Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags

Sandro J. de Souza[a], Anamaria A. Camargo[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], Marcio Acencio[d], Fernando L. Alberto[c], Maria Elisabete J. Amaral[aa], Ivy Aneas[j], Mário Henrique Bengtson[e], Dirce M. Carraro[h], Alex F. Carvalho[a], Lúcia Helena Carvalho[e], Janete M. Cerutti[p], Maria Lucia C. Corrêa[k], Maria Cristina R. Costa[f], Cyntia Curcio[o], Tsieko Gushiken[t], Paulo L. Ho[r], Elza Kimura[z], Luciana C. C. Leite[r], Gustavo Maia[f], Paromita Majumder[j], Mozart Marins[l], Adriana Matsukuma[e], Analy S. A. Melo[b], 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 Gilbert Pessoa[v], Paula Rahal[aa], Claudia A. Rainho[x], Nancy da Ro's[d], Renata G. de Sá[w], Magaly M. Sales[t], Neusa P. da Silva[g], Tereza C. Silva[n], Wilson da Silva, Jr.[f], Daniel F. Simão[a], Josane F. Sousa[i], Daniella Stecconi[e], Fernando Tsukumo[u], Valéria Valente[i], Heloisa Zalcberg[a], Ricardo R. Brentani[a], Luis F. L. Reis[a], Emmanuel Dias-Neto[a], and Andrew J. G. Simpson[a,cc]

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

Communicated by George Klein, Karolinska Institute, Stockholm, Sweden, August 30, 2000 (received for review August 24, 2000)

Transcribed sequences in the human genome can be identified with confidence only by alignment with sequences derived from cDNAs synthesized from naturally occurring mRNAs. We constructed a set of 250,000 cDNAs that represent partial expressed gene sequences and that are biased toward the central coding regions of the resulting transcripts. They are termed ORF expressed sequence tags (ORESTES). The 250,000 ORESTES were assembled into 81,429 contigs. Of these, 1,181 (1.45%) were found to match sequences in chromosome 22 with at least one ORESTES contig for 162 (65.6%) of the 247 known genes, for 67 (44.6%) of the 150 related genes, and for 45 of the 148 (30.4%) EST-predicted genes on this chromosome. Using a set of stringent criteria to validate our sequences, we identified a further 219 previously unannotated transcribed sequences on chromosome 22. Of these, 171 were in fact also defined by EST or full length cDNA sequences available in GenBank but not utilized in the initial annotation of the first human chromosome sequence. Thus despite representing less than 15% of all expressed human sequences in the public databases at the time of the present analysis, ORESTES sequences defined 48 transcribed sequences on chromosome 22 not defined by other sequences. All of the transcribed sequences defined by ORESTES coincided with DNA regions predicted as encoding exons by GENSCAN. (http://genes.mit.edu/GENSCAN.html).

Complete bacterial genome sequences allow a relatively precise and complete analysis of constituent genes and coding regions by means of direct computational analysis (1). In complex eukaryotic genomes, however, it is proving considerably more difficult to identify genes because of their fragmentation into multiple small exons divided by often considerably larger introns. In this context, the determination of the complete sequence of the human chromosome 22 allowed a detailed appraisal of the efficacy of gene prediction methodologies (2). It was noted that when known genes (where complete cDNA sequences have been determined) were compared with an *ab initio* prediction of the same region by using the best computational methods available, only 94% of annotated genes were detected. More importantly, in only 20% of cases were all exons exactly predicted, and 16% of all known exons were entirely missed. On the other hand, almost 40% of GENSCAN-predicted genes did not form part of any gene confirmed by other means and include an unknown proportion of false positives (2).

In the absence of adequate computational approaches, gene identification will depend on the alignment of finished genomic sequence with sequences from experimentally validated transcripts. Following this approach, Dunham and colleagues (2) were able to identify 247 genes corresponding to fully sequenced transcripts on chromosome 22 that they have denominated

Abbreviations: EST, expressed sequence tag; ORESTES, ORF ESTs.

[cc]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.

known genes. In addition, they annotated 150 genes that were identified by their similarity to transcripts from other organisms or to other human genes that they have denominated related genes. Finally, a further 148 genes were annotated on the basis of alignment with expressed sequence tags (ESTs) that were denominated predicted genes and that we will refer to in this paper as EST-predicted genes, to avoid misleading interpretation with the outputs of gene prediction programs (2). The final number, 545, extrapolates to a total number of expressed human genes of 36,000 (excluding pseudogenes). This value is surprisingly low but has been supported by two further studies by using comparisons of genomic DNA sequences and ESTs (3) and comparison of conserved sequences between human and *Tetraodon nigroviridis* genomic DNA (4), which have estimated values of 35,000 and 28,000–34,000, respectively. In contrast, an independent assessment of the same ESTs and chromosome 22 data, by using different computational approaches and assumptions, led to an estimated of 120,000 genes in the human genome (5). The discrepancy of these estimates clearly demonstrates that the considerable cDNA and EST data already available in the public databases are insufficient to define human transcripts, on the basis of genome sequence, with any degree of certainty. Thus, more expressed gene sequence data are required to eventually complete the definition of all transcripts and in turn to permit the definition of all genes.

At present, the publicly available expressed gene sequence data consist of either complete cDNAs or ESTs derived from the 5′ or 3′ ends of cDNA clones. We have taken the view that the value of the EST data set can be significantly enhanced by the generation of what we have termed ORF ESTs (ORESTES), partial expressed gene sequences that are derived from the central portions of human transcripts (6). Because these sequences complement the predominantly 5′ and 3′ sequences generated in other EST sequencing projects, the net result is the shotgun-like compilation of complete human transcripts. To date, we have generated 250,000 ORESTES derived from a variety of human tumors. We have compared the chromosome 22 sequence with this database. This analysis led to the confirmation of a large percentage of the genes identified with existing ESTs and cDNAs and also to the identification of 219 unannotated transcribed sequences on this chromosome. Updated searches showed that 171 of these 219 sequences are also now represented by ESTs from other projects deposited in dbEST (http://www.ncbi.nlm.nih.gov/dbEST/).

Materials and Methods

Biological Samples, Template Preparation, and Sequencing. The biological samples selected for RNA extraction were derived from tumor and surrounding normal tissues excised from cancer patients after surgery at the Hospital do Câncer A. C. Camargo, São Paulo. All specimens were collected only after explicit informed consent was received. Tissue samples were frozen in liquid nitrogen immediately after resection and allowed to partially thaw to −20°C for microdissection to enrich for tumor cells in the sample. Breast cell lines were kindly provided by Michael O'Hare, University College, London. RNA extraction and template preparation were performed as described by Dias-Neto *et al.* (6), with some minor modifications as follows: 1 μl of neat cDNAs were PCR amplified with the same primers used for cDNA synthesis.

Amplification profiles were generated by using different cycling conditions. A touch-down PCR was introduced after cDNA denaturing at 75°C. The annealing temperatures varied from 60°C to 41°C (with progressive reductions of 1–2°C per cycle), and the number of cycles was increased to 45. Profiles composed of a DNA smear were size selected to separate amplification products of distinct size ranges, varying from 0.3 to 1.5 kb.

The size-selected fragments were cloned into pUC18 by using the Sureclone kit (Amersham Pharmacia). The resulting colonies were grown overnight in liquid media and used as templates for PCR by using vector primers. One microliter of the resulting PCR product was used for DNA sequencing by using standard protocols of the ThermoSequenase II dye terminator cycle sequencing kit (Amersham Pharmacia Biotech). Sequencing reactions were analyzed by using the capillary sequencer MegaBACE 1000 (Amersham Pharmacia Biotech). In general, 150–1,000 sequences were determined from each profile.

Computational Analysis. All ESTs were trimmed to exclude primer and vector sequences, as well as low-quality regions. Clustering of the 250,000 ORESTES was performed by using CAP3 (7). For the determination of sequences matching chromosome 22, the ORESTES contigs were first masked by using REPEATMASKER. The masked contigs were then compared by BLAST against the chromosome sequence published by Dunham and colleagues (2) and were kindly provided by the authors. To be considered a significant hit, an ORESTES contig was required to have at least 94% identity with chromosome 22 sequence over at least 80% of its length.

Results and Discussion

A set of 250,000 ORESTES was generated from mRNA primarily derived from human breast, colon, stomach, and head and neck tumors. A compilation of the sequences used can be obtained from GenBank by using the keyword ORESTES and from on-line supplementary material available at www.ludwig.org.br/chr22. A preliminary analysis of these sequences showed 18% to be derived from rRNA and mtDNA transcripts or to be almost entirely composed of repetitive sequences. These sequences were excluded from further analysis. The remaining sequences were processed by using the assembly tool CAP3, resulting in the construction of 81,429 contigs. Of these, 1,181 (1.45%) were found to match sequences in chromosome 22. The coordinates of the positions of chromosome 22 sequences corresponding to ORESTES contigs that either confirmed previously annotated genes or were judged to contain *bona fide* transcribed sequences are listed in the on-line supplementary information available at http://www.ludwig.org.br/chr22.

The chromosome 22 sequence available represents approximately 1.1% of the total human genome. The high percentage of ORESTES that match chromosome 22 sequence is consistent with the previously detected high gene density in this chromosome (8–10) and also with the fact that this chromosome contains a number of highly expressed genes. Indeed, 66.6% of the known genes on chromosome 22 can be ranked among the top 10% most highly expressed genes in the human genome on the basis of UNIGENE cluster size. We have previously shown that the ORESTES protocol partially compensates for unequal transcript abundance, reducing the proportion of sequences derived from highly expressed genes and increasing the proportion of rare transcripts. In this respect, analysis of the frequencies of the resequencing of the highly abundant transcripts on chromosome 22 showed that only 0.6% of the informative ORESTES sequences matched these transcripts as compared with 1.6% of sequences in UNIGENE.

We compared the alignments between ORESTES and chromosome 22 with the position of the genes annotated by Dunham *et al.* (2) and identified at least one ORESTES contig for 162 (65.6%) of the 247 known genes on chromosome 22. It is noteworthy that comparison with the one-third complete genome of *T. nigroviridis* detected a similar level (66.8%) of the same gene set (4). There was an average of 2.1 ORESTES contigs per gene that covered approximately 25% of the total known gene transcript sequence on the chromosome. Despite the generally elevated redundancy of EST sequencing of these

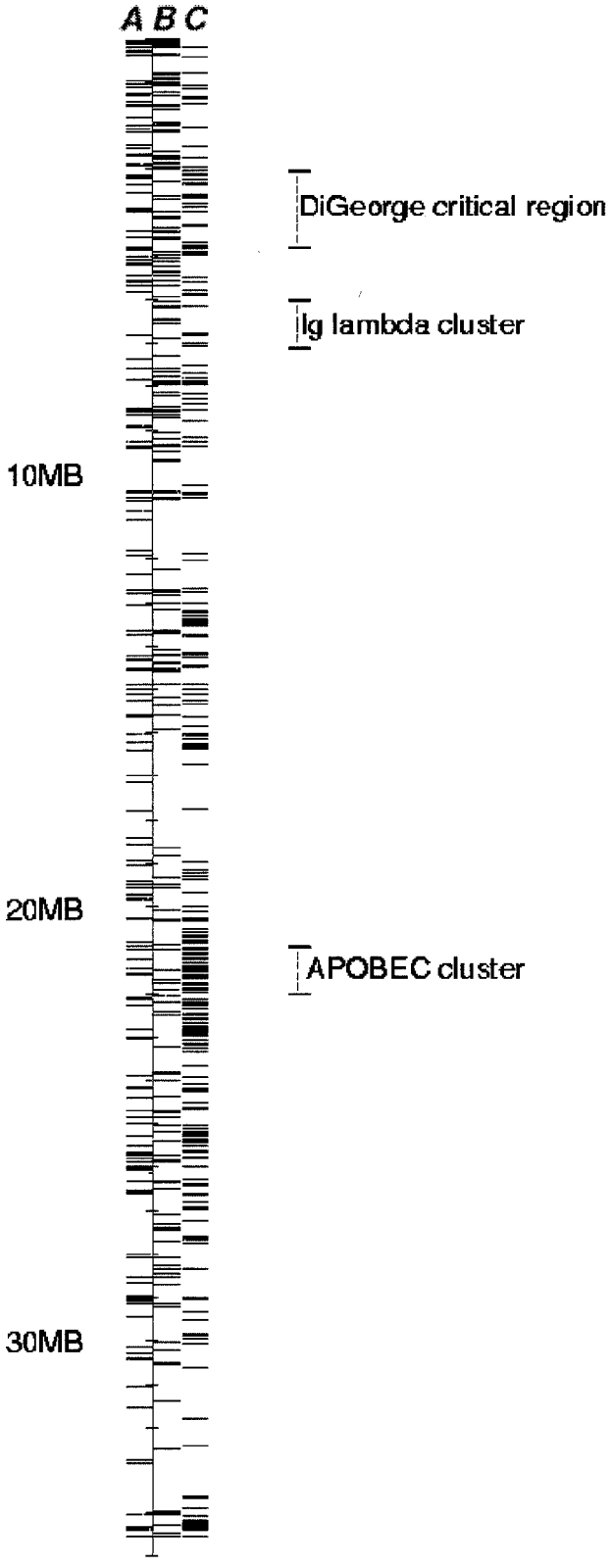

Fig. 1. Schematic diagram showing the relative position of known (*B*) and related and EST-predicted (*C*) genes as annotated by Dunham and colleagues (2). ORESTES-predicted transcribed regions identified in this report are shown in *A*. Marks in *A–C* represent only the initial position of genes and transcribed regions.

known genes on chromosome 22, because of their high expression levels, two new cases of alternative splicing were found within the ORESTES contigs. In one case (phosphatidylinositol 4-kinase), one exon is missing in the ORESTES derived sequence. In the second case (AK000625 unnamed protein), a cryptic donor site was identified by means of the ORESTES sequence that reduced the respective exon by 168 nucleotides. Thus even in cases where transcripts are highly characterized, further resequencing can still permit new insights into the complexity of the human transcriptome.

We identified at least one ORESTES contig for 67 (44.6%) of the 150 related genes identified on chromosome 22 for which confirmed expression in the humans was not provided in the original annotation of the chromosome. Of the total transcribed sequence defined by related genes, 15% was covered by ORESTES sequences. Comparison of the exons defined by ORESTES to those defined by orthologs revealed a single discrepancy that represents either an alternative splicing variant or a species-specific difference in gene structure.

Of the 148 EST-predicted genes on chromosome 22, 45 (30.4%) were confirmed by ORESTES sequences as judged by the overlap of the EST used in the original annotation and ORESTES sequences. The relatively low percentage of EST-predicted genes that match ORESTES sequences is to be expected, as the genes were predicted from ESTs derived from the extremities of transcripts, whereas ORESTES are distributed around the center. Thus, in partially sequenced transcripts there is a tendency for the sequences to be complementary rather than overlapping. Indeed, a proportion of the transcribed sequences identified are likely to represent central regions of the already predicted genes. When there is a superposition of the ORESTES sequences with conventional ESTs, this is often partial, allowing an extension of the sequence. Indeed, the average length of transcribed sequences derived from the originally EST-predicted genes is 1,022 bp but when supplemented by ORESTES, the average length of confirmed transcript increases to 1,153 bp, a 13% extension. In all, ORESTES contributed to a total of 17 kb of confirmed transcript sequence to the originally annotated EST-predicted genes on chromosome 22.

Overall, 50.5% of the annotated genes had significant similarity with ORESTES. In comparison, when all conventional ESTs were assembled into contigs by using the same CAP3 program used here, only 48.8% of genes were found to be represented. Thus, the apparently 10-fold smaller ORESTES database appears to be as informative as the entire dbEST collection of conventional ESTs. The high capacity of ORESTES to identify different genes can be attributed both to the range of tissues that we have exploited and to the tendency of ORESTES to be derived from relatively rare transcripts (6).

ORESTES sequences that match unannotated regions of chromosome 22 potentially represent regions of possibly unannotated human genes or unannotated regions of predicted genes. To consider a previously unannotated region of chromosome 22 that matched an ORESTES contig as being transcribed, we opted to impose a stringent set of criteria to reduce the likelihood of misassignment because of the presence of intronic or intergenic sequences in the ORESTES dataset. Thus, an ORESTES contig that matched an unannotated region was considered a *bona fide* transcribed sequence, (*i*) if it revealed the unambiguous presence of a splicing site on alignment to the genomic sequence, (*ii*) if it exhibited coding potential as measured by ESTSCAN (11), (*iii*) if two or more ORESTES from separate libraries formed the contig, or (*iv*) if the ORESTES sequence overlapped with another cDNA sequence present in

dbEST. (For a specific description of the criteria used for the validation of each ORESTES contig that corresponded to a previously unannotated region on chromosome 22, see Table 1 of the on-line supplementary information available at www.pnas.org and www.ludwig.org.br/chr22.) Using these four criteria, we were able to identify 219 ORESTES contigs that matched originally unannotated regions from chromosome 22. Of these, 171 match an EST sequence available at dbEST, reconfirming our prediction. Of the remainder, 38 have coding potential as measured by ESTSCAN, six have a splicing site on alignment to the genomic sequence, and four were composed of two or more ORESTES from separate libraries. All of these ORESTES-predicted transcribed regions corresponded to GENSCAN-predicted exon sequences. A total of 220 ORESTES contigs were thus excluded for which we have no supporting evidence, to date, that they represent transcribed sequences.

Because the ORESTES methodology is not indexed to a specific region of the transcript, unlike 3′ESTs, it is very likely that the number of genes we have identified on chromosome 22 is less than 219. To confirm and completely define the genes, it will be necessary to build complete transcript contigs from ORESTES or to generate full-length cDNAs corresponding to these genes. Nevertheless, homology-based searches against sequences in public databases provided hits for putative functions for some of the genes identified on chromosome 22 by ORESTES. Among the functional classes characterized here, there are three putative kinases (ORESTES contigs 31730, 20887, 32449), one putative dehydrogenase (ORESTES contig 31061), four putative transcription factors (ORESTES contigs 27759, 19960, 8918, 36344), two cell-surface receptors (ORESTES contigs 27497, 16842), two zinc-finger proteins (ORESTES contigs 27617, 2495), and two cytoskeletal proteins (ORESTES contigs 32132, 39028). The distribution of the identified transcribed regions is shown in Fig. 1 in comparison with the previously reported positions of known and predicted genes on chromosome 22. It is noteworthy that the distribution of the exons identified here generally reflects the distribution of known genes and also that even within highly annotated regions such as the DiGeorge critical region, transcribed regions were identified.

The range of estimates for the number of human genes that have recently been reported based on the extrapolation of data from existing ESTs and genomic sequences clearly demonstrates that the existing information is inadequate. The availability of the essentially complete human genome sequence now puts the onus on the further determination of expressed gene sequences to permit the accurate annotation of this valuable resource. The analysis reported here demonstrates that ORESTES are a highly informative source of expressed gene sequences and are capable of making an important contribution to the identification of all human genes.

Although we cannot extrapolate the number of ORESTES-predicted transcribed sequences to gene number, the data do clearly indicate that the estimate of a total of 36,000 human genes is likely to be a significant underestimate. Because of the ease of their generation, we now propose to extend the ORESTES database to include one million sequences derived from as wide a range of tissues as possible to complement the existing 5′ and 3′ ESTs and to extend coverage of human transcripts on this shotgun basis. This should significantly reduce the estimated 60% shortfall in coverage of human coding regions in those ESTs and cDNAs in Unigene of December, 1999 (4). Further detailed comparison between the contigs constructed from the transcribed sequences and genomic sequences should then allow the identification of the majority of human genes and their variants.

We thank Elisangela Monteiro, Anna Christina M. Salim, Anna Izabel R. de Mello, Rui C. Serafim, João P. T. Benedetti, Helena P. Chiebao, Katucha W. Luchesi, Marcia M. Piucci, Janaina R. Gusmao, Miriam L. Sarmazo, Beatriz Schnabel, Fabiola Villanova, Patricia V. Serafin, Silene K. Peres, Cássia C. Villela, Hellen T. Fuzii, Andréia Andrade, Adriana C. Carloni, Waleska K. Martins, Magnus R. D. Silva, Liliane Arnaldi for dedicated and expert technical assistance and Juçara Parra for acting as the administrative coordinator of this project. The work was equally supported by the Ludwig Institute for Cancer Research and the Fundação de Amparo à Pesquisa do Estado de São Paulo and is known as the FAPESP/LICR Human Cancer Genome Project. It is being undertaken within the auspices of ONSA, the Organization for Nucleotide Sequencing and Analysis.

1. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. (2000) *Nucleic Acids Res.* **28,** 33–36.
2. 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.
3. Ewing, B. & Green, P. (2000) *Nat. Genet.* **25,** 232–234.
4. 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.
5. Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L. & Quackenbush, J. (2000) *Nat. Genet.* **25,** 239–240.
6. Dias-Neto, E., Garcia, C. R., Verjovski-Almeida, S., Briones, M. R., Nagai, M. A., da, S. W. J, Zago, M. A., Bordin, S., Costa, F. F., Goldman, G. H., Carvalho, A. F., *et al.* (2000) *Proc. Natl. Acad. Sci. USA* **97,** 3491–3496.
7. Huang, X. & Madan, A. (1999) *Genome Res.* **9,** 868–877.
8. Craig, J. M. & Bickmore, W. A. (1994) *Nat. Genet.* **7,** 376–382.
9. Deloukas, P., Schuler, G. D., Gyapay, G., Beasley, E. M., Soderlund, C., Rodriguez-Tome, P., Hui, L., Matise, T. C., McKusick, K. B., Beckmann, J. S., *et al.* (1998) *Science* **282,** 744–746.
10. Saccone, S., Caccio, S., Kusuda, J., Andreozzi, L. & Bernardi, G. (1996) *Gene* **174,** 85–94.
11. Iseli, C., Jongeneel, C. V. & Bucher, P. (1999) *ISMB.* **1999,** 138–148.