The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags

Helena Brentani[a], Otávia L. Caballero[a], Anamaria A. Camargo[a], Aline M. da Silva[b], Wilson Araújo da Silva, Jr.[c], Emmanuel Dias Neto[d], Marco Grivet[e], Arthur Gruber[f], Pedro Edson Moreira Guimaraes[d], Winston Hide[g], Christian Iseli[h], C. Victor Jongeneel[h], Janet Kelso[g], Maria Aparecida Nagai[i], Elida Paula Benquique Ojopi[d], Elisson C. Osorio[a], Eduardo M. R. Reis[b], Gregory J. Riggins[j], Andrew John George Simpson[a,k], Sandro de Souza[a], Brian J. Stevenson[h], Robert L. Strausberg[l], Eloiza H. Tajara[m], Sergio Verjovski-Almeida[b], The Human Cancer Genome Project/Cancer Genome Anatomy Project Annotation Consortium*, and The Human Cancer Genome Project Sequencing Consortium[†]

[i]Laboratorio de Genética Molecular do Cancer, Departmento de Radiologia, Universidade de São Paulo, Travessa da Rua Dr. Ovídeo Pires de Campos S/N, 4° andar, 05403-010, São Paulo, SP, Brazil; [b]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, 05508-900, São Paulo, SP, Brazil; [m]Departamento de Biologia, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, 15054, São José do Rio Preto, SP, Brazil; [a]Ludwig Institute for Cancer Research, Rua Professor Antonio Prudente 109 4° andar, 01509-010, São Paulo, SP, Brazil; [f]Departamento de Patologia, Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, Avenida Professor Orlando Marques de Paiva 87, 05508-000, São Paulo, SP, Brazil; [c]Fundação Hemocentro de Ribeirão Preto, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Rua Tenente Catão Roxo 2501, 14051-140, Ribeirão Preto, SP, Brazil; [d]Laboratory of Neurosciences (LIM-27), Instituto de Psiquiatria, Faculdade de Medicina, Universidade de São Paulo, Rua Dr. Ovidio de Campos, S/N, 05403-010, São Paulo, SP, Brazil; [l]National Cancer Institute, Bethesda, MD 20892; [j]Duke University Medical Center, Durham, NC 27710; [g]South African National Bioinformatics Institute, University of the Western Cape, Private Bag X17, 7535 Bellville, South Africa; [h]Office of Information Technology, Ludwig Institute for Cancer Research, CH-1066 Épalinges, Switzerland; and [e]Centro de Estudo de Telecomunicações-PUC, Rua Marquês de São Vicente, 225, 22453-900, Rio de Janeiro, RJ, Brazil

Contributed by Walter Bodmer, June 12, 2003

Whereas genome sequencing defines the genetic potential of an organism, transcript sequencing defines the utilization of this potential and links the genome with most areas of biology. To exploit the information within the human genome in the fight against cancer, we have deposited some two million expressed sequence tags (ESTs) from human tumors and their corresponding normal tissues in the public databases. The data currently define ≈23,500 genes, of which only ≈1,250 are still represented only by ESTs. Examination of the EST coverage of known cancer-related (CR) genes reveals that <1% do not have corresponding ESTs, indicating that the representation of genes associated with commonly studied tumors is high. The careful recording of the origin of all ESTs we have produced has enabled detailed definition of where the genes they represent are expressed in the human body. More than 100,000 ESTs are available for seven tissues, indicating a surprising variability of gene usage that has led to the discovery of a significant number of genes with restricted expression, and that may thus be therapeutically useful. The ESTs also reveal novel nonsynonymous germline variants (although the one-pass nature of the data necessitates careful validation) and many alternatively spliced transcripts. Although widely exploited by the scientific community, vindicating our totally open source policy, the EST data generated still provide extensive information that remains to be systematically explored, and that may further facilitate progress toward both the understanding and treatment of human cancers.

Human cancer results from the accrual of genetic mutations or epigenetic changes in the genomes of individual somatic cells. These exert their effect via alterations in the structure and abundance of individual mRNA molecules that, in turn, alter crucial protein-mediated cellular functions. One step in the path toward building a comprehensive molecular portrait of human cancer is the definition of the genes actively expressed in specific tumors and corresponding normal tissues. It is within these sets of genes that we must search for cancer-defining mutations and epigenetic changes and delineate the extent of the molecular milieu within which we will deepen our understanding of cancer. Very short sequence tags of transcripts and hybridization techniques such as microarrays identify many of the genes expressed in tumors and are widely used for measuring the relative levels of gene expression (1, 2). However, longer transcript sequences such as ESTs permit genes expressed in individual cells and tissues to be identified in a completely unambiguous manner, provide additional data on transcript and gene variants, and represent a key source for the search for as yet incompletely characterized genes.

In two large projects, extensive EST sequencing of human tumor tissues has been undertaken: the Cancer Genome Anatomy Project (CGAP) (3) and the Fundaçao de Amparo à Pesquisa do Estado de Sao Paulo/Ludwig Institute for Cancer

Abbreviations: EST, expressed sequence tag; CGAP, Cancer Genome Anatomy Project; HCGP, Human Cancer Genome Project; SNP, single-nucleotide polymorphism; CR, cancer-related.

*The Human Cancer Genome Project/Cancer Genome Anatomy Project Annotation Consortium: Marcio Luis Acencio[n], Mário Henrique Bengtson[o], Fabiana Bettoni[p], Walter F. Bodmer[q], Marcelo R. S. Briones[r], Luiz Paulo Camargo[s], Webster Cavenee[t], Janete M. Cerutti[u], Luís Eduardo Coelho Andrade[v], Paulo César Costa dos Santos[n], Maria Cristina Ramos Costa[w], Israel Tojal da Silva[w], Marcos Roberto H. Estécio[x], Karine Sa Ferreira[w], Frank B. Furnari[t], Milton Faria, Jr.[s], Pedro A. F. Galante[p], Gustavo S. Guimaraes[y], Adriano Jesus Holanda[w], Edna Teruko Kimura[z], Maarten R Leerkes[p], Xin Lu[aa], Rui M. B. Maciel[u], Elizabeth A. L. Martins[bb], Katlin Brauer Massirer[o], Analy S. A. Melo[r], Carlos Alberto Mestriner[cc], Elisabete Cristina Miracca[n], Leandro Lorenco Miranda[s], Francisco G. Nobrega[dd], Paulo S. Oliveira[p], Apuã C. M. Paquola[ee], José Rodrigo C. Pandolfi[cc], Maria Inês de Moura Campos Pardini[ff], Fabio Passetti[p], John Quackenbush[gg], Beatriz Schnabel[r], Mari Cleide Sogayar[o], Jorge E. Souza[p], Sandro R. Valentini[cc], and Andre C. Zaiats[p].

[†]The Human Cancer Genome Project Sequencing Consortium: Elisabete Jorge Amaral[x], Liliane A. T. Arnaldi[u], Amélia Goes de Araújo[w], Simone Aparecida de Bessa[n], David C. Bicknell[q], Maria Eugenia Ribeiro de Camaro[y], Dirce Maria Carraro[p], Helaine Carrer[hh], Alex F. Carvalho[p], Christian Colin[o], Fernando Costa[ii], Cyntia Curcio[z], Ismael Dale Cotrim Guerreiro da Silva[w], Neusa Pereira da Silva[v], Márcia Dellamano[p], Hamza El-Dorry[kk], Enilza Maria Espreafico[ll], Ari José Scattone Ferreira[kk], Cristiane Ayres Ferreira[w], Maria Angela H. Z. Fortes[mm], Angelita Habr Gama[nn], Daniel Giannella-Neto[mm], Maria Lúcia C. C. Giannella[mm], Ricardo R. Giorgi[mm], Gustavo Henrique Goldman[oo], Maria Helena S. Goldman[pp], Christine Hackel[y], Paulo Lee Ho[bb], Elza Myiuki Kimura[qq], Luiz Paulo Kowalski[rr], Jose E. Krieger[ss], Luciana C. C. Leite[bb], Ademar Lopes[rr], Ana Mercedes S. C. Luna[mm], Alan Mackay[tt], Suely Kazue Nagahashi Mari[n], Adriana Aparecida Marques[w], Waleska K. Martins[p], André Montagnini[rr], Mario Mourão Neto[rr], Ana Lucia T. O. Nascimento[bb], A. Munro Neville[uu], Marina P. Nobrega[dd], Mike J. O'Hare[tt], Audrey Yumi Otsuka[vv], Anna Izabel Ruas de Melo[p], Maria Luisa Paçó-Larson[ww], Gonçalo Guimarães Pereira[ii], Neusa Pereira da Silva[v], João Bosco Pesquero[jj], Juliana Gilbert Pessoa[jj], Paula Rahal[x], Claudia Aparecida Rainho[xx], Vanderlei Rodrigues[yy], Silvia Regina Rogatto[xx], Camila Malta Romano[zz], Janaína Gusmão Romeiro[x], Benedito Mauro Rossi[rr], Monica Rusticci[n], Renata Guerra de Sá[yy], Simone Cristina Sant' Anna[qq], Míriam L. Sarmazo[x], Teresa Cristina de Lima e Silva[y], Fernando Augusto Soares[rr], Maria de Fátima Sonati[qq], Josane de Freitas Sousa[ll], Diana Queiroz[y], Valéria Valente[ww], André Luiz Vettore[p], Fabiola Elizabeth Villanova[vv], Marco Antonio Zago[w], and Heloisa Zalcberg[p].

www.pnas.org/cgi/doi/10.1073/pnas.1233632100

Research–Human Cancer Genome Project (HCGP) (4, 5). CGAP, launched in 1997 by the National Cancer Institute, has used single-pass sequencing from the 5′ and/or 3′ extremities of cDNA clones for sequence generation (6). The HCGP project adopted an alternative EST-based strategy, termed ORESTES, which generates sequences biased toward the central coding regions of transcripts (7). The data gathered by these two projects are thus complementary and have been combined into an International Database of Cancer Gene Expression (5), available at http://cgap.nci.nih.gov. They also constitute the basis of the Human Cancer Index at TIGR (www.tigr.org/tdb/tgi/hcgi).

These sequencing initiatives are unique in providing a very large disease-oriented transcriptional database that contains an unprecedented amount of information on expressed human genes. To maximize the benefit of these data, we have made them all publicly available immediately on generation. We here provide a description of these data, as well as their utility for the identification of human genes, the tissue specificity of their expression, and the structure of some of their transcript variants. We find that, although the data have been widely used by the scientific community, much additional information remains untapped.

Materials and Methods

Transcript Sequencing. ORESTES sequences were generated as previously described (7). Tumors and corresponding normal tissues were mostly obtained from the Hospital de Câncer A.C. Camargo, São Paulo, with the exception of purified breast tissue samples that were obtained from University College London (7). In addition, extensive use was made of both breast and colon cell lines. Identification of the source of mRNA for all libraries is available at the CGAP homepage. CGAP sequences were generated from the 3′ and 5′ extremities of both standard and normalized cDNA libraries from a variety of sources as described elsewhere (5).

Transcript to Genome Mapping. A comprehensive reconstruction of human transcripts based on genome data were undertaken based on two datasets: (*i*) a set of alignments between transcripts and genome regions, thoroughly filtered to eliminate the effects of pseudogenes, highly conserved gene families, repetitive elements and EST sequencing errors; (*ii*) a mapping onto the genome of all polyadenylation sites that could be extracted from the chromatograms of the EST sequencing projects, thus marking sites where polyadenylation has been experimentally documented (8). To be included as a spliced cluster, we stipulated that canonical splice sites must be present on at least one of the transcript sequences. ORF identification required that for the longest ORF predicted by ESTScan (www.ch.embnet.org/software/ESTScan.html) for each transcript cluster, the ESTScan score divided by the ORF length had to be >1 (9), an empirical measure that covers >99% of human SwissProt entries. In addition, the ESTScan-predicted ORFs had to have at least three nucleotides 5′ and 3′, as a measure of having at least some 5′ and 3′ UTR.

Generation of a Representative Set of CR Genes. This manually curated compilation was drawn up during a week-long meeting of the Annotation Consortium in August of 2001.[‡] It is a nonredundant list comprising 1,127 human cancer-associated genes based on initial querying of GenBank (www.ncbi.nlm.nih.gov/GenBank/index.html), GenCard (http://bioinfo.weizmann.ac.il/cards/index.html), and Harvard University (http://sbweb.med.harvard.edu/research/breast_cancer/currentlistofgenes.htm) with the words "cancer" and "tumor." The list is available at http://bit.fmrp.usp.br/jamborestes.

Detection of Single-Nucleotide Polymorphisms (SNPs). The transcript sequences were aligned against one another by using FASTA and base quality values files generated by using PHRED (10, 11). A SNP was considered for further analysis only if indicated by reads from at least two different ORESTES libraries, at least two different CGAP reads, or one read from each source. All selected SNPs were compared with SNPs already deposited in dbSNP (www.ncbi.nlm.nih.gov/SNP) by BLASTN. Only newly identified human SNPs were included in experimental validation, by PCR with DNA from a panel of 150 Brazilian individuals from three different ethnic backgrounds: 50 whites (mostly of Western and Southern European ancestry), 50 blacks (mulattos

[n]Laboratorio de Genética Molecular do Cancer, and [nn]Department of Gastroenterology, Faculdade de Medicina, Universidade de São Paulo, Travessa da Rua Dr. Ovídeo Pires de Campos S/N, 4° andar, 05403-010, São Paulo, SP, Brazil; [ee]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, 05508-900, São Paulo, SP, Brazil; [bb]Instituto Butantan, Avenida Vital Brazil 1500, 05503-900, São Paulo, Brazil; [ww]Departamento de Biologia Celular e Molecular e de Bioagentes Patogênicos, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, 14049-900, Ribeirão Preto, SP, Brazil; [u]Laboratory of Molecular Endocrinology, Department of Medicine, Federal University of São Paulo, Rua Pedro de Toledo 781, 12th Floor, 04023-039, São Paulo, SP, Brazil; [o]Chemistry Institute, University of São Paulo, 05513-970, São Paulo, SP, Brazil; [yy]Departamento de Bioquimica e Imunologia, and [ll]Departamento de Biologia Celular e Molecular e Bioagentes Patogênicos, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, 14049-900, Ribeirão Preto, SP, Brazil; [jj]Department of Biophysics, and [r]Departamento de Microbiologia, Imunologia, e Parasitologia, Universidade Federal de São Paulo, Rua Botucatu, 862 3° andar, 04023-062, São Paulo, SP, Brazil; [ff]Laboratório de Biologia Molecular, Hemocentro, Faculdade de Medicina Universidade Estidual Paulista, 18618-000, Botucatu, SP, Brazil; [s]Departamento de Bioinformática–UNAERP, Universidade de Ribeirao Preto, 14096-380, Ribeirão Preto, SP, Brazil; [z]Departamento de Histologia e Embriologia, Instituto de Ciencias Biomedicas, Universidade de São Paulo, Avenida Prof Lineu Prestes 1524, 05508-900, São Paulo, Brazil; [ss]Department of Medicine, Laboratory of Genetics and Molecular Cardiology, Heart Institute (InCor), Universidade de São Paulo, 05403-000, São Paulo, SP, Brazil; [x]Departamento de Biologia, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, 15054, São José do Rio Preto, SP, Brazil; [qq]Department of Clinical Pathology, School of Medical Sciences, State University of Campinas–UNICAMP, 13083-970 Campinas, SP, Brazil; [p]Ludwig Institute for Cancer Research, Rua Professor Antonio Prudente 109, 4° andar, 01509-010, São Paulo, SP, Brazil; [rr]Fundação Antonio Prudente, Hospital do Câncer, Rua Professor Antonio Prudente 211, 01509-900, São Paulo, SP, Brazil; [zz]Departamento de Patologia, Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, Avenida Professor Orlando Marques de Paiva 87, 05508-000, São Paulo, SP, Brazil; [w]Fundação Hemocentro de Ribeirão Preto, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Rua Tenente Catão Roxo 2501, 14051-140, Ribeirão Preto, SP, Brazil; [dd]Instituto de Pesquisa e Desenvolvimento, Universidade do Vale do Paraíba, 12244-000, São José dos Campos, SP, Brazil; [hh]Department of Biological Sciences, Escola Superior de Agricultura ''Luiz de Queiroz,'' Universidade de São Paulo, 13418-900, Piracicaba, SP, Brazil; [v]Departamento de Reumatologia, Universidade Federal de São Paulo, Rua Botucatu 740, 04023-062, São Paulo, SP, Brazil; [kk]Department of Biochemistry, Institute of Chemistry, University of Sao Paulo, Avenida Professor Lineu Prestes 748, 05508-900, São Paulo, SP, Brazil; [cc]Departamento de Ciências Biológicas, Faculdade de Ciências Farmacêuticas de Araraquara, Universidade Estadual Paulista, 14801-902, São Paulo, Brazil; [xx]Departamento de Genética, Instituto de Biociências, Universidade Estadual Paulista, 18618-970, Botucatu, SP, Brazil; [vv]Gynecology Department, Federal University of Sao Paulo, Rua Botucatu 740, 04023-062, São Paulo, SP, Brazil; [oo]Faculdade de Ciencias Farmaceuticas de Ribeirao Preto, Universidade de São Paulo, Avenida do Cafe S/N, 14040-903, Ribeirao Preto, SP, Brazil; [pp]Faculdade de Filosofia, Ciências e Letras, Universidade de São Paulo, Avenida Bandeirantes, 3900, 14040-901, Ribeirão Preto, SP, Brazil; [mm]Laboratory for Cellular and Molecular Endocrinology (LIM-25/HCFMUSP), School of Medicine, Universidade de São Paulo, Avenida Dr. Arnaldo, 455 no. 4305, 01246-903, São Paulo, SP, Brazil; [y]Departamento de Genética e Evolução, State University of Campinas–UNICAMP, 13083-970, Campinas, SP, Brazil; [ii]National Cancer Institute, Bethesda, MD 20892; [q]Cancer Research U.K. Cancer and Immunogenetics Laboratory, Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, United Kingdom; [t]Ludwig Institute for Cancer Research, Department of Medicine, Center for Molecular Genetics, University of California at San Diego, La Jolla, CA 92093-0660; [aa]Ludwig Institute for Cancer Research, Imperial College School of Medicine, St. Mary's Campus, Norfolk Place, London W2 1PG, United Kingdom; [gg]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850; [tt]Ludwig Institute for Cancer Research, Breast Cancer Laboratory, Department of Surgery, Royal Free and University College Medical School, 67-73 Riding House Street, London W1W 7EJ, United Kingdom; and [uu]Ludwig Institute for Cancer Research, Horatio House, 77-85 Fulham Palace Road, London W6 8JC, United Kingdom.

[k]To whom correspondence should be addressed. E-mail: asimpson@licr.org.

[‡]The Human Cancer Genome Project/Cancer Genome Anatomy Project Annotation Consortium, Jamborestes, Aug. 20–25, 2001, São Paulo, Brazil.

Table 1. HCGP and CGAP transcript sequence generation and clustering

Form of gene representation	Number of sequences, clusters, or genes
ORESTES submitted to GenBank	823,121 sequences
CGAP EST submitted to GenBank	1,214,358 sequences
TOTAL EST submitted to GenBank	2,037,479 sequences
Total clusters	32,129 clusters
Total clusters with known genes	22,152 clusters
Clusters without known genes	9,977 clusters
Clusters without known genes but with coding potential	1,285 clusters
Estimated total genes based on HCGP and CGAP data	23,437 genes

Table 2. Novel genes and definition of genes with restricted expression patterns utilizing EST clusters containing sequences from the HCGP and CGAP databases

Gene type	No. of genes
Positionally cloned genes	5
Genes cloned on the basis of homology	16
Tissue restricted expression	
Brain	9
Breast	14
Colon	11
Prostate	15
Cancer/testis	4
Various	13
Total	87

Full details are provided in Tables 6 and 7.

excluded), and 50 Japanese. All groups reported no racial admixture in their four grandparents.

Alternative Splicing. Two approaches were used. In the first, exon skipping was detected by using J_EXPLORER (available for download from www.sanbi.ac.za/exon_skipping), which reduces the complexity of the gene sequences to a set of possible splice junctions used to search for ESTs spanning the annotated exon–exon junctions (12). The second approach listed all variant exons as revealed by transcript to genome alignments by using only those cDNAs that span at least two exons. Variants are represented by a binary matrix where each row corresponds to a sequence, and each column corresponds to an exon (33).

Results and Discussion

Gene Identification. Collectively, the Fundaçao de Amparo à Pesquisa do Estado de Sao Paulo/Ludwig Institute for Cancer Research–Human Cancer Genome Project and the Cancer Genome Anatomy Project have deposited over two million sequences from tumors and normal tissues in GenBank (Table 1). To give perspective, this number of sequences is greater than that required to complete the high-quality shotgun sequencing of ≈40 bacterial genomes. The two projects are the largest individual contributors to the public human EST database and are responsible for >40% of all publicly available human EST data available (dbEST release 122002, www.ncbi.nlm.nih.gov/dbEST).

The first, and most obvious, use of ESTs is to define genes within the human genome. The precise experimental definition of these structures is crucial not only for CR research but also for research undertaken in all aspects of human biology. To achieve this, we have organized our EST data by aligning them, together with all other publicly available transcript sequences, to the human genome assembly. This allows an accurate organization of sequences into clusters corresponding to individual genes even in cases of minimal overlap or high similarity between paralogs. The transcript to genome alignment also provides an arbiter of quality, in that ESTs and EST clusters that span splice sites (referred to as spliced ESTs and spliced clusters, respectively) but correspond to still incompletely defined genes can be unambiguously distinguished from EST sequences derived from contaminating DNA or immature mRNA molecules.

We identified transcript clusters that contain at least one EST derived from one of the two cancer transcript sequencing projects, and which we judge to have a high probability of being derived from authentic expressed genes due to the presence of a reportedly full-length cDNA sequence, evidence of splicing or the presence of a predicted ORF. A total of 22,152 of the clusters contain full-length cDNAs. We use the presence of such full-length cDNA-containing transcript clusters as the gold standard for identification of human genes and refer to the genes they define as known genes. The known genes mapped by our ESTs comprise a subset of the total of the 29,332 known genes currently contained within the human genome, indicating that ≈75.5% of known human genes are expressed in the tumors and normal tissues we have studied. The 25% of genes that were not represented by our ESTs are defined by ESTs from other projects or full length cDNAs from various sources. We assume that these genes are either not expressed or are expressed at low levels in the cells and tissues that we have studied. Some 4,000 of the known genes to which our ESTs map have been defined since the publication of the draft human genome in 2001 (13, 14). In the original published draft, these genes were represented only by the EST clusters. Interestingly, this number is 75% of all novel genes defined in this period, demonstrating that our ESTs equally well represent previously known genes and those still being defined. The preexistence of the EST clusters led directly to the generation of full-length cDNAs of a number of CR genes. The ESTs served either to indicate the existence of genes within defined regions of the human genome or as evidence of previously unknown members of paralogous families (Table 2 and cited in detail in Table 6, which is published as supporting information on the PNAS web site).

To investigate the extent to which our ESTs represent known CR genes, we compiled a list of 1,127 human genes known or presumed to play a role in the process of transformation to malignancy, which we refer to as CR (cancer-related) genes (see http://bit.fmrp.usp.br). The CR set contains extensively studied CR genes such as *TP53*, *RB1*, *BRCA1*, *CDKN2*, and *ERBB2*, as well as members of paralogous gene families that function in critical signal-transducing pathways, such as cadherins, integrins, and mitogen-activated protein kinases. Although incomplete, we believe the CR set is representative of well characterized genes relevant to the development of human cancer. Most importantly, many of these have been cloned on the basis of strategies, such as positional cloning, that do not depend on transcript abundance. Of the genes in the list, we found that 1,009 (89%) have at least one corresponding ORESTES sequence; 1,099 genes (97%) have at least one CGAP sequence; and 1,102 genes (97%) have EST sequences derived from at least one of the two projects. Of the 25 genes for which we have not generated ESTs, 18 have no EST coverage at all, indicating that their overall expression is at very low levels in the human body.

The 9,977 clusters composed only of ESTs represent an unknown number of additional genes for which a full-length cDNA is not yet available (Table 1). A particular difficulty in relating these clusters to genes lies in the very high level of heterogeneity at their 3′ ends (8). Thus, noncoding, spliced clusters may be extensions of known genes; those that are not

might represent multiple different forms of as-yet-undefined genes. To avoid the complications of the 3′ heterogeneity of transcripts, and also to exclude clusters that represent noncoding transcripts, one approach is to use only clusters that contain an identifiable ORF (9). Although we identified 9,977 EST clusters with splicing, only 1,285 are apparently coding sequences. We take these as each representing an individual human gene in the present discussion, although this one-to-one correlation remains to be demonstrated by full-length cDNA sequencing. Nevertheless, on the basis of these criteria, the two million or so EST sequences that we have generated translate into 23,437 genes, of which 1,285 are currently not defined by full-length cDNAs. Taking our data to include 75.5% of all genes, on the basis of the coverage of known genes described above, this value would predict a total of 31,042 genes in the human genome, in line with other recent estimates (13–16). It should be noted, however, that this value is likely to steadily increase as more small, nonspliced, and nonprotein-coding genes are discovered (16–18). These analyses indicate, however, that about the same number of genes confirmed by full-length cDNA sequencing since the publication of the draft genome sequence could be added to the genome immediately on the basis of the EST clusters with predicted ORFs to which we have contributed. These data are likely to contain a rich variety of novel genes relevant to cancer. For example, we found in September 2001 that among clusters of this kind, there were 19 that apparently encode novel paralogs of genes in the CR list. Of these, five have already been deposited by others in the GenBank database : CAMK1 (accession no. BC032726), MBD2 (accession no. AY038022), TLN2 (accession no. NM_015059), BCHE (accession no. BC028713), and CNK (accession no. AK054808). The remainder will be described in detail elsewhere as the determination of the full-length sequences is completed.

We have previously shown that the number of ESTs corresponding to a gene is an indication of its level of expression (19). Thus, genes more frequently expressed in tumors than in normal tissues, for example, can be identified from the EST data. Several examples of genes with differential gene expression identified from our ESTs are included in Table 2 and are listed in detail in Table 7, which is published as supporting information on the PNAS web site. It should be pointed out, however, that in general we have rather less data from normal than tumor-derived tissues, which limits the power of this approach at present. Moreover, both normal and tumor samples can be highly heterogeneous, with normal tissues in particular being mixtures of different cell types. Thus, all conclusions drawn from analysis of our sequencing data require confirmation both in subsequent RT-PCR experiments and ultimately by histochemical analysis at the protein level. Nevertheless, EST cluster size appears to be a useful general indicator of at least overall gene expression and comparison of the number of CGAP and ORESTES sequences with serial analysis of gene expression (SAGE) tags for the same genes are positively correlated ($r = 0.6$), despite the very different sets of tissues that have been accessed by the two databases. The average cluster size for all of the known human genes for which we have generated ESTs to date is 606, whereas the spliced EST clusters with predicted ORFs contain an average of only 19 ESTs. Thus, the latter represent genes much less frequently expressed than the average in the tissues we have studied. The average cluster size of the novel known genes added to the databases since the publication of the draft genome is 55, indicating that with time genes with lower levels of expression are being defined.

Tissue Specificity of Gene Expression. The genomes of all cells in the human body are copied from that in the fertilized egg, although they are continually modified over the lifetime of the individual due to the relentless acquisition of somatic mutations. Although it is these somatic alterations that are ultimately responsible for human cancer, their relative importance will depend on the transcriptional programming of the cell in which they occur. Thus, the careful documentation of the tissue specificity of gene expression is vital to understanding the genetic basis of cancer. For this purpose, the availability of details of the tissue origin of our ESTs represents a powerful resource. All information on the tissue origins of our ESTs is available on the CGAP web site (http://cgap.nci.nih.gov). For each of seven human tissues (brain, head and neck, colon, lung, breast, uterus, and kidney), we have generated >100,000 EST sequences providing, in each case, a deep survey of gene expression. This value is comparable to the number of serial analysis of gene expression tags typically generated in individual experiments for the detection of differential gene expression and far in excess of any other EST compilation for individual tissues (20).

Table 3. Documentation of gene expression in individual human tissues and estimation of tissue transcriptome complexity

Tissue	ESTs	Total clusters with known genes	Clusters with splicing and coding potential	Estimated total tissue transcriptome
Brain	185,193	12,746	282	13,028
Breast	137,867	10,153	227	10,380
Colon	186,870	12,218	326	12,544
Head/neck	186,298	11,698	261	11,959
Kidney	115,096	12,368	341	12,709
Lung	166,469	13,076	314	13,390
Ovary	53,070	7,472	137	7,609
Prostate	85,910	10,004	224	10,228
Uterus	125,079	11,614	244	11,858
Others	795,627	19,818	1,079	20,897

On the basis of ESTs that correspond to known genes and spliced EST clusters with predicted ORFs, we have evidence for the expression of between 10,000 and 13,500 genes for the seven tissue types that we have explored in depth, with lung having the highest number of expressed genes so far (Table 3). This indicates that not >57% of all of the genes defined by our ESTs genes are expressed in any one tissue type. It would thus appear that gene utility is quite variable between tissues. For example, if we take the 16,084 genes collectively expressed in breast, colon, and head and neck, there is evidence for expression for only 7,500 (47%) in all three tissues, whereas 4,785 (30%) are expressed in only one of the three tissues (Fig. 1). These numbers are arrived at despite the undoubted impurity of the tissue

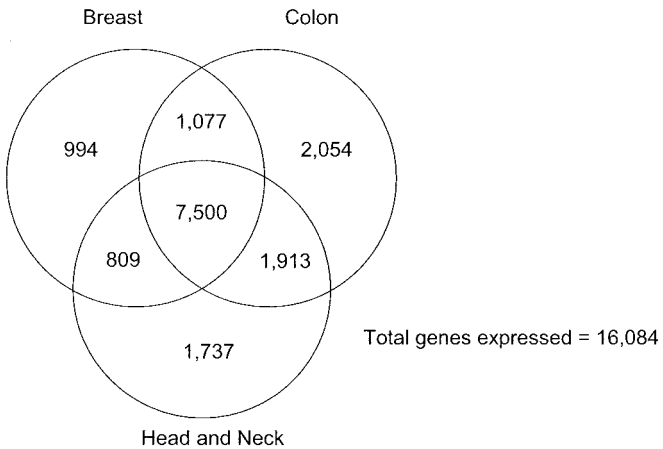

Fig. 1. Coexpression of genes among breast, colon, and head/neck.

Table 4. Pairwise comparison of gene expression in human tissues

Tissue	Brain	Breast	Colon	Head/neck	Kidney	Lung	Uterus
Total genes	13,028	10,380	12,544	11,959	12,709	13,390	11,858
Brain	X	8,484	9,745	9,370	9,797	10,337	9,502
Breast		X	8,577	8,309	8,434	8,715	8,354
Colon			X	9,413	9,806	10,176	9,610
Head/neck				X	9,316	9,771	9,155
Kidney					X	10,299	9,621
Lung						X	9,922
Uterus							X

The numbers in each box refer to the number of genes expressed in common by the two tissues that head the rows and columns that intersect at that box. The number of genes that each of these tissues does not express in common with the other can be determined by subtracting the value in the box from the total number of genes expressed by each of the tissues, as listed in Table 3.

samples, in particular those derived from bulk tumors. Indeed, such impurity is readily detected by comparing ESTs present in libraries derived from bulk tumors but not corresponding cell lines, as can be easily accomplished by using the tools at the CGAP home page (http://cgap.nci.nih.gov). In the case of colon, for example, genes represented by ESTs from tumors but not cell lines include many examples (see Table 8, which is published as supporting information on the PNAS web site) that would not be expected to be present in epithelial cells. Given this tissue contamination, we might thus presume tissue specificity to be even higher than that measured, because the same contaminating lymphoid, fibroblast, endothelial, and smooth muscle cells are likely to be contaminants common to tumors from most locations.

On the other hand, we wondered whether the apparent high level of differential gene expression between tissues and between bulk tumors and cell lines might be exaggerated due to a still-insufficient level of sequence coverage. To address this possibility, we used data from the recently completed generation of several million transcript tags by Massively Parallel Signature Sequencing (MPSS) (21) from two human cell lines, one from breast and one from colon. In the MPSS experiment, the depth of sequencing guarantees the identification of all transcripts present at the level of at least one copy per cell (22). We found that in the MPSS experiment, ≈10,000 and 15,000 genes were expressed in the breast and colon cells, respectively, of which only 8,500 were expressed in both cell lines (21). In addition, inspection of the MPSS data also revealed that the great majority of the ESTs that we hypothesized were not derived from tumor epithelial cells and were also not detected in the colon cell line sequenced by MPSS, and those that were, were at very low levels (Table 8). Thus on both counts, the levels of EST coverage that we have achieved appear to provide a view of gene expression not significantly altered by very much deeper levels of transcript tagging.

A pairwise analysis of the tissues for which we have generated >100,000 ESTs indicates a consistency of shared and specific gene expression, with ≈70% of genes being expressed in common by any given pair (Table 4). These findings are consistent with the structure and function of human tissues being defined by the usage of highly variable permutations of genes, with almost all being expressed in more than one tissue. This would contrast, for example, with a model that requires a fixed set of ubiquitously expressed housekeeping genes with a substantial number of entirely tissue specific genes for individual tissues.

The EST data have proved extremely robust when used for the identification of genes with defined patterns of tissue specificity. Table 2 indicates examples of published studies that have identified CR genes with particular expression profiles and where the clusters representing the genes contained ESTs from our projects. The details of these studies are listed in Table 7. Of particular interest has been the identification of genes that are restricted to organs such as the breast and prostate that could serve as therapeutic targets for cancers in these organs. In addition, growing interest is being focused on genes restricted in expression to normal testis and tumors (CT-antigens) that are important potential targets of therapeutic vaccines. The relatively sparse use of the data for tissue types other than breast and prostate indicates there are likely to be many more genes with restricted expression still to be identified by using our data.

Identification of Gene and Transcript Variants. In addition to gene definition and identification of the tissues in which a gene is expressed, ESTs also provide information concerning the existence of germline variants and alternatively spliced transcript forms. The former can have a significant impact on an individual's predisposition to cancer. Nevertheless, because such variants take the form of SNPs for the most part, the relatively high sequencing error rate (≈1%) of the single-pass transcript sequences we have generated requires that potential variants have to be carefully evaluated. Nevertheless, many studies have reported large-scale analyses of SNPs that have incorporated the ESTs we have generated (23, 24).

By using the original chromatograms, we identified a total of 237 SNPs for genes in the CR list, which are already listed in dbSNP. Of these, 47 were identified only with ORESTES data, 72 only with CGAP data, and 118 with both. In addition, a total of 210 further putative novel SNPs were identified by using the ORESTES data, and 295 were identified by using the CGAP data. Other studies have indicated that a high percentage of SNPs based on conventional ESTs, such as those generated in the CGAP project, can be subsequently verified (24). We suspected that, because ORESTES is a PCR-based methodology, a lower percentage of potential variants may be real. To test this, we examined experimentally 20 of the nonsynonymous SNPs identified with the ORESTES data alone and found that three (15%) were present in a bank of 100 normal human DNA sequences (Table 5). Although this percentage is (as expected) low, the ORESTES-based variants tend to be of particular

Table 5. Validated new human SNPs based on ORESTES clusters

Gene name	Accession no.	Variation	Frequency of the variant allele in 150 Brazilian individuals		
			Whites	Blacks	Japanese
TM4SF3	NM_004616	G/C G73A	38% C	8% C	27% C
CGM2	X98311	T/A F1201	39% A	63% A	62% A
KISS1	U43527	A/G Q36R	8% G	—	—

—, Not yet performed.

interest, because they most frequently lie within the coding regions, thus potentially generating altered protein molecules. On the basis of our validation rate, we predict there are likely to be >30 true novel nonsynonymous SNPs identifiable by the ORESTES data alone within the CR list.

Alternative splicing generates variability at the transcriptome level and is conceivably of direct relevance to the generation of the malignant phenotype (25–27). Extensive utilization has been made of the ESTs we have generated for the identification of alternatively spliced genes within the human genome (12, 28–32). We explored the degree of variability due to alternative exon usage in the set of 1,124 CR genes by using two different approaches. The first approach, which detects transcripts displaying exon skipping through the use of a software called J-EXPLORER (12), predicts possible splice junctions and then searches the EST databases to identify transcripts spanning the annotated exon–exon junction. The second approach lists all variants based on transcript to genome mapping and is thus less stringent (33). The former found evidence of alternative splicing for 21.3% of all CR genes, with an average of 1.4 variants per gene, and the latter found that 47.5% of all genes in the list undergo alternative splicing with a total of 3,179 variants, and an average of 3.17 variants per gene. A total of 210 genes were found by both approaches as having more than one splicing variant. We sought to validate exon-skipping events identified in three of the genes: the skipping of exon 7 in CD 53 (NM_00560), the skipping of exon 26 in PTPN13 (NM_006264), and the insertion of an additional exon between denominated exons 1 and 2 in NM23A (NM_000269). We had detected each of these only in ESTs derived from tumors, thus suggesting that the truncated transcripts might be tumor specific. However, by using a strategy that enhanced the amplification of the transcript with the skipped exon with an RT-PCR primer that spanned the novel splice site formed, we were able, in each case, to detect the alternative transcript in both normal and tumor derived samples. Thus, although carefully documented alternative splicing events can almost always be readily validated, tumor-specific alternative splicing may be rare.

Conclusion

The data we have generated have contributed to the identification of genes within the human genome, the definition of the tissue specificity of their expression, and the discovery of SNPs and alternatively spliced transcripts. It has been our intent that this should speed progress in the diagnosis, treatment, and understanding of human cancer. The number and variety of published studies in which the data we have generated appear to have played a significant role attest to the value of the open-source approach. Undoubtedly, vastly more progress has been made than if we had opted to concentrate on in-house utilization of our findings. Additionally, the thousands of high-quality EST clusters that are both spliced and contain predicted ORFs, the hundreds of genes that appear to have strongly tissue-restricted expression, as well as the thousands of potential SNPs and alternative splice forms that our data define, suggest that much more remains to be discovered within the existing publicly available data. We aim to continue generating sequence-based CR data with the aim of further improving the accurate transcriptional description of human tumors and the normal tissues from which they arise. We believe this will not only provide an important underlying support for broad-based cancer research but will also serve as the source for novel discoveries and hypotheses that will ultimately lead to improved cancer care.

We are indebted Lloyd J. Old, Scientific Director of the Ludwig Institute for Cancer Research, for support and stimulation; to J. Fernando Perez, Scientific Director of the Fundaçao de Amparo à Pesquisa do Estado de Sao Paulo (FAPESP); and to Richard Klausner, Director of the National Cancer Institute, without whom these studies would not have been possible. We are also greatly indebted to Dr. Jucara Parra for dedication to the successful execution of these studies through her role as Project Manager for the FAPESP/Ludwig Institute for Cancer Research–Human Cancer Genome Project.

1. Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., *et al.* (1999) *Cancer Res.* **59,** 5403–5407.
2. DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A. & Trent, J. M. (1996) *Nat. Genet.* **14,** 457–460.
3. Strausberg, R. L., Dahl, C. A. & Klausner, R. D. (1997) *Nat. Genet.* **15,** Spec. No., 415–416.
4. Bonalume, N. R. (1999) *Nature* **398,** 450.
5. Strausberg, R. L., Camargo, A. A., Riggins, G. J., Schaefer, C. F., De Souza, S. J., Grouse, L. H., Lal, A., Buetow, K. H., Boon, K., Greenhut, S. F., *et al.* (2002) *Pharmacogenomics. J.* **2,** 156–164.
6. Strausberg, R. L., Buetow, K. H., Emmert-Buck, M. R. & Klausner, R. D. (2000) *Trends Genet.* **16,** 103–106.
7. Dias, N. E., Garcia, C. R., Verjovski-Almeida, S., Briones, M. R., Nagai, M. A., da Silva, W., Jr., Zago, M. A., Bordin, S., Costa, F. F., Goldman, G. H., *et al.* (2000) *Proc. Natl. Acad. Sci. USA* **97,** 3491–3496.
8. Iseli, C., Stevenson, B. J., De Souza, S. J., Samaia, H. B., Camargo, A. A., Buetow, K. H., Strausberg, R. L., Simpson, A. J., Bucher, P. & Jongeneel, C. V. (2002) *Genome Res.* **12,** 1068–1074.
9. Iseli, C., Jongeneel, C. V. & Bucher, P. (1999) *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 138–148.
10. Ewing, B. & Green, P. (1998) *Genome Res.* **8,** 186–194.
11. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998) *Genome Res.* **8,** 175–185.
12. Hide, W. A., Babenko, V. N., van Heusden, P. A., Seoighe, C. & Kelso, J. F. (2001) *Genome Res.* **11,** 1848–1853.
13. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) *Nature* **409,** 860–921.
14. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001) *Science* **291,** 1304–1351.
15. Roest, C. H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., *et al.* (2000) *Nat. Genet.* **25,** 235–238.
16. Ewing, B. & Green, P. (2000) *Nat. Genet.* **25,** 232–234.
17. Reymond, A., Camargo, A. A., Deutsch, S., Stevenson, B. J., Parmigiani, R. B., Ucla, C., Bettoni, F., Rossier, C., Lyle, R., Guipponi, M., *et al.* (2002) *Genomics* **79,** 824–832.
18. Eddy, S. R. (2002) *Cell* **109,** 137–140.
19. Camargo, A. A., Samaia, H. P., Dias-Neto, E., Simao, D. F., Migotto, I. A., Briones, M. R., Costa, F. F., Nagai, M. A., Verjovski-Almeida, S., Zago, M. A., *et al.* (2001) *Proc. Natl. Acad. Sci. USA* **98,** 12103–12108.
20. Boon, K., Osorio, E. C., Greenhut, S. F., Schaefer, C. F., Shoemaker, J., Polyak, K., Morin, P. J., Buetow, K. H., Strausberg, R. L., De Souza, S. J. *et al.* (2002) *Proc. Natl. Acad. Sci. USA* **99,** 11287–11292.
21. 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., *et al.* (2003) *Proc. Natl. Acad. Sci. USA* **100,** 4702–4705.
22. 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.
23. Clifford, R., Edmonson, M., Hu, Y., Nguyen, C., Scherpbier, T. & Buetow, K. H. (2000) *Genome Res.* **10,** 1259–1265.
24. Irizarry, K., Kustanovich, V., Li, C., Brown, N., Nelson, S., Wong, W. & Lee, C. J. (2000) *Nat. Genet.* **26,** 233–236.
25. Caballero, O. L., De Souza, S. J., Brentani, R. R. & Simpson, A. J. (2001) *Dis. Markers* **17,** 67–75.
26. Chun, S. Y., Bae, O. S. & Kim, J. B. (2000) *J. Korean Med. Sci.* **15,** 696–700.
27. Sanchez, L. M., Hajos, S. E., Basilio, F. M., Mongini, C. & Alvarez, E. (2001) *Oncol. Rep.* **8,** 145–151.
28. Brett, D., Pospisil, H., Valcarcel, J., Reich, J. & Bork, P. (2002) *Nat. Genet.* **30,** 29–30.
29. Kan, Z., States, D. & Gish, W. (2002) *Genome Res.* **12,** 1837–1845.
30. Modrek, B., Resch, A., Grasso, C. & Lee, C. (2001) *Nucleic Acids Res.* **29,** 2850–2859.
31. Xie, H., Zhu, W. Y., Wasserman, A., Grebinskiy, V., Olson, A. & Mintz, L. (2002) *Genomics* **80,** 326–330.
32. Xu, Q., Modrek, B. & Lee, C. (2002) *Nucleic Acids Res.* **30,** 3754–3766.
33. Sakabe, N. J., Souza, J. E. S., Galante, P. F. A., Oliveira, P. S. L., Passetti, F., Brentani, H., Osorio, E. C., Zaiats, A., Leerkes, M. R., Kitajima, J. P., *et al.* (2003) *Proc. Fr. Acad. Sci.*, in press.