Identification of 4370 expressed sequence tags from a 3'-end-specific cDNA library of human skeletal muscle by DNA sequencing and filter hybridization.

G Lanfranchi, T Muraro, F Caldara, et al.

Genome Res. 1996 6: 35-42
Access the most recent version at doi:10.1101/gr.6.1.35

LETTER

Identification of 4370 Expressed Sequence Tags from a 3′-End-Specific cDNA Library of Human Skeletal Muscle by DNA Sequencing and Filter Hybridization

G. Lanfranchi, T. Muraro, F. Caldara, B. Pacchioni, A. Pallavicini, D. Pandolfo, S. Toppo, S. Trevisan, S. Scarso, and G. Valle[1]

Centro di Ricerca Interdipartimentale per le Biotecnologie Innovative (CRIBI) Biotechnology Centre, Università degli Studi di Padova, 35121 Padova, Italy

A systematic study on the mRNA species expressed in the human skeletal muscle is presented in this paper. To carry on this study, a new method has been developed for the construction of unbiased cDNA libraries specially designed for the production of ESTs corresponding to the 3′-end portion of the mRNAs. The method has been applied to human skeletal muscle, where the analysis of the transcription profile is particularly difficult for the presence of several very abundant transcripts. To detect and quantify high-level mRNAs, the first 1054 ESTs were obtained from randomly selected clones. The 10 most abundant transcripts accounted for >45% of the clones. Subsequently, these transcripts were identified by filter hybridization, thus making DNA sequencing more productive. Overall, 4370 clones were identified: 3372 by DNA sequencing and 998 by filter hybridization. The number of groups of sequences identifying individual transcripts was relatively low compared with other tissues, resulting in a total of 934 groups out of 4370 ESTs. Of these, 719 groups were represented by only one sequence.

The identification of human genes by systematic sequencing of genomic DNA is hindered by the dispersion of the genes among large noncoding regions and by the presence of introns within genes. Systematic sequencing of cDNA libraries is an alternative approach that offers several advantages: First, it should allow the identification of most human genes within a reasonable time (Adams et al. 1991; Sikela and Auffrey 1993); second, cDNA libraries can be prepared from different human tissues, allowing the construction of maps of tissue-specific and stage-specific genes (Okubo et al. 1992; Adams et al. 1993a); furthermore, in some cases the frequency of a given sequence in the cDNA library can be related to the relative abundance of the corresponding mRNA, giving an indication of the level of gene expression (Okubo et al. 1992).

The general strategy of such studies is based on the analysis by a "single pass" systematic sequencing of random cDNA clones, which results in the production of short partial sequences, generally referred to as expressed sequence tags (ESTs). Several laboratories have recently produced many thousands ESTs from different human tissues and cell lines (Adams et al. 1991, 1992, 1993a,b; Gieser and Swaroop 1992; Okubo et al. 1992; Sikela and Auffrey 1993; Takeda et al. 1993; Liew et al. 1994; Frigerio et al. 1995) and from other organisms (Waterston et al. 1992; Höfte et al. 1993; Wan Kim et al. 1993).

Most protocols result in the production of ESTs that correspond to random positions within the original mRNAs, thus allowing the assembly of contigs belonging to the same mRNA and the definition of full-length transcripts. This is done either by random priming (Adams et al. 1991; Wan Kim et al. 1993) or by oligo(dT) priming followed by directional sequencing from the 5′ end (Adams et al. 1993b). However, owing to the different abundance of the mRNAs, some sequences will be done many times, whereas the uncommon sequences will remain identified only by partial tags or will not be identified at all.

[1]**Corresponding author.**
E-MAIL valle@eos.bio.unipd.it; FAX + 39-49-8276280.

This is particularly critical in tissues like skeletal muscle, where some mRNAs are known to be expressed at very high levels. Therefore, it is not surprising that most human ESTs currently available in GenBank were obtained from tissues displaying a broad range of transcripts such as brain, liver, and tumor cells. However, a considerable number of ESTs from skeletal muscle have been produced (Sikela and Auffrey 1993), but there is little accompanying documentation.

For an investigation on tissues like skeletal muscle, it might be advisable to aim first at the construction of a reliable catalog of transcripts and only later to the full characterization of the unknown sequences. In this respect Okubo et al. (1992) proposed a strategy where the main objective of EST sequencing is not the identification of full-length mRNAs but rather the production of a catalog in which each transcript is identified only by its 3′ end. Thus, systematic random sequencing can be restricted to a relatively small portion of each mRNA, which avoids dispersion and allows the definition of a more accurate catalog. However, the method of Okubo et al. (1992) requires the presence of particular restriction sites in the 3′-end portion of the mRNAs, which could result in the loss of some transcripts and in the production of biased catalogs.

In this paper we describe a new method for the construction of 3′-end cDNA libraries, specially designed for EST sequencing and for further studies of genome mapping (Wilcox et al. 1991; Khan et al. 1992). Furthermore, this new method was combined with a hybridization procedure for identifying the most abundant mRNAs, and used for a systematic analysis of transcription in human skeletal muscle.

RESULTS

Construction and Verification of the cDNA Library

A schematic representation of the strategy used for the construction of the cDNA library is shown in Figure 1. A more detailed technical description can be found in Methods.

After the production of the first 1054 ESTs, the 10 most abundant transcripts were analyzed to verify the percentage of the cDNA clones that were actually corresponding to the 3′-end portions of the mRNAs. The large majority (96%) of

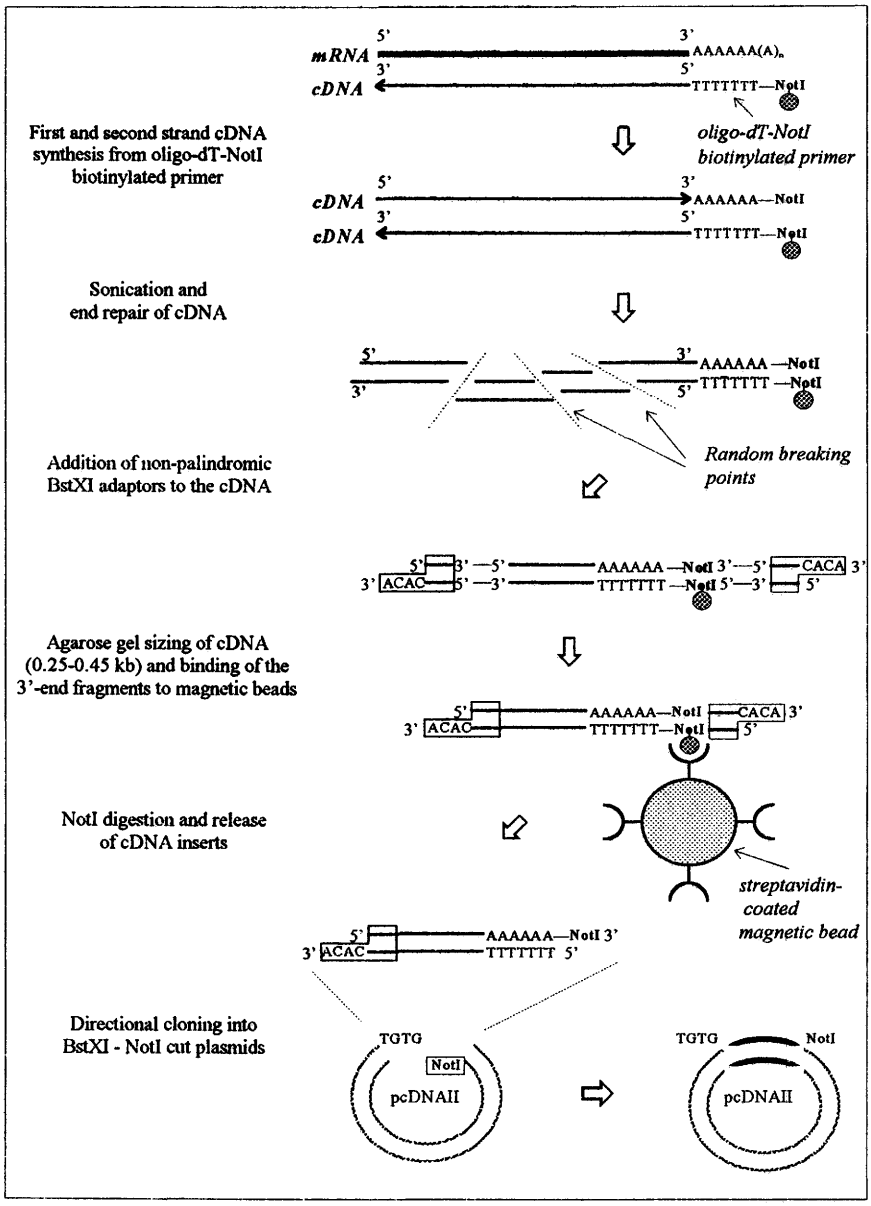

Figure 1 Schematic diagram of the method used for the construction of the cDNA library. More details can be found in the text.

the sequences matched to the putative 3′ end of the corresponding mRNA with an approximation of <10 bases. Most of the remaining sequences did not reach the 3′ end owing to a premature stop of the electrophoretic run, resulting in termination within 100 bases from the 3′ end. Only 0.8% of the cDNA inserts, all corresponding to mitochondrial transcripts, did not match the orthodox 3′ end of the mRNA. However, these odd-terminating sequences could be attributable to variable ends of the mRNAs after processing of the polycistronic mitochondrial transcript rather than to a cloning artifact.

To verify whether the method described in Figure 1 produces cDNA clones with a frequency related to the abundance of the corresponding mRNAs, we performed a Northern blot analysis on five transcripts found at different frequencies in our EST catalog. The results are shown in Figure 2 and confirm the validity of the method.

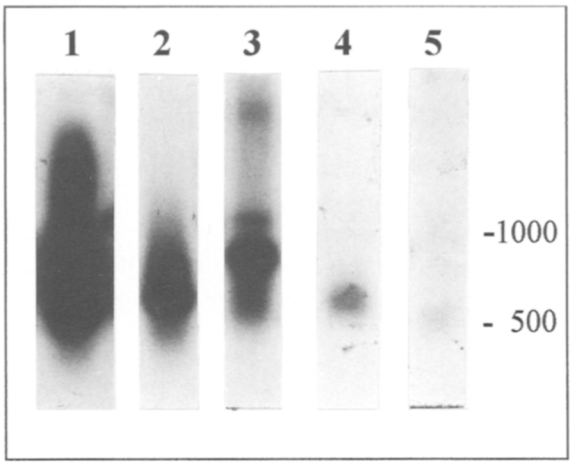

Figure 2 Northern blot analysis of five mRNAs that were found with different abundance in our EST catalog. The five lanes were loaded with the same amount of human skeletal muscle mRNA and were hybridized with the following probes: (Lane *1*) ND-3, found in our catalog 443 times; (lane *2*) α-globin, found 112 times; (lane *3*) myosin LC-2, found 91 times; (lane *4*) transcript 387, found 3 times; (lane *5*) transcript 545, found 2 times. Although the autoradiography does not allow a precise quantitation owing to saturation of the film, it can be easily appreciated that there is a very good correlation between the intensity of the bands and the relative abundance of the ESTs. The length (bases) of the markers is indicated at *right*.

Computer Analysis and Construction of a Catalog of Transcripts

Following the criteria described in Methods, the 4370 ESTs were analyzed, and as a result, they could be arranged into 934 groups, each corresponding to a different putative transcript. The 215 transcripts that were found at least twice are listed in Figure 3, where the best similarities to already known sequences are also indicated. As can be seen, most of them correspond to known genes. However, the relative frequency of new genes is much higher in the 719 sequences that were found only once and that account for 146 sequences corresponding to already known genes, 103 corresponding to ESTs already present in GenBank, 273 showing some similarity to already known sequences (68 to ESTs, 82 to Alu elements, and 123 to other genes), and 197 with no significant similarity to any sequence of GenBank.

Overall, the 934 groups correspond to 284 known human sequences, 133 ESTs that were already present in GenBank, 296 sequences showing some similarity to already known sequences (72 to ESTs, 86 to Alu elements, and 138 to other genes), and 221 with no significant similarity to any sequence of GenBank.

Occurrences of Known Transcripts and Comparison with Other EST Catalogs

One of the most noticeable features of our EST catalog is that mitochondrial transcripts are extremely abundant. They account for 1082 ESTs (24.8%), all but one corresponding to the H-strand of the mitochondrial DNA; the exception being an EST of cytochrome oxidase I, that most probably entered into the vector in the wrong orientation.

It was particularly interesting to observe that mitochondrial transcripts behave quite differently from nuclear transcripts (Attardi and Schatz 1988) as the 3′ polyadenylated ends often occur at unexpected locations. This is particularly true for certain regions of the mitochondrial DNA. For instance, the ND-4 gene was detected 102 times, 25 of which terminate at variable early positions along the gene. Many ESTs also match the mitochondrial rRNA region (group HSM00007 of Fig. 3) and randomly terminate within a wide range of 2600 bases, always corresponding to the H-strand of the mitochondrial genome. Because most of these odd-terminating RNAs have long poly(A) stretches at their 3′ end, it is conceivable that they actually correspond to real RNA terminations.

The catalog of transcripts from human skeletal muscle was compared with other EST catalogs such as heart (Liew et al. 1994), fetal brain (Adams et al. 1993b), hippocampus (Adams et al. 1993b), pancreatic islet cells (Takeda et al. 1993), and HepG2 cells (Okubo et al. 1992). The comparison is quite interesting and indicative of major differences; however, it should be taken into account that in all these projects the cDNA libraries were produced by a variety of methods, often aiming at different goals. However, in all the above catalogs, the frequency of mitochondrial ESTs is much lower: 12% in heart, 10.3% in hippocampus, 1.7% in fetal brain, 1% in pancreatic islet cells, and <0.2% in HepG2 cells.

As expected, transcripts specific for contractile elements are also very frequent in our EST catalog, in particular (α-actin that was found 373 times (8.5%). This abundance is remarkably high compared with heart, where no nuclear transcripts were found to be >1% of the total, the two most abundant being (β-myosin H-chain and (α-tropomyosin (0.87% and 0.42%, respectively).

The frequency of transcripts involved in protein synthesis is also relatively high, with 199 ESTs (4.5%) matching mRNAs for ribosomal proteins. This percentage is higher than that found in most other tissues.

DISCUSSION

Libraries of short cDNA fragments corresponding to the 3′-end regions of the mRNAs have the inconvenience of not being very informative in terms of coding sequence; however, they offer several advantages. First, the resulting sequencing projects are restricted to a relatively small region of the mRNAs allowing a better and faster compilation of catalogs of transcripts (Okubo et al. 1992). Second, all the mRNAs, regardless of their length, have an equal opportunity of being

Code	Times	Best similarity	Description
HSM00002	443*	MIHSGENOM	NADH dehydrogenase chain 3
HSM00026	373*	HUMSAACT	Skeletal alpha-actin
HSM00012	319*	MIHSGENOM	ATPase 6
HSM00007	278	MIHSGENOM	Mit.-DNA (1-2600; H strand)
HSM00008	195*	MIHSGENOM	Cytochrome C oxydase II
HSM00018	157*	MIHSGENOM	NADH dehydrogenase chain 2
HSM00028	137*	MIHSGENOM	Cytochrome C oxydase III
HSM00017	112*	HSAGLO1	Alpha globin
HSM00049	102	MIHSGENOM	NADH dehydrogenase chains 4-4L
HSM00014	91*	HSMYLC2	Myosin LC-2
HSM00030	87*	HUMTNTS	Slow skeletal muscle troponin T
HSM00009	79*	HUMCKMM8	Creatine kinase (CKMM)
HSM00050	51	MIHSGENOM	Cytochrome b
HSM00013	46	HUMGAPDH	Glyceraldehyde-3-P dehydrogen.
HSM00034	45	MIHSGENOM	Cytochrome C oxydase I
HSM00258	41	MIHSGENOM	NADH dehydrogenase chain 1
HSM00064	37	HUMTROPI07	Troponin I
HSM00015	33		- NEW SEQUENCE -
HSM00099	32	OCMYLC22	Similar to rabbit myosin L-2-2
HSM00127	30	HUMALDOAA	Fructose 1,6-diphosphate aldolase A
HSM00055	26	HUMTROPSK	Skeletal beta-tropomyosin
HSM00461	25	HUMCYTVIIA	Cytochrome c oxidase subunit VIIa
HSM00029	24	HSMHCBR	Beta-myosin H-chain (slow)
HSM00340	23	HSMG03	Myoglobin exon 3
HSM00178	21	HSRPL41	Human homologue to yeast rpL41
HSM00302	21	HUMTROPNIN	Troponin I fast-twitch isoform
HSM00743	20	HSRPS11	Ribosomal pS11
HSM00016	20	HUMDES	Desmin
HSM00019	20	HSB24B071	EST (skeletal muscle)
HSM00060	17	HSFMHC	Fast 2a myosin H-chain
HSM00700	17	HSU14973	Ribosomal pS29
HSM00292	17	M24906	Rosenthal fiber prot. (α-B-crystallin)
HSM00451	16	HSTC2	Fast skeletal troponin C
HSM00187	15	HSRPL37A	Ribosomal pL37a
HSM00023	15	HUMCOXVIM	Cytochrome c oxidase subunit VIa
HSM00778	14	HSLLREP3	LLRep3, repetitive DNA
HSM00512	13	HSBGL1	Beta-globin
HSM00351	13	HSSMYBPC	Slow-type myosin binding protein C
HSM00105	13	HUMTROPA2	Skeletal alpha-tropomyosin
HSM00460	12	MIHSGENOM	Mit.-DNA (13573-14171, H strand)
HSM00574	11	HUMPPARP1	Acidic ribosomal phosphoprotein P1
HSM00100	11	HUMPYGM20	Glycogen phosphorylase
HSM00096	10	HSTNCS	Slow skeletal troponin C (TnC)
HSM00950	10	HSTROPSR	Skeletal tropomyosin
HSM00189	10	HUMPPARP2	Acidic ribosomal phosphoprotein P2
HSM00294	10	HUMQM	Wilm's tumor-related prot. QM
HSM00899	9	MIHSGENOM	NADH dehydrogenase chain 5
HSM00124	9	HSTITIN	Titin
HSM00264	9	HUMEF2A	Elongation factor 2 (EF-2)
HSM00445	9	HUMMLC1V7	Ventric. slow twitch myosin L-chain
HSM00003	9	HUMSRAA	Ribosomal pS16
HSM00341	8	HSENO3BE	Muscle beta-enolase
HSM00110	8	HSHSP27L	Heat shock protein HSP27
HSM00459	8	HSMYH1R	Myosin H-chain light meromyosin
HSM00128	8	HUMAK1	Cytosolic adenylate kinase (AK1)
HSM00402	8	HUMMHBA123	GTP binding prot. related to MHC
HSM00355	8	HSB76I1082	EST (skel. muscle)
HSM00577	7	HSRPL32	Ribosomal pL32
HSM00826	7	HUMCOXCA	Cytochrome c oxidase VB
HSM00575	7	HUMRPL37Z	Ribosomal pL37
HSM00242	7	HUMTCG	D loop region (1-580 H-strand)
HSM01684	7	HUMRPS24A	Ribosomal pS24
HSM00848	7	RABATPAD	Similar to rabbit fast skel. ATPase
HSM00401	6	HSTPI1G	Triosephosphate isomerase
HSM00072	6	HUMCYC1A	Cytochrome c-1

Figure 3 Catalog of transcripts, obtained from the analysis of 4370 ESTs. A total number of 934 distinct sequences were identified, each corresponding to a different transcript. However, only the 215 transcripts with at least two ESTs are listed in the catalog, leaving out the 719 that occurred only once. The catalog is ordered according to the times that each transcript was found (column 2). The first column indicates the code of the transcript. The best similarity found with sequences in GenBank or SWISS-PROT data bases, with the corresponding descriptions are reported in columns 3 and 4. The asterisks (*) indicate the 10 sequences that were also identified by dot–blot hybridization.

Figure 3 (Continued on facing page.)

Code	Times	Best similarity	Description
HSM00282	5	HSRPS18	Ribosomal pS18
HSM01174	5	HUMCH13C4A	Translat. controlled tumor prot.
HSM01289	5	HUMRPS21X	Ribosomal pS21
HSM00771	5	HUMRPS6	Ribosomal pS6
HSM01257	5	HUMTRT	Skeletal troponin T isoform
HSM00686	5	HSB96B062	EST (skel. muscle)
HSM00067	5	HSBA7C011	EST (skel. muscle)
HSM01284	5	T07882	EST (fetal brain)
HSM01705	5		- NEW SEQUENCE -
HSM01267	4	HS23KDHBP	23 kD highly basic protein
HSM01578	4	HSDF1F05	F1-F0 ATP synthase
HSM00331	4	HSUBPSE1	Ubiquitin pseudogene EHB4
HSM00795	4	HUMMLC1SA	Myosin LC-1s (slow)
HSM01280	4	HUMMPSI	Metallopanstimulin
HSM00594	4	HUMORFA	Cardiac Autoantigen
HSM01738	4	MIHSGENOM	Cytochrome c oxidase subunit I
HSM01391	4	HUMGS00977	EST, Promyelocytes
HSM01696	4	T31711	EST 5' end (embryo)
HSM00760	4	HSBAT2	Sequence with Alu class A
HSM00245	4		- NEW SEQUENCE -
HSM01423	4		- NEW SEQUENCE -
HSM01192	3	HSABLGR3	Proto-oncogene tyrosine kinase
HSM00120	3	HSEF1DELA	Elongation factor-1-delta
HSM01598	3	HSHA44G	Alpha-tubulin, exons 1-3
HSM00349	3	HSIGF27	Insulin-like growth factor IGF-2
HSM01331	3	HSU02032	Ribosomal pL23a
HSM01879	3	HSU12404	Csa-19
HSM00246	3	HSU14969	Ribosomal pL28
HSM00446	3	HSU16660	Peroxisomal enoyl CoA hydratase
HSM01116	3	HSUBA52P	Ubiquitin-52 aa fusion protein
HSM02210	3	HSUBA80R	Uba80 mRNA for ubiquitin
HSM01457	3	HUMFERH	Ferritin H chain
HSM00062	3	HUMFERL	Ferritin L chain
HSM01652	3	HUMGSTM4A	Glutathione transferase
HSM00252	3	HUMH19	H19 RNA gene
HSM01501	3	HUMMTATP3	ATP synthase
HSM00022	3	HUMNRF1A	NRF1 protein
HSM00190	3	HUMRPS20	Ribosomal pS20
HSM00299	3	HUMRYR	Ryanodine receptor
HSM01040	3	HUMSAPC1	Cerebroside sulfate activator
HSM00338	3	HUMSCAR	Scar protein
HSM02185	3	HUMTBP1	Human virus tat-binding protein-1
HSM02283	3	S42658	Ribosomal pS3
HSM00869	3	HSAFIF085	EST (fibroblast)
HSM01327	3	HUM0000A25	EST 3' end (HepG2 cell line)
HSM00075	3	T03744	EST (infant brain)
HSM00119	3	T07953	EST (brain)
HSM01519	3	T18875	EST 5' end (testis)
HSM02227	3	T29985	EST 3' end (adipose tissue)
HSM00194	3	T30025	EST 3' end (adipose tissue)
HSM01448	3	T30687	EST 5' end (spleen)
HSM01092	3	T31434	EST 5' end (embryo)
HSM00517	3	HSAAACTPB	Similar to EST (placenta)
HSM01265	3	HSB68E072	Similar to EST (skeletal muscle)
HSM00290	3	BOVNADHURD	Similar to bovine NADH dehydr.
HSM00385	3	BTCIB8	Similar to bovine ubiq. reductase
HSM00138	3	HSRPL19	Ribosomal pL19
HSM00547	3	HUMTM1E	Similar to epithelial tropomyosin
HSM00149	3		- NEW SEQUENCE -
HSM00387	3		- NEW SEQUENCE -
HSM01533	3		- NEW SEQUENCE -
HSM00304	2	HSAT3	Antithrombin III
HSM00897	2	HSCOVIC	Cytochrome c oxidase subunit VIc
HSM00838	2	HSCOX7BM	Cytochrome c oxidase (coxVIIb)
HSM02146	2	HSCYCR	T-cell cyclophilin
HSM01900	2	HSHSC70	Heat shock cognate protein
HSM02261	2	HSHSP70B	Heat shock protein 70
HSM01636	2	HSL31	Ribosomal pL31
HSM01987	2	HSLEC14K	Beta-galactoside-binding lectin
HSM00886	2	HSMLC3F	Myosin LC-3f (fast)
HSM00137	2	HSPAG	Proliferation-associated gene
HSM00697	2	HSPEABP	Phosphatidylethanolamine binding
HSM00698	2	HSPMIPR	PMI (putative receptor protein)
HSM02247	2	HSRP26AA	Ribosomal pL26
HSM01235	2	HSRPL38	Ribosomal L38
HSM01132	2	HSRPS12	Ribosomal pS12
HSM00781	2	HSRPS8	Ribosomal pS8
HSM00610	2	HSU12465	Ribosomal pL35
HSM01699	2	HSU14966	Ribosomal pL5
HSM02436	2	HSU14972	Ribosomal pS10
HSM02522	2	HSU16738	Clone containing trinucl. repeat
HSM00378	2	HUMALBP	Adipocyte lipid-binding protein
HSM00557	2	HUMANCDA	Adipsin/complement factor D
HSM02228	2	HUMCD63	Lysosomal glycoprotein CD63
HSM00125	2	HUMCOX4AA	Cyt-c oxidase subunit IV
HSM00998	2	HUMCPB	Calphobindin II
HSM01064	2	HUMCRP04	Cysteine-rich protein (CRP)
HSM00417	2	HUMFIXG	Factor IX antihemophilic factor B
HSM02784	2	HUMGLGGA	Unusual fetal A-gamma-globin
HSM00274	2	HUMGLPEX	Se-dependent glutathione perox.
HSM01760	2	HUMHMGYD	HMG-Y protein isoform
HSM00581	2	HUMORF	ORF, T-lymphocyte cells
HSM00493	2	HUMPAIA	Plasminogen activator inhibitor-1
HSM01105	2	HUMRIBPROC	Ribosomal p L11
HSM02268	2	HUMRIBPROD	Ribosomal pL18a
HSM00793	2	HUMRPL30A	Ribosomal pL30
HSM00514	2	HUMRPS13A	Ribosomal pS13
HSM01285	2	HUMRPS17	Ribosomal pS17
HSM02547	2	HUMSET	Set gene (putative oncogene)
HSM00094	2	HUMTCBA	TCB (thyroid horm.-bind. prot.)
HSM00001	2	HUMTCRADCV	T-cell receptor C-alpha V-delta
HSM01306	2	HUMTLCA	ADP/ATP translocase
HSM01564	2	HUMUBI13	Ubiquitin
HSM00525	2	S73035S9	Guanosine-P reductase
HSM01449	2	HSB35C032	EST (skeletal muscle)
HSM03052	2	HSB95E072	EST (skeletal muscle)
HSM01240	2	HSDHII070	EST, (heart)
HSM01328	2	M91220	EST (retinal pigment)
HSM00498	2	T03426	EST (infant brain)
HSM02958	2	T05586	EST (brain)
HSM00746	2	T06948	EST (fetal brain)
HSM01827	2	T10004	EST (normalized cDNA; brain)
HSM02603	2	T10156	EST (normalized cDNA; brain)
HSM01342	2	T11087	EST (human pancreatic islet)
HSM01483	2	T16091	EST 3' (infant brain)
HSM02700	2	T24043	EST 3' (brain)
HSM02171	2	T32287	EST (brain)
HSM02095	2	T34729	EST (brain)
HSM00202	2	S75201	Sequence with Alu class A
HSM00384	2	HSU04737	Sequence with Alu class C
HSM00759	2	HSRSPAC	Sequence with Alu class E
HSM02710	2	M62107	Similar to EST (hippocampus)
HSM02912	2	T24920	Similar to EST (colorectal cancer)
HSM01782	2	BTCIKFYI	Similar to bovine NADH dehydr.
HSM02395	2	BTCIMNLL	Similar to bovine NADH dehydr.
HSM00797	2	CVGEM5ZFM	Similar to vector pGEM-5Zf(-)
HSM01293	2	HS165	Similar to 165kD titin-associated
HSM02386	2	HSL35A	Similar to ribosomal pL35a
HSM03020	2	HSU14970	Similar to ribosomal pS5
HSM01089	2	HUMRPS14	Similar to ribosomal pS14
HSM00041	2	HUMRPZH21	Similar to ribosomal pL44
HSM01429	2	MIBTCIB22	Similar to bovine NADH dehydr.
HSM00101	2	RATGAP01	Similar to rat GAP-43 gene
HSM00005	2		- NEW SEQUENCE -
HSM00150	2		- NEW SEQUENCE -
HSM00275	2		- NEW SEQUENCE -
HSM00284	2		- NEW SEQUENCE -
HSM00288	2		- NEW SEQUENCE -
HSM00315	2		- NEW SEQUENCE -
HSM00374	2		- NEW SEQUENCE -
HSM00545	2		- NEW SEQUENCE -
HSM00769	2		- NEW SEQUENCE -
HSM00836	2		- NEW SEQUENCE -
HSM00875	2		- NEW SEQUENCE -
HSM00957	2		- NEW SEQUENCE -
HSM00964	2		- NEW SEQUENCE -
HSM01737	2		- NEW SEQUENCE -
HSM01758	2		- NEW SEQUENCE -
HSM02190	2		- NEW SEQUENCE -
HSM03043	2		- NEW SEQUENCE -

represented. Third, the cDNA inserts can be sequenced in the direction 5′ → 3′, thus avoiding passing through the poly(A) that generally results in very poor or unreadable DNA sequences.

Our strategy of random fragmentation of the cDNA by sonication, followed by a very stringent selection of the 3′-end fragments, produced very satisfactory results as the great majority of our cDNA inserts were found to match the 3′-end region of their corresponding mRNAs (see Results) and transcripts such as titin and ryanodine receptor (14,985 and 15,345 bases long) were detected nine and three times, respectively, confirming that also very large transcripts could be processed to 3′-end inserts similar to those derived from small transcripts.

A method for the production of 3′-end-specific cDNA libraries had been described previously by Okubo et al. (1992). They made use of restriction enzymes such as *Mbo*I, for the double function of fragmenting the cDNA and producing sticky ends useful for directional insertion into the vector. We consider our new protocol for the construction of the cDNA library an improvement on the general strategy, as sonication allows a more random fragmentation of the cDNA than restriction enzymes. For instance, among the 10 most abundant transcripts, 3 do not have an *Mbo*I site (ND-2, Cox-3, and α-globin) and 1 (myosin LC-2) has an *Mbo*I site at <20 bases from the poly(A); therefore, 4 transcripts amongst the 10 most abundant would not be cloned using that procedure.

A relevant feature of our 3′-end-restricted cDNA library is that the frequency of an EST gives a good indication of the relative abundance of the corresponding mRNA. This is an important piece of information for expression studies, but it must be considered together with the inconvenience that the most abundant ESTs could be resequenced many times, wasting time and effort. In principle, our method for the production of 3′-end-restricted cDNA libraries could be used after normalization of the library (Ko 1990; Patanjali et al. 1990), but the quantitative information would be lost. In this work we used the dot–blot hybridization as an alternative approach to overcome the problem of abundant transcripts, while maintaining the quantitative information. This was facilitated by having a cDNA library restricted to only the 3′-end regions, which allowed an easier construction of specific probes. However, at a later stage of the work, we may consider to continue the search for the rarest transcripts using a normalized library restricted to the 3′-end regions.

As mentioned in the introductory section, 3′-end ESTs are very suitable for mapping studies (Wilcox et al. 1991; Khan et al. 1992). In this respect, we have started a preliminary project in collaboration with other laboratories, using a panel of rodent/human somatic cell hybrids (Wilcox et al. 1991). We found that in >90% of the cases the differences with the rodent sequences were such that a PCR product was only observed from human DNA, making the analytical procedures very simple.

The ESTs described in this paper have been submitted to the EMBL data base (accession nos. F15505–F155554 and F15586–F19692). Further information on this work, including the ABI-chromatogram images, are available at the web site http://eos.bio.unipd.it.

METHODS

Construction of the 3′-End-specific cDNA Library

A 12-gram sample of human pectoral muscle was obtained from a mastectomy of an adult woman and, after cleaning out the tumor tissue and the infiltrated lymph nodes, it was prepared as described by Chomczynski and Sacchi (1987). Polyadenylated RNA was selected by oligo(dT) column chromatography using standard techniques (Aviv and Leder 1972).

The construction of the cDNA library was based on the Librarian II kit (Invitrogen), with some major changes in the procedure (Fig. 1). The first cDNA strand was synthesized from a specially designed oligo(dT)–*Not*I primer in which the 5′-terminal nucleotide was biotinylated. After completion of the second strand, the cDNA was sonicated for 10 sec with a Branson probe set at 100 W. Then cDNA fragments were repaired with T4 DNA polymerase, ligated to *Bst*X1 nonpalindromic adaptors, and size-fractionated on low-melting-point agarose (SeaPlaque, FMC) gels. The cDNA fragments of 250–450 bp were extracted with β-agarase (New England Biolabs), dissolved in TTL buffer [0.3 M Tris at pH 8.0, 6 M LiCl, and 0.3% (vol/vol) Tween 20], and incubated with 20 µl of avidin-coated paramagnetic beads (Dynabeads, Dynal) at 42°C for 30 min, with gentle mixing every 5 min. After three washes with 300 µl of TTL and two washes with water, the 3′ cDNA ends were released from the beads by *Not*I digestion. The cDNA was directionally cloned into a *Bst*X1, *Not*I-digested pcDNAII plasmid (Invitrogen), and used for transformation of the *Escherichia coli* TOP10F′ strain using an electroporator.

DNA Sequencing

Single bacterial colonies were collected with sterile toothpicks, transferred into 50 µl of PCR buffer [20 mM Tris-HCl at pH 8.3, 50 mM KCl, 2 mM $MgCl_2$, and 0.1% (vol/vol) Tween 20], lysed at 95°C for 10 min in 96-well microtiter

plates, and processed as described by Hultman et al. (1991) using paramagnetic beads (Dynal). Single-stranded templates were processed in a sequencing reaction by the dye-deoxy terminator chemistry developed by Applied Biosystems, using a sequencing primer (5′-CTCGGATCCACTAGTAACG-3′) located 21 bases upstream from the first nucleotide of the cDNA insert. Sequencing gels were run on Applied Biosystems DNA sequencers.

Northern Blotting

Skeletal muscle mRNA was denatured with glyoxal–DMSO buffer (McMaster and Carmichael 1977) and separated on agarose gels using 10 mM sodium phosphate (pH 7.0) as the running buffer. The mRNA was then transferred to nylon sheets (Hybond N+, Amersham) in 40 mM NaOH. Probes were obtained by PCR amplification of microlysate colonies with specific internal primers and labeled using the random priming technique (Feinberg and Vogelstein 1983). Standard protocols were used for hybridization and high stringency washing of Northern blots (Thomas 1980).

Preparation of Specific Probes and Dot–blot Analysis

To allow an easy identification of the most frequent ESTs without the need of DNA sequencing, a filter hybridization procedure (dot–blot) was implemented as a standard step of our protocol after the first 1054 ESTs had already been sequenced. The dot–blot procedure involved the transfer of the amplified cDNA samples from 96- to 384-well plates, followed by the spotting of 200-nl aliquots from each sample onto nylon filters (Hybond N+, Amersham) using 384-teeth disposable devices (Genetix). Filters were then processed for hybridization using standard methods. The following 10 sequences were analyzed: ND-3, α-actin, ATPase-6, Cox-2, ND-2, Cox-3, β-globin, myosin LC-2, troponin T, and creatine kinase, hereafter referred to as abundant ESTs.

Probes specific for the 10 abundant ESTs were obtained by PCR amplification of the regions corresponding to the terminal 200–250 bases at the 3′ end of the mRNAs, using specific internal primers. The amplified DNA fragments were cloned into plasmid vectors, and the inserts were purified and labeled using the nonradioactive ECL system (Amersham).

The reliability of this hybridization method was checked by performing a double analysis on a series of 384 samples, both by dot–blot hybridization and by DNA sequencing. The results of this comparison showed that under the stringency conditions used, ~3% of false negatives resulted from dot–blot hybridization.

Because some ESTs were identified both by dot–blot hybridization and by DNA sequencing, some criteria had to be defined to avoid counting twice the same ESTs. Furthermore, it had to be taken into account that some samples were lost during the procedures for DNA sequencing, whereas the ESTs identified by hybridization did not need any further analysis and could not be lost. Therefore, a simple counting of the samples positive in the dot–blot assay would have led to an overevaluation of those samples. To avoid these problems, the occurrences of the ESTs identified by dot–blot were calculated as follows. First, the percentage of the most frequent ESTs was calculated on the first 1054 samples (i.e., before the filter hybridization procedure was implemented): For instance, α-actin was found 90 times and the 10 abundant ESTs together were found 481 times, against 573 other ESTs. Second, it was considered that a total number of 3372 ESTs had been sequenced, of which 996 were abundant and 2376 were other ESTs. Finally, the occurrences of each abundant EST were calculated using the ratio observed in the first 1054. For instance, the occurrences of α-actin resulted $90 \times 2376/573 = 373$ occurrences.

Computer Management of the Data

Each new EST was first corrected, and any residual sequence of the vector or poly(A) was removed. Each EST was then analyzed with the program FASTA (Pearson and Lipman 1988), against the other ESTs already present in our data base. Two ESTs were considered to match each other when they shared at least 50 bases with at least 98% identity. For longer sequences, the threshold was decreased 1% every 25 bases. If with the above criteria a new EST was matching at the same time two different existing groups, then the two groups were joined together. In this case a manual check was performed to verify whether any EST was a recombinant with a double insert. Only two double inserts were identified, which allowed us to estimate the percentage of bad recombinant cDNA clones to <0.1%.

To attribute to each group a possible identity, each EST was searched both against GenBank and SWISS-PROT, using BLAST-N and BLAST-X, respectively (Altschul et al. 1990). The best similarities of each EST were systematically analyzed to define the comment lines shown in Figure 3. The criteria used to define whether a sequence is new, identical, or similar to other sequences were arbitrary and served only to give an approximate indication. In general, sequences with a BLAST-N probability >e-30 were considered new sequences. Scores between e-30 and e-60 were considered indicative of similarity, whereas scores <e-60 were considered indicative of identity.

ACKNOWLEDGMENTS

We thank Professor G.A. Danieli and Professor A. Ballabio for critical reading of the manuscript. We also thank Donata Belvini, Vincenzo Favino, Sara Gomirato, Cristina Potrich, Lara Stevanato, Natascia Tiso, and Rosanna Zimbello for assistance in DNA sequencing and Giorgio Rossi for the preparation of acrylamide gels. This work was financed by Telethon Italy (grant B30).

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. 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., M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelley, T.R. Utterback, J.W. Nagle, C. Fields, and J.C. Venter. 1992. Sequence identification of 2375 human brain genes. *Nature* **355:** 632–634.

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

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

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Attardi, G. and G. Schatz. 1988. Biogenesis of mitochondria. *Annu. Rev. Cell Biol.* **4:** 289–333.

Aviv, H. and P. Leder. 1972. Purification of biologically active globin messenger RNA by chromatography on oligothymidilic acid-cellulose. *Proc. Natl. Acad. Sci.* **69:** 1408–1412.

Chomczynski, P. and N. Sacchi. 1987. Single-step method of RNA isolation by acid guanidium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162:** 156–159.

Feinberg, A.P. and B. Vogelstein. 1983. A technique for radiolabeling DNA restriction endonuclease fragment to high specific activity. *Anal. Biochem.* **132:** 6–13.

Frigerio, J.M., P. Berthézène, P. Garrido, E. Ortiz, S. Barthellemy, S. Vasseur, B. Sastre, I. Seleznieff, J.C. Dagorn, and J.L. Iovanna. 1995. Analysis of 2166 clones from a human colorectal cancer cDNA library by partial sequencing. *Hum. Mol. Genet.* **4:** 37–43.

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

Höfte, H., T. Desprez, J. Amselem, H. Chiapello, M. Caboche, A. Moisan, M.F. Jourjon, J.L. Charpenteau, P. Berthomieu, D. Guerrier et al. 1993. An inventory of 1,152 expressed sequence tags obtained by partial sequencing of cDNA from *Arabidopsis thaliana*. *Plant J.* **4:** 1051–1061.

Hultman, T., S. Bergh, T. Moks, and M. Uhlén. 1991. Bidirectional solid phase sequencing of *in vitro* amplified plasmid DNA. *BioTechniques* **10:** 84–93.

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 cDNA. *Nature Genet.* **2:** 180–185.

Ko, M.S.H. 1990. An "equalized cDNA library" by the reassociation of short double-stranded cDNAs. *Nucleic Acids Res.* **18:** 5705–5711.

Liew, C.C., D.M. Hwang, Y.W. Fung, C. Laurenssen, E. Cukerman, S. Tsui, and C.Y. Lee. 1994. A catalogue of genes in the cardiovascular system as identified by expressed sequence tags. *Proc. Natl. Acad. Sci.* **91:** 10645–10649.

McMaster, G.K. and G.G. Carmichael. 1977. Analysis of single-stranded and double-stranded nucleic acids on polyacrylamide and agarose gels by using glyoxal and acridine orange. *Proc. Natl. Acad. Sci.* **74:** 4835–4838.

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. 1990. Construction of a uniform-abundance (normalized) cDNA library. *Proc Natl. Acad. Sci.* **88:** 1943–1947.

Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc Natl. Acad. Sci.* **85:** 2444–2448.

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

Takeda, J., H. Yano, S. Eng, Y. Zeng, and G.I. Bell. 1993. A molecular inventory of human pancreatic islets: Sequence analysis of 1000 cDNA clones. *Hum. Mol. Genet.* **2:** 1793–1798.

Thomas, P.S. 1980. Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. *Proc. Natl. Acad. Sci.* **77:** 5201–5205.

Wan Kim C., P. Markiewicz, J.J. Lee, C.F. Schierle, and J.H. Miller. 1993. Studies of hyperthermophile *Termotoga maritima* by random sequencing of cDNA and genomic libraries. *J. Mol. Biol.* **231:** 960–981.

Waterston, R., C. Martin, M. Craxton, C. Huynh, A. Coulson, L. Hillier, R.K. Durban, P. Green, R. Shownkeen, N. Halloran, T. Hawkins, R. Wilson, M. Berks, Z. Du, K. Thomas, and J. Thierry-Mieg. 1992. A survey of expressed genes in *Caenorhabditis elegans*. *Nature Genet.* **1:** 114–123.

Wilcox, A.S., A.S. Khan, J.A. Hopkins, and J.M. Sikela. 1991. The use of 3′ untranslated sequences of human cDNA for rapid chromosome assignment and conversion to STSs: Implication for an expression map of the genome. *Nucleic Acids Res.* **19:** 1837–1843.

Received July 26, 1995; accepted in revised form December 14, 1995.