BMC Genomics

BioMed Central

Research article

Open Access

Outliers involving the Poly(A) effect among highly-expressed genes in microarrays

Shujia J Pan*[1], David R Rigney[2] and John L Ivy[3]

Address: [1]Department of Kinesiology and Health Education, Bellmont Hall, Room 222, University of Texas at Austin, Austin TX 78712 U.S.A, [2]Department of Research and Development, GENETWORKS Inc., P.O. Box 33296, Austin TX 78764-0296 U.S.A and [3]Department of Kinesiology and Health Education, Bellmont Hall, Room 222, University of Texas at Austin, Austin TX 78712 U.S.A

Email: Shujia J Pan* - psj@mail.utexas.edu; David R Rigney - drigney@genetworks.com; John L Ivy - johnivy@mail.utexas.edu

* Corresponding author

Published: 12 December 2002

BMC Genomics 2002, **3**:35

Received: 23 July 2002

Accepted: 12 December 2002

This article is available from: http://www.biomedcentral.com/1471-2164/3/35

Abstract

Background: The Poly(A) effect is a cross-hybridization artifact in which poly(T)-containing molecules, which are produced by the reverse transcription of a poly(A)$^+$ RNA mixture, bind promiscuously to the poly(A) stretches of the DNA in microarray spots. It is customary to attempt to block such hybridization by adding poly(A) to the hybridization solution. This note describes an experiment intended to evaluate circumstances under which the blocking procedure may not have been successful.

Results: The experiment involves a spot-by-spot comparison between the hybridization signals obtained by hybridizing a microarray to: (1) end-labeled oligo(dT), versus, (2) cDNA prepared from muscle tissue. We found that the blocking appears to be successful for the vast majority of microarray spots, as evidenced by the weakness of the correlation between signals (1) and (2). However, we found that for microarray spots having oligo(dT) hybridization levels greater than a certain threshold, the blocking might be ineffective or incomplete, as evidenced by an exceptionally strong signal (2) whenever signal (1) is greater than the threshold.

Conclusion: The PolyA effect may be more subtle than simply a hybridization signal that is proportional to the PolyA content of each microarray spot. It may instead be present only in spots that hybridize oligo(dT) greater than some threshold level. The strong signal generated at these "outlier" spots by cDNA probes might be due to the formation of hybridization heteropolymers.

Background

Microarrays are tools for functional genomics research that are manufactured by depositing spots of DNA onto a glass or nylon substrate [1,2]. The spots' DNA are often obtained from many different clones in a cDNA library, each of which may contain a stretch of polyA originating from the mRNA in the pool from which the library was derived. Because first strand cDNA probes that are hybridized to such microarrays usually consist of poly(T)-containing molecules, which are produced by the reverse transcription of some tissue's poly(A)$^+$ RNA mixture, there is the potential for promiscuous polyA/polyT cross-hybridization between the first strand cDNA probes and the microarray DNA spots. In fact, when microarrays were first being developed, investigators described situations in which their microarray hybridization data were dominated by this artifactual "polyA effect" (or its converse when polyA tails are in the labeled probe) [3,4]. To reduce the polyA effect, it is therefore common practice to add unlabeled polyA to the hybridization mixture to bind to the

polyT segments of the hybridization probe, thus competing with the polyA segments in the microarray spots.

Despite the use of unlabeled polyA as a blocking agent, when microarray hybridization results obtained using probes derived from mRNA from two different sources are nearly the same, one cannot be certain that the similarity of the results is due to the similarity of the different mRNA source material, rather than to incomplete blockage of the polyA effect in both of the hybridizations. We were faced with this issue in connection with the similarity of microarray results that we obtained in a comparison between skeletal muscle mRNA from two related strains of rats. As a result, we undertook a search for the presence of the polyA effect, in which we made a spot-by-spot comparison of a microarray that was hybridized with an oligo(dT) probe, versus the same microarray hybridized using a first strand cDNA probe produced from skeletal muscle mRNA. Our expectation was that if the polyA effect was present, there would be a strong spot-to-spot correlation between the signals generated with these two probes. We were surprised to find that there appear to be prominent artifacts related to promiscuous polyA/polyT hybridization *only with very strongly expressed genes*, which might be due to the formation of hybridization heteropolymers at the corresponding microarray spots.

Results

Hybridization to the microarray

Fig. 1 shows background-subtracted images of a microarray that has been hybridized with first strand cDNA derived from skeletal muscle mRNA. Also shown in Fig. 1 are the background-subtracted images of the array after hybridization with end-labeled oligo(dT), performed in order to measure accessible PolyA in the spots.

Search for artifacts from the PolyA effect

As shown in the scatterplot provided in Fig. 2, there is only a weak correlation between the hybridization signals of spots with first strand cDNA from rat muscle, versus hybridization of the spots with oligo(dT). Among the 567 microarray spots having muscle gene-expression signals in the range that could be detected (as defined in the Methods section), the Pearson correlation between the muscle signals and the signals obtained using the oligo(dT) probe is only 0.46, and the correlation is even weaker (0.37) when the correlation calculation is restricted to microarray spots having a muscle signal greater than the median value of 7682 phosphorimager units. We therefore conclude that a polyA effect is not readily apparent in these data, presumably because we had added unlabeled polyA to the hybridization mixture as a blocking agent.

Spots at random locations throughout the array did not bind significantly to the oligo(dT) probe, almost all of which correspond to spots at which there was also no binding of muscle cDNA. Such spots do not appear to represent defects in the manufacturing of the array filter at which no DNA was deposited, because at the end of the experiment, we looked for DNA in those spots directly by staining them with the fluorescent nucleic acid dye acridine orange, followed by imaging of the spots with a fluorimager. Spots that gave no signal with the oligo(dT) probe nevertheless had a corresponding fluorescence signal, except for the deliberately vacant positions in the right hand side of each of the microarray matrices shown in Fig. 1. Furthermore, there was one spot at which binding to the oligo(dT) probe was undetectable, but at which there was nevertheless a very strong binding to the muscle probe. That spot (field 1d, column 11, row 21 in Fig. 1 and position 14511,0 in Fig. 2) corresponds to the enteric smooth muscle form of gamma actin (Accession number T60048), which according to GenBank's annotation for that accession number is a possible reversed clone because polyT (or polyA in its complementary strand) was not found within it.

Observation of outlier microarray spots

Although there is generally little correlation between the hybridization values obtained with oligo(dT) and muscle cDNA, this may not be true for spots that show the greatest binding to oligo(dT). For the spots showing oligo(dT) hybridization values greater than 12000 (phosphorimager units), the corresponding value of the muscle cDNA hybridization is always 24000 or more (phosphorimager units). This result was obtained for the same eight microarray spots, when microarray hybridizations were performed using muscle mRNA from both fa/fa and Fa/Fa Zucker rats. In contrast, spots with oligo(dT) hybridization values less than the threshold value of 12000 exhibit corresponding values for the muscle probe hybridization that may be anywhere between zero and the upper range of values.

The finding is not likely to be a chance happenstance for the following reason. If the null hypothesis is that the muscle-probe hybridization values for spots with polyA values > 12000 have the same statistical distribution as spots with polyA values < 12000, then the probability that all eight of them have values greater than the median is $(1/2)^8 = 0.0039$, and the probability that all eight of them have a value greater than 24000 is less than $(52/567)^8$ which is less than 10^{-8}, because only 52 of the 567 detectable spots have values greater than 24000. Thus, we reject the null hypothesis and conclude that the statistical distribution of muscle-probe values among spots having polyA values greater than the threshold of 12000 is significantly different than that for the remainder of the spots.

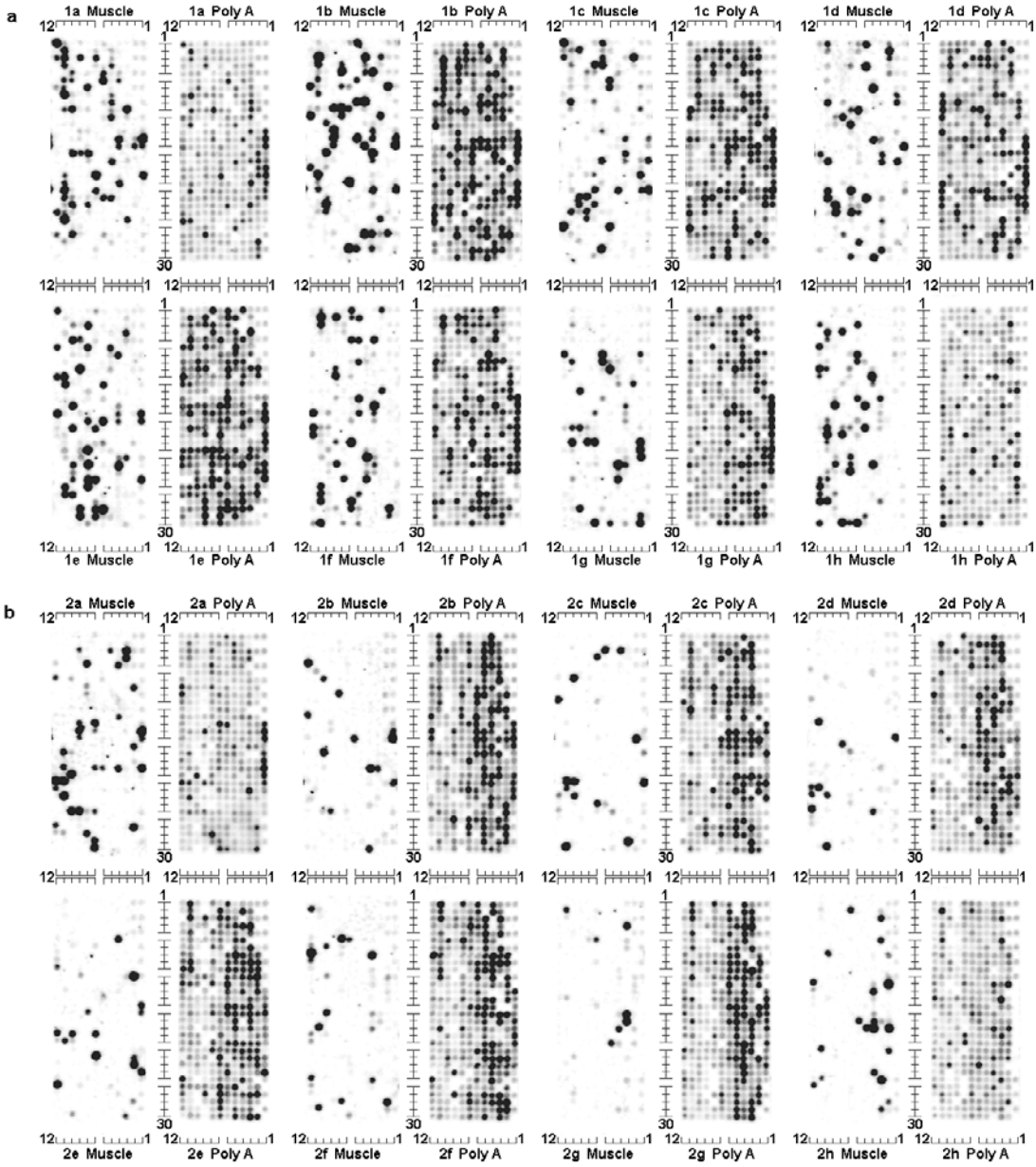

Figure 1

Microarray hybridized to muscle first strand cDNA vs. end-labeled oligo(dT) This figure shows the background-subtracted images of a single Research Genetics #GF200 array membrane, hybridized successively with two different probes. The first hybridization image, labeled as "Muscle", was performed using a first strand cDNA probe prepared from the skeletal muscle mRNA of a Zucker rat. The second hybridzation image, labeled as "PolyA", was performed after stripping the filter and rehybridizing it with end-labeled oligo(dT), in order to measure the spots' polyA content. The array consists of sixteen individual matrices. They are arranged in two fields (1 and 2, shown in Fig. 1(a) and Fig. 1(b), respectively), each of which contains eight matrices labeled with a letter from a to h. Each matrix consists of 12 columns and 30 rows of spots, which are indexed as indicated. The identity of the gene at each spot, indexed as described here, is available at http://www.resgen.com.. Spots in column 1 of each array matrix contain reference genomic DNA (in rows 1, 3, 5, 7, 9, 11, 25, 27, and 29) and reference cDNA for "housekeeping" genes (in rows 13 through 24). Spots in rows 1, 3, and 5 of column 2 of each array matrix also contain reference genomic DNA. Spot locations between the genomic DNA are empty, resulting in the landmark patterns seen in the upper-right and lower-right corners of each of the matrices.

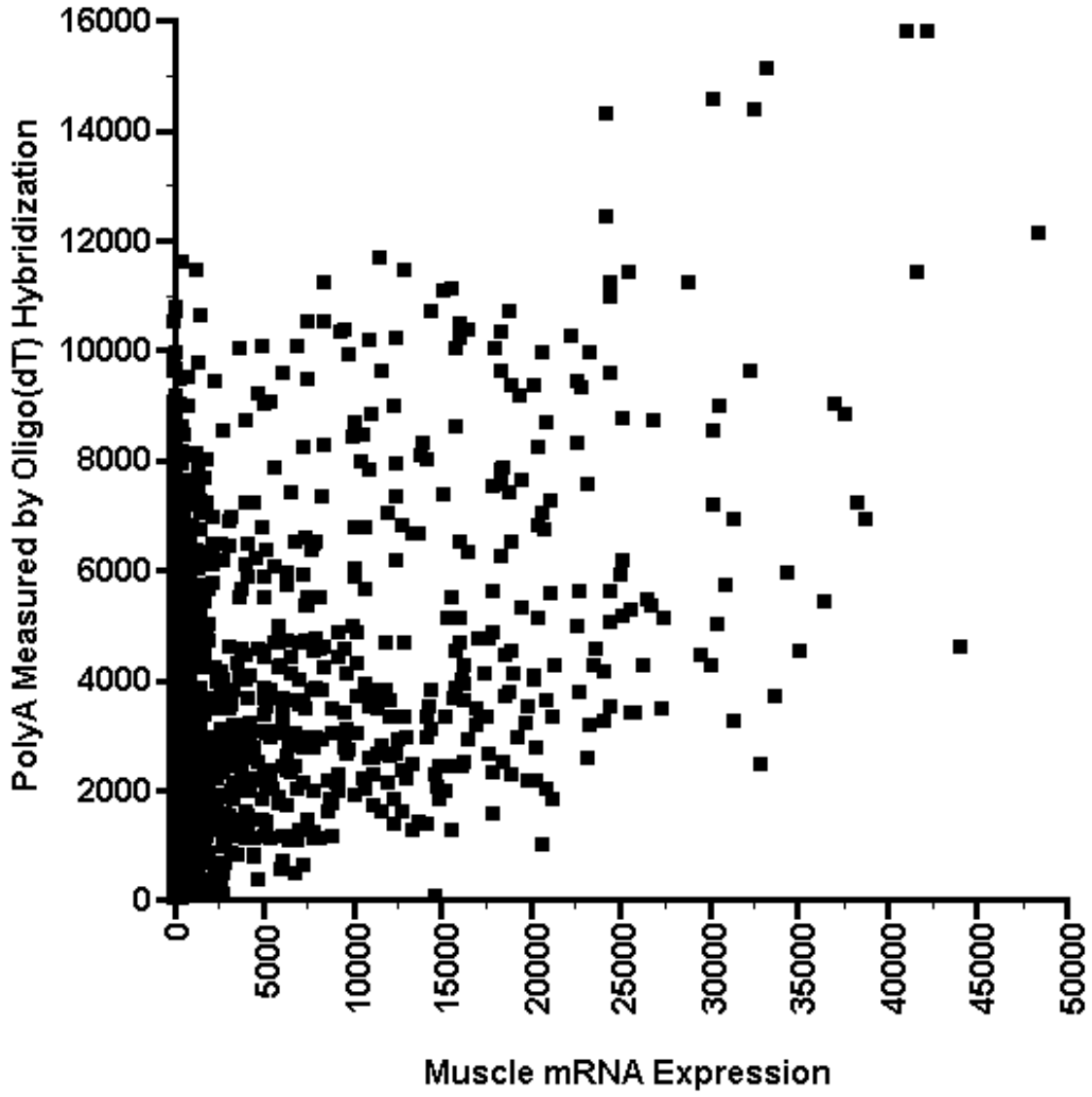

Figure 2
Scatterplot of muscle vs. oligo(dT) hybridization results. This is a scatterplot of the magnitude of hybridizations in the array membrane shown in Fig. 1, with each point indicating the intensities of a spot's hybridization with a first strand cDNA probe made from muscle mRNA, as well as an end-labeled Oligo(dT) probe used to measure PolyA. The intensities are given in phosphorimager units. Spots with Muscle values greater than 24000 and polyA values greater than 12000 are outliers that may be due to the formation of hybridization heteropolymers.

The question then arises as to whether there is anything unusual about the clones used to generate the DNA for those eight outlier microarray spots that *a priori* would have allowed us to flag them as potential concerns? As now described, distinguishing features of the clones might be that (1) they have unusually short inserts or (2) they may contain an unusually large fraction of bases that might be expected to bind to oligo(dT), namely, the bases that lie within a sequence run of AAAAAA... .

Relation between clone sequence and oligo(dT) probe hybridization signal

A search of GenBank reveals that only partial sequences are available for the clones corresponding to the eight outlier spots. One of the clones has a partial sequence containing a polyA stretch consisting of 41 bases (Accession H94857 (3') and H94912 (5') = GCN5 general control of amino-acid synthesis 5-like 1 = spot 1e6,28 in Fig. 1 = point 33211,15102 in Fig. 2). Five of the others have stretches of A or T between 15 and 26 bases, which are not unusual (Accession R52548 (3') and R52604 (5') = Superoxide dismutase 1 = spot 1c1,21 in Fig. 1 = point 24216,12384 in Fig. 2; AA460830 (3') and AA461132 (5') = RNA polymerase II polypeptide J = spot 1b11,3 in Fig. 1 = point 48345,12055 in Fig. 2; AA434115 (3') and AA434048 (5') = Cartilage glycoprotein 39 = spot 1b7,20 in Fig. 1 = point 30221,14554 in Fig. 2; AA490256 (3') and AA490356 (5') = Guanine nucleotide-binding protein G(K) alpha subunit = spot 1f,7,19 in Fig. 1 = point 24195,14273 in Fig. 2; and H93118 (3') and H93246 (5') = EST = spot 1c12,26 in Fig. 1 = point 42177,15785 in Fig. 2). The other two clones contain stretches of A no longer than 6 bases (Accession AA456352 (5') and AA454703 (3') = DEAD/H Asp-Glu-Ala-Asp/His box peptide 38 = spot 1f4,14 in Fig. 1 and point 32545,14353 in Fig. 2; and N92646 (3') and N99582 (5') = granulocyte-macrophage colony stimulating factor receptor 1 = spot 1f5,21 in Fig. 1 = point 41079,15760 in Fig. 2). However, N92646 contains an A-rich stretch of 30 base pairs with 23 A bases, and AA454703 contains more than one stretch of 6-A bases. Because stable hybridization duplexes have been reported with oligonucleotides as short as 7 bases [5], and because such stable duplexes may be formed even if there are base-pair mismatches [6], it is conceivable that the oligo(dT) probe could have hybridized to the A-rich regions in the 3' or 5' ends of the DNA in each the outlier spots.

The oligo(dT) probe could also have hybridized to the unknown sequences between the 3' and 5' ends of these clone inserts, so we sought their complete sequences. For purposes of comparison, we also sought the complete sequences of all clone inserts for which the muscle cDNA gave a hybridization signal in the range exhibited by the outlier clones (>24000 phosphorimager units). We sought the full sequences by using the clustering tool "IMAGEne" available at the IMAGE Consortium web page http://image.llnl.gov. This tool attempts to match the 5' and 3' partial sequences of a clone insert with the sequences of all other IMAGE clone inserts, as well as with all RefSeq reference sequences [7]. If matches are found, the entire clone insert sequence may then be constructed by piecing together overlapping matches of the contiguous sequences. Among the 52 clones for which the muscle cDNA gave a hybridization signal in the range exhibited by the outlier clones (>24000 phosphorimager units), it was possible to construct the entire insert sequence of 18 clones (see Additional File 1). Three of these clones are among the outlier group for which the polyA signal is greater than 12000 phosphorimager units.

We then noted two features of the sequence data that may explain in part the large signals that the outlier microarray spots exhibited when hybridized to the oligo(dT) probe. The first feature is that the hybridization signal is generally inversely proportional to sequence length, and two of the outlier clones have unusually short lengths. As seen in Fig. 3, the two outlier clones having insert lengths of less than 400 base pairs give the strongest hybridization signals. Because each microarray spot has the same total amount of DNA (5 ng), the microarray spots associated with these clones must have an unusually large number of short, densely deposited target molecules, and should therefore produce a proportionately large signal. The third outlier spot had an oligo(dT) hybridization signal only barely larger than the threshold of 12000 phosphorimager units and has a somewhat longer insert length (622 base pairs). The second noteworthy feature of the sequence data is that outlier spots may contain an unusually large fraction of bases that might be expected to bind to oligo(dT), namely, the bases that lie within a sequence run of seven or more As (AAAAAAA). As seen in Fig. 4, for two of the outlier clones, the fraction of such bases have relatively high values of 0.028 and 0.040. The third outlier spot has an oligo(dT) hybridization signal only barely larger than threshold of 12000 phosphorimager units and has a more typical fraction of 0.020.

Discussion

When microarrays were first being developed, investigators described situations in which their microarray hybridization data were dominated by the "polyA effect", in which poly(T)-containing molecules, which are produced by the reverse transcription of a poly(A)$^+$ RNA mixture, bind promiscuously to the poly(A) stretches of the DNA in microarray spots. As illustrated by Nguyen et al. [4], when this artifact is pronounced, the amount of hybridization at a particular microarray spot is an indication of the accessible polyA content of the DNA at the spot, rather than of the amount of the particular nucleic acid species in the hybridization solution that one intends to hybridize gene-specifically to that spot. Thus, when the polyA effect is pronounced, the hybridization of the reverse transcription product made from any poly(A)$^+$ RNA mixture will be essentially the same, namely, the hybridization that would be obtained with an oligo(dT) probe alone. This is because the binding of poly(T)-containing molecules in the hybridization solution to the polyA stretches in the microarray spots dominates the hybridization results, overwhelming the signal produced by gene-specific hybridization [4].

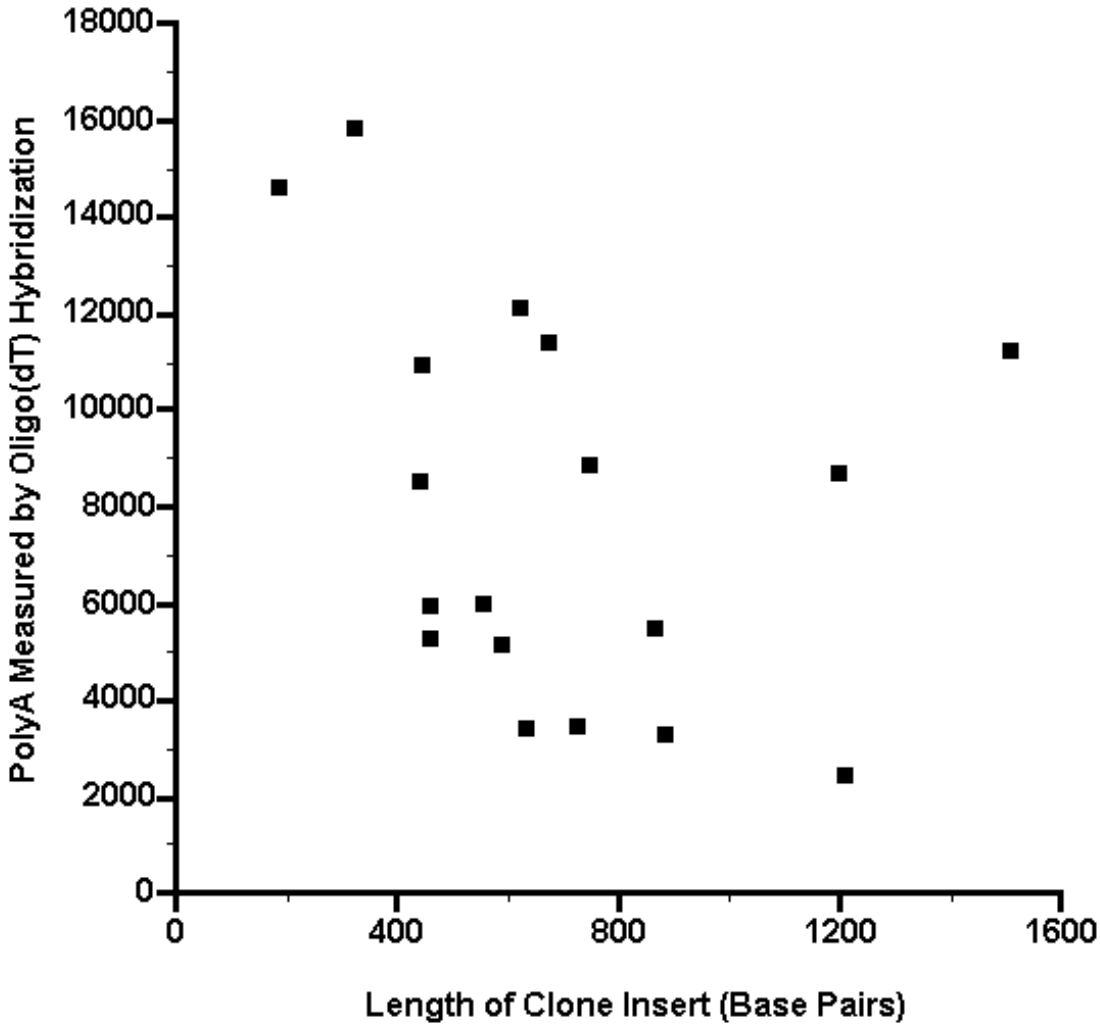

Figure 3
Hybridization intensity from oligo(dT) probe, as a function of the length of the clone insert for different microarray spots Oligo(dT) was end labeled with T4 kinase, then hybridized to a microarray. The complete sequences of 18 clone inserts were then determined, corresponding to microarray spots that had a signal of >24000 phosphorimager units with a muscle cDNA probe (see Additional File 1). The lengths of the insert sequences are plotted here versus the hybridization intensity from the oligo(dT) probe.

To reduce the polyA effect, it has become common practice to add unlabeled polyA to the hybridization mixture, to bind to the polyT segments of the hybridization probe, thus competing with the polyA segments in the microarray spots. Despite the use of this blocking procedure, when microarray hybridization results obtained using probes derived from mRNA from two different sources are nearly the same, one cannot be certain that the similarity of the results is due to the similarity of the different mRNA sources, rather than to incomplete blockage of the polyA effect in both of the hybridizations. We were faced with this issue when we obtained nearly identical microarray hybridizations using skeletal muscle mRNA from lean (Fa/Fa) versus obese (fa/fa) Zucker rats at an age of 6

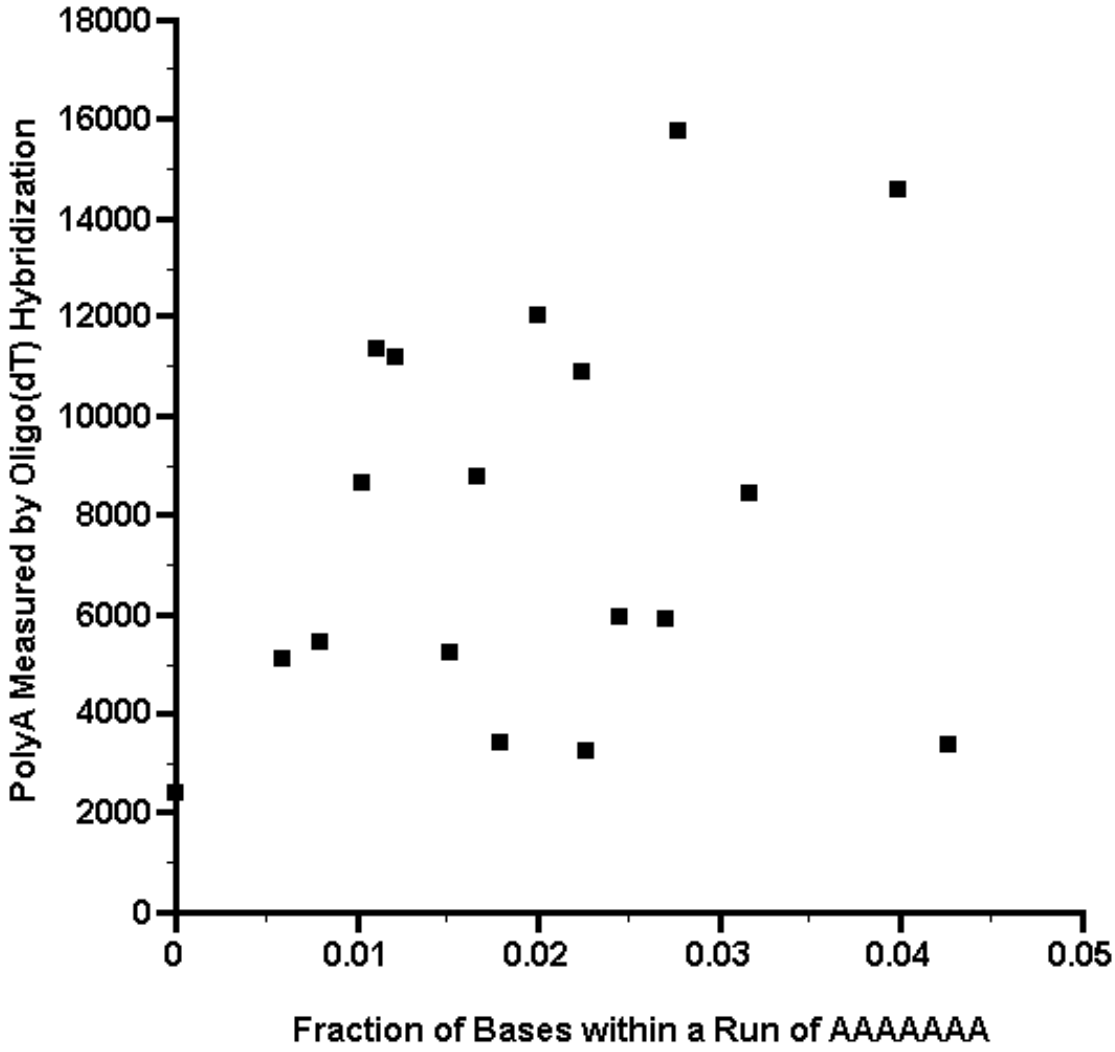

Figure 4
Hybridization intensity from oligo(dT) probe, as a function of the fraction of bases that are within a sequence run of AAAAAAA Oligo(dT) was end labeled with T4 kinase, then hybridized to a microarray. The complete sequences of 18 clone inserts were then determined, corresponding to microarray spots that had a signal of >24000 phosphorimager units with a muscle cDNA probe (see Additional File 1). Each sequence was examined on both strands to determine the number of bases that lie within a run of seven or more As (AAAAAAA), which is a run length that might form a stable hybrid with oligo(dT). The total number of such bases was then divided by twice the sequence length to obtain the fraction of bases that are within a sequence run of AAAAAAA. The fractions are plotted here versus the hybridization intensity from the oligo(dT) probe.

weeks (the age at which phenotypic differences between fa/fa and Fa/Fa are thought to begin). Among the more than 5000 genes on the microarray, only one was possibly expressed differentially (the sarcomeric isoform of the mitochondrial creatine kinase gene, sMtCK, which appeared to have higher expression in the lean muscle). We there-

fore conducted the experiment, with the results shown in Figs. 1 and 2, in order to see whether we would obtain the same hybridization results using only end-labeled oligo(dT) as a probe, thereby providing evidence whether the polyA effect caused the two muscle cDNA hybridizations to produce similar results.

Our main results are the scatterplot showing microarray hybridization signals using a probe derived from muscle mRNA, versus the signals using only an oligo(dT) probe (Fig. 2). The scatterplot exhibits only a weak correlation between these two hybridization results, for the vast majority of microarray spots. Furthermore, we did not see a simple additive effect, which might have been identified as follows. Let the signal obtained at spot number *i* from hybridization with the probe derived from muscle mRNA be denoted as Hi. It may be represented as the sum of a background signal Bi due to hybridization solely to the polyA segment in the spot, plus the gene-specific hybridization Hi', i.e., Hi = Bi + Hi'. If we knew the background polyA hybridization Bi, for example, by inferring its value from hybridization with the oligo(dT) probe, we could subtract the polyA background to obtain a corrected gene-specific hybridization signal, i.e. Hi' = Hi - Bi. However, looking along the vertical axis of the scatterplot, we were surprised to see that there are many spots at which Hi is nearly zero, indicating that Bi must also be nearly zero, even when there is a significant polyA signal.

The unexpected exception to this observation occurs when the polyA signal is greater than a threshold of about 12000 phosphorimager units. As seen in Fig. 2, for those eight microarray spots, the value of the background polyA hybridization Bi appears to be much larger than the median value of Hi among all spots, i.e., if the polyA signal is greater than 12000, then the value of Hi will always be greater than ~24000 phosphorimager units. Furthermore, as described in the Results section, the statistical distribution of Hi among spots having polyA values greater than the threshold of 12000 is significantly different than that for the remainder of the spots.

Regarding the question of whether there is anything unusual about the clones used to generate the DNA for those eight outlier microarray spots, we found that distinguishing features of the clones might be that (1) they have unusually short inserts of less than 400 base pairs, and (2) they may contain an unusually large fraction of bases that might be expected to bind to oligo(dT), namely, the bases that lie within a sequence run of polyA. Investigators might therefore use these as criteria for the potential existence of a polyA effect artifact, and manufacturers of microarrays may use these criteria in deciding which clones not to use.

However, apart from using empirical observations like those shown in Figs. 3 and 4, we have no suggestion for how to predict the magnitude of the oligo(dT) hybridization signal from the known sequences. According to Southern and colleagues [8], the factor that determines the magnitude of oligonucleotide hybridization signals is not the stability of the duplexes that may be formed through hybridization, but rather the rate of forward reaction, which is determined by the rate of hybrid nucleation. Their work indicates that the availability of nucleation sites in the immobilized target molecule will be determined by secondary structure, which is not predictable in any obvious way from the primary DNA sequences. Ongoing research into this problem interprets hybridization data not only in terms of Watson-Crick base pairing, but also in terms of base stacking interactions, loops, bridges, and dangling ends – and in the case of DNA immobilized on nylon membranes – in terms of diffusion of solvents into and out of membrane pores, multiple interactions within pores, and details of the way in which the DNA is attached to its membrane support [8].

Finally, consider why there is a threshold oligo(dT) probe signal, above which hybridization by an mRNA-derived probe invariably produces a signal much larger than the median among the remainder of the microarray spots. We suspect that the most important reason for the existence of a threshold signal is related to the formation of hybridization networks at these spots, which are also known as hyperpolymers [9]. If the polyT tail of a reverse-transcribed probe molecule binds promiscuously to the polyA in one of these spots, the remainder of the anchored probe molecule may hybridize to another labeled or unlabeled probe molecule in the hybridization solution, which in turn may hybridize to yet another probe molecule in the solution, etc., forming a network of hybridized probe molecules. If a threshold density of accessible polyA in the microarray spot's DNA is needed in order to form a stable anchor for such a network, this would provide a mechanistic explanation for our observation.

Conclusions

The PolyA effect may be more subtle than simply a hybridization signal that is proportional to the PolyA content of each microarray spot. It may instead be present only in spots that hybridize oligo(dT) greater than some threshold level. The strong signal generated at these "outlier" spots by cDNA probes might be due to the formation of hybridization heteropolymers.

Methods

Housing and care of animals

Female lean (Fa/Fa) and obese (fa/fa) Zucker rats, 5 wk of age, were purchased from the Animal Model CORE Facility of the University of California at Davis. They were

housed in the Animal Resource Center of the University of Texas at Austin under standard laboratory conditions. All procedures were approved by the University of Texas at Austin Animal Use Committee.

Tissue samples

Rats were quickly anesthetized by intravenous injection of sodium pentobarbital (50 mg/kg). Red quadriceps muscles from one leg of each animal were removed and freeze clamped as rapidly as possible with tongs that had been cooled in liquid N2. Samples were stored at -70°C for later use.

Extraction of total and messenger RNA

Tissue samples were homogenized in an acid-phenol reagent, with the volume of the tissue not exceeding 10 percent of the volume of the reagent used. Total RNA was then obtained from the homogenate by the procedure recommended by the reagent's supplier (TRI-reagent, Sigma #T9424). This was followed by selection of poly(A) mRNA from the total RNA by hybridizing the RNA with oligo(dT) magnetic microparticles, then isolating the mRNA magnetically (mRNA Isolation Kit, Miltenyi Biotec #751-02). The spectrophotometric 260/280 ratios for mRNA from the obese and lean tissues were 1.95 and 1.99, respectively, which were sufficiently close to 2.0 that an additional round of oligo(dT) selection was not performed.

Preparation of first strand cDNA

Complementary DNA was made using 200 ng of purified, oligo(dT) primed mRNA and SuperScript II RNase H- reverse transcriptase (BRL Life Technologies #18064-014). The cDNA was made radioactive by incorporation of ^{33}P-dCTP (ICN #58201) using the procedure described at http://www.resgen.com/gf200pro.html, except that 33 mM of dATP, dTTP, and dGTP (Pharmacia #27-2035-01) and 1 U/μl of a ribonuclease inhibitor (Ambion #2690) were added. Unincorporated nucleotides were separated from the labeled cDNA using a push column (Stratagene #400701).

Hybridization to array membrane

A high density DNA array membrane was used to measure the level of expression of each of 5,184 genes (Research Genetics, GENEFILTER #GF200, Lot #980611-21). Hybridization of the membrane with radioactive cDNA was performed by the procedures described at http://www.resgen.com/gf200pro.html, except that the last room temperature wash was performed in 2X SSC. The hybridization solution included 0.5 μg/ml poly(dA) (Research Genetics #POLYA.GF) and 0.5 μg/ml cot1 DNA (BRL #15279-011) as blocking agents. Stripping the filter, in order to reprobe it, was done by placing the filter in a boiling 1% SDS solution, as recommended by the manufacturer. The stripped filter was used to expose a phosphorimager plate, which was then scanned. The resulting image had < 0.005 remaining from the image that was obtained before stripping. We obtained nearly identical microarray hybridizations using cDNA prepared from skeletal muscle mRNA from lean (Fa/Fa) versus obese (fa/fa) Zucker rats at an age of 6 weeks (the age at which phenotypic differences between fa/fa and Fa/Fa are thought to begin). Only one gene was possibly expressed differentially (the sarcomeric isoform of the mitochondrial creatine kinase gene, sMtCK, which appeared to have higher expression in the lean muscle). We confirmed the sMtCK result with a Northern blot, but did not do so for any of the other genes, which may be considered a limitation of our experiment.

Quantifying the hybridization results

Radioactivity in hybridized filters was imaged using a phosphorimager (Model 445 SI, Molecular Dynamics). The resulting image files were then processed using custom software, as shown in Fig. 5. Each of the membrane's sixteen 12 × 30-spot arrays was extracted from the original image file, along with a border of pixels corresponding to one spot width. A background value was then estimated for each pixel in each of these images, as follows. If a pixel was situated in the region between or adjacent to DNA spots, and if the spatial derivative of the image at that pixel indicated that it was not part of the overlap region between two intense spots, that pixel was defined to be a background pixel. If a pixel was situated within the region corresponding to a spot of DNA, or if it was situated in the region between two overlapping intense spots, the background value for that pixel was set equal to the value of the nearest background pixel, as defined above. A background-subtracted image of the radioactivity was then obtained by subtracting the background value for each pixel from the original image.

The intensity of spots in an image is a function of how long the phosphorimager plate was exposed, as well as factors such as the efficiency of probe labeling. Therefore, in order to be able to compare background-subtracted images of the same 12 × 30 spot array for successive hybridizations, normalization was performed as follows. The image for one of the hybridizations was selected to be the reference, and corresponding pixels in the other image were normalized by substituting them into a quadratic polynomial normalization function (of the pixel value), the coefficients of which were estimated by singular value decomposition (SVD) [10]. This quadratic model allows for some compensation in the event that the response of the phosphorimager changes between measurements. We found that the fitted second order coefficients were always very close to zero, and the constant offset coefficients were also very close to zero. Thus, normalization of back-

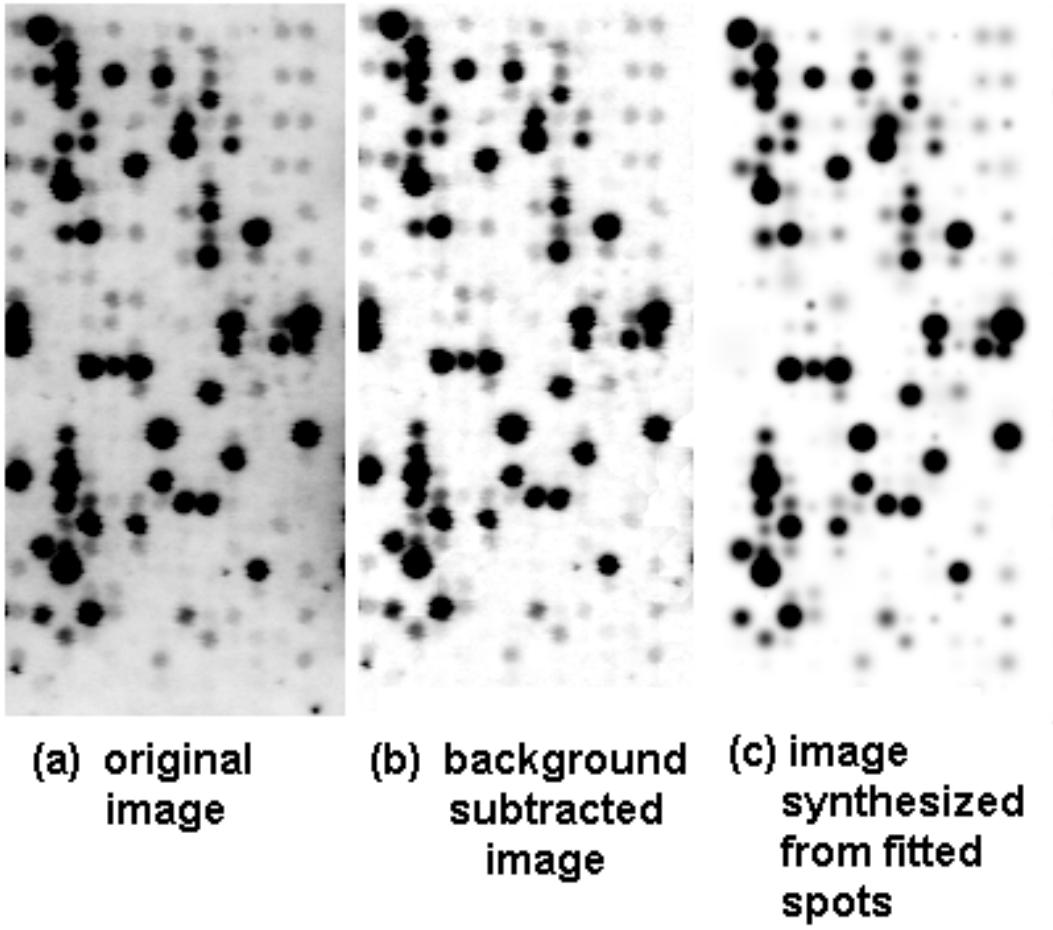

Figure 5
Processing of images to obtain quantitative estimates of microarray spot hybridization intensities. (a) Original phosphorimager image of the radioactivity in one of the array's sixteen 12 column × 30 row matrices of spots. **(b)** Subtraction of background from the array image, as described in Methods. **(c)** Image synthesized by constructing two-dimensional gaussian functions with locations, amplitudes, and standard deviations that were estimated by fitting the spots in the image (b).

ground-subtracted images of the same array for different hybridizations was accomplished essentially by multiplying all pixels in the non-reference image by a constant scale-factor, the value of which had been estimated by SVD. Note that the phosphorimager data have units related to the voltage of its phototube, which are not related in any obvious way to nucleic acid concentrations or amounts. Some investigators rescale the data such that the average signal per microarray spot equals 1, but we did not do so.

The distance between spots on the array is only 0.75 mm, and as a consequence, many of the adjacent intense spots overlap one another. When integrated intensity within a specified region about the center of the spot is used to represent the magnitude of hybridization, it is then useful to model the image as the sum of superimposed two-dimensional spot functions. This was done using two-dimensional gaussian functions to represent the spots, as shown in Fig. 5, with the location, amplitude, and standard deviation coefficients all estimated automatically from the background-subtracted image in two dimensions, using the Levenberg-Marquardt method [10]. The similarity between background-subtracted and fitted spot images (Fig. 5b vs. 5c) indicates that much of the spread of the intense spots appears to be gaussian, which might be attributed to the gaussian shape of the laser beam that scans the phosphorplate. When the spot intensity was close to the noisy background level, this model cannot be used to fit the data, due to near singularity of the matrix equations that had to be solved to perform the fitting. In that case, the method that we used to estimate the intensity of each spot was to sum the nine adjacent pixels in the center of each spot of the background-subtracted image. The correlation between the values so obtained and the corresponding value of the integrated spot intensity from successful gaussian curve fitting was 0.99.

A histogram was constructed from the logarithm of the values of all the spot values. As observed in [11], the histogram for our muscle data was bimodal, consisting of a gaussian-like distribution of low-intensity spots, overlapping an adjacent distribution of moderate and high-intensity spots. The transition between these two distributions occurred within a clearly recognizable range of values, 1550 to 1800 phosphorimager units. We therefore followed the conservative practice recommended in [11] by considering all spot values less than a value of 1800 to define undetectable hybridizations. According to this criterion, 567 of the spots were detectable.

Search for promiscuous hybridization to polyA segments

To investigate the polyA effect as a potential artifact, we performed the following experiment, using essentially the approach described in [4]. Oligo(dT) (10–20 mer mixture, Research Genetics cat. # POLYT.GF) was end labeled with T4 kinase (Life Technologies #10476-018) and (γ-^{33}P)ATP (ICN #58000). The array filter was then hybridized with this probe in Hybrisol I (Oncor #S4040) for 10 hours at 42°C, then washed twice at room temperature for two minutes in 6X SSC and 0.1% SDS. It was then used to expose a phosphorplate for 12 hours, which was scanned using a phosphorimager (Model 445 SI, Molecular Dynamics) to identify the array spots having long stretches of poly(dA). Quantifying of the spot intensities was performed as indicated above.

Staining of the microarray filter with a fluorescent nucleic acid indicator

After all hybridizations of the microarray filter had been performed, we stained the filter with a fluorescent dye that is specific for nucleic acids. We did so in order to determine whether the absence of hybridization signals for some microarray spots was due to the absence of any DNA there, which would indicate a defect in the manufacturing of the microarray at those spots. The filter was stained with 1 μg/ml acridine orange (Sigma A6014) for 10 minutes at room temperature, rinsed in distilled water, then imaged with a fluorimager (Molecular Dynamics). The dye was excited by the 488 nm argon laser line of the fluorimager, and emission was detected with a narrowband filter centered at 530 nm.

Authors' contributions

SJP participated in the conception, design, and coordination of the study; conducted the animal, hybridization, and fluorescence studies; assisted with the computer-related aspects of the study; and wrote parts of the manuscript. DRR participated in the conception, design, and coordination of the study; conducted the computer-related aspects of the study; assisted with the hybridization and fluorescence studies; and wrote parts of the manuscript. JLI participated in the conception, design, and coordination of the study and edited the manuscript. All authors read and approved the final manuscript.

Additional material

Additional File 1

Complete DNA sequences of 18 clone inserts. *This file contains complete DNA sequences of 18 clone inserts. The clones correspond to microarray spots for which the muscle cDNA probe signal is greater than 24000 phosphorimager units. The sequences were constructed using the clustering tool IMAGEne at the IMAGE Consortium web site http://image.llnl.gov. For each sequence, we also provide the IMAGE clone ID, Genbank accession numbers for the 5' and 3' partial sequences, the RefSeq or IMAGE Consortium cluster ID that is used to join the 5' and 3' partial sequences, the microarray location of the corresponding spot, the strength of the corresponding signal from hybridizing to the oligo(dT) probe, the strength of the corresponding signal from hybridizing to the muscle cDNA probe, and the sequence length.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-3-35-S1.txt]

Acknowledgements

This work was supported by an NRSA Individual Fellowship awarded to S. Pan from the U.S. National Institutes of Health (F32 AT000051).

References

1. Duggan DJ, Bittner M, Chen Y, Meltzer P and Trent JM **Expression profiling using cDNA microarrays.** *Nat Genet* 1999, **21(Suppl 1):**10-14

2. Jordon B and ed **DNA Microarrays. Gene Expression Applications.** *Berlin: Springer-Verlag* 2001,
3. Gress TM, Hoheisel JD, Lennon GG, Zehetner G and Lehrach H **Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues.** *Mamm Genome* 1992, **3:**609-619
4. Nguyen C, Rocha D, Granjeaud S, Baldit M, Bernard K, Naquet P and Jordan BR **Differential gene expression in the murine thymus assayed by quantiative hybridization of arrayed cDNA clones.** *Genomics* 1995, **29:**207-216
5. Maskos U and Southern EM **Parallel analysis of oligodeoxynucleotide (oligonucleotide) interactions. I. Analysis of factors influencing oligonucleotide duplex formation.** *Nucleic Acids Res* 1992, **20:**1675-1678
6. Maskos U and Southern EM **A study of oligonucleotide reassociation using large arrays of oligonucleotides synthesized on a glass support.** *Nucleic Acids Res* 1993, **21:**4663-4669
7. Pruitt KD and Maglott D **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29:**137-140
8. Southern E, Mir K and Shchepinov M **Molecular interactions on microarrays.** *Nat Genet* 1999, **21(Suppl 1):**5-9
9. Meinkoth J and Wahl G **Hybridization of nucleic acids immobilized on solid supports.** *Anal Biochem* 1984, **138:**267-284
10. Press WH, Teukolsky SA, Vetterling WT and Flannery BP **Numerical Recipes in C** *New York: Cambridge University Press* 1992, 676-688
11. Pietu G, Alibert O, Guichard V, Lamy B, Bois F, Leroy E, Mariage-Samson R, Houlgatte R, Soularue P and Auffray C **Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array.** *Genome Res* 1996, **6:**492-503