An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression

David N. Messina
Washington University School of Medicine in St. Louis

Jarret Glasscock
Washington University School of Medicine in St. Louis

Warren Gish
Washington University School of Medicine in St. Louis

Michael Lovett
Washington University School of Medicine in St. Louis

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs

Recommended Citation
https://digitalcommons.wustl.edu/open_access_pubs/2092

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact engeszer@wustl.edu.
An ORFeome-based Analysis of Human Transcription Factor Genes and the Construction of a Microarray to Interrogate Their Expression


Genome Res. 2004 14: 2041-2047
Access the most recent version at doi:10.1101/gr.2584104

Supplemental Material
http://genome.cshlp.org/content/suppl/2004/09/30/14.10b.2041.DC1.html

References
This article cites 39 articles, 23 of which can be accessed free at:
http://genome.cshlp.org/content/14/10b/2041.full.html#ref-list-1

Creative Commons License
This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

Email Alerting Service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.
An ORFeome-based Analysis of Human Transcription Factor Genes and the Construction of a Microarray to Interrogate Their Expression

David N. Messina,¹ Jarret Glasscock,¹ Warren Gish, and Michael Lovett²

Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110, USA

Transcription factors (TFs) are essential regulators of gene expression, and mutated TF genes have been shown to cause numerous human genetic diseases. Yet to date, no single, comprehensive database of human TFs exists. In this work, we describe the collection of an essentially complete set of TF genes from one depiction of the human ORFeome, and the design of a microarray to interrogate their expression. Taking 1468 known TFs from TRANSFAC, InterPro, and FlyBase, we used this seed set to search the ScriptSure human transcriptome database for additional genes. ScriptSure’s genome-anchored transcript clusters allowed us to work with a nonredundant high-quality representation of the human transcriptome. We used a high-stringency similarity search by using BLASTN, and a protein motif search of the human ORFeome by using hidden Markov models of DNA-binding domains known to occur exclusively or primarily in TFs. Four hundred ninety-four additional TF genes were identified in the overlap between the two searches, bringing our estimate of the total number of human TFs to 1962. Zinc finger genes are by far the most abundant family (762 members), followed by homeobox (199 members) and basic helix-loop-helix genes (117 members). We designed a microarray of 50-mer oligonucleotide probes targeted to a unique region of the coding sequence of each gene. We have successfully used this microarray to interrogate TF gene expression in species as diverse as chickens and mice, as well as in humans.

[Supplemental material is available online at www.genome.org.]

1These authors contributed equally to this work.
2Corresponding author.
E-MAIL Lovett@genetics.wustl.edu; FAX (314) 747-2489.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2584104.
We were able to match 1361 of our 1468 TF sequences with ScriptSure clusters (92.7%). Thus, we failed to identify 107 genes out of our seed set within ScriptSure. This occurred because the version of ScriptSure we used was built on a draft genome assembly (International Human Genome Sequencing Consortium 2001) in which there were still gaps. However, this number is useful, because it provides us with an estimate of the rate of false negatives in our overall BLASTN analysis (7.3%).

Two methods of searching ScriptSure were used: a high-stringency BLASTN (http://blast.wustl.edu) of ScriptSure with each seed TF, and a query of each ScriptSure transcript, conceptually translated, against a collection of TF DNA-binding hidden Markov models (HMMs) extracted from Pfam (Fig. 2). We reasoned that the combination of these two approaches would provide a balance between sensitivity and specificity. The BLASTN search alone would yield false positives, and requiring each candidate to have a TF’s DNA-binding domain would alleviate that problem. The conservative combination of these two search methods will, however, result in some false negatives (discussed below) and an overall slight underestimate of the total TF gene content.

A repeat-masked version of the 1468 TF seed sequences was used as a BLASTN query against the complete ScriptSure database. The query identified 5130 potential new TF candidate clusters that met our criteria. The bit score cutoff used in the analysis was determined from our analysis of a bit score distribution of HOX gene family members (see Methods). We next filtered out clusters that contained no introns (possible spurious alignments to processed pseudogenes) and clusters that were represented by only one or two aligned ESTs. The number of newly identified clusters that passed through this filter was 3338.

Further Selection Through TF Pfam Signatures

We were concerned that our ScriptSure BLASTN searches would yield false positives from homologies within conserved domains that are not TF-specific, such as protein–protein interaction domains. Therefore, we sought to select the clusters that contained bona fide TF protein motifs. The database used in this analysis was a subset of Pfam, containing DNA-binding domains found in TF genes (see Methods). We searched the entire set of “Spliced / Multiple” ScriptSure clusters (i.e., >2,200 clusters and not just the 3338 clusters identified by BLASTN) for Pfam TF DNA-binding motifs and found 3748 clusters that contained at least one TF protein motif. These clusters included 1236 out of the 1369 seed set present in ScriptSure, indicating that this Pfam method has a false-negative rate of ~10%. Therefore, the Pfam search identified 2512 new putative TFs. Because the HMM used in our Pfam search used a relatively low cutoff for motif similarities, we expected to also detect false positives by this route. To
remove these and the BLASTN false positives, we tested for overlaps between the results from the two search methods. The overlap between the 3338 BLASTN clusters and the 2512 Pfam clusters comprised 494 clusters (Fig. 2). These 494 newfound clusters, plus the 1468 seed set constituted a total of 1962 potential TF gene sequences. Interestingly, when all of the ESTs comprising the 1962 potential TF sequences were compared with all spliced EST clusters from ScriptSure, we found that the medians of the two distributions were significantly different. The putative TF set had a threefold higher level of alternative splice forms when compared with the non-TF set, suggesting that extensive potential isoform diversity may be encoded by this set as a whole. These genes and their accession numbers are listed in the online Supplemental material.

**Genomic Localizations of TF Genes**

Tight clustering of genes is sometimes indicative of coregulated gene expression (Boutanaev et al. 2002; Lercher et al. 2002). For TFs in particular, there are precedents for biologically significant clusters of HOX genes on human chromosomes 7, 17, 12, and 2. We sought to determine if the newly identified set of putative TF genes were distributed randomly throughout the human genome or were found to cluster at discrete chromosomal locations.

Because heterochromatic and euchromatic regions of the genome are known to be relatively gene-poor and gene-rich, respectively, apparent clustering of genes was expected (International Human Genome Sequencing Consortium 2001). Taking this into consideration, we measured how often we observed three or more TF genes appearing in a window of eight clusters, translating to a probability of 0.37 under a binomial model (Fig. 3). This analysis identified 29 regions that passed the criteria. Four of these regions were Hox clusters, and another nine clusters were attributed to zinc finger clusters on chromosome 19 (19p12, 19q13.2, and 19q34; Eichler et al. 1998). The remaining 16 clusters had little underlying annotation of their transcript members. Interestingly, a comparison of the 29 clusters with the completed mouse genomic DNA sequence revealed that 18 out of the 29 were conserved in the mouse, supporting the notion that there may be functional reasons for some of this clustering. For a list of the genes comprising each cluster, see Supplemental materials.

**Microarray Design**

From the set of 1962 TF genes, we designed a microarray of 50-mer oligonucleotide probes with which to interrogate their expression. One of the major issues in designing a microarray of TFs is that many of these genes fall into families that share significant regions of conserved sequence homology. For example, there are >500 TFs that contain zinc finger domains (Eichler et al. 1998; this study, see below). To design a probe that will measure the expression of only one gene, it was necessary to identify sequence regions in each gene that were unique to it. The obvious choice in this situation is to target the 3′ untranslated region (3′ UTR), which is usually the most evolutionarily divergent region in a transcript. However, our intention in building a TF microarray was to use it across species, at least for organisms that are evolutionarily close enough to retain a high degree of sequence similarity, such as the mouse and the chicken. Therefore, we chose to design 50-mer oligonucleotide probes from within each coding region (determined by conceptual translation of each putative TF) and as 3′ as possible within the coding sequence. It should be noted that designing probes far away from the 3′ ends of genes may result in a significant loss of sensitivity when used with 3′ biased amplification protocols. One way to circumvent this potential limitation is to use alternative amplification methods such as full-length amplification (Castle et al. 2003). We also selected these probes to be matched for Tm (Li and Stormo 2001). These probes are listed in Supplemental materials. This array has been successfully used to interrogate TF gene expression across species as distant as chicken, mouse, and man (Hawkins et al. 2003).

**DISCUSSION**

In this study we collected a set of known human TFs and used two complementary computational methods to search the hu-
man transcriptome for the entire set of human TF genes. Our analysis identified 1962 putative TF genes, a number that correlates well with previous estimates (International Human Genome Sequencing Consortium 2001; Venter et al. 2001). This number is an estimate, but our seed set searches provide us with some idea of the error rates in that estimate. To count a gene as a TF, we required that it must either be previously annotated or described as such in the literature (the seed set), or be paralogous to a known TF gene and contain a DNA-binding domain known to occur exclusively or primarily in TFs (the found set). The term “transcription factor” encompasses many types of proteins. Those factors that do not bind directly to DNA are likely to be underrepresented in our analysis. However, our seed set did contain some non–DNA-binding TFs that had been experimentally verified and described in the literature.

There are four variables that influence our estimate of total TF genes: gaps in the human genome sequence, the incompleteness of Pfam, the degree of comprehensiveness of dbEST, and the exclusion of single-exon predicted genes from our analysis. Our finding that ~7% of our 1468 input seed set failed to find matches in the gene set is the second largest family, with over one-third of all human TF genes. Homeobox-containing Zinc-binding TFs constitute the largest class, amounting to 130 TFs that have been expounded upon in many recent studies in which "evolvability" (Kirschner and Gerhart 1998) and the evolution of development (Jacob 2001; Revilla-I-Domingo and Davidson 2003; Wray 2003) have been linked to increased complexity in regulatory networks. Our observation that humans devote at least 8% of their ORFeome to primary regulators of transcription (a number that is probably an underestimate given the levels of alternative splicing in TF genes) is consistent with the idea that developmental and body plan complexity are related to complexity in transcriptional regulatory networks.

The TF microarray we describe here is a versatile tool widely applicable to many areas of biological research. Two key features distinguish it from other microarrays. First, no other array described to date has probes for measuring the expression of as many TF genes. The Affymetrix U133 Genechip set, a large and commonly used human gene microarray, contains probes for ~85% of the TF genes represented on this array (C. Helms, D. Messina, and M. Lovett, unpubl.). Second, although our probes were designed by using human DNA sequences, they represent coding sequences and not 3' UTR. Thus, this microarray can (and has) been used to successfully measure TF gene expression in other species, including mouse and chicken (Hawkins et al. 2003). The evolutionary distance to the last common ancestor of human and chicken is ~310 million years (Ureta-Vidal et al. 2003). Based on this distance, we would expect the TF array to be useful for studies in many vertebrate species, including chimpanzee, rhesus monkey, rat, dog, cat, horse, pig, cow, and sheep. A comparison of a random sampling of 50 TF genes from chicken and zebrafish to our collection of oligonucleotides revealed an average of 84% nucleotide identity for chicken and 79% nucleotide identity for zebrafish (diverged by 450 million years from human). Thus, this array may also prove useful for more divergent species such as zebrafish, pufffish, and frog. However, we would urge caution in applying this tool to species more diverged than chicken. In these cases the rate of false negatives will increase (i.e., 50-mer oligonucleotides that fail to match their orthologous gene) and decreased hybridization stringencies will lead to an overall compression of dynamic range. Careful validation steps, and tuning of hybridization conditions, are required in all of these cross-species applications.

METHODS

Seed List

To build an initial set of TFs for the array, we gathered records from TRANSFAC (version 4.09-public; Wingender et al. 2000). The version of TRANSFAC we used did not have references to commonly used sequence identifiers, such as SWISS-PROT or GenBank sequence records. Therefore, we took gene names and descriptions from TRANSFAC records and correlated them by hand with GenBank records, from which we were able to obtain a mixture of genes, mostly members of the core transcriptional machinery, and the “structural” category includes genes that are thought to regulate transcription by altering chromatin structure, such as the HMGA family (for review, see Reeves 2001).

Looking at the data listed in Table 1 in another way, we can estimate the proportion of the total transcriptome that encodes TFs in each of the four species (Table 2). We do not know the precise gene content of all of these species, but assuming that current numbers are approximately correct, it appears that TFs account for between 8% and 9% of all human genes. As one might expect, we see an upward trend in the proportion of TFs as the complexity of the organism increases. The development of more, and more finely tuned, regulatory mechanisms in higher eukaryotes has been hypothesized to explain their greater biological sophistication (Huang et al. 1999; Claverie 2001). This idea has been expounded upon in many recent studies in which "evolvability" (Kirschner and Gerhart 1998) and the evolution of development (Jacob 2001; Revilla-I-Domingo and Davidson 2003; Wray 2003) have been linked to increased complexity in regulatory networks. Our observation that humans devote at least 8% of their ORFeome to primary regulators of transcription (a number that is probably an underestimate given the levels of alternative splicing in TF genes) is consistent with the idea that developmental and body plan complexity are related to complexity in transcriptional regulatory networks.
mRNA sequences. It should be noted that the latest version of TRANSFAC (TRANSFAC 6.0; Matys et al. 2003) contains more sequences than the version we used, but when different splice forms are eliminated, these all appear to overlap with our final set of genes. Additional human TFs were extracted from InterPro (Apweiler et al. 2001) and FlyBase Consortium (1999). We searched InterPro for all records annotated as “human” and occurring at or below the “transcription factor” node of the GO hierarchy (GO ID 0003700). InterPro records contain protein, not mRNA sequence. We therefore used LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/) to identify the GenBank mRNA records that correspond to the EMBL protein identification in the InterPro record. If available, we chose mRNA sequences from the RefSeq database (Pruitt and Maglott 2001); RefSeqs comprise the majority of our set (1270/1468). Otherwise, we took the most complete GenBank mRNA sequence or EST representing that gene (198/1468). For four genes (GSH1, HMX3, DLX1, and CHX10), the best available sequences were RefSeq gene models (see http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html for description). Thirteen genes were extracted from genomic sequence, and two (GSC and HMX2) were obtained from an Ensembl gene model (http://www.ensembl.org). For zinc finger genes, we manually removed subclasses of zinc finger proteins that are known not to bind DNA. However, it is possible that some of the zinc binding TFs in our final set will later be determined to be non-DNA-binding.

Additional known TF family members not identified by the above procedures were identified by using LocusLink or extensive literature searches and added to our database. Once data from these multiple sources have been collected, duplicates were eliminated, yielding a set of 1468 known human TFs (http://hg.wustl.edu/lovett/projects/nohr/Tfarray.html/).

### Homologous Transcript Clusters

Each initial seed list member (Fig. 1) was matched to its best ScriptSure cluster, requiring a lower bound of 90% coverage of the seed sequence and 80% identity in the alignment between the seed sequence and the ScriptSure cluster. The 1369 seed sequences found to have a matching ScriptSure cluster were then masked for repetitive sequence. RepeatMasker with the parameters “-w -s -no_is -xsmall” was used for one round of masking, identifying interspersed repeats. RepeatMasker was used again in a separate round of masking by using the parameters

### Table 1. A Comparison of Transcription Factors in Selected Eukaryotes

<table>
<thead>
<tr>
<th>Gene family</th>
<th>Homo sapiens</th>
<th>C.e.</th>
<th>D.m.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zinc binding&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>422</td>
<td>340</td>
<td>762</td>
</tr>
<tr>
<td>Homeobox</td>
<td>186</td>
<td>13</td>
<td>199</td>
</tr>
<tr>
<td>BHLH</td>
<td>92</td>
<td>25</td>
<td>117</td>
</tr>
<tr>
<td>β-Scaffold&lt;sup&gt;d&lt;/sup&gt;</td>
<td>77</td>
<td>10</td>
<td>87</td>
</tr>
<tr>
<td>BZip</td>
<td>59</td>
<td>13</td>
<td>72</td>
</tr>
<tr>
<td>NHR</td>
<td>49</td>
<td>0</td>
<td>49</td>
</tr>
<tr>
<td>Trp cluster</td>
<td>38</td>
<td>8</td>
<td>46</td>
</tr>
<tr>
<td>Forkhead</td>
<td>36</td>
<td>4</td>
<td>40</td>
</tr>
<tr>
<td>Bromodomain&lt;sup&gt;*&lt;/sup&gt;</td>
<td>14</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>T-box</td>
<td>16</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>Jumonji</td>
<td>6</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>E2F</td>
<td>9</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Dwarfin</td>
<td>9</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Paired box</td>
<td>9</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Heat shock</td>
<td>6</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Tubby&lt;sup&gt;a&lt;/sup&gt;</td>
<td>5</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>AF-4&lt;sup&gt;a&lt;/sup&gt;</td>
<td>7</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>RFX</td>
<td>6</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Methyl-CpG-binding&lt;sup&gt;+&lt;/sup&gt;</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>AP-2</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>TEA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Pocket domain (Rb)&lt;sup&gt;+&lt;/sup&gt;</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>GCM&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Other</td>
<td>214</td>
<td>14</td>
<td>228</td>
</tr>
<tr>
<td>Coactivators and corepressors</td>
<td>111</td>
<td>11</td>
<td>122</td>
</tr>
<tr>
<td>Structural</td>
<td>80</td>
<td>39</td>
<td>119</td>
</tr>
<tr>
<td>Total</td>
<td>1468</td>
<td>494</td>
<td>1962</td>
</tr>
</tbody>
</table>

The set of transcription factors is shown for four species, divided into families by the type of DNA-binding domain present and sorted by abundance in human. The data for human transcription factors show the seed and found set numbers separately, as well as the total number from the two sets added together. The Homo sapiens data are from this study. Unless otherwise specified, the data for S. cerevisiae, C. elegans, and D. melanogaster are from Riechmann et al. (2000); bHLH indicates basic helix-loop-helix; bZip, basic leucine zipper; C.e., Caenorhabditis elegans; D.m., Drosophila melanogaster; GCM, glial cell missing; NHR, nuclear hormone receptor; RFX, regulatory factor X; S.c., Saccharomyces cerevisiae; TEA, transcriptional enhancer activator (TEA/ATTs); and Trp cluster, tryptophan cluster.

<sup>a</sup>S. cerevisiae, C. elegans, and D. melanogaster data on zinc binding subfamilies AN1, BTB/POZ-containing, MYND, and PHD, β-scaffold subfamily cold shock, bromodomain family, Tubby family, methyl-CpG-binding family, TEA family, and pocket domain (Rb) family from Rubin et al. (2000); supplemental information (http://www.sciencemag.org/feature/data/1049664.shl).

<sup>b</sup>S. cerevisiae, C. elegans, and D. melanogaster data on zinc binding subfamily MIZ and AF-4 family from “species distribution” feature of Pfam Web site (http://pfam.wustl.edu).

<sup>c</sup>D. melanogaster data on GCM family from Akiyama et al. (1996).

<sup>d</sup>S. cerevisiae and C. elegans data on GCM family from “species distribution” feature of Pfam Web site (http://pfam.wustl.edu).
Table 2. Transcription Factors as a Proportion of Total Gene Content

<table>
<thead>
<tr>
<th>Organism</th>
<th>Approximate number of genes (TFs/total)</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. cerevisiae</td>
<td>224/6569*</td>
<td>3.4%</td>
</tr>
<tr>
<td>C. elegans</td>
<td>824/19546*</td>
<td>4.2%</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>744/13525*</td>
<td>5.5%</td>
</tr>
<tr>
<td>H. sapiens (NCBI)</td>
<td>1962/24652*</td>
<td>8.0%</td>
</tr>
<tr>
<td>H. sapiens (Ensembl)</td>
<td>1962/21787*</td>
<td>9.0%</td>
</tr>
</tbody>
</table>

The number of transcription factor genes for *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *H. sapiens* is shown, divided by the approximate total number of genes for each organism, to estimate the percentage of the total gene content that transcription factor genes represent. Two estimates are given for *H. sapiens*, based on current gene predictions from Ensembl and NCBI. Transcription factor gene counts are the same as in Table 1. Total gene count sources are as follows: *SGD* (July 2003), http://www.yeastgenome.org/FAQ; *Ensembl* *C. elegans* v19.102.1 (December 16, 2003), excluding 442 pseudogenes, http://www.ensembl.org/Caeorhabditis_elegans/stats/; *Ensembl* *D. melanogaster* v19.3a.1 (January 7, 2003), http://www.ensembl.org/Drosophila_melanogaster/stats/; NCBI human genome assembly build 34 (July 2003), http://www.ncbi.nlm.nih.gov/mapview/stats/BuildStats.cgi?taxid=9606&build=34&ver=1, and *Ensembl* v19 build 34a (December 16, 2003), excluding 1744 pseudogenes, http://www.ensembl.org/Homo_sapiens/stats/.

\( * = \text{no hit} \quad \text{no isolation test} \); this works best for masking simple repeats. Lastly, low complexity sequence was masked by using nseg with the parameters \(-s=200 \quad S=100 \quad \text{gapS}=200 \quad X=26 \quad gappx=55 \quad \text{gapW}=18 \quad \text{gapQ}=11 \quad R=11 \quad \text{N}=11 \quad \text{Z}=3000000000 \quad \text{V}=10000 \quad B=1000 \quad \text{gi novalidcontig}=0.0001 \); 2512 of these were clusters that did not match the Pfam database with a \( P \) value of \( <0.0001 \); 2512 of these were clusters that did not match our seed list.

**Mapping TFs to Genomic Contigs**

The Human Genome Consortium’s June 2002 draft of the human genome was used as the template for the version of ScriptSure we used in our analysis (June 2002b). Therefore, this same draft of the genome was used to place our identified clusters back onto the genome. The TF genes in our seed set as well as those identified in the overlap of the BLASTN and Pfam analysis were mapped back to the genome (1968 clusters total). Because ScriptSure reports its cluster coordinates relative to genomic contigs (rather than chromosome coordinates), UCSC’s “lift” file was used to translate between contig and chromosome coordinates (http://genome.ucsc.edu). Loci were considered significant if three TF clusters were found in a collinear cluster of eight total clusters (\( P = 0.37 \) under binomial model).

**Oligonucleotide Probe Design**

For each TF we identified, we designed a 50-mer to represent that gene on our array. We designed probes with four criteria: (1) the probe must be from a unique region of the sequence of a gene to eliminate potential cross-hybridization to other genes; (2) to allow use of the array on nonhuman samples, the probe must be from protein coding sequence (CDS); (3) the design was targeted to a region of coding sequence as 3’ as possible; and (4) the probes were matched for melting temperatures (\( T_m \)). The vast majority of probes had a \( T_m \) of 72°C, with very occasionally a probe varying by as much as 3°C when severe design constraints existed. We were able to automate the selection of probes meeting criteria 1 and 4 with a microarray probe design program, Probes2 (Li and Stormo 2001); the other steps were semiautomated with custom Perl scripts. The Sanger Centre/Ensembl set of 27,395 verified human cDNA sequences (downloaded on July 14, 2001, current version available at ftp://ftp.ensembl.org/pub/current_human/dna/fasta/cdna/) was used in conjunction with Probes2 to identify unique regions of each gene and design 50-mer probes. After candidate probes meeting these criteria were generated, we performed BLASTN similarity searches (default parameters) against the human genome sequence and inspected the results manually to confirm all criteria were met. Probes were synthesized (Sigma Genosys), resuspended at 60 \( \mu \)M in 1.5 M betaine and 6% DMSO, and spotted in duplicate on poly-L-lysine coated microscope slides with a GMS-417 arrayer (Affymetrix).

**ACKNOWLEDGMENTS**

This work was supported in part by a grant from the National Organization for Hearing Research Foundation (M.L.) and a training grant from the National Institutes of Health National Human Genome Research Institute (supporting D.M. and J.G.).

**REFERENCES**


amplification protocols for analysis of transcript structure and alternative splicing. Genome Biol. 4: R466.


WEB SITE REFERENCES

http://www.ensembl.org; Ensembl genome browser.
http://www.gene-regulation.com; TRANSFAC, the transcription factor database.
http://www.sapiens.wustl.edu/ScriptSure/; ScriptSure homepage.
http://www.ebi.ac.uk/interpro/; InterPro.
http://www.gene-regulation.com; TRANSFAC, the transcription factor database.
http://genome.ucsc.edu/; UCSC genome Web site.
http://blat.wustl.edu; WU-BLAST Web site.
http://www.sciencemag.org/feature/data/1049664.shl; Eukaryote Genome Projects and Community Literature.
http://www.ebi.ac.uk/interpro/; InterPro.
http://www.sapiens.wustl.edu/ScriptSure/; ScriptSure homepage.
http://www.ebi.ac.uk/interpro/; InterPro.
http://www.gene-regulation.com; TRANSFAC, the transcription factor database.
http://genome.ucsc.edu/; UCSC genome Web site.
http://blat.wustl.edu; WU-BLAST Web site.
http://www.sciencemag.org/feature/data/1049664.shl; Eukaryote Genome Projects and Community Literature.
http://www.gene-regulation.com; TRANSFAC, the transcription factor database.
http://www.sapiens.wustl.edu/ScriptSure/; ScriptSure homepage.
http://www.gene-regulation.com; TRANSFAC, the transcription factor database.
http://genome.ucsc.edu/; UCSC genome Web site.