RESEARCH

# Novel Gene Transcripts Preferentially Expressed in Human Muscles Revealed by Quantitative Hybridization of a High Density cDNA Array

Geneviève Piétu,[1-3] Olivier Alibert,[2] Valérie Guichard,[1,2] Bernard Lamy,[2] Florence Bois,[2] Elisabeth Leroy,[1] Régine Mariage-Samson,[1] Rémi Houlgatte,[1] Pascal Soularue,[2] and Charles Auffray[1,2]

[1]Genexpress, Centre National de la Recherche Scientifique UPR 420, BP 8, 94801 Villejuif and [2]Généthon, BP60, 91002 Evry, France

A set of 1091 human skeletal muscle cDNA clone inserts representing more than 800 human gene transcripts were spotted as PCR products at high density on nylon membranes. Replicas of the filters were hybridized in stringent conditions with $^{33}$P-radiolabeled cDNA probes transcribed from skeletal muscle poly(A)$^+$ RNA. Hybridization signals were collected on phosphor screens and processed using a software specifically adapted for this application to identify and quantitate each spot. Parameters likely to influence the hybridization signal intensity were assessed to eliminate artifacts. Each clone was assigned to one of four intensity classes reflecting the steady-state level of transcription of the corresponding gene in skeletal muscle. Differential expression of specific gene transcripts was detected using complex cDNA probes derived from nine different tissues, allowing assessment of their tissue specificity. This made it possible to identify 48 novel gene transcripts (including 7 homologous or related to known sequences) with a muscle-restricted pattern of expression. These results were validated through the analysis of known muscle-specific transcripts and by Northern analysis of a subset of the novel gene transcripts. All these genes have been registered in the Genexpress Index, such that sequence, map, and expression data can be used to decipher their role in the physiology and pathology of human muscles.

The ultimate goal of the human genome project is an integrated physical, genetic, and genic map that combines information on the sequence, the chromosomal position, and the expression pattern of each of the ~60 to 80,000 genes expressed in humans (Antequera and Bird 1994; Fields et al. 1994). Systematic analysis of gene transcript repertoires has been conducted by a number of laboratories that have isolated and partially sequenced cDNA clones from libraries derived from a variety of specific human tissues (for review, see Houlgatte et al. 1995). This approach has proven to be a rapid and efficient method to identify human gene transcripts on a large scale, an important proportion of which represent novel genes showing limited or no similarity to those previously known in humans or other species.

Therefore, it is expected that the combined efforts of many laboratories will provide rapidly an almost complete catalog of human gene transcripts. To decipher the biological function of the novel genes, it may be useful to know in which tissues, cell types, and circumstances they are transcriptionally active, and how their expression changes as a result of developmental processes, cellular differentiation, or disease. In this respect, collecting expression profiles from a large collection of genes, including those that are expressed in a restricted tissue-specific manner or ubiquitously, may be more informative than analyses limited to specific selected cases (Okubo et al. 1992, 1994; Matsubara and Okubo 1993; Adams et al. 1995; Lee et al. 1995).

[3]Corresponding author.
E-MAIL pietu@infobiogen.fr; FAX 33.1.49583509.

A variety of methods are available for measuring the steady-state transcriptional level of individual genes such as Northern blot, in situ hybridization, and RT–PCR.

An alternative approach that can be applied to entire collections of cDNA clones is the collection of hybridization signatures from a larger number of gene transcripts by analysis of arrayed cDNA clones, dotted on a membrane at high density by a robotic device, followed by hybridization with complex cDNA probes obtained by reverse transcription of mRNA from different tissues (Friemert et al. 1989; Lennon and Lehrach 1991; Gress et al. 1992; Meiert-Ewert et al. 1993; Auffray et al. 1995; Nguyen et al. 1995; Takahashi et al. 1995; Zhao et al. 1995). As the intensity of the hybridization signal can be demonstrated to be, in appropriate conditions, proportional to the relative abundance of the complementary species in the probes, it is then possible to obtain qualitative and quantitative data, the value of the intensity of hybridization reflecting the steady-state level of transcription of the corresponding genes.

As part of our integrated molecular analysis of the human genome and its expression (IMAGE; Auffray et al. 1995), we have collected 4648 sequence signatures derived from both ends of skeletal muscle cDNA clones. Comprehensive annotations and classification of data base similarities allowed to cluster together overlapping sequences derived from the same transcript into sequence families, and the development of expressed sequence-tagged site (eSTS) markers allowed to define the chromosomal localization of most of the corresponding genes using panels of human–rodent somatic cell hybrids (Auffray et al. 1995; Houlgatte et al. 1995).

To characterize the expression profiles of these gene transcripts, we have conducted a study based on the hybridization of a subset of the initial collection of skeletal muscle cDNA clones arrayed on high density filters, with radio-labeled complex probes obtained by reverse transcription of mRNA from various tissues into cDNA. The extent of hybridization was estimated by quantitation of the signal using phosphor screens, and a detailed analysis of the parameters influencing the specificity of hybridization was conducted. Application of the technology to study differential expression identified novel genes with a muscle-restricted expression profile and the results were confirmed by Northern blot analysis.

## RESULTS

### Analysis of the Hybridization Signals

All the dots were positive when hybridization of the filter was performed with the vector probe (universal primer) present in the PCR product (Fig. 1A). Quantitation of the hybridization signal intensity allowed estimation of the amount of DNA accessible to the probe contained in each individual dot on the filter. Hybridization of the filter with the muscle cDNA probe produces signals of variable intensity (Fig. 1B).

We have used a statistical approach allowing to distinguish between positive and negative hybridization signals in each experiment. The population of dots with the lowest signal intensity followed a Normal distribution (0,1) and was considered to represent the clones that do not hybridize significantly above background (Fig. 2A). The dots with a hybridization signal intensity (rIm) that significantly deviates from this distribution ($\alpha = 5 \times 10^{-2}$, with a value of
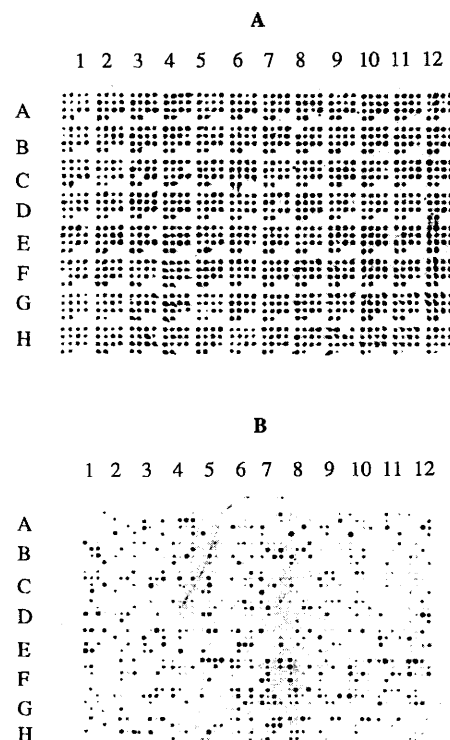


**Figure 1** Hybridization signals of 1339 human skeletal muscle cDNA clone inserts spotted on a high density filter. (A) Hybridization with $\alpha$-$^{33}$P-labeled oligonucleotide corresponding to the vector probe. (B) Hybridization with the $^{33}$P-labeled muscle cDNA probe. Exposure time was overnight using phosphor screens.
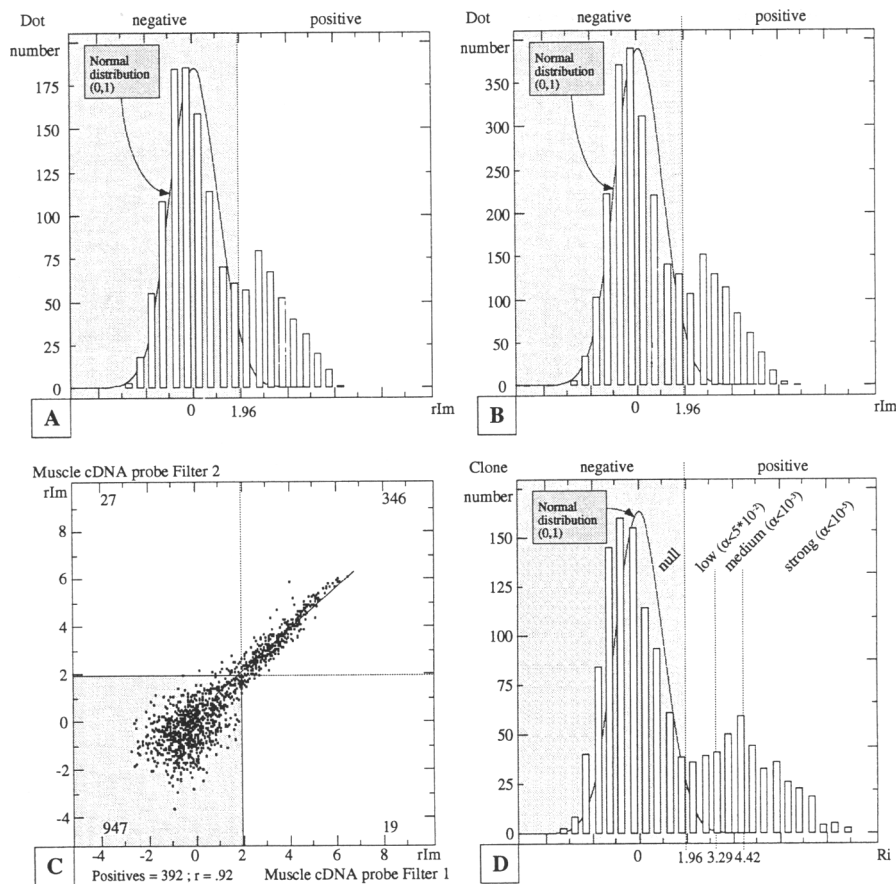
PIÉTU ET AL.



**Figure 2** Distribution of the signal intensities after hybridization with the muscle cDNA probe. Gray areas correspond to the "negative" populations. (A) Hybridization signal intensity (rIm) distribution of cDNA dots present on the filter presented in Fig. 1B. The number of dots in each interval is represented as a histogram. The population of lowest intensities is compared to the theoretical Normal distribution (0,1). An rIm value of 1.96 ($\alpha = 5 \times 10^{-2}$) is used as a threshold to separate the negative and positive populations as described in the Results section. (B) Hybridization signal intensity (rIm) distribution of cDNA dots present on two filters hybridized with the same muscle cDNA probe. (C) Comparison of the signal intensity rIm of dots for each cDNA clone on two filters hybridized with the same muscle cDNA probe. Lines indicate the separation of positive and negative dots for each filter. The correlation factor (r) and the regression line were calculated for the clones with at least one positive signal (white area). The number of clones in each sector is indicated. (D) Registered hybridization signal intensity Ri based on the results shown in B. Values of Ri of 3.29 ($\alpha = 10^{-3}$) and 4.42 ($\alpha = 10^{-5}$) were used to statistically delineate the classes of signal intensity.

(Fig. 2B). A good correlation between the two hybridizations was observed (r = 0.92) (Fig. 2C).

The use of registered intensity (Ri) values as defined in Materials and Methods made it possible to obtain a better discrimination between the two populations when the results of the two hybridizations were taken into account (Fig. 2D). Clones with Ri > 1.96 were considered as positive. Three values of $\alpha$ ($5 \times 10^{-2}$, $5 \times 10^{-3}$, and $5 \times 10^{-5}$) were then used to define four classes of hybridization signal intensity and the corresponding clones were classified as null, low, medium, or strong.

Using filters hybridized with the same probe, experiments were compared two by two to test the reproducibility of the method. For each clone in the different hybridizations, the coefficient of variation (cv) was estimated to measure the variation of the signal intensity. For all the positive clones of a filter, using the average of the cv, the global variation of intensity (CV) could then be measured. It ranged from 20% to 35% using the muscle cDNA probe and 9% to 11.5% for the vector probe.

rIm > 1.96) were considered as "positive" (Fig. 2A).

Comparison of the signal intensity for filters hybridized in duplicate with the same probe is reported in Figure 2, B and C. The fitting between the population of lowest intensity signal and the theoritical Normal distribution (0,1) is better when the results from two filters are summed up

## Analysis of Parameters Influencing the Hybridization Signal Intensity

We performed a series of experiments to analyze parameters that could influence the Ri, such as the length of the insert, the length of poly(A) stretches, the method of priming, and the presence of repetitive sequences.

The effect of the length of the inserts on the

intensity of the hybridization signal was assessed. There is no global correlation (r = 0.07) between the intensity of the hybridization signal obtained with the muscle cDNA probe and the length of the cDNA clone inserts (Fig. 3A). However, analysis of five clones corresponding to the titin gene transcript with insert length ranging from 0.3 to 1.9 kb demonstrated a twofold increase of the rIm (data not shown).

The effect of the length of the poly(A) tail in the cDNA clones was tested to determine whether it could interfere with the specificity of the hybridization signal, as hybridization could occur with the poly(T) in the first strand cDNA probe. In the hybridization conditions described in Materials and Methods, however, no correlation (r = 0.01) between the length of the poly(A) stretches and the signal intensity of each clone was observed when the filters were hybridized with the muscle cDNA probe (Fig. 3B). On the contrary, a good correlation was observed (r = 0.80) when the filters were hybridized with an oligo(dT)(37) probe (data not shown). In addition, we observed that there is no correlation between the clones hybridizing with an oligo(dT)(37) probe or with the muscle cDNA probe, and distinct populations of clones are clearly distinguished (Fig. 3C). Finally, competition during the synthesis of the first strand of the muscle cDNA probe with long poly(A) homopolymers [poly(A) 80] did not change the Ri (Fig. 3D), and the correlation of the signal intensity obtained with the two probes in the presence or in the absence of competitor was excellent (r = 0.95).

Two methods of priming for the synthesis of the probe were compared. Priming was usually performed in the presence of random hexamers.

An alternative approach is the synthesis of the first strand cDNA using oligo(dT) primers anchored at the boundary between the poly(A) tail and the 3' untranslated region of the mRNAs by adding an A, G, or C to the 3' end of the primer. A correlation (r = 0.43) was observed between the Ri obtained using the two methods of priming (Fig. 3E). cDNA clones that hybridized with the anchored oligo(dT) primed probe and not with the random primed probe corresponded to clones with long poly(A) stretches according to the sequencing data collected from the 3' end of the cDNA clones.

To detect clones harboring repetitive sequences, hybridizations of the filters were performed with oligonucleotides corresponding to Alu and Line sequences or with total genomic DNA. When the population of clones hybridizing with Alu and Line oligonucleotides on one hand and with the muscle cDNA probe on the other hand were compared, they overlapped by only 15%, corresponding to the proportion of mRNA harboring repetitive sequences (Fig. 3F). Similar results were obtained when hybridizations with the muscle cDNA and genomic DNA probes were compared (data not shown).

## Hybridization Signatures Obtained With the Muscle cDNA Probe

Quantitation of the hybridization signal intensity allowed to assign each clone to a class of intensity. The distribution of the cDNA clones in each group of signal intensity was analyzed and correlated with the analysis of their sequence signatures (Table 1). Among the 1091 cDNA clones, 80% were not detected with the muscle cDNA

**Table 1. Global Correlation Between Sequence and Hybridization Signatures**

| Intensity class | Known sequences | Novel sequences | | Total |
| | | homologous or related | unrelated | |
| --- | --- | --- | --- | --- |
| Null | 109 | 137 | 627 | 873 (80.0%) |
| Low | 15 | 13 | 56 | 84 (7.7%) |
| Medium | 15 | 9 | 38 | 62 (5.7%) |
| Strong | 8 | 8 | 56 | 72 (6.6%) |
| Total | 147 (13.5%) | 167 (15.3%) | 777 (71.2%) | 1091 (100%) |

Sequence similarity classification is based on the Genexpress Index (Houlgatte et al. 1995). Classes of intensity were defined as described in Results and in the legend to Fig. 2 (D).
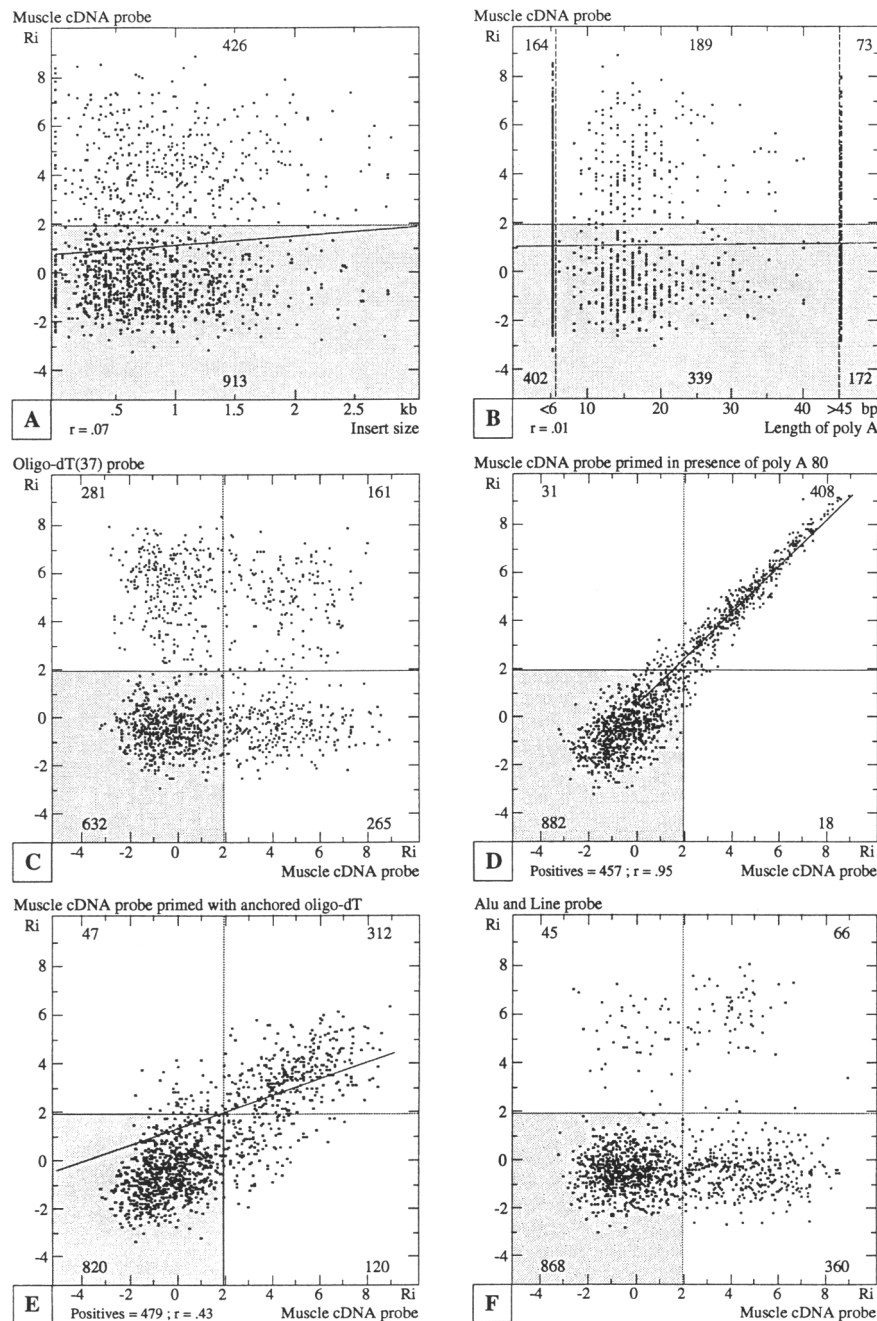
**Figure 3** Parameters influencing the hybridization signal. The Ri values presented are based on hybridization of two membranes for each experiment. For details of the presentation, see legend to Fig. 2. (*A*) Effect of the length of the insert (in kb). (*B*) Effect of the length of poly(A) stretches (in bp). Clones with poly(A) stretches smaller than 6 bp or longer than 45 bp are grouped. (*C*) Comparison of hybridizations performed with the muscle cDNA probe or with oligo(dT)(37). (*D*) Comparison of hybridizations performed with the muscle cDNA probe primed in the presence or absence of poly(A) (80). (*E*) Comparison of hybridizations performed with muscle cDNA probes primed using random hexamers or one base-anchored oligo(dT). (*F*) Comparison of hybridizations with the muscle cDNA probe and oligonucleotide probes corresponding to Alu and Line repetitive sequences.

probe (null) and the remaining 20% could be classified in low (7.7%), medium (5.7%), and strong (6.6%) signal intensity classes (Table 1). As expected, clones corresponding to known genes and with strong signal intensity represented members of gene families previously reported to be highly expressed in muscle, including muscle glycogen phosphorylase, adult skeletal muscle $\alpha$-actin, titin, glyceraldehyde-3-phosphate dehydrogenase, creatine kinase, calpain small subunit, and nebulin. Conversely, none of the negative clones corresponded to such transcripts. Among the cDNA clones with novel related or unrelated sequences, 8 and 56, respectively, produced a strong hybridization signal suggesting that they correspond to novel transcripts highly expressed in muscle.

## Hybridization Signatures Obtained With Complex cDNA Probes Derived From Various Tissues

Four filters were hybridized with cDNA probes derived from skeletal muscle, fetal heart, heart, brain, liver, testis, placenta, uterus, and thymus mRNA. Differences in hybridization signal intensity for specific clones with different probes are anticipated to reflect differences in the steady-state level of transcription and thus the tissue specificity of the transcript.

For each clone, values of the hybridization signal intensity Ri were compared two by two for all the probes tested.

Ninety-eight cDNA

clones that hybridized very significantly above background with the skeletal and heart cDNA probes, (Ri > 3.29, $\alpha = 10^{-3}$) and not with all the other probes (Ri < 1.96, $\alpha = 5 \times 10^{-2}$) were considered to have a muscle-restricted expression profile. These 98 clones correspond to 62 gene transcripts (Table 2), of which 14 correspond to known transcripts encoding well-characterized muscle-specific proteins that are major products of muscle cells, 7 correspond to novel transcripts homologous or related to known sequences, and 41 to novel transcripts with no data base similarity. In addition, 12 cDNA clones hybridized only with the placenta cDNA probe and 51 clones hybridized with 3–7 probes. Finally, 31 clones that hybridized with 8 or 9 of the cDNA probes tested were considered to have an ubiquitous pattern of hybridization.

To confirm and further investigate the specificity of transcription of the genes corresponding to some of these clones, a series of Northern blots containing a panel of RNA from various human tissues were probed with a set of cDNA clones. For this purpose, 21 probes corresponding to known or novel gene transcripts were chosen at random in the categories with muscle-restricted (14) or ubiquitous patterns of hybridization (7). In the first group, Northern analysis confirmed a muscle-restricted pattern for 13 clones, whereas 1 clone was found to be expressed ubiquitously (Fig. 4). For each of the seven clones of the second group, an ubiquitous pattern of transcription was confirmed (Fig. 4).

## DISCUSSION

We have used cDNA clone inserts arrayed on high density filters to monitor the expression profile of the corresponding gene transcripts by quantitative hybridization with a series of cDNA probes transcribed from mRNA of various human tissues.

With the goal in mind of identifying with the highest degree of confidence the gene transcripts that are expressed preferentially in muscle when compared to other tissues, these experiments were conducted using stringent conditions of hybridization and strict criteria for data analysis.

In the initial phase of these experiments, using more relaxed conditions of hybridization and washing, we observed a high proportion of the clones hybridizing with the skeletal muscle cDNA probe (Auffray et al. 1995). The patterns observed were similar with cDNA probes derived from various tissues, suggesting that spurious signals due to the presence of relatively long poly(A/T) stretches known to be present in the cDNA clones obscured the specific signals. This was confirmed by hydridization with oligo(dT) probes, and led us to implement the conditions of high stringency during hybridization and washing that were used throughout the present study. Similar observations made in parallel studies (Nguyen et al. 1995) led to the conclusion that contamination of the mRNA preparation used to prepare the cDNA probe by oligo(dT) can also be a source of spurious hybridization signals, and the demonstration that this can be overcome by annealing to poly(A)(80). Such contamination was not observed in the mRNA preparations we used.
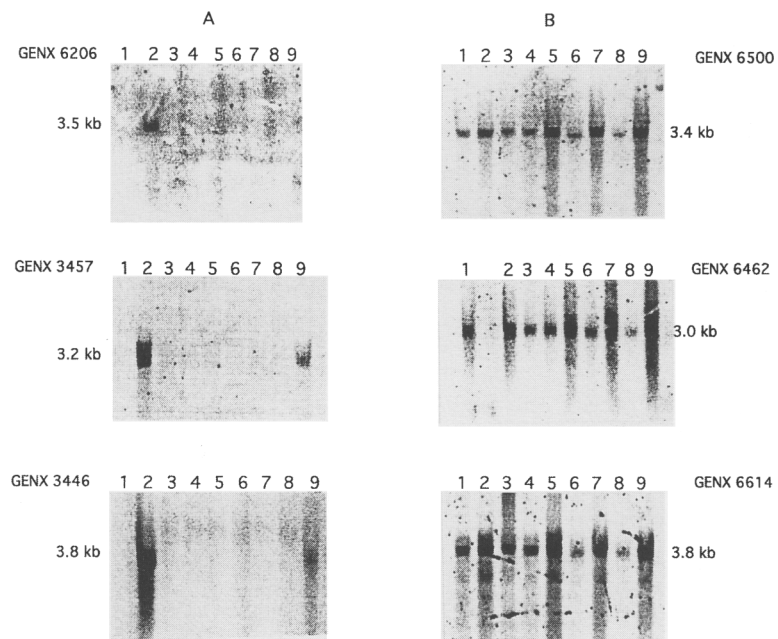


**Figure 4** Northern blot analysis of total RNA (10 μg) from various human tissues probed with cDNA inserts of clones with muscle-restricted (A) or ubiquitous (B) patterns of hybridization. (A) Clones b-17a03 (GENX-6206), b-24b07 (GENX-3457), and b-18e04 (GENX-3446); (B) clones b-97c03 (GENX-6500), b-90g09 (GENX-6462), and b-c0h09 (GENX-6614). The GENX number is from the Genexpress Index (Houlgatte et al. 1995). (Lane 1) Liver; (lane 2) muscle; (lane 3) lung; (lane 4) placenta; (lane 5) kidney; (lane 6) spleen; (lane 7) brain; (lane 8) thymus; (lane 9) heart. Size of the transcript detected with each probe is indicated in kilobases.

PIÉTU ET AL.

**Table 2. Gene Transcripts with a Muscle-restricted Pattern of Hybridization**

| GENX | Frequency in muscle (%) | Species | Gene product | Tissues | Genexpress eSTS | GDB map location |
|---|---|---|---|---|---|---|
| **Known gene transcripts** | | | | | | |
| 2965 | 5.10 | human | adult skeletal muscle α-actin | FB | | 1q42.1–q42.3 |
| 2995 | 0.04 | human | HIN-2* | | | |
| 3195 | 3.38 | human | creatine kinase M | FB | 19 | 19q13.3 |
| 3237 | 1.19 | human | muscle glycogen phosphorylase | | | 11q13.1 |
| 3338 | 2.20 | human | titin | HE, HA | 2 | 2q31–q32 |
| 3560 | 0.66 | human | desmin | | 2 | 2q35 |
| 3565 | 0.09 | human | skeletal muscle 190-kD titin-associated protein | BR | 18 | |
| 3614 | 1.85 | human | myoglobin | | 22 | 22q12 |
| 3906 | 0.31 | human | β-tropomyosin | | 9 | |
| 4532 | 1.05 | human | nebulin | | 2 | 2q31–q32 |
| 4549 | 0.04 | human | MN1* | FB | 22 | 22q11 |
| 4705 | 0.04 | human | L40401 mRNA* | FLS | | 14q24.3 |
| 6274 | 0.04 | human | antisecretory factor-1* | | | |
| 6530 | 0.04 | human | sialyltransferase* | | | |
| **Homologous gene transcripts** | | | | | | |
| 3570 | 0.09 | rat | fast skeletal muscle troponin T | | | |
| 4600 | 0.13 | chicken | cofilin | HG, HA | 14 | |
| 6277 | 0.04 | bovine | zeta-Cop* | | | |
| **Related gene transcripts** | | | | | | |
| 3122 | 0.04 | human | cystein-rich protein | BR | X, 3 | |
| 3463 | 0.09 | human | heat shock protein HSP27 | CC | | |
| 4507 | 0.57 | Drosophila | ring canal protein KELCH | | 2 | |
| 5260 | 0.04 | mouse | MAP4* | | 3 | |
| **Unrelated gene transcripts** | | | | | | |
| 3414 | 0.09 | | | | 4 | |
| 3432 | 0.13 | | | | 10 | |
| 3446 | 0.09 | | | HE | 5 | |
| 3457 | 0.09 | | | | 19 | |
| 3471 | 0.09 | | | | 1 | |
| 3531 | 0.09 | | | | 4 | |
| 3537 | 0.04 | | | | | |
| 3554 | 0.09 | | | | 1 | |
| 3587 | 0.09 | | | | 19 | |
| 3594 | 0.09 | | | HL, FB, HA | 13 | |
| 3600 | 0.09 | | | | 13 | |
| 3649 | 0.04 | | | | | |
| 3733 | 0.26 | | | | | |
| 4491 | 0.04 | | | | | |
| 4534 | 0.04 | | | | | |

**Table 2.** *(Continued)*

| GENX | Frequency in muscle (%) | Species | Gene product | Tissues | Genexpress eSTS | GDB map location |
|------|------|------|------|------|------|------|
| 4535 | 0.04 | | | | 12 | |
| 4568 | 0.04 | | | | | |
| 4637 | 0.04 | | | | 3 | |
| 4656 | 0.04 | | | | 6 | |
| 4657 | 0.04 | | | | | |
| 4667 | 0.04 | | | | 11 | |
| 4679 | 0.04 | | | | | |
| 4687 | 0.04 | | | | 18 | |
| 4745 | 0.04 | | | | 14 | |
| 4760 | 0.04 | | | | | |
| 4789 | 0.04 | | | | 7 | |
| 4794 | 0.04 | | | FB | 8 | |
| 5434 | 0.04 | | | | 20 | |
| 5809 | 0.09 | | | | 2 | |
| 6080 | 0.04 | | | | 5 | |
| 6085 | 0.04 | | | | 7 | |
| 6162 | 0.04 | | | | 17 | |
| 6163 | 0.04 | | | | 22 | |
| 6166 | 0.04 | | | | 19 | |
| 6206 | 0.04 | | | | 1 | |
| 6268 | 0.04 | | | | | |
| 6269 | 0.04 | | | | | |
| 6278 | 0.04 | | | | 6 | |
| 6476 | 0.04 | | | | 5 | |
| 6488 | 0.04 | | | | | |
| 6904 | 0.18 | | | HA | 4 | |

Information available for each clone was extracted from the Genexpress Index (Houlgatte et al. 1995): GENX number, clone frequency in the muscle library, species for the homologous sequence, gene product name, tissues in which the same gene transcripts have been registered by partial cDNA sequencing [(BR) brain; (CC) colorectal cancer; (FB) fetal brain; (FLS) fetal liver and spleen; (HA) heart atrium; (HE) heart; (HG) HepG2 liver cell line; (HL) HL60 cell line; (HI) hippocampus; (PL) placenta], Genexpress eSTS marker chromosomal assignment, and GDB map location. (*) Sequence similarities have been updated with GenBank release 92.0.

As the result of using highly stringent conditions, the number of clones hybridizing back to the skeletal muscle cDNA probe was limited to 20% of the entire array. Two factors in the design of the experiments contributed to this low figure. One is the use of $^{33}$P-labeled probes that provide signals of reduced intensity when compared to those obtained with $^{32}$P-labeled probes. However, the spots obtained have a better spatial resolution, facilitating their unambiguous identification and the future implementation of higher density arrays. The second factor is the low level of redundancy of the array in which the number of clones representing the same gene transcript was limited, based on the results of previous DNA sequencing. Therefore, it was expected that the clones corresponding to the abundant or moderately abundant transcripts, whose number was diminished when compared to the initial library, would be those hybridizing preferentially to the skeletal muscle cDNA probe, and this was indeed the case. As a large number of the clones detected were present only once in the initial collection of 3596 clones, it appears that the conditions used provided a level of sensitivity that is within the range of 0.01%–0.1% as reported in similar studies (Gress et al. 1992; Nguyen et al. 1995; Zhao et al. 1995).

We further took advantage of the population of clones that do not appear to hybridize to the

skeletal muscle cDNA probe to assess statistically parameters that may affect the significance of the hybridization signals recorded when using different filters, different probes, or at different times. As this population appeared to follow a normal distribution, it was possible to characterize the positive clones by their low probability of belonging to this distribution.

We then used this statistical approach to test several parameters that could further affect the specificity of hybridization and hinder their proper interpretation. As the array has been previously characterized extensively, it was possible to demonstrate that the length of the insert (on average 0.8 kb) and of the poly(A/T) tract were not influencing significantly hybridization with the skeletal muscle cDNA probe. Furthermore, our observations validated random priming as the most reliable method to generate the probe. The value of this approach is further demonstrated by the fact that all the clones corresponding to genes known to be highly expressed were among those classified with the highest degree of confidence, and none belonged to the negative population.

Having established that the method used provides an adequate level of specificity and sensitivity, we used it to monitor the differences in transcription levels of the corresponding genes in various human tissues. To identify the clones corresponding to human gene trancripts that are expressed preferentially in muscle tissues, we used a high threshold of significance to delineate the positive signals with the skeletal and cardiac muscle cDNA probes, combined with signals in the negative population with all the other probes. This is likely to result in a limited number of false positives to the cost of ignoring the clones that correspond to the rarer transcripts and therefore are intermingled in the negative population. Relaxing the parameters used and further statistical analysis should help rescue some of these false-negative clones.

The validity of the selection process was established independently in two ways. First, clones corresponding to known transcripts were those encoding proteins reported previously to be expressed specifically in muscle. Electronic profiles based on systematic cDNA sequencing indicate that such transcripts are also detected in other tissues, such as brain, albeit infrequently (Houlgatte et al. 1995; Table 2). Second, the results of Northern blot analyses were in agreement with the hybridization results except in one case.

RT–PCR experiments using the eSTS markers for the corresponding genes (Auffray et al. 1995) should help in further assessing the specificity and sensitivity of the two methods.

Altogether, the results obtained indicate that we have identified 48 novel gene transcripts (including seven homologous or related to known sequences) preferentially, if not specifically, expressed in skeletal and cardiac muscles. They are likely to be involved in important aspects of the physiology of these tissues and the pathologies that affect them. Although we have focused our attention on transcripts that are expressed preferentially in muscles, we have also collected a wealth of information that provides hints on genes that are expressed ubiquitously and could be involved in housekeeping functions, as well as genes that have a high level of transcription in tissues other than muscles, such as placenta. For all the genes included in this study, further investigation of the biological functions they control will be facilitated by the fact that all the clones in the array have been characterized extensively in terms of sequence and map location as registered in the Genexpress Index (Houlgatte et al. 1995). In other similar studies (Nguyen et al. 1995; Takahashi et al. 1995; Zhao et al. 1995) as the DNA sequence is determined only for those clones that are selected on the basis of their hybridization profile, a large fraction of the information collected is not readily available for integration with other data.

We have started to extend this approach to study differential gene expression in other tissues by applying the same methodology to a normalized infant brain cDNA array (Soares et al. 1994) that has been characterized and sequenced by us and our colleagues within the frame of the IMAGE Consortium (Auffray et al. 1995; Houlgatte et al. 1995). The avaibility of other similar libraries in which the majority of all human gene transcripts are represented, and the accelerating speed at which the corresponding genes are being placed on the genome maps with ever increasing resolution, will provide the resource with which to monitor gene expression profiles in a broad range of spatiotemporal, physiological, and pathological conditions and across species.

Currently, we are investigating further developments and refinements of the methodology that should help in its widespread use, including the use of microarrays and dual fluorescence hybridization that holds the promise of increased speed and sensitivity (Schena et al. 1995).

## METHODS

### cDNA Clone Selection and Amplification

From the 4648 sequences derived from both ends of 3596 skeletal muscle cDNA clones, 2904 sequences appeared unique after primary redundancy analysis (Auffray et al. 1995). A subset of 1339 clones was rearrayed in 96-well microtiter plates and used for filter construction. Two hundred forty-eight clones were not taken into consideration in the present study because they were demonstrated subsequently to be chimeric (112), failed to pass resequencing quality controls to confirm their identity (116), or correspond to mitochondrial transcripts (20). The selection process reduced the proportion of clones with sequences already known in humans from 26% to 13.5%, retained clones with novel sequences homologous or related to known sequences in various species (15.3%) so that the set was composed mainly (71.2%) of clones with novel sequences unrelated to sequences registered previously in data bases (also see Table 1). These 1091 selected clones correspond to 910 sequence families registered in the Genexpress Index after detailed sequence clustering and data base analyses (Houlgatte et al. 1995). Of these, 832 are classified as novel gene transcripts and 78 correspond to previously known genes. Redundancy is limited so that 807 gene transcripts are represented by a single clone and 74, 16, 8, 2, and 1 are represented by 2, 3, 4, 5, or 6 clones, respectively. The transcripts encoding titin and nebulin are represented by 30 and 10 clones, respectively.

The cDNA clone inserts were amplified using PCR after direct inoculation of 0.1 μl of culture medium in 50 μl of PCR mixture containing PCR buffer [50 mM KCl, 10 mM Tris-HCl (pH 8), 1.5 mM MgCl$_2$, 0.01% (wt/vol) gelatin], 0.2 mM dNTP, 1 unit of AmpliTaq polymerase (Perkin-Elmer) and 5 pmoles of each primer. Vector primers 5'-GCGGATAACAATTTCACACAGG-3' and 5'-GGTTT-TCCCGTACACGACGTTG-3'(Genset, Paris, France) flanking the cloning site of the pBluescript SK + plasmid were used to amplify the cDNA inserts in 96-well microtiter plates. The process was started by a hot-start of 5 min at 92°C, followed by 35 cycles including denaturation at 92°C for 1 min, annealing to primers for 1 min at 50°C and elongation at 72°C for 1 min using a Techne PHC-3 thermal cycler. The presence of the amplified product was assessed by standard 0.8% agarose gel electrophoresis followed by ethidium bromide staining, allowing estimation of its concentration and of the insert size.

### High Density Filter Construction

Fifty to 100 ng of PCR products from cDNA inserts in 96-well microtiter plates prepared as described above were dotted using the "Saturnin" robot of Généthon onto nylon filters of 8 × 12 cm (Hybond N+, Amersham Corporation, Amersham, UK), at a density of 16 microtiter plates arrayed in a 4 × 4 format. DNA was cross-linked to the membrane by ultraviolet radiation (120 mJ/cm$^2$) using the Stratagene UV-Stratalinker 2400 (Stratagene, LaJolla, CA).

Control clones were spotted at specific positions on the filter: 5 cDNA clones corresponding to the human titin gene transcript containing inserts of various size (0.3 to 1.9 kb). Membranes were prepared in batches and stored at 4°C before use.

### Probe Preparation and Labeling

An oligonucleotide corresponding to the universal primer (vector probe) was labeled with 50 μCi of [γ-$^{33}$P]ATP (Amersham) (3000 Ci/mmole) with T4 polynucleotide kinase (10 units) (Boehringer Mannheim, Mannheim, Germany) for 30 min at 37°C. Oligonucleotides corresponding to Alu and Line sequences (PdJ34, 517, 559, and L1HS) were from ATCC (Rockville, MD).

Single strand cDNA probes were derived from human skeletal muscle of thigh poly(A)$^+$ RNA of a 45-year-old male or poly(A)$^+$ RNA from human heart, liver, placenta, testis, uterus, brain, thymus, or fetal heart (Clontech). One microgram of poly(A)$^+$ mRNA was reverse transcribed using Superscript II reverse transcriptase (RNase H-) (GIBCO-BRL, Gaitherburg) as described in the manufacturer's protocol, using random hexamers for priming. The poly(A)$^+$ RNA preparations were demonstrated to be free of contaminating genomic DNA because primers specific for the GAPDH gene yielded a PCR product only after reverse transcription. Labeling was performed during reverse transcription by incorporating 50 μCi of [α-$^{33}$P]dATP (Amersham; 3000 Ci/mmole) in the presence of 2.4 μM cold dATP. Unincorporated radioactive nucleotides were separated from the labeled probe by filtration on a Nuc Trap Push column (Stratagene) according to the instruction manual. Approximatively 60% of the radioactive label was incorporated into cDNA and the specific activity of the probes ranged from 10 to 30 × 10$^6$ cpm/μg. The length of the single strand cDNA probe ranged from 0.5 to 3.5 kb as assessed by 1% alkaline agarose gel electrophoresis and autoradiography. The probes were stored at −20°C before use.

### Hybridization

Filters were incubated within glass tubes in an hybridization oven (Appligène, Strasbourg, France) in a volume of 15 ml. Membranes in duplicate were prehybridized for 5 hr at 42°C in a 15-ml solution containing final concentrations of 4× SSC (1× SSC = 150 mM NaCl, 15 mM sodium citrate), 50% formamide, 10× Denhart's, 0.1% sodium dodecyl sulfate (SDS), 8% dextran sulfate, 50 mM phosphate buffer (pH 7.2), 1 mM EDTA. Hybridization of the replicate set of filters was performed overnight at 42°C in the same solution with 15 to 20 × 10$^6$ cpm of $^{33}$P-radiolabeled probes in the presence of 100 μg/ml of denatured herring sperm DNA.

The membranes were washed twice for 10 min in 2× SSC/0.1% SDS followed by washing once for 15 min in 1× SSC/0.1% SDS and twice for 15 min in 0.1× SSC/0.1% SDS. All washes were carried out at 65°C. Exposure to phosphor screens was for 1 to 3 days.

Stripping of hybridized membranes was performed by two successive immersions in a solution of 0.4 M NaOH, 0.1% SDS at 65°C for 30 min. Membranes were rinsed in 0.2 M Tris-HCl (pH 8.0), 1× SSC/0.1% SDS for 10 min at room temperature. Membranes were used a maximum of five to seven times.

Hybridization of the membranes with the $^{33}$P-radiolabeled oligonucleotide probe was performed in 7% SDS, 0.5 M phosphate buffer (pH 7.2), 1 mM EDTA (Church and Kieffer-Higgins 1988) for 15 hr at 50°C followed by washing in 2× SSC for 15 min at 50°C followed by 15 min at room temperature and a final wash in 1× SSC/0.1% SDS for 15 min at 37°C.

## Hybridization Signal Analysis

Filters were scanned on the PhosphorImager imaging plate system (Molecular Dynamics, Sunnyvale, CA) for quantitative analysis of signal intensities. After image acquisition, the scanned 16-bit images were imported on a Sun workstation and image analysis was performed using the XdotsReader software (Cose, Le Bourget, France).

The software processes the results of an exposure into images of individual filters. Then it translates the hybridization signal coordinates into dot localization on the filter using a reference grid for the arrangement of the dots. It takes into account slight variations in dot position attributable to filter deformation by assigning the signal detected to the nearest position expected. The software quantifies each dot individually after local background subtraction. These tasks, image cutting, dot identification, and dot quantitation are processed sequentially and automatically. The results are validated interactively and a table is generated that contains for each dot its reference number and the experimental values.

Different types of values are obtained for the quantitation of the dot intensity: the radius of the dot, the mean of the dot pixel intensities for one dot, the maximal intensity of the pixels of the dot, the sum of the pixel intensities of the dot, the average of the pixel intensities of the dot weighted by the distance to the center of the dot. We have chosen to use the mean of the pixel intensities for each dot or intensity mean, noted as Im. To take into account experimental variations in specific activity of the cDNA probe preparations or exposure time that might alter the signal intensity, the data obtained from different hybridizations were normalized by dividing the Im for each dot by the average of the intensities of all the dots present on the filter to get a normalized Im value (nIm).

The values of nIm were distributed in two populations when recorded on a $\log_e$ scale. The population corresponding to the lowest intensities appeared to follow a Normal distribution, and was assimilated to the standard Normal distribution (0,1) after calculation of the reduced centered value rIm for each dot. Thus, when n identical filters were hybridized with the same probe, a registered intensity (Ri) value was calculated for each clone, as $Ri = mi \times \sqrt{n}$ where mi corresponds to the average of the value rIm obtained for all dots resulting from hybridization to the same clone. A probability $\alpha$ to belong to the Normal distribution (0,1) was then determined using the reference table of reduced centered values.

To estimate the variation of measured intensities for each clone, the cv was defined as $cv = s \times 100/m$ where s is the standard deviation and m the mean of the values obtained in several hybridizations.

Computing tools have been developed to obtain a work sheet allowing management of the data obtained from a large number of clones and hybridization experiments. Files are generated synthesizing results from several filters hybridized with the same probe or comparing hybridizations of identical filters with different probes.

## Northern Blot Analysis

Northern blots were prepared using the Labimap Multiblotter MB-96 system (Labimap, France). Ten micrograms of total RNA (Clontech) from various tissues were dena-tured, electrophoresed on 1% agarose gels, transferred onto an Hybond N+ membrane (Amersham), and hybridized according to the manufacturer's instructions. We also used Northern blots from Clontech containing 2 μg of mRNA. Hybridizations were performed following the manufacturer's instructions. Probes derived from the cDNA clones were prepared by PCR using the same primers as described above for the amplification of the clone inserts and labeled with [$\alpha$-$^{33}$P]dATP by random priming. After hybridization, the membranes were exposed to phosphor screens for 5 hr. The size of the transcripts detected was estimated using the GIBCO-BRL 0.24- to 9.5-kb RNA Ladder.

## REFERENCES

Adams, M.D., A.R. Kerlavage, R.D. Fleischmann, R.A. Fuldner, C.J. Bult, N.H. Lee, E.F. Kirkness, K.G. Weinstock, J.D. Gocayne, O. White, et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377:** 3–174.

Antequera, F. and A. Bird. 1994. Predicting the total number of human genes. *Nature Genet.* **8:** 114.

Auffray, C., G. Béhar, F. Bois, C. Bouchier, C. Da Silva, M.D. Devignes, S. Duprat, R. Houlgatte, M.N. Jumeau, B. Lamy, F. Lorenzo, H. Mitchel, R. Mariage-Samson, G. Piétu, Y. Pouliot, C. Sébastiani-Kabatchis, and A. Tessier. 1995. IMAGE: Integrated molecular analysis of the human genome and its expression. *C.R. Acad. Sci.* **318:** 263–272.

Church, G. and A. Kieffer-Higgins. 1988. Multiplex DNA sequencing. *Science* **240:** 185–188.

Fields, C., M.D. Adams, O. White, and J.C. Venter. 1994. How many genes in the human genome? *Nature Genet.* **7:** 345–346.

Friemert, C., V. Erfle, and G. Strauss. 1989. Preparation of radiolabeled cDNA probes with high specific activity for rapid screening of gene expression. *Methods Mol. Cell. Biol.* **1:** 143–153.

Gress, T.M., J.D. Hoheisel, G.G. Lennon, G. Zehetner, and H. Lehrach. 1992. Hybridization fingerprinting of

high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm. Genome* **3:** 609–619.

Houlgatte, R., R. Mariage-Samson, S. Duprat, A. Tessier, S. Bentolila, B. Lamy, and C. Auffray. 1995. The Genexpress Index: A resource for gene discovery and genic map of the human genome. *Genome Res.* **5:** 272–304.

Lee, N.H., K.G. Wienstock, E.F. Kirkness, J.A. Earle-Hughes, R.A. Fuldner, S. Marmaros, A. Glodek, J.D. Gocayene, M.D. Adams, A.R. Kerlavage, C.M. Fraser, and J.C. Venter. 1995. Comparative expressed-sequence-tag analysis of differential gene expression profiles in PC-12 cells before and after nerve growth factor treatment. *Proc. Natl. Acad. Sci.* **92:** 8303–8307.

Lennon, G.G. and H. Lehrach. 1991. Hybridization analyses of arrayed cDNA libraries. *Trends Genet.* **7:** 314–317.

Matsubara, K. and K. Okubo. 1993. cDNA analyses in the human genome project. *Gene* **135:** 265–274.

Meier-Ewert, S., E. Maier, A. Ahmadi, J. Curtis, and H. Lehrach. 1993. An automated approach to generating expressed sequence catalogues. *Nature* **361:** 375–376.

Nguyen, C., D. Rocha, S. Granjean, M. Baldit, K. Bernard, P. Naquet, and B. Jordan. 1995. Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. *Genomics* **29:** 207–216.

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

Okubo, K., J. Yoshii, H. Yokouchi, M. Kameyama, and K. Matsubara. 1994. An expression profile of active genes in human colonic mucosa. *DNA Res.* **1:** 37–45.

Schena, M., D. Shalon, R.W. Davis, and P.O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270:** 467–470.

Soares, M.B., M. de Fatima Bonaldo, P. Jelenc, L. Su, L. Lawton, and A. Efstratiadis. 1994. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci.* **91:** 9228–9232.

Takahashi, N., H. Hashida, N. Zhao, Y. Misumi, and Y. Sakaki. 1995. High-density cDNA filter analysis of the expression profiles of the genes preferentially expressed in human brain. *Gene* **164:** 219–227.

Zhao, N., H. Hashida, N. Takahashi, Y. Misumi, and Y. Sakaki. 1995. High density cDNA filter analysis: A novel approach for large scale quantitative analysis of gene expression. *Gene* **156:** 207–213.

# Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array.

G Piétu, O Alibert, V Guichard, et al.

| | |
|---|---|
| **References** | This article cites 19 articles, 5 of which can be accessed free at:<br>**http://genome.cshlp.org/content/6/6/492.full.html#ref-list-1** |
| **License** | |
| **Email Alerting Service** | Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.** |