Chromosomal localization and expressed sequence tag generation of clones from a normalized human adult thymus cDNA library.

J E Lamerdin, R S Athwal, M S Kansara, et al.

Genome Res. 1995 5: 359-367
Access the most recent version at doi:10.1101/gr.5.4.359

RESEARCH

Chromosomal Localization and Expressed Sequence Tag Generation of Clones from a Normalized Human Adult Thymus cDNA Library

Jane E. Lamerdin,[1,4] Raghbir S. Athwal,[2] Mala S. Kansara,[2] Arbansjit K. Sandhu,[2] Sankhavaram R. Patanjali,[3] Sherman M. Weissman,[3] and Anthony V. Carrano[1]

[1]Human Genome Center, Biology and Biotechnology Research Program, L-452, Lawrence Livermore National Laboratory, Livermore, California 94550; [2]Fels Institute for Cancer Research & Molecular Biology, Temple University School of Medicine, Medical Research Building, Philadelphia, Pennsylvania 19140; [3]Boyer Center for Molecular Medicine, Yale University School of Medicine, New Haven, Connecticut 06536-0812

Expressed sequence tags (ESTs) from 298 clones have been generated from a randomly primed, normalized human adult thymus cDNA library. We describe the chromosomal localization of 136 of these ESTs by PCR-based mapping to a human monochromosomal somatic cell hybrid panel. Data base similarities to known genes are also described. A subset ($n = 18$) of these randomly primed ESTs extended the sequence of ESTs from other tissues currently in dbEST. Of the nonrepetitive human adult thymus ESTs generated in this study, 237 (79.5%) have no similarity to current data base entries. This would suggest that our collection contains ~100 new coding regions from thymus tissue, a large proportion of which likely will represent the middle regions of genes. The mapped ESTs should prove useful as new gene-based markers for mapping and candidate gene hunting, particularly when anchored to a well-developed physical map of the human genome.

The utility and popularity of expressed sequence tags (ESTs) in the development of an expression map of the human genome has been well established (Sikela and Auffray 1993). International efforts such as the I.M.A.G.E. Consortium (Lennon et al. 1995), and more recently, the Washington University–Merck EST Project (Boguski et al. 1995), have been organized to sequence selected cDNA libraries and to share sequence and map information from these libraries. Other efforts, both in the commercial and public sectors, have long been under way to sequence cDNAs from various tissue-specific libraries (Adams et al. 1991, 1993a,b; Khan et al. 1992; Okubo et al. 1992). However, efforts to map the huge numbers of ESTs generated by such projects generally have not kept pace because of the labor-intensive nature of mapping (Durkin et al. 1992; Geiser and Swaroop 1992; Polymeropolous et al. 1992, 1993), although PCR-based mapping strategies readily are amenable to automation. Recently, a consortium of laboratories including the Whitehead Institute at the Massachusetts Institute of Technology Center for Genome Research, Stanford, Oxford, and Cambridge universities, the Sanger Centre, Cambridgeshire, UK, and Genethon (Evry CEDEX, France) has begun mapping unique public domain ESTs to radiation hybrid panels and Centre d'Etude du Polymorphisme Humain (CEPH) mega-yeast artificial chromosome (YAC) clones (Boguski and Schuler 1995) using PCR-based methods. With the promise of large-scale genomic sequencing projects on the horizon, for use in the quick generation of low redundancy sequences of well-mapped regions of the human genome, many partially sequenced cDNAs may be mapped by virtue of sequence identity. In the interim, map location information for cDNAs from new tissue-specific libraries would aid researchers in the identification of candidate genes for specific heritable diseases. In ad-

[4]**Corresponding author.**
E-MAIL lamerdin@llnl.gov; FAX (415) 422-2282.

5:359–367 ©1995 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/95 $5.00

Table 1. Chromosomal localization of ESTs with data base similarities

Clone[a]	PCR[b]	GSDB no.[c]	CL[d]	D-segment[e]	Putative identification	DB: accession no.[f]	Percent identity/ length (nucleotide or aa)[g]
3C5	B	L44288	1	D1S2233E		db: H02524	99% over 182
3F5	A	L44411	1	D1S2237E	human transmembrane glycoprotein CD53	gb: L11670	97% over 299
4B7*	B	L44292	1	D1S2238E	human ATP-dependent RNA helicase A	gb: L13848	98% over 404
4C5	B	L44294	1	D1S2240E	human CD46 exon 7	gb: A18595	100% over 45
4E1	A	L44296	1	D1S2242E		db: T23836	98% over 266
1B6	A	L44303	2	D2S1872E	human integrin α4 subunit	gb: X16983	99.6% over 279
4E12*	B	L44311	2	D2S1880E	mouse DIFF6 or CDC3-like protein	gb: D26128	91% over 89
4F10*	B	L44312	2	D2S1881E		db: T11610	100% over 91
4G1	B	L44313	2	D2S1882E	human leucine-rich protein	gb: M92439	88% over 197
5A5	A	L44315	2	D2S1884E	human germ line T-cell receptor β gene region	gb: L36092	99% over 236
1D10	A	L44319	3	D3S3241E		db: F12581	96.4% over 197
3B5	B	L44325	4	D4S2700E	human annexin V (ANX5) gene	gb: UO5770	98% over 293
5A1*	B	L44327	4	D4S2702E	human chromosome 4 STS4-125	gb: M95213	93% over 290
6E3	A	L44337	5	D5S1786E		db: D19622	99% over 204
4H6*	B	L44346	6	D6S1356E	human nuclear pore complex protein hnup153	gb: Z25535	99% over 302
5F3	A	L44354	7	D7S2262E		db: T32402	99.4% over 167
6E1*	A	L44369	11	D11S3937E	human AMP deaminase (AMPD3)	gb: M84722	98% over 130
1D4	A	L44373	13	D13S1104E		db: Z43749	99.4% over 311
5G2*	B	L44374	13	D13S1105E		db: R38203	99% over 243
5A2	A	L44380	14	D14S793E		db: T35430	98.7% over 300
5A8	A	L44381	14	D14S794E	human protein phosphatase 2c, α isoform	sp: P35813	96% over 30 aa
4E2	A	L44387	15	D15S857E		db: Z39194	94% over 116
5D11*	A	L44395	17	D17S1600E	human clathrin heavy chain	gb: D21260	98.6% over 361
6H4	A	L44396	17	D17S1601E		db: T06472	100% over 63
1G11	A	L44408	X	DXS7425E		db: F13179	76% over 168
3G4*	B	L44501	2, X			db: T78352	89% over 139
5H1	A	L44547	13, 21			db: T95189	94% over 346
6F4*	A	L44571	13, 21			db: T23158	97% over 236
6H6	A	L44576	2, 19			db: T52615	77% over 189
1C2	A	L44422	HR			db: Z21094	88% over 220
1G2*	A	L44439	HR			db: R77232	97% over 294

Table 1. (*Continued*)

Clone[a]	PCR[b]	GSDB no.[c]	CL[d]	D-segment[e]	Putative identification	DB: accession no.[f]	Percent identity/ length (nucleotide or aa)[g]
2E7*	A	L44462	HR		human Ran binding protein	gb: U19248	95% over 344
3F2	A	L44495	HR		rat eukaryotic initiation factor 5 (eIF-5)	gb: L11651	87.7% over 285
4C8*	B	L44517	HR		human ATP synthase A chain (pro.6); mitochondrial	gb: X78136	97.5% over 202
4G8	B	L44525	HR		human MacMarcks mRNA	gb: X70326	98.2% over 391
4H9	A	L44527	HR			db: T39728	98.9% over 276
5C7*	A	L44536	HR		human liver glycogen phosphorylase type IV	gb: M36807	97.5% over 324
5D2	A	L44538	HR		mouse Hox 2.7 homeo box protein	gb: X62311	90.4% over 167
6C3*	A	L44559	HR		human ubiquitin-conjugating enzyme	gb: X78140	79.5% over 195
6D7*	A	L44565	HR		human CCAAT-box-binding factor	gb: M37197	98% over 349

[a] Asterisks (*) indicate ESTs exhibiting coding potential as predicted by GRAIL 1 or 2.
[b] PCR conditions are as described in Methods.
[c] Accession number in Genome Sequence Data base.
[d] (CL) Chromosomal location; (HR) chromosomal assignment not determined because of same size amplicon in human and rodent DNAs.
[e] GDB accession number.
[f] (DB) Data base; (gb) GenBank; (db) dbEST; (sp) SWISSPROT.
[g] (aa) Amino acids. Nucleotide or residue identity over the length indicated.

dition, PCR-ready markers will be useful for the integration of these clones into well-developed physical maps of the human genome. We report here the development of 298 ESTs and the chromosomal localization of 136 of these ESTs from a randomly primed, normalized human adult thymus library.

RESULTS AND DISCUSSION

EST Generation

Single-pass sequencing (1 × redundancy) was performed on a total of 500 cDNA clones from the NSTHII library; 333 clones (67%) produced usable sequence. Nearly 20% of the failures were caused by mobility problems associated with the custom λgt10c dye primers. With the recent release of a wider variety of "base callers" to analyze anomalously migrating sequencing ladders (ABI Prism Sequence Analysis software, V2.11), such problems associated with custom dye primers should be diminished. The remainder of the poor data corresponded to noisy sequence, presumably caused by inadequate quantitation of the templates or artifacts generated by PCR amplification. The average insert size of the library was determined to be 570 bp. Clones <400 bp were not sequenced.

Of the clones with usable sequence data, 72 (21.6%) were partially or completely comprised of one or more repetitive element sequences: 38 clones contained L1, 30 contained Alu, and 8

Table 2. Chromosomal localization of ESTs with no data base similarities

Clone[a]	PCR[b]	GSDB no.[c]	CL[d]	D-segment[e]	Clone[a]	PCR[b]	GSDB no.[c]	CL[d]	D-segment[e]
1A3*	A	L44279	1	D1S2232E	4A11	B	L44351	7	D7S2259E
1B10	A	L44280	1	D1S2244E	5D3*	A	L44352	7	D7S2260E
1B9	A	L44281	1	D1S2248E	5D9	A	L44353	7	D7S2261E
1H1	A	L44282	1	D1S2249E	1D2*	A	L44355	8	D8S1535E
2D1*	B	L44283	1	D1S2250E	1G4	A	L44356	8	D8S1536E
2F11	A	L44284	1	D1S2251E	2A8	A	L44357	8	D8S1537E
2F12	A	L44285	1	D1S2252E	6A7	A	L44358	8	D8S1538E
2G3	A	L44286	1	D1S2253E	1A12	A	L44359	9	D9S1628E
3B4	A	L44287	1	D1S2254E	4F9*	B	L44360	9	D9S1629E
3E3	A	L44289	1	D1S2234E	6H7	A	L44361	9	D9S1630E
4A4	B	L44290	1	D1S2235E	1B1	B	L44362	10	D10S1519E
4B6	B	L44291	1	D1S2236E	1F5	B	L44363	10	D10S1520E
4C2	A	L44293	1	D1S2239E	1G5	A	L44364	10	D10S1521E
4C9	B	L44295	1	D1S2241E	4C1	A	L44365	10	D10S1522E
5G11	A	L44297	1	D1S2243E	4E9*	B	L44366	10	D10S1523E
6A5	A	L44298	1	D1S2245E	5D5	A	L44367	10	D10S1524E
6B2	A	L44299	1	D1S2246E	1A6	A	L44368	11	D11S3936E
6B6	A	L44300	1	D1S2247E	1E2	B	L44370	12	D12S1355E
1A4*	C	L44301	2	D2S1870E	1H9*	B	L44371	12	D12S1356E
1B4	A	L44302	2	D2S1871E	4G12	B	L44372	12	D12S1357E
1D1	A	L44304	2	D2S1873E	2D7	A	L44375	14	D14S796E
1G12	A	L44305	2	D2S1874E	4A8	B	L44376	14	D14S797E
2C10	A	L44306	2	D2S1875E	4B11	B	L44377	14	D14S798E
3H4	B	L44307	2	D2S1876E	4G7*	A	L44378	14	D14S791E
4A10	B	L44308	2	D2S1877E	4H1	B	L44379	14	D14S792E
4B1	B	L44309	2	D2S1878E	5F1	A	L44382	14	D14S795E
4D3	B	L44310	2	D2S1879E	1H10	B	L44383	15	D15S853E
4G11*	B	L44314	2	D2S1883E	2C4	A	L44384	15	D15S854E
5G10	A	L44316	2	D2S1885E	2G6	A	L44385	15	D15S855E
5G3	A	L44317	2	D2S1886E	4A3	A	L44386	15	D15S856E
6F3	A	L44318	2	D2S1887E	4F2	B	L44388	15	D15S858E
1D5	A	L44320	3	D3S3242E	5E1	A	L44389	15	D15S859E
2F8*	A	L44321	3	D3S3243E	6F5	A	L44390	15	D15S860E
4H5	B	L44322	3	D3S3244E	2B4	A	L44391	16	D16S2915E
5E6	A	L44323	3	D3S3245E	4C12*	C	L44392	16	D16S2916E
6G6*	A	L44324	3	D3S3246E	1C3	B	L44393	17	D17S1598E
4F12*	B	L44326	4	D4S2701E	5C2	A	L44394	17	D17S1599E
2C12	A	L44328	5	D5S1777E	1E1	A	L44397	19	D19S800E
3F6	A	L44329	5	D5S1778E	1E3	A	L44398	19	D19S801E
4G9	B	L44330	5	D5S1779E	3F4	A	L27444	19	D19S468E
5C3	A	L44331	5	D5S1780E	5F7	A	L44399	19	D19S802E
5D12*	A	L44332	5	D5S1781E	1C7*	A	L44400	20	D20S687E
5F5	B	L44334	5	D5S1783E	1H3	A	L44401	20	D20S688E
5F11	A	L44333	5	D5S1782E	2C8*	A	L44402	20	D20S689E
5G1	A	L44335	5	D5S1784E	4E7	A	L44403	20	D20S690E
6B5	A	L44336	5	D5S1785E	4E8	B	L44404	20	D20S691E
6G7	A	L44338	5	D5S1787E	5F6	A	L44405	20	D20S692E
1F12	A	L44339	6	D6S1349E	1C11	A	L44406	22	D22S1077E
1H2	A	L44340	6	D6S1350E	4D7*	B	L44407	22	D22S1078E
2B11	A	L44341	6	D6S1351E	4G2	B	L44409	X	DXS7426E
2C6	A	L44342	6	D6S1352E	4F11	B	L44410	Y	DYS396E
2H10	A	L44343	6	D6S1353E	3D10	A	L44491	5 or 8	
3D4	A	L44344	6	D6S1354E	5F8*	A	L44544	6,7	

Table 2. (*Continued*)

Clone[a]	PCR[b]	GSDB no.[c]	CL[d]	D-segment[e]	Clone[a]	PCR[b]	GSDB no.[c]	CL[d]	D-segment[e]
4A7	B	L44345	6	D6S1355E	2C1	A	L44452	HR	
6A4	A	L44347	6	D6S1357E	2F2	A	L44465	HR	
1A8	A	L44348	7	D7S2257E	3D2	A	L44492	HR	
3E4	A	L44349	7	D7S2278E	4A2*	B	L44509	HR	
4A1	A	L44350	7	D7S2258E	4E10	B	L44521	HR	

[a] Asterisks (*) indicate ESTs exhibiting coding potential as predicted by GRAIL 1 or 2.
[b] PCR conditions are as described in Methods.
[c] Accession number in Genome Sequence Data base.
[d] (CL) Chromosomal location; (HR) chromosomal assignment not determined because of same size amplicon in human and rodent DNAs.
[e] GDB accession number.

contained THE, MER, long terminal repeat (LTR)-like, or other repeats. Ten of the clones containing L1 also encoded some unique sequence, and 25 of the 30 clones containing Alu elements also included presumptive coding regions. Of those clones containing no repetitive elements, only two were represented twice. Four clones were identified as ribosomal or mitochondrial in origin (two of each category). One EST is a near perfect match to the *Escherichia coli* pantothenate kinase gene and, as such, is probably a result of *E. coli* contamination. One other clone appeared to be derived from the vector. No other perfect matches to nonhuman sequences were observed. For the most part, these results concur with the quantitative hybridization studies performed earlier on this cDNA library (Patanjali et al. 1991), in which ~36% of the clones represented repetitive DNA, and the average abundance of individual cDNAs was estimated to be <1%.

Chromosomal Localization

Primer pairs ($n = 274$) were designed and tested initially against the originating cDNA, human genomic DNA, two monochromosomal hybrid cell lines containing either human chromosome 19 or 2, and hamster genomic DNA. Of the 274 primer pairs, 199 (72.6%) were successful, in that they amplified a single fragment of the correct size in the cDNA and human genomic DNA (and occasionally one of the monochromosomal hybrid cell lines). Another 39 primer pairs (14%) amplified the correct size fragment in the cDNA only and thus presumably spanned an intron. Some of these ($n = 10$) produced a detectable size fragment in human genomic DNA on agarose gels and so still may be useful for mapping. This success rate of 72.6% is lower than that observed for mapping of ESTs derived from the 3′-untranslated repeats (UTRs) of directionally cloned cDNAs (~90%; Berry et al. 1995). This lower success rate presumably is caused by the use of randomly primed cDNAs, which are more likely to originate from portions of the coding regions containing splice junctions.

Then, 168 primer pairs were selected for chromosomal localization by mapping to the somatic cell hybrid panels. The results of this mapping, along with D-segment designations, are indicated in Table 1, with accompanying data base similarity information. Table 2 lists putative new genes. A total of 136 ESTs (81%) were assigned to a single human chromosome and 6 (3.6%) mapped to >1 human chromosome. As seen in Tables 1 and 2, three of these six ESTs have good coding potential, as predicted by GRAIL. One of these ESTs (3D10) has similarity to a region just distal to the *macaca mulatta* haptoglobin-related gene, which contains an LTRV repeat. Potentially, the remaining five ESTs represent members of gene families, pseudogenes, or low-copy-number repeat elements. Sixteen primer pairs (9.5%) amplified the same size fragment in the rodent background as in human genomic DNA (designated HR in Tables 1 and 2), and as such could not be localized. As indicated in Table 1, some of these may represent amplification of human and rodent homologs of discrete genes, such as that encoding a GTP-binding protein, Ran1,

Table 3. Data base similarities for unmapped adult thymus ESTs

Clone[a]	GSDB no.[b]	Putative identification	DB: accession no.[c]	Percent identity/length (nucleotide or aa)[d]
1A10	L44413	human haptoglobin-related protein	gb: M69197	73% over 213
1A7	L44415	human Krüppel-related zinc finger protein (H-plk)	gb: M55422	89% over 187
1A9	L44416		db: T67003	98% over 283
1C1*	L44420		db: F07556	98% over 99
1C10	L44421	human MHC class II HLA–DRw–β (DR4,w4) gene	gb: M20555	91% over 329
1C6	L44424		db: D25606	94% over 269
1D12	L44427		db: Z42694	97% over 154
1E11	L44428		db: T50014	95% over 237
1E4*	L44430	human ribosomal protein RS23	gb: D14530	98% over 252
1F10	L44433	human thymocyte antigen CD1a	gb: M28825	92% over 344
1G1	L44437		db: T34543	98.9% over 275
1H11	L44440	*Saccharomyces cerevisiae* calcium transporting ATPase, ATC1	sp: P13586	70% over 44 aa
2D3*	L44456		db: T41172	80.7% over 295
2F1*	L44464	*Drosophila melanogaster* dosage compensation regulator (MLE)	sp: P24785	44.5% over 128 aa
2G12	L44468	human Tcr-C–Δ gene, exon 4	gb: M94081	98% over 277
3C1*	L44485	human arylhydrocarbon receptor	sp: P35869	94% over 41 aa
3E11	L44493	human thromboxane synthase TBXAS1	gb: D34613	85% over 159
4B10	L44510	human MHC class II HLA–DRw–β (DR4,w4) gene	gb: M20555	91% over 333
4D8	L44520		db: T81092	99% over 215
5A4	L44529	human ribosomal protein L1	gb: D23660	100% over 40
5C7	L44536	human glycogen phosphorylase PHS1	sp: P06737	94% over 71 aa
5E11*	L44541	*S. cerevisiae* proteasome component pup1 (PRCU)	sp: P25043	64% over 126 aa
5G4*	L44546		db: T10722	91% over 68
5H8*	L44551		db: Z28756	98.6% over 217
5H12*	L44548		db: T80259	93.6% over 361
6A2*	L44553	human MAR/SAR DNA-binding protein (SAT1)	gb: M97287	98.9% over 362
6B1	L44556		db: T60839	99% over 287
6B4*	L44557	human replication protein A 32-kD subunit (RFA2)	gb: J05249	99% over 332
6D5*	L44563	human membrane cofactor gene (MCP)	gb: Y00651	94% over 56
6E6*	L44569	human *dek* proto-oncogene	gb: X64229	98% over 284

[a] Asterisks (*) indicate ESTs exhibiting coding potential as predicted by GRAIL 1 or 2.
[b] Accession number in Genome Sequence Data base.
[c] (DB) Data base; (gb) GenBank; (db) dbEST; (sp) SWISSPROT.
[d] (aa) Amino acids. Nucleotide or residue identity over the length indicated.

the CCAAT-box binding factor and liver glycogen phosphorylase type IV loci, or of multigene family members, such as the ubiquitin-conjugating enzyme, or pseudogenes. The chromosomal location of some of the genes identified by data base similarity is known (e.g., annexin V, which maps to 4q28–q31), and in all cases confirmed the assigned location of the EST. Nine clones whose chromosomal location could not be determined because of amplification of all human chromosomes, or amplification of human and rodent background, contained repetitive elements not picked up by the initial repeat screen (data not shown).

Identification of ESTs

The results of data base comparisons of ESTs from adult human thymus are indicated in Tables 1 (for mapped ESTs) and 3 (for unmapped ESTs). In cases where multiple hits were observed, only the top scoring alignment is recorded in the Tables. Human adult thymus ESTs that matched ESTs in dbEST only with unknown function are also included, but the putative identification is left blank. Notable hits are those loci involved in cell-mediated immune response, such as the wide variety of cell surface antigens, including T-cell and major histocompatibility complex (MHC) antigens, and binding molecules in the complement system. The SAT1 (MAR/SAR) DNA-binding protein, which is expressed predominantly in thymus, is present, as well as various other proteins involved in DNA replication, transcription, and translation. Interestingly, two ESTs were near exact matches to different regions of a human ATP-dependent RNA helicase A, a homolog of the maleless (MLE) protein in *Drosophila melanogaster*, which functions as a regulator of dosage compensation. One additional human thymus EST (2F1, Table 3) was more similar to the MLE protein than to the human homolog. Several ESTs similar to the human blood components haptoglobin-related protein, annexin V (an anticoagulant protein), and platelet thromboxane synthase were also identified. Also of interest are similarities to genes potentially involved in developmental regulation, such as the mouse homeo box *Hox2.7* and *diff6* loci, and another EST with near perfect sequence identity to the human *dek* proto-oncogene.

Eighteen of the ESTs generated in this study extended ESTs deposited in dbEST either 5′- or 3′-ward by as much as 350 bp. Most of the clones extended were derived from directionally cloned or oligo(dT)-primed libraries. Furthermore, 7 of the 44 human adult thymus ESTs that had matches to known genes in SWISSPROT or GSDB did not have a significant match to an EST in dbEST. This would indicate that the thymus ESTs delimit regions of these genes not detected by single-pass sequencing from clones in the directionally cloned or oligo(dT)-primed libraries, or isoforms of genes not predominantly expressed in the tissues of choice for current large-scale EST projects (e.g., brain). The former is certainly true, as some of the genes detected in human adult thymus, such as the MacMarcks gene involved in signal transduction and the MAR/SAR DNA-binding protein, have been seen in other tissues (Adams et al. 1993a), but the thymus ESTs from this study did not match the ESTs identified previously.

Approximately 83% of the ESTs with similarity to known genes match within the coding regions as opposed to untranslated regions, with the largest proportion of the matches present in the middle and 5′ end of the genes. Approximately 6% of the ESTs matched 5′ UTR regions, and 8.5% represented 3′ UTRs. Only two instances of matches to introns were observed: within the transmembrane glycoprotein CD53 and the MHC class II HLA–DRw–β loci. We would certainly expect both of these genes to be expressed in lymphocytes present in thymus tissue; thus, these ESTs may represent heterogenous nuclear RNAs (hnRNAs), which, because they are polyadenylated, would have copurified with mRNA from the total RNA isolated by poly(A)$^+$ fractionation.

Of the nonrepetitive human adult thymus ESTs generated in this study, 237 (79.5%) have no similarity to any current data base entries. This number is higher than that reported by Adams et al. (1993b) for human fetal brain (65%). Whereas ~15% of these randomly primed ESTs contain some amount of repetitive sequence (usually on one end), 64% of the ESTs generated in this study had no detectable repeats. Only 25% of the 298 ESTs examined contained regions of marginal to excellent coding potential as predicted by GRAIL 1 or 2. Of these, only 18 of the 27 clones exhibiting data base matches to coding regions (not 3′ or 5′ UTRs) of known genes were predicted to contain coding regions. This would indicate that the success rate of GRAIL in this data set is only 67%. If we extrapolate this success rate to the remainder of the data set, it is likely that ⩾50% of the 298 clones are good reagents for new gene-based markers for thymus tissue. This would suggest also that our collection contains ~100 new coding regions from human adult thymus, some of which may be different regions of the same gene. Despite our limited data set, it appears that ESTs from a randomly primed library may prove useful as new gene-based markers for mapping, particularly as a large proportion of them may represent the middle regions of genes. These ESTs will aid in the virtual assembly of coding regions using existing data. Also, they will provide additional coding information that eventually can be placed in its proper context by genomic sequencing.

METHODS

cDNA Sequencing and Primer Design

The NSTHII cDNA library used for this work was a randomly primed, normalized library constructed from poly $(A)^+$-fractionated total RNA from human adult thymus tissue and cloned into λgt10. Contaminating genomic DNA was removed from the RNA by differential centrifugation with LiCl (Patanjali et al. 1991). Individual clones were amplified by PCR with primers flanking the *Eco*RI cloning site of λgt10 (Patanjali et al. 1991). The PCR products were purified by centrifugation in Centricon-100s (Amicon, Danvers, MA) and sequenced directly with a custom dye-labeled λgt10-specific primer and *Taq* cycle sequencing chemistry. Data were collected on an Applied Biosystems model 370A DNA Sequencer [(AB) Applied Biosystems Division, Perkin-Elmer, Foster City, CA], and edited manually using AB Se qEd software.

The sequence data set obtained was screened against itself to eliminate redundant clones and queried against a subset of human repetitive elements (Alu, L1, THE, MER) using fastdb (Intelligenetics, Mountain View, CA). This work was performed using a SUN workstation. PCR primers were designed from nonredundant clones in regions that contained no repetitive elements using PRIMER (The Whitehead Institute for Biomedical Research, Cambridge, MA). Primers were synthesized at 40-nmole scale on an AB 380B Nucleic Acid Synthesizer and used unpurified in subsequent PCR reactions. All sequences were queried using BLASTP against the SWISSPROT (Release 31), and BLASTN against the Genome Sequence Data Base (GSDB: daily update) and dbEST (weekly update) data bases using the GenQuest server at Oak Ridge National Laboratory and custom scripts. The data base hits represented in Tables 1 and 3 were obtained from searches performed between July 12 and 13, 1995. Data base matches with a score of $\geqslant p\text{Val} = 1 \times 10e^{-5}$ were examined manually. Percent identity and match length for data base hits are shown in Tables 1 and 3. Analysis of the ESTs for coding potential was performed using Grail 1 and 2 (Uberbacher and Mural 1991) via the e-mail server at Oak Ridge National Laboratory. In all tables, ESTs with coding potential are indicated with an asterisk. The 298 ESTs containing unique sequences were deposited in GSDB under accession numbers L44279–L44576.

Mapping of ESTs to a Somatic Cell Hybrid Panel

The mapping of ESTs was performed in the laboratory of R.S. Athwal using PCR against a panel of rodent/human hybrid cell lines derived from various sources. Primers were initially tested against human placental DNA (Sigma, St. Louis, MO) and monochromosomal hybrids containing either human chromosome 19 (UV5HL9-5B; L.H. Thompson, Lawrence Livermore National Laboratory or 2 (CHH-2; R.S. Athwal), as well as hamster genomic DNA (UV5; L.H. Thompson) in the laboratory of A.V. Carrano. Monochromosomal hybrids for 15 human chromosomes (1, 2, 5, 6, 7, 8, 9, 12, 13, 14, 15, 16, 17, 21, and X) were produced in the laboratory of R.S. Athwal as described (Athwal et al. 1985). Hybrids containing human chromosomes 4, 10, 18 19, 20, 22, and Y were purchased from the Human Genetic Mutant Cell Repository (Camden, NJ); hybrids for chromosomes 3 and 11 were obtained from D. Patterson (Eleanor Roosevelt Institute for Cancer Research, Denver, CO). In addition, four ESTs were mapped in the lab of A.V. Carrano using the BIOS panel (BIOS Laboratories, New Haven, CT) combined with the CHH-2 and UV5HL9-5B monochromosomal hybrids.

EST-specific primers were tested using PCR against 100 ng of DNA from the individual hybrid cell lines in a 100-μl reaction volume containing 1 μM each primer, 200 μM each dNTP, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.001% gelatin, and 1 unit of *Taq* DNA polymerase (Perkin-Elmer). Three different sets of cycling conditions were used in the Perkin-Elmer 480. Initial denaturation was for 4 min at 94°C followed by 35 cycles of: (set A) 94°C for 1 min, 58°C for 2 min, 72°C for 2 min; (set B) 94°C for 1 min, 60°C for 1 min, 72°C for 1 min; and (set C) 94°C for 1 min, 62°C for 1 min, 72°C for 1 min. The set of PCR conditions used for primers derived from each clone are indicated in Tables 1 and 2.

ACKNOWLEDGMENTS

We thank David Ow and Dr. Annette Schwartz for assistance with the data base submissions; and Mishelle Montgomery, Lorie Devlin, and Melissa Ramirez for excellent technical assistance. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno, A.R. Kerlavage, W. R. McCombie, and J.C. Venter. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252:** 1651–1656.

Adams, M.D., A.R. Kerlavage, C. Fields, and J.C. Venter. 1993a. 3,400 New expressed sequence tags identify diversity of transcripts in human brain. *Nature Genet.* **4:** 256–267.

Adams, M.D., M.B. Soares, A.R. Kerlavage, C. Fields, and J.C. Venter. 1993b. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nature Genet.* **4:** 373–380.

Athwal, R.S., M. Smarsh, B.M. Searle, and S.S. Deo. 1985. Integration of a dominant selectable marker into human chromosomes and transfer of marked chromosomes to mouse cells by microcell fusion. *Somat. Cell Mol. Genet.* **11:** 177–187.

Berry, R., T.J. Stevens, N.A.R. Walter, A.S. Wilcox, T. Rubano, J.A. Hopkins, J. Weber, R. Goold, M.B. Soares, and J.M. Sikela. 1995. Gene-based sequence-tagged-sites (STSs) as the basis for a human gene map. *Nature Genet.* **10:** 415–423.

Boguski, M.S. and G.D. Schuler. 1995. ESTablishing a human transcript map. *Nature Genet.* **10:** 369–371.

Durkin, A.S., D.R. Maglott, and W.C. Nierman. 1992. Chromosomal assignment of 38 human brain expressed sequence tags (ESTs) by analyzing fluorescently labeled PCR products from hybrid cell panels. *Genomics* **14:** 808–810.

Gieser, L. and A. Swaroop. 1992. Expressed sequence tags and chromosomal localization from a subtracted retinal pigment epithelium library. *Genomics* **13:** 873–876.

Khan, A.S., A.S. Wilcox, M.H. Polymeropoulos, J.A. Hopkins, T.J. Stevens, M. Robinson, A.K. Orpana, and J.M. Sikela. 1992. Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nature Genet.* **2:** 180–185.

Lennon, G.G., C. Auffray, M. Polymeropoulos, and M.B. Soares. 1995. The I.M.A.G.E. Consortium: An integrated molecular analysis of genomes and their expression. *Genomics* (in press).

Okubo, K., N. Hori, R. Matoba, T. Niiyama, A. Fukushima, Y. Kojima, and K. Matsubara. 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet.* **2:** 173–179.

Patanjali, S.R., S. Parimoo, and S.M. Weissman. 1991. Construction of a uniform-abundance (normalized) cDNA library. *Proc. Natl. Acad. Sci.* **88:** 1943–1947.

Polymeropoulos, M.H., H. Xiao, A. Glodek, M. Gorski, M.D. Adams, R.F. Moreno, M.G. FitzGerald, J.C. Venter, and C.R. Merril. 1992. Chromosomal assignment of 46 brain cDNAs. *Genomics* **12:** 492–496.

Polymeropoulos, M.H., H. Xiao, J.M. Sikela, M. Adams, J.C. Venter, and C.R. Merril. 1993. Chromosomal distribution of 320 genes from a brain cDNA library. *Nature Genet.* **4:** 381–386.

Sikela, J.M. and C. Auffray. 1993. Finding new genes faster than ever. *Nature Genet.* **3:** 189–191.

Uberbacher, E.C. and R.J. Mural. 1991 Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci.* **88:** 11261–11265.

Received August 9, 1995; accepted in revised form October 16, 1995.