BMC Bioinformatics

BioMed Central

Research article

Development and production of an oligonucleotide MuscleChip: use for validation of ambiguous ESTs

Rehannah HA Borup[1], Stefano Toppo[2], Yi-Wen Chen[1], Tanya M Teslovich[1], Gerolamo Lanfranchi[2], Giorgio Valle[2] and Eric P Hoffman*[1]

Address: [1]Research Center for Genetic Medicine, Children's National Medical Center, 111 Michigan Avenue N.W, Washington, DC 20010, USA and [2]CRIBI Institute, University of Padova, Padova, Italy

E-mail: Rehannah HA Borup - rborup@cnmcresearch.org; Stefano Toppo - stefano@cribi.unipd.it; Yi-Wen Chen - ychen@cnmcresearch.org; Tanya M Teslovich - tteslovich@cnmcresearch.org; Gerolamo Lanfranchi - lanfra@cribi.unipd.it; Giorgio Valle - giorgio.valle@unipd.it; Eric P Hoffman* - ehoffman@cnmcresearch.org

*Corresponding author

Published: 29 October 2002

BMC Bioinformatics 2002, **3**:33

Received: 16 July 2002
Accepted: 29 October 2002

This article is available from: http://www.biomedcentral.com/1471-2105/3/33

Keywords: Expression profiling, oligonucleotide microarrays, Affymetrix, muscle, EST

Abstract

Background: We describe the development, validation, and use of a highly redundant 120,000 oligonucleotide microarray (MuscleChip) containing 4,601 probe sets representing 1,150 known genes expressed in muscle and 2,075 EST clusters from a non-normalized subtracted muscle EST sequencing project (28,074 EST sequences). This set included 369 novel EST clusters showing no match to previously characterized proteins in any database. Each probe set was designed to contain 20–32 25 mer oligonucleotides (10–16 paired perfect match and mismatch probe pairs per gene), with each probe evaluated for hybridization kinetics (Tm) and similarity to other sequences. The 120,000 oligonucleotides were synthesized by photolithography and light-activated chemistry on each microarray.

Results: Hybridization of human muscle cRNAs to this MuscleChip (33 samples) showed a correlation of 0.6 between the number of ESTs sequenced in each cluster and hybridization intensity. Out of 369 novel EST clusters not showing any similarity to previously characterized proteins, we focused on 250 EST clusters that were represented by robust probe sets on the MuscleChip fulfilling all stringent rules. 102 (41%) were found to be consistently "present" by analysis of hybridization to human muscle RNA, of which 40 ESTs (39%) could be genome anchored to potential transcription units in the human genome sequence. 19 ESTs of the 40 ESTs were furthermore computer-predicted as exons by one or more than three gene identification algorithms.

Conclusion: Our analysis found 40 transcriptionally validated, genome-anchored novel EST clusters to be expressed in human muscle. As most of these ESTs were low copy clusters (duplex and triplex) in the original 28,000 EST project, the identification of these as significantly expressed is a robust validation of the transcript units that permits subsequent focus on the novel proteins encoded by these genes.

Background

There are three platforms used for expression profiling with microarrays: spotted cDNA arrays [1,2], spotted oligonucleotide arrays [3], and *in situ* synthesized oligonucleotide arrays (Affymetrix) [4]. There are inherent advantages and disadvantages to each approach. Spotted arrays (both cDNA and oligonucleotide) are more easily customizable and can be considerably less expensive to produce and use. However, the manipulation of many thousands of solutions has led to contamination problems with commercially available clone sets [5]. The inherent flexibility of the spotted microarrays has the added liability of difficulties in standardization of arrays and array data; the resulting data is not considered highly transportable [6].

Affymetrix arrays are produced directly from nucleotide sequences in databases, and do not include any liquid handling of specific clones. 25 mer oligonucleotides are designed against selected genes/ESTs as perfect-match and mis-match probe pairs tiled across each gene, with subsequent *in situ* synthesis of each probe set on a solid glass support. This bio-informatics-driven probe design and probe sequence extraction process, combined with the photolithographic probe synthesis method, provides standardized arrays, with resulting data that is inherently transportable. Typically, 30–40 probes are synthesized per gene, and the highly redundant data allows the implementation of algorithms that computationally compensate for cross hybridization, experimental variability across a surface, and poorly performing probes. An advantage of the factory-produced Affymetrix GeneChips is that they are inherently standardized, and thus permit easy comparisons between profiles using the same chip. However, this is also a disadvantage, as chip design and production of novel photolithography arrays is quite labor intensive and expensive. Indeed, most reports to date have utilized only "stock chips" produced and marketed by Affymetrix (see [www.affymetrix.com]), with little flexibility as to the specific genes under study.

Here, we describe the design, production, and use of a custom Affymetrix array, based upon a non-normalized EST sequence resource from human skeletal muscle [7], with additional probe sets selected from preliminary data on human muscle using Affymetrix stock chips [8]. A goal of producing the described MuscleChip was to use expression profiling as a means of validating novel ESTs. It is commonly acknowledged that ESTs represented by only singletons or small clusters in dbEST may represent artifact, and thus require verification. We felt that the robust analytical protocol of Affymetrix chips would permit accurate verification of large numbers of potentially novel ESTs in a highly parallel manner.

Results

Design and production of custom Affymetrix MuscleChip

Three resources were used to select probe sets for the MuscleChip: a non-normalized muscle EST resource [7], "diff calls" from stock chip (Affymetrix HuFL) comparisons of normal and Duchenne muscular dystrophy muscle [8], and 172 gene sequences of interest from a number of investigators in muscle research (Table 1). From the EST resource, approximately 30,000 3' ESTs were sequenced from a non-normalized normal human muscle cDNA library ([7]; and unpublished), and sequences resolved into 2,052 clusters which were placed onto the MuscleChip. From the normal/DMD HuFL data, 1,120 genes were selected as showing significant differences between normal and dystrophic muscle. Of these 1,120 probe sets 1,052 were drawn directly from the Affymetrix HG-U95A stock chip, and the remaining 68 probe sets were drawn from the original HuFL chip. The lists of genes from these three sources were combined and then resolved into 4,601 probe sets representing 3,344 sequences that were chosen from the different sources. Many of the genes/ESTs were designed with multiple probe sets to enable intra-chip verification of expression data (Table 1); the redundant probe sets included _at (complete rule set; default probe set), _f (functional consensus), _g (possible cluster group), _i (incomplete set), _r (rules dropped), and _s (similar to other clusters/sequences) (See Methods, MuscleChip production).

The MuscleChip was originally designed with 2,052 EST sequences, 734 of which showed no match with UniGene clusters using 1999 builds. To update the sequence definitions, those EST clusters or singletons previously showing no high similarity to databases were BLAST searched against the most recent nr database release, using the NetBlast stand alone search tool (Table 2). There were only 26 EST clusters that showed no high similarity in either cDNA/RNA sequence resources, or genomic DNA resources, with an additional 343 clusters showing homology to non-characterized genomic or EST cluster sequence data (Table 2). 365 sequences showed highly significant alignments with recently characterized proteins, which were then updated in the sequence definition file [http://microarray.cnmcresearch.org/musclechipindex.asp].

Validation of MuscleChip

To verify the MuscleChip, we used the same biotinylated cRNA that we had recently reported from muscular dystrophy patients and normal controls [8]. These pooled samples were derived from 5 Duchenne muscular dystrophy patients and 5 age and sex-matched controls, as we have previously described [8]. The same hybridization solutions were applied to both MuscleChips and to the Affymetrix HG-U95A human stock GeneChip, and the results compared. Two profiles of each mixed sample were

Table 1: Sequences and probe sets on the MuscleChip

Sequences Chosen	
2,052 ESTs & 172 collaborator sequences (custom):	2,224
U95A sequences (commercial):	1,052
Hu6800/HuFL sequences (commercial):	68
Total custom and commercial sequences:	**3,344**
Sequences and probe sets on the MuscleChip:	
Total number of probe sets on the Muscle Chip:	4,601
Total number of custom and commercial sequences:	3,344
Redundancy within sequence representation (custom set):	
sequences represented by 1 probe set:	1,209
sequences represented by 2 probe sets:	773
sequences represented by 3 probe sets:	242

Table 2: Distribution of BLAST results of 734 novel ESTs on MuscleChip. BLAST search with 734 query EST sequences

No high similarity found in either RNA or genomic DNA:	26
No match to known gene. but matches to undefined genomic or EST sequences:	343
Total number of non hits	369
Significant alignments with following distribution:	
- Low significant matches:	10
- High significant matches:	
- Matches to ref	291
- Matches to emb	14
- Matches to gb	50
Total number of hits:	365
Total number of sequences	734

generated from different aliquots of RNA, derived from different regions of the biopsies, for a total of four HG-U95A profiles, and four MuscleChip profiles. Fluorescent images were scaled to target intensity of 800. Scaling factors for the MuscleChip (0.7–1.17) were similar to HG-U95A (0.78–1.67). "Present calls", which are generally independent of scaling factors, were approximately 25–30% of genes on the HG-U95A, and approximately 40–50% of genes on the MuscleChip. This reflects the greater representation of muscle-specific transcripts on the MuscleChip, relative to HG-U95A.

We then compared profiles for the same cRNA hybridization sample on the MuscleChip and HG-U95A stock chips. The correlation between replicate samples (Con1 and Con2) on the HG-U95A and the MuscleChip for 1,052 shared probe sets was 0.98 and 0.99 respectively (data not shown). The high reproducibility of Affymetrix stock chip and MuscleChip results is consistent with other data in our laboratory [14].

Absolute intensities (Avg Diff) for each probe set shared between the two chips were plotted against each other, using both "absent" and "present" calls (Figure 1. Panel A; Data Analysis in Methods). Overall, there was an excellent correlation between values for the two different chips, verifying the performance of the MuscleChip (correlation coefficient, r = 0.89).

From the graph in Figure 1, the slope of the comparison of shared probe sets in HG-U95A vs. the MuscleChip deviated from 1. This was due to the fact that the MuscleChip was partially constructed from probe sets on the HuFL chip showing high expression in muscle. This leads to a type of ascertainment bias, where the overall hybridization to the MuscleChip is higher than U95A, given the same amount of muscle cRNA hybridized to the chip. To correct for this, one simply takes shared probe sets, and considers the Avg Diff (hybridization intensities) equal between the two array types (e.g. normalize data using shared genes, rather than a fixed target intensity). This leads to a MuscleChip-specific scaling factor, that, when applied, makes U95A and MuscleChip results directly comparable (Figure 1, panel B). This MuscleChip-target intensity was 800 times 3.18 (target intensity of 1970 for MuscleChip = target intensity of 800 for U95A).

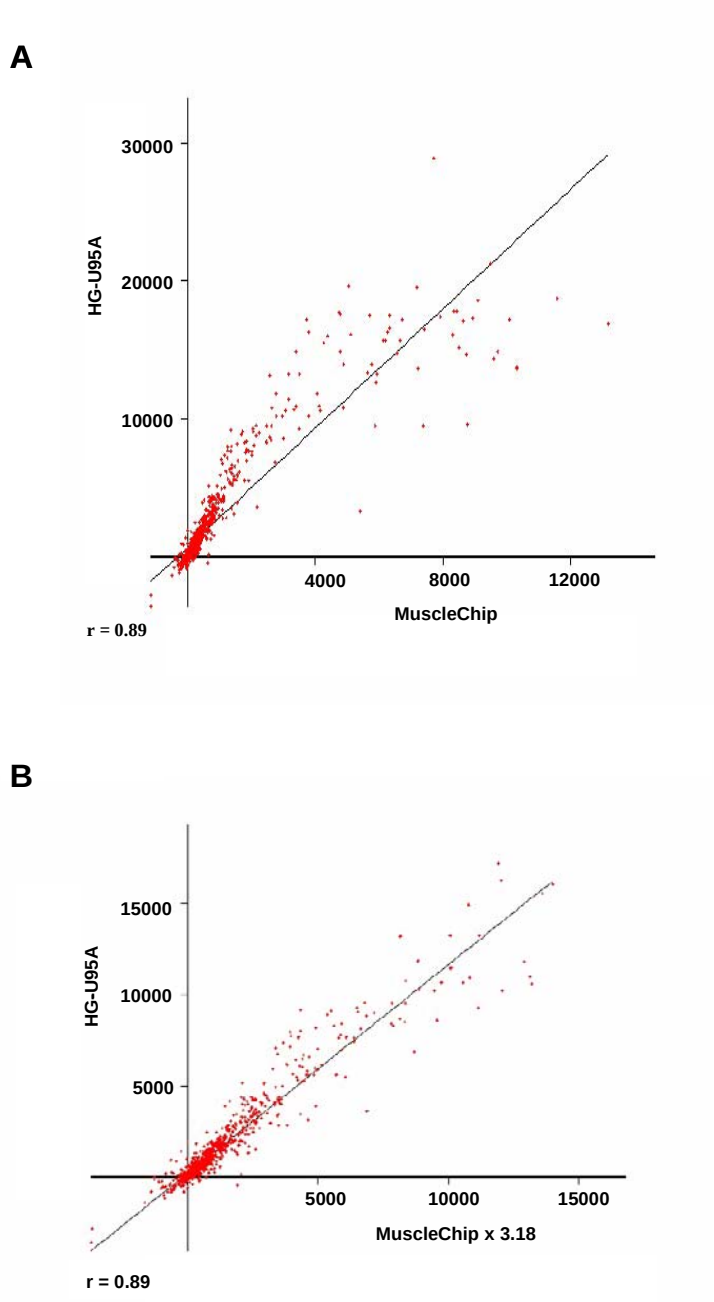

Figure 1
Validation of MuscleChip by comparison to shared probe sets on HG-U95A stock chip, and the chip to chip variability of custom probe sets on the MuscleChip. Panel A. Shown is correlation of absolute analysis values (Avg Diff) for probe sets shared by the two chips, after hybridization with the same human normal muscle cRNA. There is an excellent correlation coefficient (0.89) between the two chips, although there is evidence of saturation of the HG-U95 chip for more abundant cRNAs. Panel B. Shown is the same data as in Panel A, but with removal of potentially saturated probe sets, and after adjustment for different scaling factors (MuscleChip Avg Diff times 3.18). A correlation coefficient of 0.89 is observed.

Another difference between the MuscleChip and U95A data is seen for probe sets with the highest levels of hybridization; namely, a plateau effect for HG-U95A, while values continued to increase for the MuscleChip (Figure 1. Panel A). This increased dynamic range was attributed to newer production facilities used for the MuscleChip, with greater numbers of oligos successfully synthesized in each feature.

As a second form of validation, we compared "present" and "absent" calls for the same probe sets, using the same cRNA, on both the MuscleChip and HG-U95A, using both MAS 4 (data above), and the newly released MAS 5 versions of the Affymetrix algorithms. MAS 5 also has the advantage of eliminating negative Avg Diff values, and replacing them with simpler positive value signal metrics. Using MAS 4, the large majority (95%) of the 1,052 probe sets shared between the two chips showed the same "call", and also showed excellent correlation between absolute intensities. Using MAS 5, the number of present calls decreased for both U95A and MuscleChip; this is expected given the greater stringency of the newer algorithms for transcripts expressed at low levels, and thus fewer expected "false positive" present calls. This analysis showed a decrease in the number of "present" calls from 5,110 assigned by MAS4 to 4,427 "present" calls assigned by MAS5, of which 4,284 "present" calls were assigned by both algorithms. However, the correlation coefficient between the signal values obtained by the two algorithms was as high as 0.95. The high correlation between the average intensity signal obtained by the two algorithms supports the expected reduced number of false "present" calls. The increased stringency is also supported by the fact that 97% of the "present" calls detected by MAS5 is detected by MAS4, but only 87% of "the present" calls detected by MAS4 is also detected by MAS5. The balance between the number of false and true "present" calls is directly influenced by the balance between stringency and sensitivity that can be modified in the MAS5 software. Although the detection calls assigned by the MAS5 algorithms were somewhat different from the calls assigned by the MAS4 software it did not affect the correlations between the shared probe sets on HG-U95A and the MuscleChip. The correlation coefficient between shared probe sets on the HG-95A and the MuscleChip was 0.89 and 0.87 for the MAS 4 and MAS 5 analysis, respectively.

Correlation of EST clone number with absolute intensity (Avg Diff)

Clone EST number (clone frequency) in EST sequencing projects of non-normalized cDNA libraries is often considered to reflect the relative abundance of the transcript (e.g. "virtual Northerns" in dbEST). We hypothesized that hybridization intensities on the redundant Affymetrix GeneChip platform could provide a second means of studying relative transcript abundance. Such correlations are difficult with Affymetrix stock chips, as the probe sets are not derived from a specific non-normalized EST source, and correlations of EST number in specific dbEST cDNA libraries would be problematic. Furthermore, the majority of EST sequencing projects have used normalized or highly subtracted libraries.

As we had printed a large number of EST clusters from a non-normalized normal muscle EST data set on the MuscleChip, we selected the subset of probe-sets from this EST database, and correlated the number of clones in each EST cluster with the absolute intensities after hybridization to normal muscle. Two datasets were used in this analysis; one set from hybridization of pooled muscle samples from normal 6–9 years old boys as described previously [8], and one data set consisting of the average value of the hybridization intensities obtained from hybridizations to the MuscleChip with six individual samples from normal adults.

We found a reasonable correlation between clone number and Avg Diff values using the pooled samples described before [8], with a correlation coefficient of 0.67 (Figure 2, Panel A), and a slightly lower degree of correlation using the averaged intensity from the individual experiments, with a correlation coefficient of 0.6 (Figure 2, Panel B).

To determine possible explanations for "outliers" between clone number and Avg Diff, we examined 11 different outliers (chosen by visual inspection), as indicated in Figure 2, panel A. ESTs 1 through 8 showed a very high expression level by hybridization, yet a low EST clone number (Figure 2, Panel A). Seven of the eight high outliers corresponded to probe sets that were designed as either _f or _s probe sets, indicating high homology to additional EST clusters. Thus, these probes would typically be expected to show a high degree of cross hybridization to additional similar genes and ESTs. We therefore conclude that these outliers likely result from cross-hybridization to more abundant clusters.

We also investigated three "low" outliers, where clone number was unexpectedly high relative to absolute intensity (Avg Diff). All of these low-value outliers were _i and _r probe sets, which typically would be expected to show poor hybridization compared to _at-only probe sets. Thus the low hybridization was explained by "incomplete" and "rules dropped" probe set design for these particular probe sets. This data suggests that analysis is more accurate and reproducible if limited to _at probe sets; this is also acknowledged on Affymetrix stock GeneChips.

We then restricted the analysis of the averaged non-pooled samples (Figure 2, Panel B) to _at only probe set,

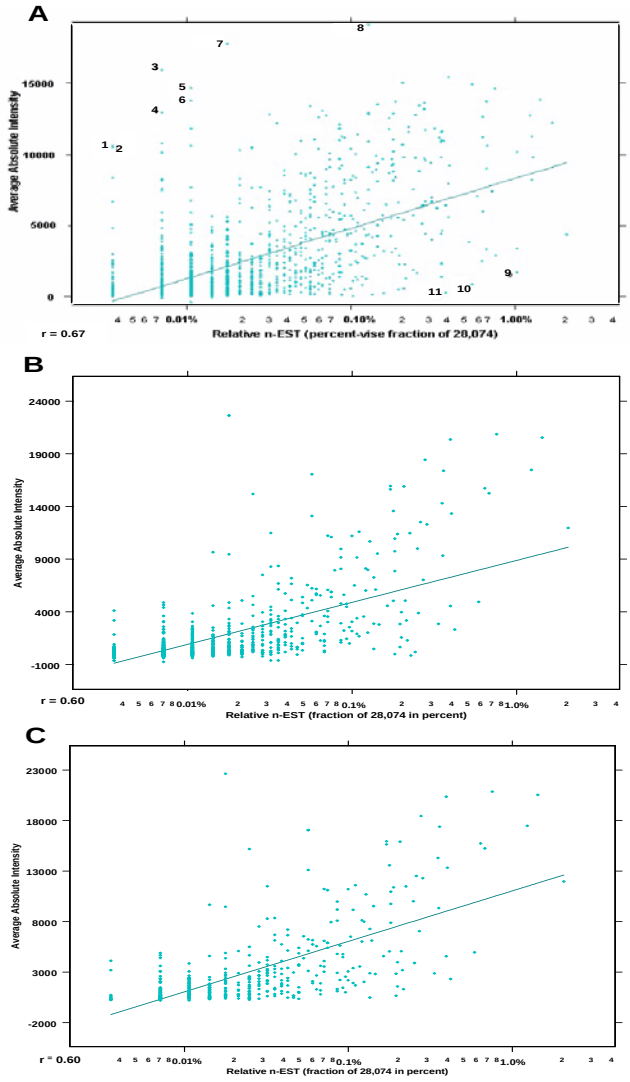

Figure 2

EST cluster member number correlation with hybridization intensity (Avg Diff) on the MuscleChip. Panel A. Shown is a log-linear scatter plot of Avg Diff (expression level of each RNA given by normalized hybridization intensity on the MuscleChip) versus the relative percentage of EST cluster members in a non-normalized normal muscle EST sequencing project (n-EST/cluster in 28,074 sequenced clones). The number of unique EST clusters shown is 2,052. The correlation coefficient of the log-linear regression is 0.67 indicating that, on average, there is a relationship between EST cluster member number and hybridization intensity. A small number of "outliers" from this analysis are indicated and numbered as described in the Results. The large majority of these outliers correspond to non-(_at) probe sets, where probe set design rules were altered or dropped. Thus, most outliers are likely due to cross-hybridization to other mRNA species. Panel B. Shown is a log-linear scatter plot of the average absolute intensity of six individual profiles (y-axis: average Avg Diff) versus the relative percentage of EST cluster members (x-axis: relative percentage of 28,074 on $\log_{10}$ scale) in a non-normalized normal muscle EST sequencing project. The number of unique EST clusters shown is 2,052. The correlation coefficient of the log-linear regression is 0.6 showing a correlation between number of sequences in EST clone and the averaged absolute intensity of individual samples hybridized to the MuscleChip. High outliers are non _at only probe sets. Panel C. Shown is a log-linear scatter plot of the average absolute intensity of six individual profiles (y-axis: average Avg Diff) versus the relative percentage of EST cluster members (x-axis: relative percentage of 28,074 on $\log_{10}$ scale) in a non-normalized normal muscle EST sequencing project. The EST clusters shown are reduced to ideal (_at only) probe sets. The correlation coefficient of the log-linear regression is 0.6. The difference in sensitivity between the two techniques may contribute to lower linear correlation in the intermediate and high abundant transcripts giving a correlation coefficient of 0.6 identical to the correlation when the non-ideal probe sets are included.

which showed an identical level of correlation with a correlation coefficient of 0.6 (Figure 2, Panel C). Indeed, the _at data shown in Figure 2, Panel C show neither high outliers for clones with low EST number or low outliers for clones with high EST numbers, suggesting that the design of "ideal" probe sets improves performance (specificity and sensitivity) as expected. Our data does not address the issue of whether EST clone number or hybridization intensity (Avg Diff chip values) is, or is not, a more accurate reflection of "true" transcript abundance.

Correlation of low abundance ESTs with "present" calls

To determine if low abundance ESTs (singletons. or duplex and triplex clusters) from a muscle EST sequencing project could be verified using expression profiling, we correlated the cluster member number (n-EST) with "present", "marginal" and "absent" calls using Affymetrix default interpretations of probe set hybridization patterns (Figure 3). For 777 clusters with 1, 2 or 3 EST clone members, we found 68% of triplex clusters to be "present" or "marginal", 46% of duplex clusters, and 30% of singletons (Figure 3, Panel A). Thus, there was a clear correlation between the clone number for relatively rare transcripts, and the ability of the custom probe sets on the MuscleChip to identify these transcripts as "present" calls based on hybridization of normal muscle cRNA (Figure 3, Panel A).

Use of the MuscleChip for verification of expression of anonymous ESTs

As the sequence list used to design and produce the MuscleChip was about a year old by the time of application, we updated all sequence definitions (Table 2), and then focused further analysis on expression data for the 369 novel EST clusters that showed no significant match to characterized genes (no name or function assigned, and no high homology to genes with function assigned). From these 369 novel EST clusters, 605 probe sets had been designed and printed. However, we were able to design probe sets fulfilling all rules (e.g. _at probe sets) for only 250/369 (68%) of the novel sequences, due to the limiting amount of sequence data for many of the singletons/clusters with low member number (Table 4). Of the 605 probe sets studied (all types), 285 were seen one or more times as "present" in the four profiles used for validation of the MuscleChip (pooled samples [8]). These 285 probe sets indicate that the corresponding genes are likely truly expressed in muscle, however this set included probe sets that could possibly cross-hybridize to other sequences (e.g. _g. _s. etc.; see Methods). Limiting this analysis to only probe sets fulfilling all stringency rules for unique probe set design (_at probe sets), we found 114 of 250 (46%) as "present" calls in all four profiles (Table 4). Thus, 114 novel EST clusters/singletons have a very high likelihood to be expressed in human muscle and represent novel expressed genes.

We correlated the 114 novel "present" calls with n-EST, and found that 89/114 (78%) were from clusters with only two or three members (duplex, triplex clusters) (data not shown). The lack of singletons in this set is due to the inability to design complete, non-overlapping, unique probe sets fulfilling all rules with only ~250 bp of sequence in these singletons.

Since we have shown that a correlation between low abundance transcripts and "present" calls exists (Figure 3, Panel A), we wanted to assess the correlation using only the subset of 250 EST with _at probe sets that showed no match to a characterized gene or protein, when analyzed by BLAST (Table 4). Out of these 250 ESTs, 195 EST clusters had 2, 3 or 4 EST clone members (Table 3). We found 81% of quadruple clusters to be "present" or "marginal", 72% of triplex clusters, and 45% of duplex (Figure 3, Panel B; Table 3), which clearly shows an increase in "present " calls with clone number.

To assign a higher level of confidence to the set of 369 novel expressed transcripts (novel hit EST clusters), the 250 ESTs associated with _at only probe sets were screened by "present" call in all of four profiles (pooled normal muscle [8]). Of the 114 ESTs that were selected in this fashion, the 102 that showed similarity to genomic, mRNA or hypothetical protein (RefSeq database) were queried against the Human Genome draft sequence by the BLAST-Like Alignment Tool (BLAT) algorithm using the Human Genome browser at UCSC [http://www.genome.ucsc.edu/][12,13]. BLAT alignments of length greater than 40 bases and identity higher than 95% were retained (Table 5). For each EST query the genomic sequences (BAC clone, contig or working draft clone) that the EST aligned to in either the BLAST or BLAT search is shown, as are BLAT identity and chromosome match for the best alignments (Table 5; for both best and the next best alignments see Additional File 1 – Table 1). Hypothetical protein mRNA hits from the RefSeq database were identified by BLAST search for 40 of the 102 ESTs and accession number is shown (Table 5; See additional file 1, Table 1, for table containing hotlinks to GenBank – ENTREZ at NCBI). 19 of the 40 RefSeq BLAST sequences did not match by BLAT search to the Genome Browser and the accession numbers are shown in parentheses (Table 5). The letters A, E, F, G in the gene prediction column indicates when one or more gene prediction programs have predicted a coding region or exon where the EST is aligned to the human genome sequence (see Methods, Validation of non characterized ESTs).

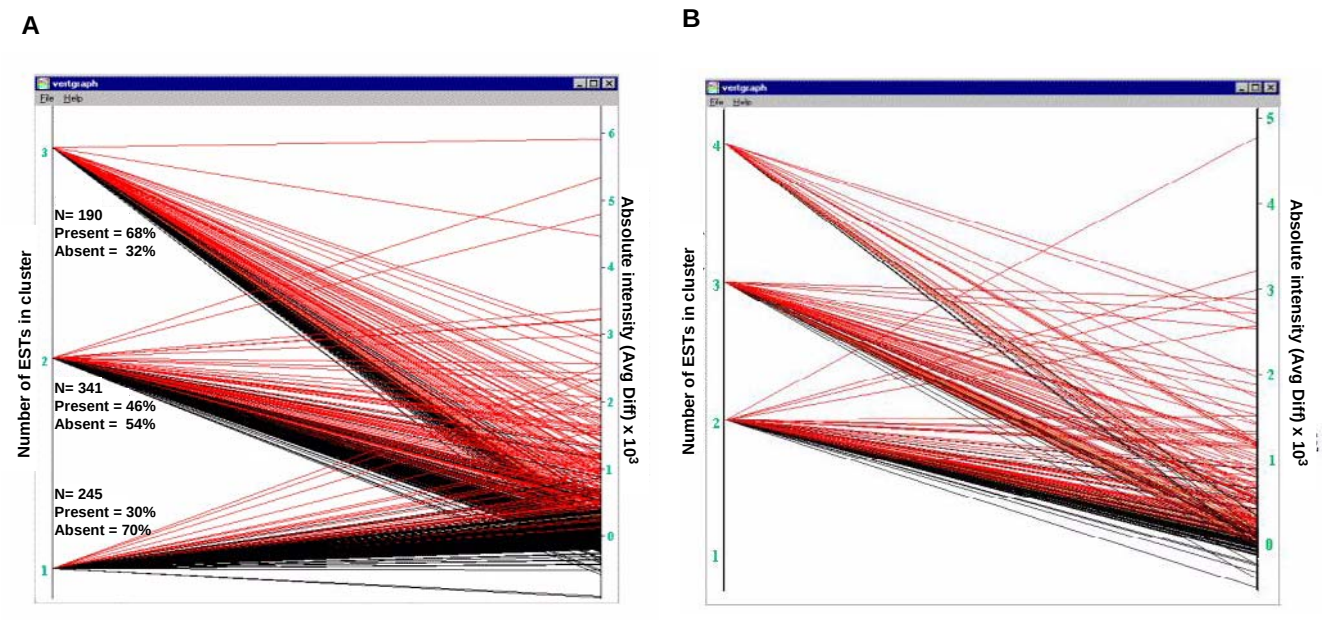

Figure 3
Hybridization to probe sets on the MuscleChip provides verification of singleton, and other low member EST clusters. Panel A. Graph output from "the Vertical Line" program showing unique EST singletons or clusters (duplex, or triplex) as individual lines emanating from the left axis (777 clusters total). Probe sets were designed for each cluster on the MuscleChip, and normal muscle cRNA hybridized to the MuscleChip. Using default Affymetrix algorithms, hybridization pattern to each probe set were assigned as "present" (red lines), "marginal" (yellow lines), or "absent" (black lines) (See Methods). There is a clear correlation with definition of "present" calls and EST cluster member number (see also Table 3). This result shows that 30% of EST singletons were verified as "present" in normal muscle RNA by this method. Panel B. Shown is a subset of EST that showed no match to a known gene, when analyzed by BLAST (Table 4). As in Panel A, "the Vertical line Program" is used to map the absolute intensity to clone number in EST cluster, using the default Affymetrix algorithms and hybridization pattern to assign each probe set a call as "present" (red lines), "marginal" (yellow lines), or "absent" (black lines). Out of 250 ESTs, 195 EST clusters had 2, 3 or 4 EST clone members. We found 81% of quadruple clusters to be "present" or "marginal", 72% of triplex clusters, and 45% of duplex showing increase in "present " calls with clone number.

Further support for the transcription of these 102 genes in muscle was provided by the analysis of an additional 33 MuscleChip expression profiles of human muscle (Table 5). Data on expression level, the average normalized hybridization intensity (Avg Diff) and the percentage of "present" and "marginal" calls of the 33 profiles are presented (Table 5). Similarly to previous described results (Figure 3, Panel A and Panel B) there was a correlation between Avg Diff hybridization intensities and percentage of "present" calls, with the higher Avg Diff values showing higher percentage of "present" calls (Table 5). This data provides robust verification of the expression of these 102 ESTs in human muscle with 40 ESTs genome anchored to transcriptional units by BLAT and BLAST search. The BLAT analysis also assigns priority of ESTs that maps to multiple regions of the genome by percent identity, suggesting that most multiple hits is a result of existence of pseudo genes or closely related gene families.

Table 3: Distribution of "present", "marginal" and "absent" calls for 195 _at "novel EST clusters"

# ESTs	Abs Call	% ESTs in cluster
n-EST = 2 (116 sequences)		
50	P	43.1 %
2	M	1.7 %
64	A	55.2 %
n-EST = 3 (58 sequences)		
39	P	67.2 %
3	M	5.2 %
16	A	27.6 %
n-EST = 4 (21 sequences)		
16	P	76.2 %
1	M	4.8 %
4	A	19.0 %

Table 4: "Present" calls for probe sets representing novel sequences.

Probe sets representing 369 novel sequences	605
Number of "Present" calls for 605 probe sets	285
_at probe sets in 369 novel sequences	250
_at probe that were "Present"	114
Number of non verified non-hit _at only probe set sequences	136

Discussion

Muscle is an ideal tissue for expression profiling in humans, as it is relatively easily accessible from both normal volunteers and neuromuscular disease patients, biopsied in relatively large quantities, and prepared for pathology in a manner that is ideal for expression profiling. However, the relatively poor sensitivity and specificity of existing expression profiling resources for human muscle are a limiting factor in conducting profiling experiments. In addition, muscle shows very highly specialized cell-type-specific expression of highly related gene families, where isoforms of specific proteins may differ by only a few amino acids. Thus, use of probe sets completely specific for individual gene family members is critical if the muscle transcriptome is to be understood (particularly with regards to different types of muscle conditioning/exercise).

For these reasons, we sought to create an expression profiling resource that would give good sensitivity and specificity for human muscle. We designed the MuscleChip so that it would provide good sensitivity in normal adult muscle for exercise and atrophy studies in volunteers (2,052 EST clusters from normal muscle), and pathological muscle (1,120 genes from differentially regulated genes in dystrophic muscle; [8]) (Table 1). We designed and printed 4,601 probe sets representing 3,344 unique genes/ESTs, and then validated this MuscleChip by comparison to HG-U95A GeneChip profiles. The MuscleChip appeared to have a greater dynamic range than the HG-U95A stock chip (Figure 1), despite the fact that they both shared the same feature size and synthesis methods. We also found that the typical target intensity of 800 used for whole-chip normalization of most Affymetrix stock chips was too high for the MuscleChip, due to the greater proportion of "present" calls (e.g. the tissue-specific nature of our chip) (Figure 1. Panel A). Consistent with previous reports, the probe sets fulfilling all previously described "rules" [4,10] performed most consistently.

We found relatively good correlation between hybridization of cRNA from normal muscle (absolute intensities), with cluster number (n-EST), both when analyzing all probe sets (Figure 2, Panel A), and then a subset of probe sets corresponding to low cluster number (singletons, duplex, and triplex clusters) (Figure 3, Panel A). Using all probe sets, there was a correlation coefficient of 0.67 between cRNA hybridization and EST clone frequency in a non-normalized cDNA library (Figure 2, Panel A). Neither method is likely entirely representative of the true frequency of each RNA species in normal tissue. EST clone number is probably skewed by the fact that the RNA was made from only a single individual (pectoral muscle from a woman undergoing a mastectomy), resulting in mRNA species differences due to polymorphic variability. The difference in sensitivity between these two techniques may contribute to lower the linear correlation, with correlation coefficient of 0.6, obtained between the averaged intensity value (intensity average from six individual profiles of normal adult muscle) and n-EST number (Figure 2, Panel B and Panel C). Also, while the cDNA was from an older female, the muscle biopsies used for expression profiling were all from males. Second, creation of cDNA libraries may skew the frequency of specific cDNA clones due to priming or cloning differences, or toxicity of certain cDNA sequences to the prokaryotic host. Hybridization intensities can vary due to differential cDNA synthesis, or cRNA amplification of muscle RNA. Moreover, incorporation of biotin labels, and hybridization kinetics, can lead to skewing of hybridization patterns. Given all these variables, the observed correlation of EST clone number in each cluster, and hybridization intensities, was considered encouraging.

One of the goals of this study using the MuscleChip was to use the hybridization patterns of probe sets as a means of confirming the expression in muscle of novel ESTs, and thereby confirming the novel ESTs as "genes". This analysis is particularly important when the EST clusters under

Table 5: Novel sequences validated by BLAT search against theworking draft of the human genome at UCSC (only first page is shown, please go to Additional Files, File 1 to see the full table)

BLAST/BLAT query Acc	BAC-/Genomic-/mRNA-clone	Hypothetical Protein mRNA (1)	Gene Prediction	BLAT 1st Identity	BLAT 1st Chr.Hit	Average value of Avg Diff	Std. Dev.	% P or M 33 arrays
F15766	AL158039.5	NM_032728.1	A, E, F, G	97.8%	9q34.2	757	239	97.0%
F15804	AC002351.1		G	98.6%	12q24.11	1570	843	93.9%
F15913	AL357374.11		A, E, F, G	99.7%	20q11.22	1611	623	100.0%
F15967	AC008763.7	(XM_058961.1)	A, E, F, G	98.2%	19p13.3	1125	285	100.0%
F16039	AP000348.1	(XM_059329.1	A, E, F, G	99.7%	22q11.23	6950	3513	100.0%
F16253	AC022307.14	NM_018045.1	A, F, G	99.1%	1p35.1	330	217	84.8%
F16384	AC096677.2		A	99.2%	1q32.3	1352	725	87.9%
F16396	AC092664.2		A	99.7%	2q13	2412	685	100.0%
F16592	AC023415.3			100.0%	13q31.2	882	614	100.0%
F16715	AL031727.42	(XM_059093.1)	A	99.5%	1p36.13	2653	532	100.0%
F16733	AC092069.2		A, E, F, G	98.6%	19p13.2	549	139	97.0%
F16774	AC092636.3		A	100.0%	2q11.2	2061	1554	100.0%
F16938	AC021165.4		F, G	99.0%	19p13.3	1722	422	90.9%
F16970	AC068700.5		A	100.0%	8q21.13	333	195	90.9%
F17157	AL512652.18		A	98.1%	13q12.11	538	297	87.9%
F17214	BC015836		A	100.0%	3q25.2	4572	2296	100.0%
F17216	AP000512.1		G	98.2%	6p21.33	473	120	97.0%
F17258	AC005568.1		A	97.7%	16p13.3	2515	553	100.0%
F17715	AL353689.18	NM_020247.1	A, E, F	99.3%	1q42.13	2744	1124	100.0%
F17797	AC013603.14	NM_024025.1	A, E, F	96.1%	8p12	1236	635	100.0%
F18177	AC004643.1	NM_017885.1	A, E, F	99.5%	16p13.3	2164	900	100.0%
F18188	Z63603.1		A	99.1%	12q24.31	618	322	100.0%
F18468	AL138796.6		A, E	99.3%	1q21.1	8300	1981	100.0%
F18648	AC004985.2	(NM_032014.1)	A, E, F, G	100.0%	7p13	2018	577	100.0%
F18791	AL359079.15	NM_017698.1	A, F	99.8%	Xq23	598	194	78.8%
F18796	AC010271.7		A	98.6%	19q13.13	308	175	81.8%
F18838	AC011380.5	NM_032412.1	A	100.0%	5q31.3	1410	550	100.0%
F18952	AC027674.9		A	99.2%	10q11.22	1207	357	97.0%
F19033	AP003419.1	(XM_040083.1)	A, E	100.0%	11q13.2	662	148	97.0%
F19110	AC079814.9		G	99.0%	3p21.31	14508	4317	100.0%
F19200	AB020863.1		A	100.0%	8p22	817	378	84.8%
F19351	AL590128.2	(AY037153.1)	A, E	99.7%	1p36.31	2326	1581	75.8%
F20210	AL137281.1		A, G	98.1%	11q21	1641	913	100.0%
F20217	AC007687.16		A	100.0%	3q26.33	1009	774	93.9%
F20427	AP000803.2		A	100.0%	11q13.1	1022	367	75.8%
F20554	AL591408.3	NM_032747.1	E, F	99.1%	10q24.33	1770	442	100.0%
F20576	AC007225.2		A	99.8%	16q11.2	48	52	39.4%
F20639	AC004230.1		A	100.0%	11q11	3225	1134	100.0%
F20736	AC027674.9		A	100.0%	10q11.22	372	159	97.0%

(1) Entries in parentheses are hypothetical protein mRNA RefSeq hits found by BLAST to nr database (ENTREZ nucleotide database, NCBI) but not by BLAT-search in the Human Genome browser [http://genome.ucsc.edu].

study contain one, or only a few, cluster members (e.g. EST singletons). Such EST clusters with very low cluster members are generally viewed with suspicion, even when confirmed in genomic sequence, due to the possibility of genomic contamination in cDNA libraries, the normalization processes used for many cDNA library constructions, and/or "illegitimate transcription" that could lead to sequencing of EST clones that are not truly "expressed" in the tissue under study. An advantage of the MuscleChip is that relatively large probe sets are designed against each gene, typically containing 16 gene-specific perfect match 25 mers tiled against the 3' end of the mRNA/EST, and an additional 16 paired mismatch probes with a single nucleotide change in the center of the oligo. This allows hybridization patterns across the entire probe set to be analyzed, leading to a more quantitative and specific analysis of expression of the corresponding mRNA in the tissue. Three different algorithms are used to analyze hybridization patterns, and a threshold for each algorithm used to determine whether the corresponding cRNA is "present" or "absent" in the RNA under study [4,10]. Using the default algorithms and associated thresholds, we showed that 285 of 605 (47%) probe sets corresponding to 369 novel EST clusters/singletons were determined to be expressed ("present") in normal muscle (Table 4). A number of these probe sets did not fulfill all "rules", due to the limit-

ing amount of sequence data available for EST singletons; limiting the analysis to ideal _at probe sets showed that 114 of 250 (46%) probe sets corresponding to novel EST clusters were "present" in 4 out of 4 control profiles (Tables 4), and 157 of 250 (63%) were "present" in 1 control out of 4 profiles (data not shown). We further limited the study set to 369 of the original 734 novel ESTs, which by recent BLAST update (March 2002) showed no high similarity to a characterized gene, according to the thresholds described in Methods. The 369 ESTs are represented by 605 probe sets on the MuscleChip, of which 250 are ideal _at probe sets. Of the 114 probe sets that were validated by "present" calls in four out of four pooled controls (Table 4), 102 showed similarity to genomic, mRNA (RefSeq database) or hypothetical protein.

To anchor the transcript units on the human genome sequence, we used the sequences of 102 "present" ESTs in a BLAT search against the human genome draft sequence (Table 5, [12,13]). The BLAT search resulted in significant alignment to "translated RNA" sequence in 40 of the cases (35%) and to genomic DNA in all but two cases. Two ESTs did not match any part of the draft sequence, although a significant alignment to IMAGE mRNA clones from the nr database at NCBI base was verified. Both ESTs showed relatively low average intensities in 33 MuscleChip experiments; however 89% and 100% of the profiles showed "present" calls for the two ESTs respectively (Table 5).

The 102 expression validated novel ESTs represent 0.31% of the ~33,000 genes on the new human genome HG-U133 Affymetrix array set. The HG-U133A and HG-U133B array together have approximately 44,000 probe sets querying about 33,000 unique genes/ESTs. We have found that the HG-U133A has a sensitivity to muscle of 40–45% "present" calls, corresponding approximately 9,000 to 10,000 probe sets, and the HG-U133B shows a sensitivity of 15–20% "present" calls, corresponding to approximately 3,800 probe sets, which in combination gives approximately 13,000 transcripts. Thus, the 102 uncharacterized or novel ESTs represent ~0.75% of transcribed muscle genes that are not represented on the stock arrays.

Conclusion

We have designed a muscle specific oligonucleotide chip containing 4,601 probe sets representing 3,344 unique genes/ESTs. The performance of the MuscleChip was validated by comparison of shared probe sets from expression profiles to HG-U95A stock chip. This comparison showed a linear correlation of intensities between the two types of arrays, although the MuscleChip was shown to have a higher dynamic range of intensities before reaching the saturation level.

The MuscleChip was used to validate ESTs from a non-normalized cDNA library of human muscle. Correlation of intensity values from the MuscleChip expression profiles with EST clone number in each EST cluster showed a correlation coefficient between 0.60 and 0.67 when restricting the analysis to probe sets fulfilling all probe design rules and representing only unique transcripts (_at probe sets).

Furthermore, we have shown that the MuscleChip can be used as a means of confirming the expression in muscle of novel ESTs, and thereby confirming the novel ESTs as "genes". We found 102 EST clusters to be expressed in human muscle, transcriptionally validated by 33 MuscleChip expression profiles, and genome-anchored by BLAST or BLAT search to the Human Genome sequence. 40 of the ESTs clusters were furthermore verified by alignment to potential transcriptional units in the RefSeq database.

Methods

Sequence definitions for MuscleChip

The MuscleChip was designed to contain 4,601 probes representing 3,344 sequences from four different sources. 2,052 sequences were from EST clusters from a non-normalized skeletal muscle sequencing project [7] (see [http://muscle.cribi.unipd.it] for description of the muscle EST sequencing project), 1,052 sequences from the HG-U95A Affymetrix stock chip, 68 sequences were from the HuFL array (the first edition human Affymetrix stock chip), and 172 sequences were of special interest to the laboratory (see [http://microarray.cnmcresearch.org/musclechipindex.asp] for spreadsheets (MuscleChip EASI) of all gene/EST names, accession numbers, and probe set designators, and (MuscleChip Probe Sequences) file for the sequences of all probe sets used). The MuscleChip was designed to be highly redundant, both between the commercial GeneChip and the custom MuscleChip, and within the MuscleChip itself, with many ESTs being represented by multiple probe sets (Table 1). This was done to facilitate validation of the MuscleChip. The 1,120 sequences from HG-U95A and HuFL respectively are categorized as "commercial sequences" and are represented on the MuscleChip with the same probes as on the Affymetrix stock chips. In order to select the commercial probes, eight expression profiles were generated using biopsies from five male patients with primary dystrophinopathy (Duchenne Muscular Dystrophy, DMD), four patients (two male and two female) with α-sarcoglycan dystrophy (α-SG), and five age-matched male controls (Con) as previously described [8] with an additional two profiles from five female age-matched controls (ConF; data not shown). Probe sets that were assigned a "present" call, showed differential expression in comparative analysis (assigned a diff call by Affymetrix GeneChip software [10]), and had an average fold change greater than 2.0 were considered

for inclusion on the MuscleChip. Duplicate records were removed, as was overlap with the custom EST sequences, resulting in 1,052 (of 5,513 tested on the HuFL Chip) commercial sequences selected for representation on the MuscleChip. Due to the upgrade from HuFL to HG-U95A (Human Genome Unigene 95 build), which resulted in new probe designs for some sequences, new probe sets on the HG-U95A array (16 probe pairs per probe set) were picked to replace each HuFL probe set (20 probe pairs per probe set). We used a subtracted non-normalized human muscle EST database with 28,074 sequenced ESTs ([7] and unpublished data) to design and print probe sets corresponding to 2,052 EST clusters (1,341 defined genes, and 734 undefined ESTs). ESTs were from the 3'end of the corresponding mRNAs. The library was subtracted using a filter hybridization procedure to identify and remove the 10 most abundant mRNAs in the library, in order to minimize the number of redundant transcripts to sequence [7]. The 10 genes used for the filter hybridization were identified from a preliminary sequencing of 1,054 randomly selected EST clones [7]. These 10 most abundant transcripts accounted for more than 45% of the clones in the non-normalized library. 28,074 additional clones were sequenced after subtraction, clustered using TIGR Assembler program, compared to the Unigene 90 build for gene identification (>95% identity for a match assignment), and then manually checked for incorrect clustering or sequence identities. Out of the 2,052 EST clusters represented on the MuscleChip, 734 represented clusters that had shown no identity to previously characterized clusters or genes in the Unigene build 90 (Table 1). These 734 clusters showed EST member numbers varying from 1 to 579 (n-EST = 1 to n-EST = 579). To update the probe descriptions, and also to correct for inappropriately defined singletons by the Assembler program [9], we tested the consensus sequence for each of the 734 previously undefined ESTs against Unigene 2002 builds using Netblast cl3 (Table 2). A query sequence was defined to have a low significant match (hit) for alignments with identities < 100 bp and a significant match for alignments with identities > 100 bp. A match was considered indicative of **identity** with the search sequence when E-score < 10^{-60}, indicative of **similarity** to the search sequence when 10^{-60} < E-score < 10^{-34}, and to show no high similarity when E >10^{-34}. New gene definitions were found for 365 of the 734 EST clusters in Unigene and RefSeq, leaving 369 as undefined, novel EST clusters (Table 2).

A list of all genes, EST clusters, and accession numbers is available on our web site ([http://microarray.cnmcresearch.org/musclechipindex.asp].

MuscleChip production

Probe sets were chosen from the 650 base pairs at the 3' end of each gene or consensus sequence. Sixteen 25-mers covering a total of 400 bp were selected and aligned against all other probes represented on Affymetrix stock arrays in order to prune the dataset and to give a "measure" of similarity and identity among Affymetrix Human Genome probe sets (~70,000). Furthermore, each probe set was characterized according to the type of sequence it "describes" by an identifying underscore label (_at, _s, _g, _f, _r, and _i; EASI™ Expression Analysis Sequence Information Database). All probe sets on Affymetrix chips are synthesized in the sense orientation, and the target cRNA in anti-sense orientation (indicated by the _at label on probe sets). If no additional underscore label has been assigned, the probe set corresponds to a single gene or the consensus sequence of a cluster of sequences (ESTs). For probe sets that could potentially cross-hybridize to a small number of sequences, a similarity constraint (_s) was assigned. For probe sets corresponding to one or more EST clusters that were similar but not identical (including polymorphisms, clusters of genes or ESTs, overlapping and non-overlapping) or from regions where it was not possible to pick a full probe set (16 probes), a _f (for sequence family) or a _g (for common groups) was assigned. For EST clusters where the number of cluster members was only one (singletons), or other low numbers, then restraints on the amount and accuracy of sequence data available often forced use of non "ideal" probe sets. In these cases it was not possible to design a unique set of probes while applying all Affymetrix probe rules. Such probe sets were assigned a _r (rules dropped) or _i (incomplete) designator if there were fewer than 15 probes in the probe set. Of the 369 novel EST clusters printed on the chip, only 250 (68%) were represented by at least one _at probe set on the MuscleChip (see Results).

In total, approximately 75,000 perfect match probes and ~75,000 mismatch probes were synthesized by photolithography and photo-activated chemistry using 24 μm^2 features. Chips were synthesized on 90 chip wafers, then divided and packaged into standard Affymetrix GeneChip format.

RNA isolation, cRNA production

RNA isolation, purification, cDNA production, and synthesis of biotinylated cRNA were as previously described [8]. Quality control measures for cRNA and chip hybridization are provided on our web site [http://microarray.cnmcresearch.org/pgaoutline-qcofsamples.asp].

Data analysis

The data analysis of the arrays was done using the standard metrics as implemented in Affymetrix MicroArray Suite 4.0 [4,10]. Briefly, the hybridization signal for each sequence is queried by probe sets of 16 probe pairs (a 25-mer perfect match (PM), and a 25-mer mismatch probe (MM)). Using the Affymetrix metrics, an absolute intensi-

ty (normalized hybridization intensity) and an absolute call ("present", "absent" or "marginal") was assigned to each probe set. The absolute intensity, in Affymetrix terminology called the Average Difference (Avg Diff), is the average value of the PM minus the MM in each probe pair in a probe set. By combining three standard algorithms used to calculate values for the internal ratios of PM to MM to signal strength in each probe pair, the absolute call of "absent", "marginal" or "present" is assigned based on whether or not any one of the three algorithms meet threshold values (Affymetrix MicroArray Suite 4.0 User Guide).

Validation of non-characterized ESTs

In order to validate the ESTs represented on the chip (2,052 probe sets), the ESTs were grouped and ranked according to the number of sequences that constituted each particular EST cluster, with the number of ESTs ranging from 1 (n-EST = 1) to 579 (n-EST = 579). To determine the distribution of absolute intensities within each of the 96 EST clusters (1 to 579 ESTs in each cluster), we designed a program in C++ called "Vertical Line" that can be used to visualize the absolute intensity as well as the absolute call, using a variety of color-coding schemes. The program is available on our website [http://microarray.cnmcresearch.org/resources.htm], click "vertical line executable" to save an executable of the program.

To assign higher confidence to the set of 369 EST clusters that showed no high similarity to genes with known name or function by BLAST search to the non redundant (nr) sequence database at NCBI, the 250 ESTs associated with _at only probe set were filtered by "present" call in all of four pooled normal muscle [8]. Of the 114 ESTs that were selected, the 102 that showed similarity to genomic, mRNA, or hypothetical protein (RefSeq) by BLAST search, were BLAT searched against the Human Genome draft sequence using the UCSC Human Genome Project Working Draft, [http://www.genome.ucsc.edu/][12,13]. BLAT alignments of length greater than 40 bases and identity higher than 95% are defined as a match [12]. For each EST query, the accession number of the best BLAT alignment to genomic and mRNA sequence is shown along with the chromosome match and percentage of identity for both the best and the next best alignments (Table 5). In the cases when the original EST BLAST hit were a genomic sequence, the hit was retained only if it was identical to the BLAT alignment, otherwise the accession number shown in the BAC-/ Genomic-/ mRNA-clone column is for the best genomic DNA BLAT hit (Table 5). In the other cases where the best original BLAST hit was a hypothetical gene or protein, the accession number is shown next to the genomic BLAT accession. Entries in parentheses are hypothetical protein mRNA RefSeq hits found by BLAST to nr database (ENTREZ nucleotide database, NCBI) but not by BLAT-search in the Human Genome browser [12,13].

When any of four gene prediction programs predicted an open reading frame or exon in the region of the human genome sequence that overlapped with the EST alignment, letter codes corresponding to the particular program used is shown in the table. An Assembly gene prediction is defined by A. Ensembl gene prediction by E, FgenesH++ prediction by F and GenScan gene predictions by G. See Additional File 1 for excel-sheet version of Table 5 containing additional results on next best BLAT hits and direct hotlink to ENTREZ nucleotide database, NCBI.

Authors' Contributions

RHAB was responsible for all microarray analysis, probe design, chip validation analysis and was the principal author. ST, GL and GV were responsible for all work regarding construction of cDNA library. YWC did all total RNA extractions, sample preparation and Chip hybridizations. TMT programmed C++ executable "Vertical Line" used in the analysis. EPH conceived of the study, and participated in its design and coordination. All authors read and approved the final manuscript.

Additional material

Additional File 1

Excel sheet showing the full version of Table 5. BLAT results for all 102 entries, plus additional columns showing the next best BLAT hits are shown. Accession numbers are formatted as direct hotlink to ENTREZ nucleotide database, NCBI.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-3-33-S1.xls]

Acknowledgements

Supported by grants from the National Institutes of Health (5RO1 NS29525-10; and a "Programs in Genomic Applications" from NHLBI [UO1 HL66614-01] and HOPGENES) to EPH, and a Telethon Italy grant to GL and GV.

References

1. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: **Expression profiling using cDNA microarrays.** *Nature Genetics* 1999, **21:**10-14
2. Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G: **Making and reading microarrays.** *Nature Genetics* 1999, **21:**15-19
3. Southern E, Kalim M, Shchepinov M: **Molecular interactions on microarrays.** *Nature Genetics* 1999, **21:**5-9
4. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: **High density synthetic oligonuclotide arrays.** *Nature Genetics* 1999, **21:**20-24
5. Knight J: **When the chips are down.** *Nature* 2001, **419:**860-861
6. Halgren RG, Fielden MR, Fong CJ, Zacharewski TR: **Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones.** *Nucleic Acids Res* 2001, **29:**582-588
7. Lanfranchi G, Muraro T, Caldara F, Pacchioni B, Pallavicini A, Pandolfo D, Toppo S, Trevisan S, Scarso S, Valle G: **Identification of 4370**

expressed sequence tags from a 3'-end-specific cDNA library of human skeletal muscle by DNA sequencing and filter hybridization. *Genome Research* 1996, **6:**35-42
8. Chen Y-W, Zhao P, Borup R, Hoffman EP: **Expression profiling in the muscular dystrophies: identification of novel aspects of molecular pathophysiology.** *Journal Cell Biology* 2000, **151:**1321-1336
9. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quakenbush J: **An optimized protocol for analysis of EST sequences.** *Nucleic Acids Research* 2000, **28:**3657-3665
10. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL: **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nature Biotechnology* 1996, **14:**1675-1680
11. **Affymetrix MicroArray Suite User Guide. Version 4.0.** *Appendix 4 and 5: 353–361. Affymetrix 2000*
12. Kent JW: **BLAT – The BLAST-like alignment tool.** *Genome Research* 2002, **12:**656-664
13. Kent JW, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The Human Genome Browser at UCSC.** *Genome Research* 2002, **12:**996-1006
14. Bakay M, Chen Y-W, Borup RHA, Zhao P, Nagaraju K, Hoffman EP: **Sources of variability and effect of experimental approach on expression profiling data interpretation.** *BMC Bioinformatics* 2002, **3:**4