicing isoforms. New events involving at least 10 nucleotides of deletions or insertions in previously known genes were considered as putative new splicing isoforms and were classified according to Wang et al. (10). Presence of conserved donor/acceptor sites as well as eventual confirmation of splicing events by other human ESTs or nonhuman transcripts in the public databases were recorded. A total of 788 new splicing events were identified in 748 different transcripts (see Supplementary Table 6).

Among the new splicing events observed here, 86% and 89% showed the canonical GT/AG donor and acceptor sites, respectively. The majority (75.3%) of new splicings was due to insertion events (Table 2). Overall, the most frequent events were exon extensions at 5′ (type II) and insertions at 3′ (type III). New splicing events could be confirmed by other human ESTs from GenBank in 28.6% of the cases and by sequences from other organisms in 6%.

New splicing isoforms were selected for experimental validation using a set of mRNAs different from that used for EST generation. From a total of 34 new splicing isoforms evaluated, the following 23 (68%) were confirmed by RT-PCR in *AF5Q31*, *CARD8*, *CCNL1*, *CTNNB1*, *CTPS*, *DNMT3A*, *EWSR1*, *MKNK1*, *NESHBP*, *PDCD8*, *PFDN5*, *PORIMIN*, *PSMD10*, *PTK2*, *PTPN18*, *RABL2A*, *SMARCA4*, *SMARCB1*, *ST5*, *SULF2*, *TRIPP11*, *TRRAP*, and *VAV2*. Examples of four novel splicing events are shown in Fig. 2. Additional examples are shown in Supplementary Fig. 1.

Many cases of splicing events affecting protein domains were observed, such as in *SMARCB1*, *PFDN5*, *AF5Q31*, and *EWSR1*. *SMARCB1* is involved in chromatin remodeling and is mutated in a variety of human tumors of diverse tissue origins, including some very aggressive pediatric cancers (22, 23). The novel exon of *SMARCB1* identified here promotes an in-frame addition of 18 amino acids to the protein, with an unpredicted effect over its function. In *PFDN5*, the newly identified exon adds 101 amino acid residues (with putative phosphorylation sites and an amidation site) to the protein. Interestingly, isoforms of *PFDN5* transcripts were detected in Northern blots by Mori et al. (24) who also showed that this heterohexameric chaperone protein binds to c-Myc, repressing its transcriptional activity. The 101-amino-acid insertion may affect PFDN5 binding capacity and interfere with its ability to repress the transcriptional activity of c-Myc.

Because the ORESTES database was shown to be enriched with rare messages (7), splicing variants described here may represent rare new splicing isoforms related to these tumors. Whereas some false-negative or positive splicing events may have occurred among the 788 new splicing events described, our validation rate (68%) suggests that true events are the majority. Indeed, the set of human ESTs available in public databases has recently been shown to provide potential markers for the classification of cancer (25). The new splicing isoforms reported here contribute to this notion and warrant further research aimed at the identification of tumor-specific transcriptional events.

Finding Tumor Markers Using Transcript Analysis. The complete HCGP data set is composed of a 9.8-fold excess of tumor-derived sequences. Despite this, a set of 8,988 genomic clusters exclusively from nontumor tissues sampled in the project was found. This set probably is enriched in genes down-regulated during tumor development and progression. Within this set, 270 genomic clusters contained sequences derived from normal head, neck, or thyroid tissues (the complete list can be searched at the project Web site). Twenty-three of these contain sequences from normal head and neck tissues and are of particular interest because they mapped to regions known to be frequently deleted in HNSCC (refs. 21, 26–28; Supplementary Table 7). Although we found several clusters composed exclusively from sequences derived from normal thyroid, only one of these (R2_CLUSTER_36019) mapped to a region (17p11) known to be deleted in thyroid tumors (21, 29, 30). Several of the genes included in this subset have antiproliferative, apoptotic, or differentiation-induction activities: *CASP10* is involved in the proximal pathway of Fas-mediated apoptosis and is down-regulated in lung cancer, neuroblastoma, and non-Hodgkin lymphomas (31–33). Two other genes are related to apoptosis: *CSEN*, a member of the neuronal calcium sensor family (34), and *MAPK8*, which is

important for the induction of apoptosis following stress (35). *RGS6* and *HHEX* are involved in neuronal or thyroid differentiation, respectively (36, 37), and *APRIN* is a putative regulator of cell proliferation (38). Altogether, these observations strongly suggest that genes with a still unknown tumor suppressor activity in HNSCC or thyroid cancer will be found within this selected data set of loss of heterozygosity candidates.

Conversely, clusters exclusively derived from HNSCC or thyroid tumor tissues suggest oncogenes from genomic regions amplified in tumors. A set of 3,776 such clusters were identified (complete lists can be searched at the project Web site), including 144 that mapped to regions previously described as frequently amplified in HNSCC (see Supplementary Table 8) and therefore are the best candidates for further validation. In fact, 12 of these clusters are transcripts from known oncogenes, whereas another subset of about 30% of these clusters are from genes of unknown functions. Genes from this list may include potential antigens that could be used in immunodiagnosis or immunotherapeutic approaches, as well as tumor markers. However, due to the biased amount of sequences from tumor samples in our project, the list of tumor-only EST clusters identified here should be analyzed with caution.

Three transcripts were further investigated for differential expression in HNSCC by quantitative real-time PCR in paired patient samples (eight oral cavity tumors and nine larynx tumors). Selected genes mapped to genomic regions that were frequently amplified (*ZRF1* and *NDRG1)* or frequently lost (*RAP140*) in HNSCC. *ZRF1* mRNA levels were increased in most samples of

Table 2. Alternative splicing statistics

	Splicing events (%)	% of all 788 events
Deletion type		
I	68 (34.9)	8.6
II	15 (7.7)	1.9
III	17 (8.7)	2.2
IV	48 (24.6)	6.1
2 events	24 (12.3)	3.1
3 events	16 (8.2)	2.0
Multiple	7 (3.6)	0.9
Total	195 (100.0)	24.8
Insertion type		
I	103 (17.4)	13.0
II	201 (33.9)	25.5
III	158 (26.6)	20.1
IV	59 (10.0)	7.5
2 events	65 (11.0)	8.3
3 events	4 (0.7)	0.5
Multiple	3 (0.5)	0.4
Total	593 (100.0)	75.3

NOTE: New splicing isoforms were classified according to Wang et al. (10). Insertions were divided into types I to IV: insertion of a new exon, 5′ extension, 3′ extension, and junction of two exons, respectively. Deletions were divided into types I to IV: exon skipping, 5′ deletion, 3′ deletion, and deletion of a central portion of the exon, respectively.

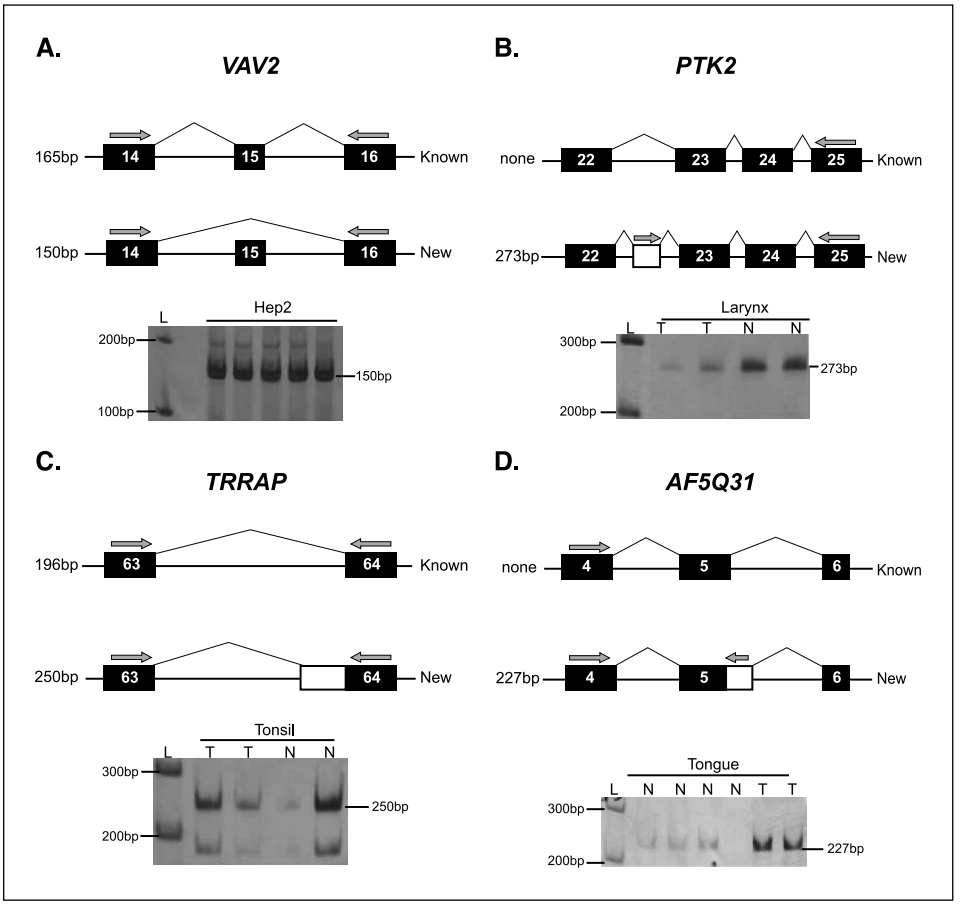

Figure 2. Experimental validation of new splicing isoforms using RT-PCR. *A,* type I deletion. *B,* type I insertion. *C,* type II insertion. *D,* type III insertion. Names of the genes are at the top of each panel. *Arrows,* primers; *solid boxes,* known exons with their corresponding exon numbers; *open boxes,* new exons. *Horizontal lines,* introns. *Diagonal lines,* alternative splicing events. For all gels, *lane L* corresponds to the molecular weight marker ladder (100 bp). The sources of template RNA samples used in the PCR reactions are indicated above the gel lanes. They are either replicates of cell line samples from Hep-2 (CCL-23, American Type Culture Collection, Manassas, VA), or normal (*N*) and tumor (*T*) tissue from *Larynx*, *Tonsil*, or *Tongue*, as indicated in each panel. All PCR products were confirmed by sequencing. Predicted band sizes (in base pairs) of PCR products amplified from *Known* and *New* mRNA splicing isoforms are indicated next to each diagram, and those predicted from genomic DNA are 2,618 bp for *VAV2*, 15,007 bp for *PTK2*, 1,560 bp for *TRRAP*, and 5,197 bp for *AF5Q31*.

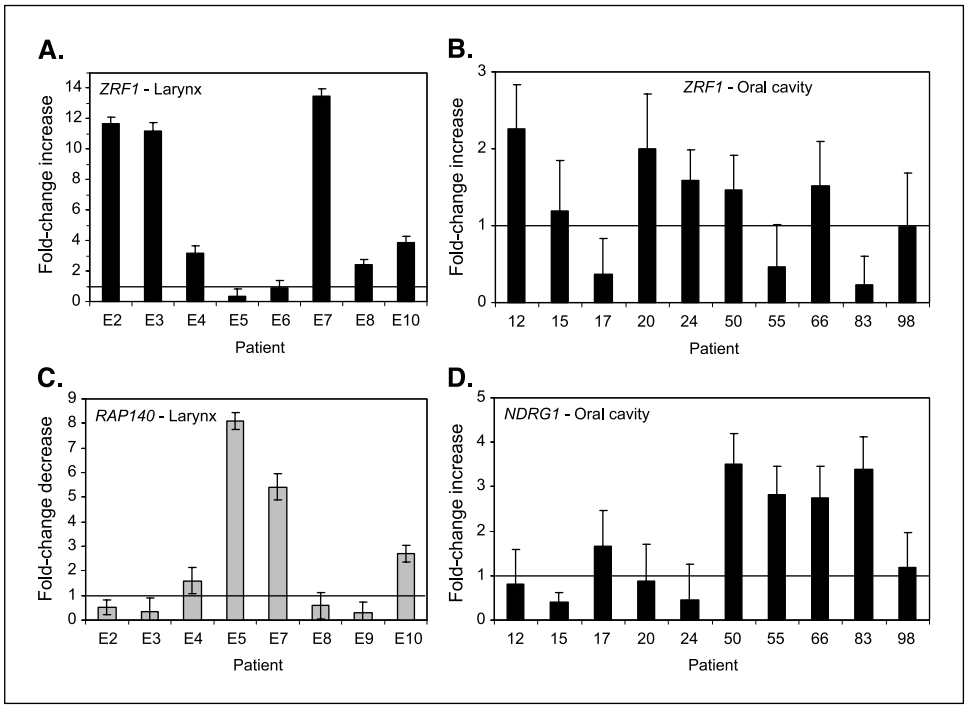

Figure 3. Expression of putative tumor markers of HNSCC. Real-time quantitative PCR was used for the measurement of mRNA levels of candidate genes in paired samples of tumor and adjacent normal tissue from larynx (9 patients) or oral cavity (10 patients). Expression of *ZRF1* (*A* and *B*), *RAP140* (*C*), and *NDRG1* (*D*) in each individual tumor sample reflects the fold-change increase (*black columns*) or decrease (*gray columns*) relative to the average expression level of each gene in the normal samples. *Bars,* SE from two replicate experiments.

larynx (2.4- to 13.4-fold change, 6 of 8 samples, i.e., 75%) and of oral cavity tumors (1.5- to 2.3-fold change, 5 of 10 samples, i.e., 50%; Fig. 3*A* and *B*). This is in agreement with a previous study that showed an increase in chromosome 7 and *ZRF1* copy numbers in HNSCC (39). *RAP140* exhibited a decreased expression in half of the patient samples of larynx tumors (1.6- to 8.1-fold change, 4 of 8 samples, i.e., 50%; Fig. 3*C*). No information is available in the literature on *RAP140* expression or function in tumors. *NDRG1* mRNA levels were increased in half of the oral cavity tumor samples (1.7- to 3.5-fold change, 5 of 10 samples, i.e., 50%; Fig. 3*D*). *NDRG1* has been referred to be either under- or overexpressed in a variety of cancers, including lung, brain, melanoma, liver, prostate, breast, and renal cancers and also in colon adenomas and adenocarcinomas (40–42). The variability observed in the expression of cancer-related genes in individual patient samples probably reflects the biological heterogeneity of HNSCC (43). Nevertheless, the 50% to 75% validation rate achieved in the present work indicates that an informative list of possible candidate markers was generated in this transcriptome analysis, for which further experimental validation is warranted.

In conclusion, two large projects have engaged in extensive EST sequencing of human tumor tissues: the Cancer Genome Anatomy Project led by the National Cancer Institute (44) and the HCGP led by Fundação de Amparo à Pesquisa do Estado de São Paulo/Ludwig Institute for Cancer Research (7), providing a very large disease-oriented transcriptional database that contains an unprecedented amount of information on expressed human genes. We show here a detailed analysis of a subset of these data, the head, neck, and thyroid sequences, documenting its utility in the identification of new human genes, candidate tumor suppressors/oncogenes, and highlighting the structure of transcript variants. Although the EST data have been widely used by the scientific community, we show that a wealth of additional information remains unexplored and may be used to further the understanding of cancer biology.

Acknowledgments

Received 9/28/2004; revised 12/14/2004; accepted 12/28/2004.

Grant support: Invitrogen, Conselho Nacional de Desenvolvimento Científico e Tecnológico, Coordenação de Aperfeiçoamento do Pessoal do Ensino Superior, and Fundação de Amparo a Pesquisa do Estado de São Paulo. E. Dias-Neto thanks Associação Beneficente Alzira Denise Hertzog da Silva for funding to the Laboratory of Neurosciences (LIM-27).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Appendix

The Head and Neck Annotation Consortium authors are the following: Mario H. Bengtson, Renata A. Canevari, Marcelo F. Carazzolle, Christian Colin, Fernando F. Costa, Maria Cristina R. Costa, Marcos R.H. Estécio, Leda Isabel C.V. Esteves, Miriam H.H. Federico, Pedro Edson Moreira Guimarães, Christine Hackel, Edna T. Kimura, Suzana G. Leoni, Ru M.B. Maciel, Simone Maistro, Flavia R.R. Mangone, Katlin B. Massirer, Silvia E. Matsuo, Francisco G. Nobrega, Marina Pasetto Nóbrega, Diana Noronha Nunes, Fabio Nunes, José Rodrigo Pandolfi, Maria Inês M.C. Pardini, Fátima Solange Pasini, Tarcisio Peres, Cláudia Aparecida Rainho, Patrícia P. dos Reis, Flávia Cristina C. Rodrigues-Lisoni, Silvia Regina Rogatto, Andrey dos Santos, Paulo C.C. dos Santos, Mari Cleide Sogayar, and Cleslei F. Zanelli.

Affiliations: Departamentos de Bioquímica, Instituto de Química (M.H.B., C.C., K.B.M., D.N.N., and M.C.S.), Neurociências (LIM-27), Instituto e Departamento de Psiquiatria (P.E.M.G.), and Biologia Celular e Desenvolvimento, Instituto de Ciências Biomédicas (E.T.K., S.G.L., S.E.M.), Disciplinas de Oncologia, Departamento de Radiologia, Faculdade de Medicina (M.H.H.F., S.M., F.R.R.M., F.S.P., P.C.C.d.S.) and Patologia Bucal, Faculdade de Odontologia (F.N.), Universidade de São Paulo; Laboratório de Endocrinologia Molecular, Departamentos de Medicina e Morfologia, Universidade Federal de São Paulo, São Paulo, SP, Brazil (R.M.B.M.);

Faculdade de Medicina (R.A.C., L.I.C.V.E., S.R.R.), Hemocentro, Faculdade de Medicina (M.I.M.C.P.), and Departamento de Genética, Instituto de Biociências (C.A.R., P.P.d.R.), Universidade Estadual Paulista, Botucatu, SP, Brazil; Departamento de Ciências Biológicas, Escola de Farmácia, Universidade Estadual Paulista, Araraquara, SP, Brazil (J.R.P., C.F.Z.); Departamento de Biologia, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto, SP, Brazil (F.C.C.R-L.); Laboratório de Genômica e Expressão, Departamento de Genética e Evolução do Instituto de Biologia (M.F.C.), Laboratório de Biologia Molecular e Genômica, Hemocentro (F.F.C., T.P.), Departamento de Genética Médica, Faculdade de Ciências Médicas (C.H., A.d.S.), Universidade Estadual de Campinas, Campinas, SP, Brazil; Universidade de Ribeirão Preto, Curso de Medicina, Ribeirão Preto, SP, Brazil (M.C.R.C.); Instituto de Pesquisa e Desenvolvimento, Universidade do Vale do Paraíba, São José dos Campos, SP, Brazil (F.G.N., M.P.N.); and Department of Leukemia, University of Texas MD Anderson Cancer Center, Houston, Texas (M.R.H.E.).

References

1. Parkin DM, Bray F, Ferlay J, Pisani P. Estimating the world cancer burden: Globocan 2000. Int J Cancer 2001;94:153–6.
2. Alho OP, Kantola S, Pirkola U, Laara E, Jokinen K, Pukkala E. Cancer of the mobile tongue in Finland—increasing incidence, but improved survival. Acta Oncol 1999;38:1021–4.
3. Myers JN, Elkins T, Roberts D, Byers RM. Squamous cell carcinoma of the tongue in young adults: increasing incidence and factors that predict treatment outcomes. Otolaryngol Head Neck Surg 2000;122:44–51.
4. Ho PS, Ko YC, Yang YH, Shieh TY, Tsai CC. The incidence of oropharyngeal cancer in Taiwan: an endemic betel quid chewing area. J Oral Pathol Med 2002;31:213–9.
5. Strausberg RL, Simpson AJ, Wooster R. Sequence-based cancer genomics: progress, lessons and opportunities. Nat Rev Genet 2003;4:409–18.
6. Brentani H, Caballero OL, Camargo AA, et al. The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. Proc Natl Acad Sci U S A 2003;100:13418–23.
7. Camargo AA, Samaia HP, Dias-Neto E, et al. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. Proc Natl Acad Sci U S A 2001;98:12103–8.
8. Dias-Neto E, Correa RG, Verjovski-Almeida S, et al. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. Proc Natl Acad Sci U S A 2000;97:3491–6.
9. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res 2002;12:656–64.
10. Wang Z, Lo HS, Yang H, et al. Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. Cancer Res 2003: 63:655–7.
11. Poli-Frederico RC, Bergamo NA, Reis PP, et al. Chromosome 22q a frequent site of allele loss in head and neck carcinoma. Head Neck 2000;22:585–90.
12. Freund JE, Simon GA. Modern elementary statistics. 9th ed. Upper Saddle River (NJ): Prentice Hall; 1997.
13. Iseli C, Jongeneel CV, Bucher P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol 1999;138–48.
14. Reis EM, Nakaya HI, Louro R, et al. Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. Oncogene 2004;23:6684–92.
15. Kapranov P, Cawley SE, Drenkow J, et al. Large-scale transcriptional activity in chromosomes 21 and 22. Science 2002;296:916–9.
16. Rinn JL, Euskirchen G, Bertone P, et al. The transcriptional activity of human chromosome 22. Genes Dev 2003;17:529–40.
17. Yelin R, Dahary D, Sorek R, et al. Widespread occurrence of antisense transcription in the human genome. Nat Biotechnol 2003;21:379–86.
18. Chen J, Sun M, Kent WJ, et al. Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids Res 2004;32:4812–20.
19. Kampa D, Cheng J, Kapranov P, et al. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res 2004;14:331–42.
20. Mattick JS. RNA regulation: a new genetics? Nat Rev Genet 2004;5:316–23.
21. Knuutila S, Autio K, Aalto Y. Online access to CGH data of DNA sequence copy number changes. Am J Pathol 2000;157:689.
22. Klochendler-Yeivin A, Muchardt C, Yaniv M. SWI/SNF chromatin remodeling and cancer. Curr Opin Genet Dev 2002;12:73–9.
23. Sevenet N, Sheridan E, Amram D, Schneider P, Handgretinger R, Delattre O. Constitutional mutations of the hSNF5/INI1 gene predispose to a variety of cancers. Am J Hum Genet 1999;65:1342-8.
24. Mori K, Maeda Y, Kitaura H, Taira T, Iguchi-Ariga SM, Ariga H. MM-1, a novel c-Myc-associating protein that represses transcriptional activity of c-Myc. J Biol Chem 1998;273:29794–800.
25. Hui L, Zhang X, Wu X, et al. Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. Oncogene 2004;23:3013–23.
26. Singh B, Wreesmann VB, Pfister D, et al. Chromosomal aberrations in patients with head and neck squamous cell carcinoma do not vary based on severity of tobacco/alcohol exposure. BMC Genet 2002;3:22.
27. Beder LB, Gunduz M, Ouchida M, et al. Genome-wide analyses on loss of heterozygosity in head and neck squamous cell carcinomas. Lab Invest 2003;83:99–105.
28. Tremmel SC, Gotte K, Popp S, et al. Intratumoral genomic heterogeneity in advanced head and neck cancer detected by comparative genomic hybridization. Cancer Genet Cytogenet 2003;144:165–74.
29. Farrand K, Delahunt B, Wang XL, et al. High resolution loss of heterozygosity mapping of 17p13 in thyroid cancer: Hurthle cell carcinomas exhibit a small 411-kilobase common region of allelic imbalance, probably containing a novel tumor suppressor gene. J Clin Endocrinol Metab 2002;87:4715–21.
30. Roque L, Rodrigues R, Pinto A, Moura-Nunes V, Soares J. Chromosome imbalances in thyroid follicular neoplasms: a comparison between follicular adenomas and carcinomas. Genes Chromosomes Cancer 2003;36: 292–302.
31. Shin MS, Kim HS, Kang CS, et al. Inactivating mutations of CASP10 gene in non-Hodgkin lymphomas. Blood 2002;99:4094–9.
32. Shin MS, Kim HS, Lee SH, et al. Alterations of Fas-pathway genes associated with nodal metastasis in non-small cell lung cancer. Oncogene 2002;21: 4129–36.
33. Takita J, Yang HW, Chen YY, et al. Allelic imbalance on chromosome 2q and alterations of the caspase 8 gene in neuroblastoma. Oncogene 2001;20:4424–32.
34. Jo DG, Kim MJ, Choi YH, et al. Pro-apoptotic function of calsenilin/DREAM/KChIP3. FASEB J 2001; 15:589–91.
35. Mingo-Sion AM, Marietta PM, Koller E, Wolf DM, Van Den Berg CL. Inhibition of JNK reduces G2/M transit independent of p53, leading to endoreduplication, decreased proliferation, and apoptosis in breast cancer cells. Oncogene 2004;23: 596–604.
36. Pellizzari L, D'Elia A, Rustighi A, Manfioletti G, Tell G, Damante G. Expression and function of the homeodomain-containing protein Hex in thyroid cells. Nucleic Acids Res 2000;28:2503–11.
37. Liu Z, Chatterjee TK, Fisher RA. RGS6 interacts with SCG10 and promotes neuronal differentiation. Role of the G γ subunit-like (GGL) domain of RGS6. J Biol Chem 2002;277:37832–9.
38. Maffini MV, Geck P, Powell CE, Sonnenschein C, Soto AM. Mechanism of androgen action on cell proliferation: AS3 protein as a mediator of proliferative arrest in the rat prostate. Endocrinology 2002; 143:2708–14.
39. Resto VA, Caballero OL, Buta MR, et al. A putative oncogenic role for MPP11 in head and neck squamous cell cancer. Cancer Res 2000;60:5529–35.
40. van Belzen N, Dinjens WN, Diesveld MP, et al. A novel gene which is up-regulated during colon epithelial cell differentiation and down-regulated in colorectal neoplasms. Lab Invest 1997;77:85–92.
41. Kurdistani SK, Arizti P, Reimer CL, Sugrue MM, Aaronson SA, Lee SW. Inhibition of tumor cell growth by RTP/rit42 and its responsiveness to p53 and DNA damage. Cancer Res 1998;58:4439–44.
42. Cangul H, Salnikow K, Yee H, Zagzag D, Commes T, Costa M. Enhanced expression of a novel protein in human cancer cells: a potential aid to cancer diagnosis. Cell Biol Toxicol 2002;18:87–96.
43. Ginos MA, Page GP, Michalowicz BS, et al. Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck. Cancer Res 2004;64:55–63.
44. Strausberg RL, Dahl CA, Klausner RD. New opportunities for uncovering the molecular basis of cancer. Nat Genet 1997;15:415–6.