The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome

Anamaria A. Camargo[a], Helena P. B. Samaia[a], Emmanuel Dias-Neto[a], Daniel F. Simão[a], Italo A. Migotto[a], Marcelo R. S. Briones[b], Fernando F. Costa[c], Maria Aparecida Nagai[d], Sergio Verjovski-Almeida[e], Marco A. Zago[f], Luis Eduardo C. Andrade[g], Helaine Carrer[h], Hamza F. A. El-Dorry[e], Enilza M. Espreafico[i], Angelita Habr-Gama[j], Daniel Giannella-Neto[k], Gustavo H. Goldman[l], Arthur Gruber[m], Christine Hackel[n], Edna T. Kimura[o], Rui M. B. Maciel[p], Suely K. N. Marie[q], Elizabeth A. L. Martins[r], Marina P. Nóbrega[s], Maria Luisa Paçó-Larson[i], Maria Inês M. C. Pardini[t], Gonçalo G. Pereira[u], João Bosco Pesquero[v], Vanderlei Rodrigues[w], Silvia R. Rogatto[x], Ismael D. C. G. da Silva[y], Mari C. Sogayar[e], Maria de Fátima Sonati[z], Eloiza H. Tajara[aa], Sandro R. Valentini[bb], Fernando L. Alberto[c], Maria Elisabete J. Amaral[aa], Ivy Aneas[j], Liliane A. T. Arnaldi[p], Angela M. de Assis[c], Mário Henrique Bengtson[e], Nadia Aparecida Bergamo[x], Vanessa Bombonato[t], Maria E. R. de Camargo[n], Renata A. Canevari[x], Dirce M. Carraro[h], Janete M. Cerutti[p], Maria Lucia C. Corrêa[k], Rosana F. R. Corrêa[j], Maria Cristina R. Costa[f], Cyntia Curcio[o], Paula O. M. Hokama[t], Ari J. S. Ferreira[e], Gilberto K. Furuzawa[p], Tsieko Gushiken[t], Paulo L. Ho[r], Elza Kimura[z], José E. Krieger[j], Luciana C. C. Leite[r], Paromita Majumder[j], Mozart Marins[l], Everaldo R. Marques[l], Analy S. A. Melo[b], Monica Melo[c], Carlos Alberto Mestriner[bb], Elisabete C. Miracca[d], Daniela C. Miranda[m], Ana Lucia T. O. Nascimento[r], Francisco G. Nóbrega[s], Élida P. B. Ojopi[x], José Rodrigo C. Pandolfi[bb], Luciana G. Pessoa[v], Aline C. Prevedel[z], Paula Rahal[aa], Claudia A. Rainho[x], Eduardo M. R. Reis[e], Marcelo L. Ribeiro[n], Nancy da Rós[d], Renata G. de Sá[w], Magaly M. Sales[t], Simone Cristina Sant'anna[z], Mariana L. dos Santos[d], Aline M. da Silva[e], Neusa P. da Silva[g], Wilson A. Silva, Jr.[f], Rosana A. da Silveira[t], Josane F. Sousa[i], Daniella Stecconi[e], Fernando Tsukumo[u], Valéria Valente[i], Fernando Soares[cc], Eloisa S. Moreira[a], Diana N. Nunes[a], Ricardo G. Correa[a], Heloisa Zalcberg[a], Alex F. Carvalho[a], Luis F. L. Reis[a], Ricardo R. Brentani[a], Andrew J. G. Simpson[a,dd], and Sandro J. de Souza[a]

[a]Ludwig Institute for Cancer Research, 01509-010, São Paulo, Brazil; [g]Departamento de Reumatologia, and [y]Departamento de Biofísica, [b]Escola Paulista de Medicina, Universidade Federal de São Paulo (UNIFESP), 04023-062, São Paulo, Brazil; [c]Hemocentro, Universidade Estadual de Campinas, 13089-970, São Paulo, Brazil; [d]Departamento de Radiologia da Faculdade de Medicina da Universidade de São Paulo, 01296-903, São Paulo, Brazil; [e]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, 05513-970, São Paulo, Brazil; [f]Departamento de Clínica Médica, [i]Departamento de Biologia Celular e Molecular e Bioagentes Patogênicos, and [w]Departamento de Bioquímica e Imunologia, Faculdade de Medicina de Ribeirão Preto, 3900 14049-900, São Paulo, Brazil; [h]Departamento de Ciências Biológicas, Escola Superior de Agricultura Luíz de Queiroz, Universidade de São Paulo, 13418-900, São Paulo, Brazil; [j]Instituto do Coração (INCOR), Faculdade de Medicina, Universidade de São Paulo, 05403-000, São Paulo, Brazil; [k]Laboratório de Nutrição e Doenças Metabólicas, and [q]Departamento de Neurolologia, Faculdade de Medicina, Universidade de São Paulo, 01246-903, São Paulo, Brazil; [l]Departamento de Ciências Farmacêuticas, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, 14040-903, São Paulo, Brazil; [m]Departamento de Patologia, Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, 05508-000, São Paulo, Brazil; [n]Departamento de Genética Médica, Faculdade de Ciências Médicas, Universidade de Campinas, 13081-970, São Paulo, Brazil; [o]Departamento de Histologia Embriologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, 05508-000, São Paulo, Brazil; [p]Departamento de Medicina, Universidade Federal de São Paulo, 04029-032, São Paulo, Brazil; [r]Centro de Biotecnologia, Instituto Butantã, 05503-900, São Paulo, Brazil; [s]Instituto de Pesquisa e Desenvolvimento, Universidade do Vale do Paraíba, 12244, São Paulo, Brazil; [t]Hemocentro, Faculdade de Medicina de Botucatu, Universidade Estadual Paulista, 18618-000, São Paulo, Brazil; [u]Departamento de Genética e Evolução, Instituto de Biologia, and [z]Departamento de Patologia Clínica, Faculdade de Ciências Médicas, Universidade de Campinas, 13083-970, São Paulo, Brazil; [x]Departamento de Genética, Instituto de Biociências, Universidade Estadual Paulista, 18618-000, São Paulo, Brazil; [y]Departamento de Ginecologia e Obstetrícia, Escola Paulista de Medicina, 04301-900, São Paulo, Brazil; [aa]Departamento de Biologia, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, 15054, São Paulo, Brazil; [bb]Departamento de Ciências Biológicas, Faculdade de Ciências Farmacêuticas de Araraquara, Universidade Estadual Paulista, 14801-902, São Paulo, Brazil; and [cc]Departamento de Anatomia Patológica, Hospital A. C. Camargo, 01509-010, São Paulo, Brazil

Edited by Robert H. Waterston, Washington University School of Medicine, St. Louis, MO, and approved August 2, 2001 (received for review April 11, 2001)

Open reading frame expressed sequences tags (ORESTES) differ from conventional ESTs by providing sequence data from the central protein coding portion of transcripts. We generated a total of 696,745 ORESTES sequences from 24 human tissues and used a subset of the data that correspond to a set of 15,095 full-length mRNAs as a means of assessing the efficiency of the strategy and its potential contribution to the definition of the human transcriptome. We estimate that ORESTES sampled over 80% of all highly and moderately expressed, and between 40% and 50% of rarely expressed, human genes. In our most thoroughly sequenced tissue, the breast, the 130,000 ORESTES generated are derived from transcripts from an estimated 70% of all genes expressed in that tissue, with an equally efficient representation of both highly and poorly expressed genes. In this respect, we find that the capacity of the ORESTES strategy both for gene discovery and shotgun transcript sequence generation significantly exceeds that of conventional ESTs. The distribution of ORESTES is such that many human transcripts are now represented by a scaffold of partial sequences distributed along the length of each gene product. The experimental joining of the scaffold components, by reverse transcription–PCR, represents a direct route to transcript finishing that may represent a useful alternative to full-length cDNA cloning.

The identification of all human genes and transcripts remains a goal of highest priority and a rate-limiting step in progress toward the exploitation of the completed draft human genome sequence. The complexity and variability of human gene structure prevents their direct identification within genome sequence, and supporting data from protein and/or transcript sequence are necessary (1–3). The range of estimates of gene numbers that emanated from the analysis of the draft sequence indicates that we were far from defining a complete catalogue of human genes based on transcript evidence available at that time (4, 5). Thus, there was a pressing need for the generation of further transcript sequence to accelerate the attainment of this goal.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: EST, expressed sequence tag; ORESTES, ORF EST; RT-PCR, reverse transcription–PCR.

See Commentary on page 11837.

[dd]To whom reprint requests should be addressed. E-mail: asimpson@node1.com.br.

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

The majority of human transcript sequences take the form of single-pass sequencing from the extremities of cDNA clones, known as expressed sequence tags (ESTs) (6–10). These data, however, can be used only to compile complete sequences of short, abundant transcripts through the generation of contigs from overlapping end sequences. In other cases, one by one full-length cDNA cloning and sequencing is required. Although full-length cDNA cloning and sequencing strategies have been developed, these have proven to be most efficient for short transcripts (10–12).

We have developed a modification of the EST strategy termed ORESTES (ORF ESTs; refs. 13 and 14), where sequences are produced along the length of transcripts rather than just from clone extremities, which can allow effective shotgun transcript sequencing, accelerating both transcript definition and genome annotation. The success of a pilot project permitted us to undertake a large-scale ORESTES program, the Fundação de Amparo à Pesquisa do Estado de São Paulo/Ludwig Institute for Cancer Research-Human Cancer Genome Project (FAPESP/LICR-HCGP; ref. 15). We have now released a dataset of 696,745 sequences to the scientific community as a contribution to the task of defining human genes and their products. Our analysis shows that the ORESTES strategy is extremely efficient in terms of transcript sequence generation and that the data accumulated to date permit the generation of transcript scaffolds from which full-length transcript data can be readily generated.

Materials and Methods

Biological Samples and Poly(A)$^+$ Extraction. The samples selected for RNA extraction were derived from tumor and surrounding normal tissues excised from patients during surgery at the Hospital do Câncer A. C. Camargo, São Paulo, Brazil. All specimens were collected after explicit informed consent. Samples were frozen in liquid nitrogen and allowed to partially thaw to −20°C for microdissection. High quality poly(A)$^+$ RNA was prepared as described previously (13, 14).

cDNA Production and Sequencing. Samples of 10 to 30 ng of purified mRNA were heated at 65°C for 5 min and subjected to reverse transcription at 37°C for 60 min in the presence of 200 units of Moloney murine leukemia virus reverse transcriptase and 15 pmols of a randomly selected primer in a final volume of 20 μl. The primers used for cDNA synthesis and amplification varied between 12 to 33mers and 20 to 80% guanine and cytosine. After cDNA synthesis, 1 μl of the single stranded cDNA was PCR amplified. A touch down PCR with 45 cycles was used after the cDNA denaturing at 75°C. Annealing temperatures varied from 60°C to 41°C (with progressive reductions of 1°C to 2°C per cycle). Profiles composed of a DNA smear were size selected, cloned, and sequenced by using standard protocols as previously described (13, 14).

Computational Analysis. The automated protocol for the analysis of the experimentally generated data has been described elsewhere (13, 14). All of the ORESTES data were loaded into a relational database (MYSQL). We have also used a locally developed database called Integrated Database of Human Transcripts (unpublished work) that integrates information on human transcripts from different sources (including UniGene build no. 128). The different sets of ORESTES sequences, the random sets of 3′ and 5′ ESTs, and the set of full-length human cDNAs were all generated from these relational databases. Data on the degree of full-length match and coverage by these EST sets were determined by using CROSS-MATCH (minmatch = 12 and minscore = 20).

Transcript Finishing. Primers for joining the ORESTES contigs were designed and used in reverse transcription (RT)-PCR reactions. Two micrograms of total RNA were reverse-transcribed by using SuperScript II and oligo(dT) in a final volume of 20 μl. RT-PCR was carried out in a 10-μl reaction mixture containing 1 μl of cDNA, 1× *Taq* DNA polymerase buffer, 200 μM dNTPs, 2 pmols of primers, and 1 unit *Taq* DNA polymerase (GIBCO/BRL). RT-PCR products were used directly in sequencing reactions with Big Dye terminator mix on an ABI377 sequencer (Perkin-Elmer) according to the manufacturer's instructions. Sequencing reactions were performed with the same primers used for RT-PCR.

Results and Discussion

ORESTES Dataset. The data set of 696,745 ORESTES sequences used in our analysis was generated in the course of the FAPESP/LICR-HCGP. A compilation of these sequences can be obtained from GenBank by using the keyword ORESTES. The sequences were produced from RNA extracted from only 24 different types of normal and malignant tissues (Table 3, which is published as supporting information on the PNAS web site, www.pnas.org) by using 3,540 minilibraries that produced an average of 197 sequences each. Of the ORESTES sequences generated, 7.8% were identified as mitochondrial transcripts or reverse transcribed copies of rRNA and 5.8% were identified as having a high similarity to known bacterial genes presumably derived from contaminants in the tissue samples. An additional 6.1% of the ORESTES sequences consisted of repetitive elements precluding their further analysis. From the remaining 559,675 sequences, around 62% showed high similarity at the nucleotide level to human transcripts for which putative full-length mRNA sequences are available or to EST sequences from other projects. The remaining 38% had no match against any publicly available human transcript sequences. Of these, 68% showed a high quality match to the draft human genome sequence. Those ORESTES that match other ESTs or have no match with other transcript sequences remain to be compiled into complete transcript sequences; as yet we have no precise way of knowing how many different genes they represent, what percentage of the derived transcripts they cover and hence what percentage of the overall transcriptome they represent. On the other hand, the exact coverage of ORESTES sequences that matches full-length mRNA sequences can be accurately determined. Thus, to judge the strengths and limitations of the ORESTES strategy we undertook a detailed analysis of those ORESTES sequences that correspond to genes for which full-length mRNA sequences are available.

ORESTES Strategy and Gene Discovery. We used a collection of 15,095 mRNA sequences identified from GenBank records that represent between 37.5% and 60% of the estimated 25,000 to 40,000 human genes (4, 5). The average length of these transcripts is 2,655 bp as compared with an estimated average length of 2,410 bp for all human transcripts (4). We grouped these into four approximate expression classes based on the number of EST sequences in the respective UniGene cluster (UniGene build no. 128) as shown in Table 1. This approach will be considered in detail elsewhere and has also been used by others as a ready means of estimating gene expression (16). Nevertheless to verify the validity of estimating relative gene expression on this basis, we compared the four abundance classes estimated from UniGene cluster size with the average number of serial analysis of gene expression (SAGE) tags in SAGEMAP for each gene. These data are also shown in Table 1, and clearly support the grouping of the genes into relative abundance classes with the provision that the lower two classes are not significantly different. In addition, we undertook a series of RT-PCR experiments where we compared the cluster size with the detection of the message in a breast cell line population. Again, a clear and positive association between the two sets of data is presented

Table 1. Full-length transcript database with expression level classification based on UniGene cluster size

UniGene cluster size*	Expression class	Full-length transcripts	UniGene clusters	% Cluster with full-length transcripts
$2 < X < 10$ (30)	Rare	1,985	38,769 (68.1%)	5.1%
$10 < X < 20$ (25)	Poor	3,265	4,871 (8.5%)	67.0%
$20 < X < 100$ (70)	Moderate	5,596	8,701 (15.3%)	64.3%
$X > 100$ (100)	High	4,249	4,572 (8.0%)	92.9%
Total		15,095	56,913	26.5%

*The average total number of SAGE tags for each or the UniGene cluster size categories is shown in parentheses in each case.

(Fig. 5, which is published as supporting information on the PNAS web site, www.pnas.org).

To estimate our rate of gene discovery, we calculated the percentage of full-length mRNAs, corresponding to the four abundance classes, represented by at least one ORESTES at different points during the course of the sequencing project (Fig. 1*A*). Fifty percent of the highly abundant full-length mRNAs were represented after only 30,000 ORESTES had been sequenced. The same representation was achieved with ≈51,000 ORESTES sequences for the moderately expressed full-length mRNAs and ≈300,000 and 450,000 sequences for the poorly and rarely expressed full-length mRNAs, respectively. After generation of nearly 700,000 sequences, more than 94% of the highly and moderately expressed mRNAs were represented and more than 50% of the weakly and rarely expressed transcripts were matched by at least one ORESTES. Based on the percentage of UniGene clusters in each size class and the proportion of these that have corresponding ORESTES sequences (Table 1), we estimate that, despite the relatively limited tissue coverage that we have so far achieved, the ORESTES dataset may represent 60% of all human genes.

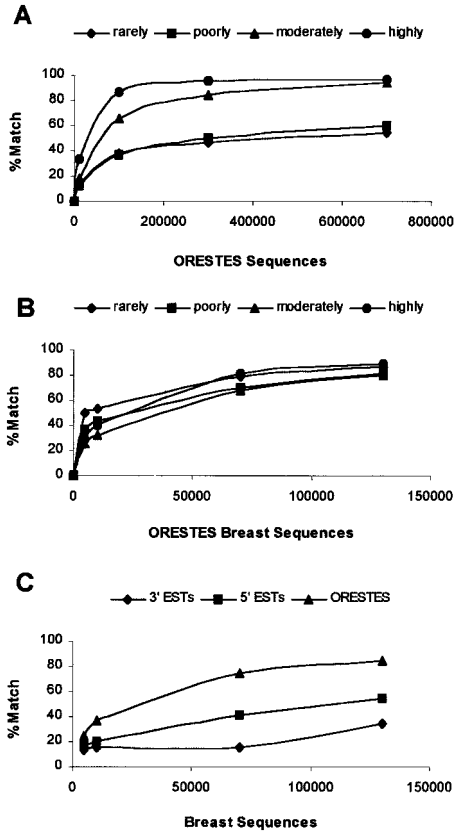

Fig. 1. Percentage of full-length transcripts with at least one sequence match (*A*) between the ORESTES sequences derived from 24 different tissues against 15,095 full-length mRNA sequences and (*B*) between the ORESTES sequences derived from breast tissue against full-length transcripts expressed in breast. (*C*) Comparison between the percentage of match for ORESTES sequences and 5′ and 3′ESTs derived from breast tissue.

The proportion of full-length mRNA sequences for which an ORESTES sequence was generated corresponded with the apparent abundance of the transcript as judged by UniGene cluster size. However, we had previously found that ORESTES at least partially compensates for transcript abundance (13). One possible explanation for this is that the genes in the low and rare classes are more likely to be tissue and stage specific than moderately and highly expressed genes. Because we sampled only a limited number of tissues, one would expect genes expressed in other tissues not to be represented in our dataset. To test this hypothesis, we repeated the analysis by using data from a single tissue, the breast.

By using tissue source information from ESTs in UniGene, we were able to identify 9,135 full-length mRNA sequences with evidence for expression in breast. We then compared these with the 133,345 ORESTES sequences derived from breast. Remarkably, we now found an approximately equal percentage of full-length mRNA sequences matched by ORESTES sequences in each abundance class (Fig. 1*B*). Thus, the inherent capability of ORESTES to generate disproportionate numbers of sequences from poorly expressed genes appears to be extremely powerful. Considering the difference in the number of sequences, the ORESTES matches against the most highly abundant breast mRNAs followed a similar curve as with the total ORESTES data set (Fig. 1*A*). This observation serves as an internal control for the comparison and supports our contention that the improved percentage of matches against the rare transcripts when only a single tissue is considered is indeed due to the common tissue source of the full-length sequences and ESTs.

We compared the percentage of full-length mRNAs matched by similar numbers of breast-derived ORESTES and breast-derived publicly available 5′ and 3′ EST sequences (Fig. 1*C*). For the purposes of this comparison, we grouped the full-length mRNAs into a single data set irrespective of their cluster size. We generated five random sets of breast-specific 5′ and 3′ ESTs for the proposed comparison. The results show that ORESTES is far more effective in generating partial sequences from transcripts. For example, with datasets of 70,000 sequences derived from ORESTES and 3′ and 5′ ESTs, 75, 34, and 41%, respectively, of the full-length mRNAs expressed in breast were sampled. Therefore, a comparison of the three distributions clearly demonstrates that ORESTES is a strategy for gene discovery significantly surpassing that of conventional ESTs. We have been unable to undertake similar direct comparison with randomly primed cDNA libraries because of lack of available data. How-

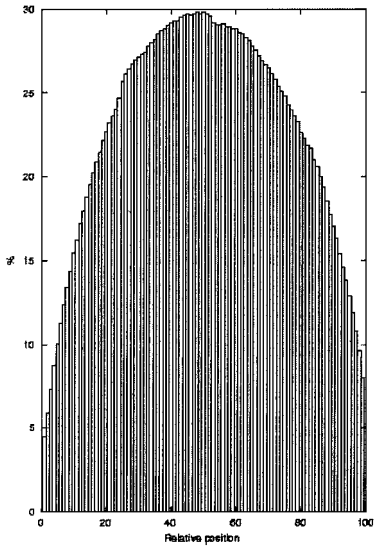

Fig. 2. Positional distribution of ORESTES sequences within full-length transcripts.

ever, a comparison with data from a single randomly primed library deposited in GenBank revealed that, although this approach also generates centrally biased transcripts, it does not have the normalizing effect that ORESTES exhibits (Fig. 7 and Table 4, which are published as supporting information on the PNAS web site, www.pnas.org).

ORESTES Coverage of the Transcriptome. The fundamental difference between ORESTES sequences and 5′ and 3′ ESTs is that the former are generated from the central region of transcripts. We found in our pilot project that ORESTES followed a mathematically predictable distribution around the midpoint of transcripts (13). We sought to verify whether the centralized position of ORESTES would be preserved within the highly dispersed and much larger dataset described here generated by many laboratories, by using minilibraries of varying quality and a large number of distinct mRNA preparations. The relative position of each ORESTES was calculated for known genes by scoring which of 100 equally spaced points along the length of full-length mRNAs were covered by ORESTES sequences. The analysis resulted in the reproduction of a remarkably symmetrical curve (Fig. 2). Furthermore, subdivision into primer category used for minilibrary preparation demonstrates that this distribution is absolutely inherent to the method and is not sequence-dependent (Fig. 6, which is published as supporting information on the PNAS web site, www.pnas.org).

The centralized distribution of ORESTES enables the partial sequences to be used not only for transcript identification but also as a means of progressing toward extended transcript compilation and eventual full-length transcript sequence determination. To investigate this progression, we calculated the percentage of nucleotides, constituting the full-length mRNA set that was matched by ORESTES, at various points along the course of sequence production (Fig. 3*A*). As expected, the percentage of all nucleotides covered is lower than the percentage of transcripts with at least one ORESTES, but 700,000 such sequences contain ≈40% of all nucleotides within full-length mRNAs with UniGene clusters of 20 or more and 20% of all nucleotides for the less abundant transcripts. Interestingly, the curves are continuing to rise for all expression level classes even at the end of sequence generation, showing that increased coverage is achieved as a function of the number of sequences generated. As expected, we found that, within an individual

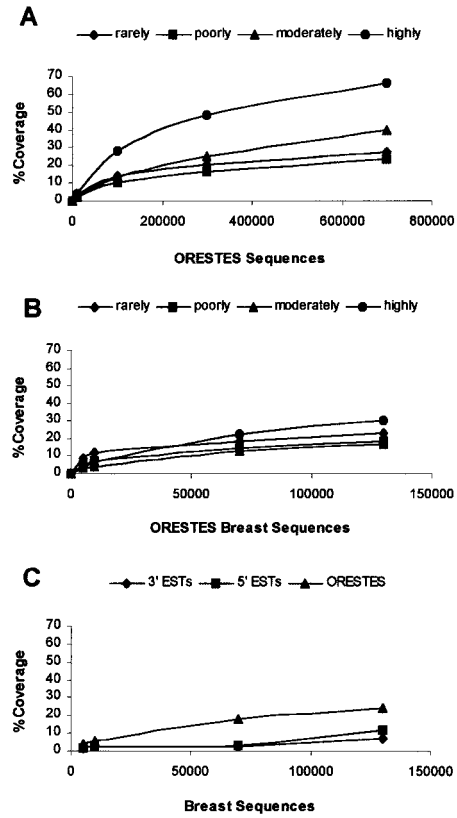

Fig. 3. Percentage of coverage of full-length transcripts by ORESTES sequences derived from 24 different tissues (*A*) and by ORESTES sequences derived from breast tissue (*B*). (*C*) Comparison between the percentage of coverage by ORESTES sequences and 5′ and 3′ ESTs derived from breast tissue.

tissue, nucleotide coverage was significantly more efficient than within the context of all human transcripts (Fig. 3*B*). In this case, after the generation of 100,000 ORESTES from breast tissue, between 10% and 20% of the total number of nucleotides in each expression class was covered. Furthermore, as would be predicted, ORESTES sequences were far more effective at generating sequence data than 5′ and 3′ ESTs (Fig. 3*C*). In this context, it should be remembered that, whereas 3′ ESTs are clustered, 5′ESTs are rather random because of premature termination of the reverse transcription reaction. Thus, we might expect the latter to give a better overall coverage; nevertheless, 5′ESTs still cover transcript sequences in a less efficient way than ORESTES.

Transcript Finishing Approach. Based on the extent of coverage observed and the characteristic distribution of the ORESTES sequences, it is possible to envisage an efficient approach to complete transcript sequence generation. The set of full-length mRNAs obtained from the public databases that we have used within the present study were all generated by first producing a full-length cDNA clone followed by its sequence determination. An alternative approach would be simply to continue to generate ORESTES, as well as 5′ and 3′ ESTs, and allow contigs to be generated that will eventually cover all transcripts. In this context, based on the trends of the curves shown in Fig. 3*B*, we extrapolate that around 800,000 ORESTES from a single tissue would allow the majority of the total length of all transcripts expressed therein to be determined. It remains to be determined which and how many different tissues must be sequenced to identify the complete set of human transcripts. However, after a partial coverage such as that achieved here with 700,000

Table 2. ORESTES contig coverage of full-length mRNAs

Sequences	Contigs	Gap	D1	D2	D3
10,000	1.4	491 bp	1,259 bp	476 bp	970 bp
70,000	2.0	554 bp	872 bp	1,131 bp	638 bp
300,000	2.5	457 bp	593 bp	1,542 bp	413 bp
700,000	2.9	353 bp	386 bp	1,814 bp	283 bp

ORESTES from all tissues, it is already possible to use these as a guide for directed gap closure, the equivalent of the finishing phase of genomic sequencing. This process is particularly viable today, within the context of the human genome, now that considerable genome sequence is available (4, 5). The finished genome data allow contigs formed from overlapping ORESTES from the same transcript to be identified and correctly ordered, thus permitting their subsequent linkage.

In order for contig joining to be an effective route to complete full-length transcript sequence generation, the contigs have to be sufficiently close to allow their joining by RT-PCR reactions and sufficiently well distributed so that their joining will allow coverage of the complete transcript. Within this context, we assessed the characteristics of contigs generated from ORESTES sequencing (Table 2). First, we calculated the number of contigs of ORESTES sequences for each full-length mRNA and the average distance in base pairs along the mRNA molecule between contigs. We found that the average number of contigs per full-length mRNA (for which at least one ORESTES sequence exists) increased from 1.4 at 10,000 ORESTES to 2.9 at 700,000 ORESTES, whereas the average size of the gaps between ORESTES decreased. The average gap size was 353 bp at 700,000 sequences, a distance compatible with the size of fragments that can be amplified by PCR and appropriate for a directed strategy of transcript finishing. We then asked what portion of the transcript would be covered if the contigs were all joined by RT-PCR. To do this, we calculated the average distance from the 5′ end of the full-length mRNA to the start of the first internal ORESTES contig (D1 in Table 2), the distance from the beginning of the first ORESTES contig to the end of the last ORESTES contig (D2 in Table 2), and the distance from the end of the last ORESTES contig to the 3′ end of the full-length mRNAs (D3 in Table 2). At 700,000 sequences, we have a span on average of 1.8 kb starting 386 bp from the 5′ end and finishing 283 bp from the 3′ end. The joining of these contigs would thus cover more than 80% of the transcript and most of the coding region. Indeed, a single 5′ and 3′ EST of 400 bp or a fragment generated by rapid amplification of cDNA ends (RACE) would then complete the remaining 20% of the sequence of the transcript.

We tested the transcript finishing approach to full-length cDNA sequence determination by "closing gaps" for four human transcripts partially represented by ORESTES sequences (Fig. 4). These human transcripts (accession nos. AF286904, AF286905, AF315356, and AF352051) correspond to orthologues of the mouse enhancer of polycomb 1 (EPC1) and 2 (EPC2), Notch 2, and proliferation potential-related protein genes and varied in size between 3.4 and 11.2 kb. The UniGene cluster sizes for these genes, according to build no. 128, are 33, 164, 164, and 91, respectively. ORESTES contigs corresponding to these genes were ordered by alignment to their corresponding orthologues and/or human genomic sequences, and 5′ and 3′ ESTs were used to define the probable extremities of the transcripts. As illustrated in Fig. 4, the extent of coverage of the full-length sequence by ORESTES varied between 47% and 71% and the number of ORESTES contigs varied from 2 to 10. The largest gap between two ORESTES contigs was 880 bp, a size that we were still able to close by RT-PCR. We thus expect that it would already be possible to complete the closure of most human transcript sequences by using this RT-PCR approach and the currently available data.

Concluding Remarks. After the generation of the human genome draft sequence, the priority is now to define all human genes and their corresponding transcripts. It is now clear that the genome sequence alone is not sufficient to allow this (4, 5). Strong evidence for genes can be gained by cross-species genome comparisons (17, 18); nevertheless, the most definitive approach to the elucidation of transcripts remains their direct sequencing. In this respect, we propose that significantly more human transcript sequencing should be undertaken because that completed to date has not been sufficient for the delineation of all human genes. Only complete sequences from full-length cDNA libraries can provide proof of definitive transcript structure because, for example, the identification of two sites of alternative splicing in the same gene with two separate ESTs does not reveal whether both occur in the same transcript variant or not. EST sequences are also of lower quality than complete double-stranded cDNA sequences, thus only allowing deduction of the

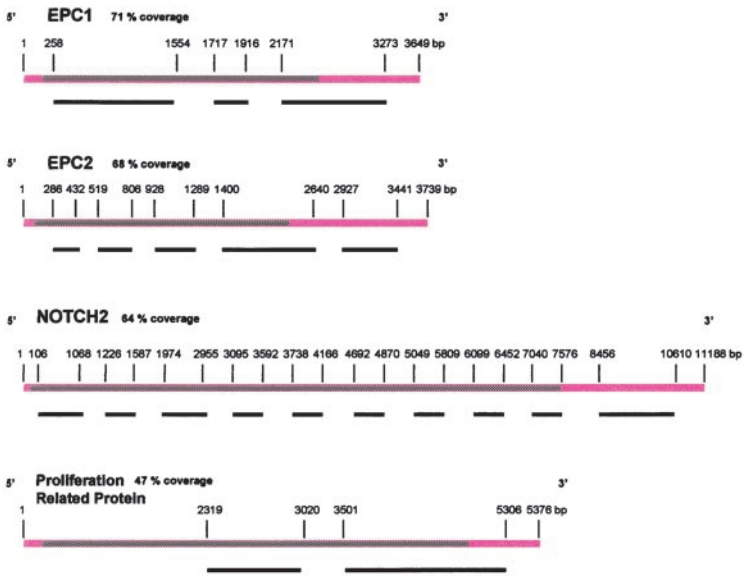

Fig. 4. Schematic representation of the transcript finishing approach. The sequence of four full-length transcripts corresponding to human orthologues of the mouse enhancer of polycomb 1 (EPC1) and 2 (EPC2), Notch 2, and proliferation potential-related protein were obtained by using ORESTES data in combination with genomic sequences available through the Human Genome Project (HGP). Coding regions for each of the four transcripts are represented as hatched bars and ORESTES contigs as solid bars below the genes.

correct sequence when multiple sequences are aligned. This latter process can be complicated because of the variability of transcripts from the same gene, the existence of closely related paralogues, and the impossibility of grouping nonoverlapping sequences from the same gene. To some extent, however, the increasing availability of finished genome sequence increases the value of ESTs, which can be aligned against finished genome sequence, largely overcoming clustering problems and where the genome serves as a source of the definitive gene sequence. We thus propose that a significant further investment in EST sequencing may also be warranted. To date, 700,000 ORESTES are available, together with almost 3 million ESTs generated by alternative strategies. The total number of ESTs represents around 3% of the sequences used to compile the draft genome. We suggest that 5–10 times more EST sequences from as wide a representation of tissues as possible would make a substantial contribution to the complete compilation of human transcripts. This number would still represent a rather small percentage of the overall genome project. Furthermore, the multiple transcript coverage that would be achieved would not only conclusively identify human genes but also provide an extensive insight into the repertoire of alternative transcripts generated from these genes. It is likely that such transcript variability will be key to understanding human biology (19, 20).

In comparison with other EST approaches, ORESTES has advantages and disadvantages, particularly related to its labor-intensive nature. The ORESTES technique is based on low-stringency RT-PCR (13, 14) and thus requires very high quality RNA, because any contaminating DNA is readily amplified. In addition, there are certainly PCR artifacts in the sequence. The latter do not preclude the eventual construction of high-quality transcript sequence from ORESTES data if use of genomic data is made. It does mean, however, that caution must be used in using ORESTES for the identification of single nucleotide polymorphisms. On the other hand, ORESTES exhibit a strong tendency to generate centrally biased regions of the transcripts favoring the incorporation of coding regions within the sequence. Furthermore, ORESTES are normalized for rare transcripts. In addition, the technique is capable of generating data from even very small amounts of starting material. Indeed, our data show that ORESTES exhibit a synergistic complementarity to other transcript sequencing strategies and are likely to continue to make a contribution to the detailed delineation of the complete repertoire of human transcripts, contained within the now sequenced human genome.

We thank Renato Alvarenga, Nelly R. C. Alves, Amélia G. Araújo, Daniela Dover Araújo, Gilson S. Baia, João P. D. Benedette, Simone A. de Bessa, Marcilei E. Buim, Valéria C. Cardoso, Helena P. Chiebao, Christian Colin, Cristiane A. Ferreira, Hellen T. Fuzii, Janaína R. Gusmão, Rafaela M. Maia, Adriano Malosso, Adriana A. Marques, Waleska K. Martins, Katlin B. Massirer, Fabiana Matioli, Adriana Matsukuma, Juliana F. Melo, Anna Izabel R. de Mello, Elisangela Monteiro, Ana P. O. Mora, Julio C. Moreira, Valentina F. M. Orelli, Audrey Y. Otsuka, Reimar Padovani, Silene K. Peres, Kamila C. Peronni, Márcia M. Picucci, Ana P. R. Reck, Tatiana I. Ricca, Anna Christina M. Salim, Míriam L. Sarmazo, Natalie Regina Leóz Schoken, Patrícia V. Serafin, Elisandra A. T. Silva, Teresa C. L. Silva, Marli H. Tavela, Olinda Mara Trevilato, Eliana Umeki, Rebecca Beolchi Vieira, Fabiola E. Villanova, Carla Vilella, Fernanda S. Zanola, Marcelo H. Zeviani, and Rui C. Serafim e Beatriz Schabel for dedicated technical assistance and Juçara Parra for acting as the administrative coordinator. The work was supported by the Ludwig Institute for Cancer Research (LICR) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). The ORESTES sequences were generated in the course of the Human Cancer Genome Project, which was undertaken by the Organization for Nucleotide Sequencing and Analysis (ONSA).

1. Dunham, I., Shimizu, N., Roe, B. A., Chissoe, S., Hunt, A. R., Collins, J. E., Bruskiewich, R., Beare, D. M., Clamp, M., Smink, L. J., *et al*. (1999) *Nature (London)* **402,** 489–495.
2. Guigo, R., Agarwal, P., Abril, J. F., Burset, M. & Fickett, J. W. (2000) *Genome Res.* **10,** 1631–1642.
3. Claverie, J. M. (1997) *Hum. Mol. Genet.* **6,** 1735–1744.
4. 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 (London)* **409,** 860–921.
5. 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.
6. Williamson, A. R. (1999) *Drug Discov. Today* **4,** 115–122.
7. Strausberg, R. L., Buetow, K. H., Emmert-Buck, M. R. & Klausner, R. D. (2000) *Trends Genet.* **16,** 103–106.
8. Quackenbush, J., Liang, F., Holt, I., Pertea, G. & Upton, J. (2000) *Nucleic Acids Res.* **28,** 141–145.
9. Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., *et al*. (1996) *Genome Res.* **6,** 807–828.
10. Strausberg, R. L., Feingold, E. A., Klausner, R. D. & Collins, F. S. (1999) *Science* **286,** 455–457.
11. Kikuno, R., Nagase, T., Suyama, M., Waki, M., Hirosawa, M. & Ohara, O. (2000) *Nucleic Acids Res.* **28,** 331–332.
12. Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., *et al*. (2001) *Nature (London)* **409,** 685–690.
13. Dias-Neto, E., Correa, R. G., Verjovski-Almeida, S., Briones, M. R., Nagai, M. A., da Silva Jr., W., Zago, M. A., Bordin, S., Costa, F. F., Goldman, G. H., *et al*. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 3491–3496.
14. de Souza, S. J., Camargo, A. A., Briones, M. R., Costa, F. F., Nagai, M. A., Verjovski-Almeida, S., Zago, M. A., Andrade, L. E., Carrer, H., El-Dorry, H. F., *et al*. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 12690–12693.
15. Bonalume-Neto, R. (1999) *Nature (London)* **398,** 450.
16. Bortoluzzi, S., d'Alessi, F., Romualdi, C. & Danieli, G. A. (2000) *Genome Res.* **10,** 344–349.
17. Batzogloum, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. (2000) *Genome Res.* **10,** 950–958.
18. Roest Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., *et al*. (2000) *Nat. Genet.* **235,** 235–238.
19. Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbruck, S., Lehmann, G., Luft, F., Reich, J. & Bork, P. (1999) *Trends Genet.* **15,** 389–390.
20. Black, D. L. (2000) *Cell* **103,** 367–370.