The transcriptional activity of human Chromosome 22

John L. Rinn,[1,2,5] Ghia Euskirchen,[1,5] Paul Bertone,[1,2,5] Rebecca Martone,[1] Nicholas M. Luscombe,[2] Stephen Hartman,[1] Paul M. Harrison,[2] F. Kenneth Nelson,[2] Perry Miller,[3] Mark Gerstein,[2] Sherman Weissman,[4] and Michael Snyder[1,2,6]

[1]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06520-8103, USA; [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520-8114, USA; [3]Department of Medical Anesthesiology, Yale University, New Haven, Connecticut 06520-8051, USA; [4]Department of Genetics, Yale University, New Haven, Connecticut 06520-8005, USA

A DNA microarray representing nearly all of the unique sequences of human Chromosome 22 was constructed and used to measure global-transcriptional activity in placental poly(A)$^+$ RNA. We found that many of the known, related and predicted genes are expressed. More importantly, our study reveals twice as many transcribed bases as have been reported previously. Many of the newly discovered expressed fragments were verified by RNA blot analysis and a novel technique called differential hybridization mapping (DHM). Interestingly, a significant fraction of these novel fragments are expressed antisense to previously annotated introns. The coding potential of these novel expressed regions is supported by their sequence conservation in the mouse genome. This study has greatly increased our understanding of the biological information encoded on a human chromosome. To facilitate the dissemination of these results to the scientific community, we have developed a comprehensive Web resource to present the findings of this study and other features of human Chromosome 22 at http://array.mbb.yale.edu/chr22.

[*Keywords*: Microarray; Chromosome 22; human genome; transcriptome; placental RNA; mouse homology]

Received October 31, 2002; revised version accepted December 24, 2002.

As the sequencing phase of the human genome project nears completion, increasingly complete and accurate nucleotide-level data are becoming available (Lander et al. 2001; Venter et al. 2001). The next major challenge is to decipher the biological information encoded by the billions of ordered nucleotides. This goal requires identifying the various genes and proteins encoded in the DNA as well as how they function, how they are regulated, and how they work together to carry out complex biological processes. An essential step toward understanding the coding information of the human genome is to obtain a detailed knowledge of human transcriptional coding sequences on a genomic scale.

Current approaches for mapping mRNA-coding regions on a genomic scale have used a variety of techniques such as serial analysis of gene expression (SAGE), sequencing of expressed sequence tags (ESTs), STS mapping, radiation hybrid mapping, and full-length cDNA analysis (Saccone et al. 1996; Deloukas et al. 1998; Dunham et al. 1999; Caron et al. 2001). However, these techniques do not comprehensively interrogate all of the genomic coding information. Furthermore, these methods are not versatile for probing many tissue types and conditions, and consequently may fail to detect alternatively spliced messages or tissue-specific alterations in transcriptional activity.

Recently, new developments in microarray technology have made it possible for high-throughput mapping of the transcriptional activity of large segments of the genome (Shoemaker et al. 2001; Kapranov et al. 2002). Oligonucleotides representing nonrepetitive segments of a chromosome can be prepared at high density and probed with labeled cDNAs prepared from various tissues (Hegde et al. 2000). In principle, this approach can be used to detect transcriptional activity of both protein-coding and non-protein-coding RNAs chromosome-wide.

This approach has been used recently in two complementary studies carried out by Shoemaker et al. (2001) and Kapranov et al. (2002). Shoemaker et al. (2001) prepared oligonucleotide arrays to represent the known and predicted genes on human Chromosome 22 and probed them with cDNA probes prepared to RNA isolated from a number of tumor cell lines. They found representative expression for a majority of the known genes and a significant fraction of predicted genes, but they did not comprehensively examine unannotated regions of the chromosome. Kapranov et al. (2002) developed a micro-

[5]These authors contributed equally to this work.
[6]Corresponding author.
E-MAIL michael.snyder@yale.edu; FAX (203) 432-3597.

Article and publication are at http://www.genesdev.org/cgi/doi/10.1101/gad.1055203.

array containing 25-bp oligonucleotides for most of the nonrepetitive DNA of human Chromosome 22, and probed with double-stranded cDNA prepared from 11 different cell lines. The investigators observed the expression of many unannotated regions. However, the expression of intron sequences (which comprise 36% of the Chromosome 22 DNA; Dunham et al. 1999) and the conservation of expressed regions in other species were not reported.

In this study, we constructed a microarray containing PCR products encoding 17.4 Mb of nonrepetitive (NR) sequence on Chromosome 22, and used this array to map transcribed regions from the entire chromosome. We found that a significant fraction of the annotated regions are expressed in placental poly(A)$^+$ RNA. Moreover, we found that (1) there are twice as many sequences expressed on Chromosome 22 than previously thought; (2) many regions with no prior annotation are expressed and highly conserved in the mouse genome; and (3) much of the transcriptional activity exists within introns of annotated genes. Our results suggest that a large fraction of the genome is expressed as mRNA, and that there are many coding sequences that have not been annotated. We have also provided a detailed map of transcription units on the chromosome and made these findings readily available to the scientific community as a Web-based resource (available online at http://array.mbb.yale.edu/chr22).

Results

Construction of the human Chromosome 22 DNA microarray

A DNA microarray comprising nearly all of the nonrepetitive sequences of human Chromosome 22 was constructed to map transcriptional activity across an entire chromosome. This array contains both coding and noncoding genomic DNA sequences. The nonrepetitive regions of human Chromosome 22 were identified using RepeatMasker (A.F.A. Smit and P. Green, unpubl.) and divided into 21,024 PCR fragments, ranging in size from 300 bp to 1.4 kb (mean size = 720 bp). PCR primer sequences were designed, and the fragments were amplified from HeLa genomic template DNA; 19,525 fragments representing 93% of the targeted sequences were successfully prepared. Fragments were printed in duplicate onto three glass slides using a contact microarrayer. A set of positive and negative control fragments was also included on each slide.

Several quality-control experiments were performed to assess the fidelity of the amplified sequence and the reproducibility of microarray hybridization results. We first sequenced 349 PCR fragments with priority placed on those fragments that hybridized to cDNA probes prepared from placental poly(A)$^+$ RNA (see below). Sequences were compared to the entire human genome using BLASTN (Altschul et al. 1997). Of the 349 fragments sequenced, 314 matched the expected Chromosome 22 sequence. For the remaining 35 fragments, 15 matched a sequence very similar to that expected on Chromosome 22 (mean = 95% sequence identity to that of Chromosome 22), and 20 were from elsewhere in the genome.

We next ascertained how many of the fragments on the array contained repetitive elements by hybridizing labeled COT1 DNA (i.e., repetitive DNA) to the arrays. Approximately 6% of the fragments hybridized to COT1 DNA. This percentage was reduced to 1% when unlabeled COT1 DNA was added to the hybridizations. Therefore, we included unlabeled COT1 DNA in all of our subsequent hybridization experiments.

Many known and predicted Chromosome 22 genes are expressed

To experimentally map the transcriptionally active regions of Chromosome 22, placental poly(A)$^+$ RNA was hybridized to the array. RNA from placenta was chosen because it is (1) a normal tissue (i.e., not cancerous or from cell lines), (2) a complex tissue composed of many cell types, and (3) easily obtained in large quantities from a single source. Each chromosome fragment was probed in six independent experiments using cDNA prepared from triple selected poly(A)$^+$ placental mRNA.

To identify fragments with significant hybridization, a statistical data analysis scheme was devised specifically for microarrays probed with a single color fluor (see Materials and Methods). A total of 2504 fragments exhibited significant hybridization to labeled placental cDNA. We carefully mapped all of the hybridizing fragments onto Chromosome 22. Figure 1 depicts the transcriptional activity and density of human Chromosome 22 in relation to Sanger Centre annotated genes, and Table 1 summarizes the annotation distribution of these fragments.

To compare our results with known features of Chromosome 22, annotated genes corresponding to the version 2.3 data release from the Sanger Centre were aligned to the sequence coordinates of the 21,024 microarray fragments. The genes in the Sanger annotation fall into three categories: (1) known genes, which are well-characterized genes with a known full-length cDNA; (2) related genes, which are homologous to other known genes; and (3) predicted genes, which are predicted by homology to EST clusters. For the 339 known genes in the Sanger annotation data, we found that at least one exon hybridized in 206 (60.8%) cases (Table 2). This result demonstrates that a majority of the Chromosome 22 genes can be detected using a single tissue type.

In addition to detecting expression of the known genes, we found that 40.2% and 35.8% of the related and predicted genes were found to be expressed, respectively (Table 2). Thus, this approach can globally detect known, related, and predicted genes simultaneously.

An equal amount of expression is detected in unannotated regions of Chromosome 22

Hybridization of cDNA probes to known and predicted exons was accompanied by an equal amount of hybrid-

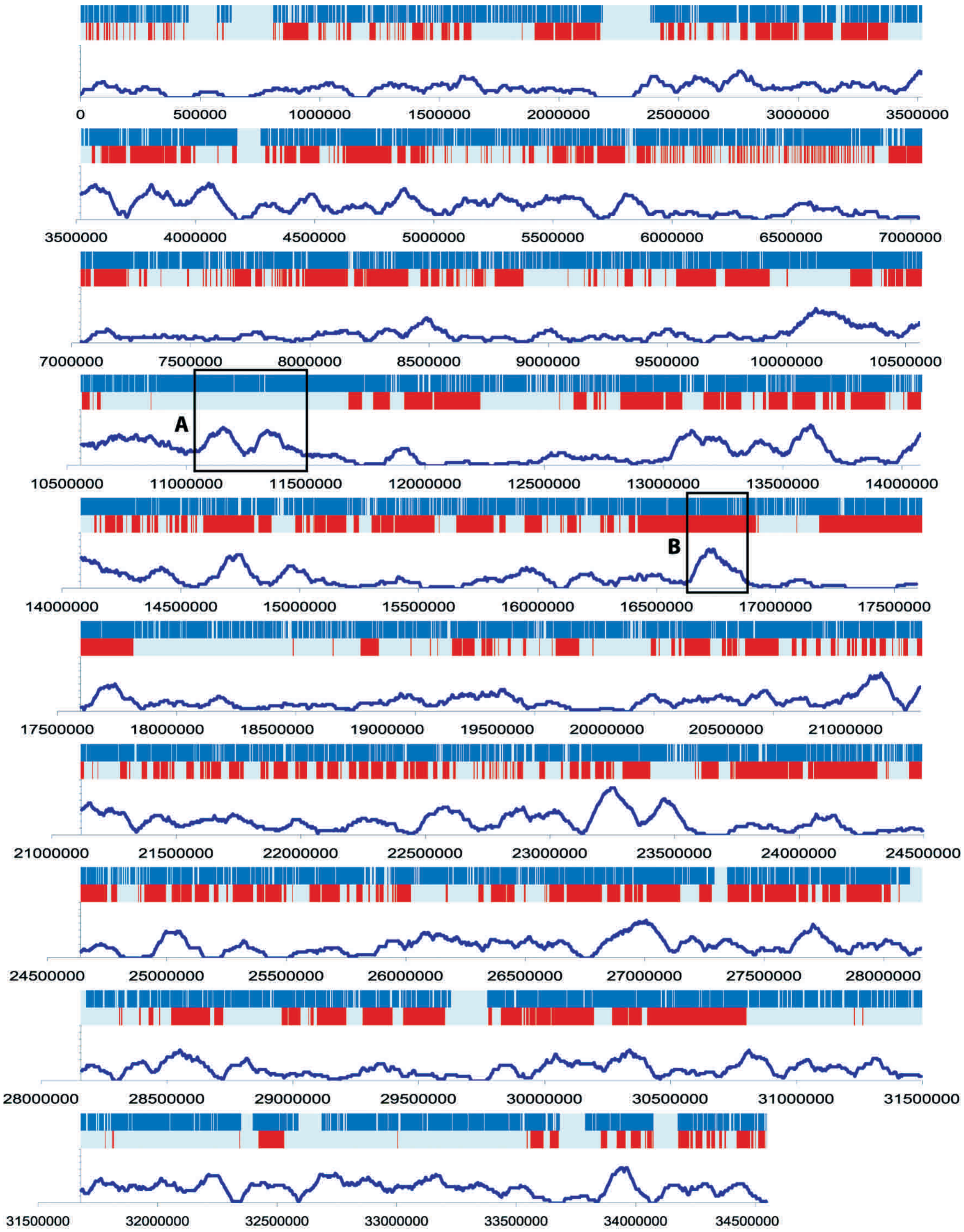

Figure 1. The human Chromosome 22 placental transcriptome. Dark blue strips indicate regions that are represented on the Chromosome 22 DNA microarray. Red strips indicate the positions of Sanger Centre release 2.3 annotated genes. The magnitude of the density plot represents the number of positive hybridizing fragments divided by the total number of fragments in a 100-kb window. (*A*) A large amount of transcriptional activity in a previously unannotated region of Chromosome 22. (*B*) A peak in transcriptional activity corresponding to known gene annotations. The window was moved fragment by fragment to give a continuous density plot. Positions with spikes in the density plot and low frequency of red strips indicate regions of novel transcriptional activity. Coordinates are given from centromere to telomere (starting at band 22q11.1) because the p arm has not been sequenced. The NCBI assembly lists sequence coordinates from the 5′ end of the p telomere to the 3′ end of the q telomere. Adding a 5′ offset of exactly 13 Mb to approximate the size of the p arm establishes a common reference frame for the NCBI/UCSC Golden Path assembly.

Table 1. *Distribution of positive hybridizing fragments and their respective gene annotations from the Sanger 2.3 data release*

Annotation type	Total	Exon-containing	Intron-containing
Gene	946 (11.9%)	428 (15.8%)	518 (9.8%)
Related	135 (11.4%)	66 (13.6%)	69 (9.9%)
Predicted	87 (9.9%)	50 (15.2%)	37 (6.8%)
Unannotated	1302 (12.2%)		

Parentheses show the percentage of total probes in the annotation category that showed positive hybridization. Fragments that were not associated with a gene hybridized with equal frequency to those intersecting annotated genes, suggesting an equal magnitude of transcription in previously unannotated regions.

ization to previously unannotated sequences. A total of 1302 (12.2%) of 10,693 fragments lacking prior annotation were observed to be expressed in placental tissue (Table 1). This amount is similar to the 946 (11.8%) of the 7967 microarray fragments intersecting known genes. Figure 1, box A, shows a large amount of transcriptional activity in a region of Chromosome 22 that was previously unannotated. Figure 1, box B, shows a peak in transcriptional activity corresponding to known gene annotations. Viewed together, these results indicate that there are as many transcribed sequences in unannotated regions as in annotated regions.

To confirm that the unannotated transcribed regions are expressed as mRNA (defined here as transcriptionally active regions, or TARs), 118 RNA blots of placental poly(A)$^+$ RNA were probed with randomly selected TARs (Fig. 2). Three fragments containing exons of known genes were also used to probe the RNA blots as a control; all three identified transcripts of the appropriate size (data not shown). Thirty (25%) unannotated fragments hybridized to mRNA transcripts ranging in size from 0.6 kb to >10 kb (Fig. 2). Several had multiple isoforms, perhaps indicating the presence of alternate splice products. Interestingly, two probes separated by 30 kb in genomic space hybridize to the same 6-kb transcript (Fig. 2, bar), further indicating that this 30-kb region encodes a gene.

To ensure that transcripts were not homologous to coding sequence elsewhere in the genome, all probes producing transcripts were searched using BLASTN (Altschul et al. 1990). This showed that 26/30 matched only the Chromosome 22 genomic sequences and 4 probes have potential homology ($E < 1e^{-5}$) to other genomic coding sequences. Thus, most of the transcribed sequences identified by the RNA blot analysis are derived solely from Chromosome 22. The lower than expected success rate of the RNA blot analysis (30/118) was also noted in a similar study (Kapranov et al. 2002). We speculate that the TARs are of low copy number, explaining why most have eluded prior detection using less sensitive methods.

To precisely map the expressed regions as well as determine the DNA strand of the hybridizing sequence, we used a novel strategy that we have termed differential hybridization mapping (DHM; Kumar et al. 2002). Briefly, a 60-nt oligomer and its complement were selected from regions within the hybridizing PCR fragments, spotted on the array, and probed with the labeled poly(A)$^+$ placental cDNAs. The cDNA will hybridize to the 60-nucleotide (nt) oligonucleotide that the message derived from and not to its complement. Thus, differential hybridization of the two oligonucleotides maps the expression to one strand.

To find potential exons in the 1302 unannotated TARs, their sequences were analyzed using four commonly used gene prediction methods (Genscan, Grail-EXP, GeneID, and by homology to known genes; Guigo et al. 1992; Burge and Karlin 1997; Xu and Uberbacher 1997). For the top 381 exon predictions (see Materials and Methods), we selected a 60-base oligonucleotide representing a unique sequence from each predicted exon and its complement. In this way, oligonucleotide selection is expected to be biased toward potential coding sequences. The oligonucleotide pairs were spotted on a separate area of the Chromosome 22 array and probed with labeled poly(A)$^+$ placental cDNAs. When one of the oligonucleotides in the pair hybridized and the complement did not, they were considered to hybridize differentially to one strand. Those pairs exhibiting differential expression on the same strand in 3 of 4 replicate experiments were scored as positive expressed sequences (see Materials and Methods).

As a control, we included multiple oligonucleotides mapping a region that contains an exon sequence on one strand representing a gene known to be expressed in placental tissue and an intron sequence on the opposite strand. As expected, only the exon strand hybridized to the poly(A)$^+$ RNA (Fig. 3C).

Significant differential hybridization was observed in 53 of the 381 pairs, indicating that the hybridizing region and strand could be identified in many cases. Presumably, in the cases that did not exhibit differential hybridization, the expressed region was not represented by the 60-nt oligonucleotides or both strands were expressed. In summary, the RNA blot analysis and oligonucleotide DHM data independently verified that a significant amount of the unannotated regions are expressed in mature mRNA transcripts.

Table 2. *Genes in the three Sanger 2.3 annotation categories that were represented by at least one hybridizing exon*

Annotation	Identified	Total	Identified/total (%)
Known genes	206	339	60.8
Related genes	45	112	40.2
Predicted genes	35	109	35.8

60.8% of the known genes were detected using only one tissue type, as well as detecting expression from a large fraction, 40.2% and 35.8%, of the related and predicted genes, respectively. This success rate is similar not only to other studies using microarrays to annotate human Chromosome 22 but also to studies using ESTs.

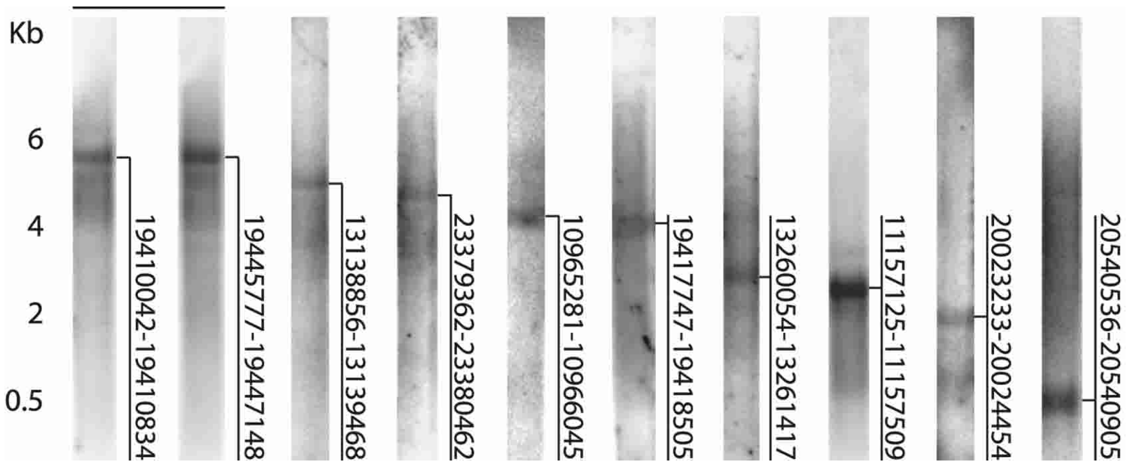

Figure 2. Northern blot analysis of 118 fragments that were expressed in previously unannotated regions of Chromosome 22. Thirty (25.4%) showed discrete bands. Ten are shown above and labeled with the corresponding chromosomal location of the probe used in the Northern hybridization. Bar indicates two probes separated by ~30 kb in genomic space that hybridize to the same 6-kb transcript.

Expression is observed from within annotated introns

Careful analysis of the hybridizing fragments that intersect annotated introns revealed the unexpected finding that many introns contain expressed sequences. In fact, of the 5264 fragments located entirely within annotated introns, 518 (9.8%) were found to be expressed in five of six experiments. There are three possible explanations for this observation: (1) a novel expressed sequence is encoded on the strand opposite the intron; (2) there is an unannotated exon located within the intron that had not been discovered previously; or (3) expressed intron sequences were detected.

To distinguish among these possibilities, we used the DHM technique as described above. For this, 119 60-nt oligonucleotides representing various intron regions and their complementary sequences were spotted onto a microarray and probed with labeled poly(A)$^+$ placental cDNAs. Of the 119 oligonucleotide pairs, 23 (19.3%) showed significant differential hybridization. Expression from the same strand as the intron was detected in 13 cases, indicating that sequences from within the intron are expressed. In 5 of these 13 cases, an exon was predicted within the intron; one example is presented in Figure 3B. In 10 cases, expression is derived from the opposite strand of the intron, suggesting that a novel expressed fragment overlaps with the intron. In total, nearly half of the hybridizing fragments that intersect intron regions were shown to contain expressed sequences antisense to their respective introns.

To thoroughly investigate this observation, we used DHM with multiple oligonucleotide probe pairs to completely cover a subset of the hybridizing fragments previously annotated as introns. In one case, 6 oligonucleotide pairs from within a 1.3-kb region showed differential hybridization to the strand antisense to an annotated intron (Fig. 3A). In another example, 2 positive 60-nt nucleotides hybridized within a 400-bp region opposite a known intron. In these cases, the regions that are transcribed on the opposite strand of introns are not short in length because multiple probes detect expression throughout the segment. In summary, we detected expressed sequences hybridizing to regions both internal to annotated introns and to the strand opposite introns.

Many unannotated expressed sequences are conserved

We hypothesize that many of the positive hybridizing fragments whose sequences lie outside those of known genes represent novel exons. It follows that a percentage of these are likely to be homologous to other mammalian genes, providing supporting evidence of putative coding regions.

A homology comparison of unannotated TARs with the mouse genome was performed using BLASTN and BLASTP with published criterion as described (see Materials and Methods). Of the 1231 positive microarray fragments intersecting Sanger-annotated genes, 541 (~44%) intersect an ortholog in the mouse genome. Interestingly, 90 (7%) positive fragments that do not intersect with annotated genes potentially encode proteins that are homologous to mouse proteins (82) or genomic sequence (8). For instance, an unannotated fragment is predicted to encode a protein with high sequence similarity to a mouse procollagen protein (Fig. 4A). Of the 90 DNA fragments that encode similar proteins to mouse sequences, 25 are located in introns, and many are on the antisense strand of the annotated introns. Two examples are presented in Figure 4B and C. Thus, it appears that a large portion of the novel TARs are evolutionarily conserved.

Discussion

In this study we used a multifaceted approach to provide a detailed transcriptional map of human Chromosome 22. A microarray containing most of the unique sequence was developed and subsequently hybridized to probes prepared from human placental poly(A)$^+$ RNA to

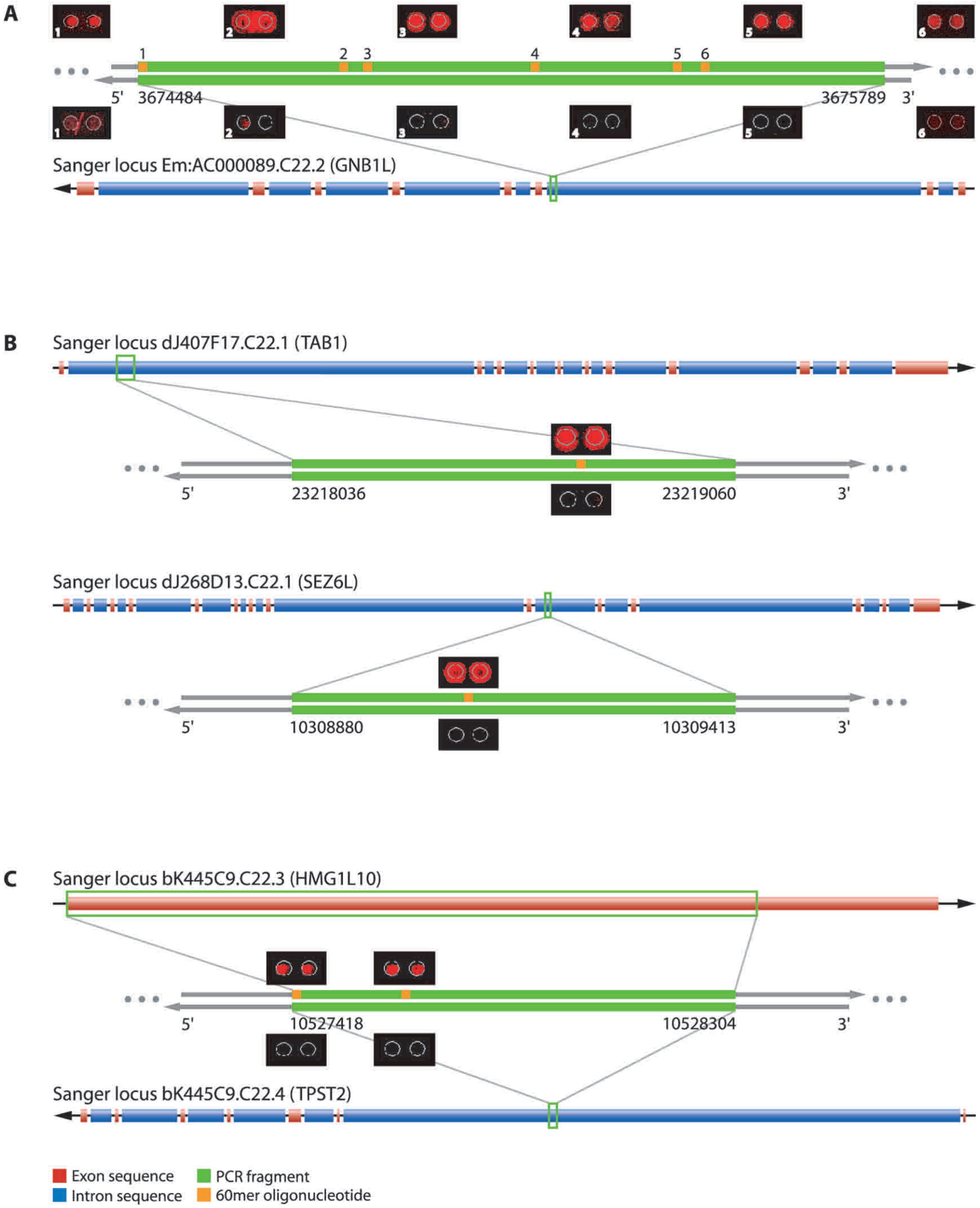
A
1
2
3
4
5
6
1
2
3
4
5
6
5'
3674484
3675789
3'
1
2
3
4
5
6
Sanger locus Em:AC000089.C22.2 (GNB1L)
B
Sanger locus dJ407F17.C22.1 (TAB1)
5'
23218036
23219060
3'
Sanger locus dJ268D13.C22.1 (SEZ6L)
5'
10308880
10309413
3'
C
Sanger locus bK445C9.C22.3 (HMG1L10)
5'
10527418
10528304
3'
Sanger locus bK445C9.C22.4 (TPST2)
Exon sequence
Intron sequence
PCR fragment
60mer oligonucleotide

Figure 3. Differential hybridization mapping within positive PCR fragment sequences. (*A*) Hybridization to multiple 60-nt oligonucleotides positioned opposite an intron sequence annotated on the antisense strand. (*B*) Hybridization to oligonucleotides representing a predicted exon within an annotated intron on the sense strand. (*C*) Control spots showing differential hybridization to a known exon (1) located on the strand opposite an annotated intron and (2) whose expression was previously verified. NCBI/UCSC sequence coordinates are offset by 13 Mb to approximate the size of the unsequenced p arm.

identify transcriptionally active regions throughout the chromosome. In addition to detecting known and predicted coding regions, we also found that an equal amount of previously unannotated regions were expressed. We verified that many novel coding segments produced bona fide messages using RNA blot analysis. A comparison of novel regions to mouse sequences revealed that many of the novel transcriptionally active

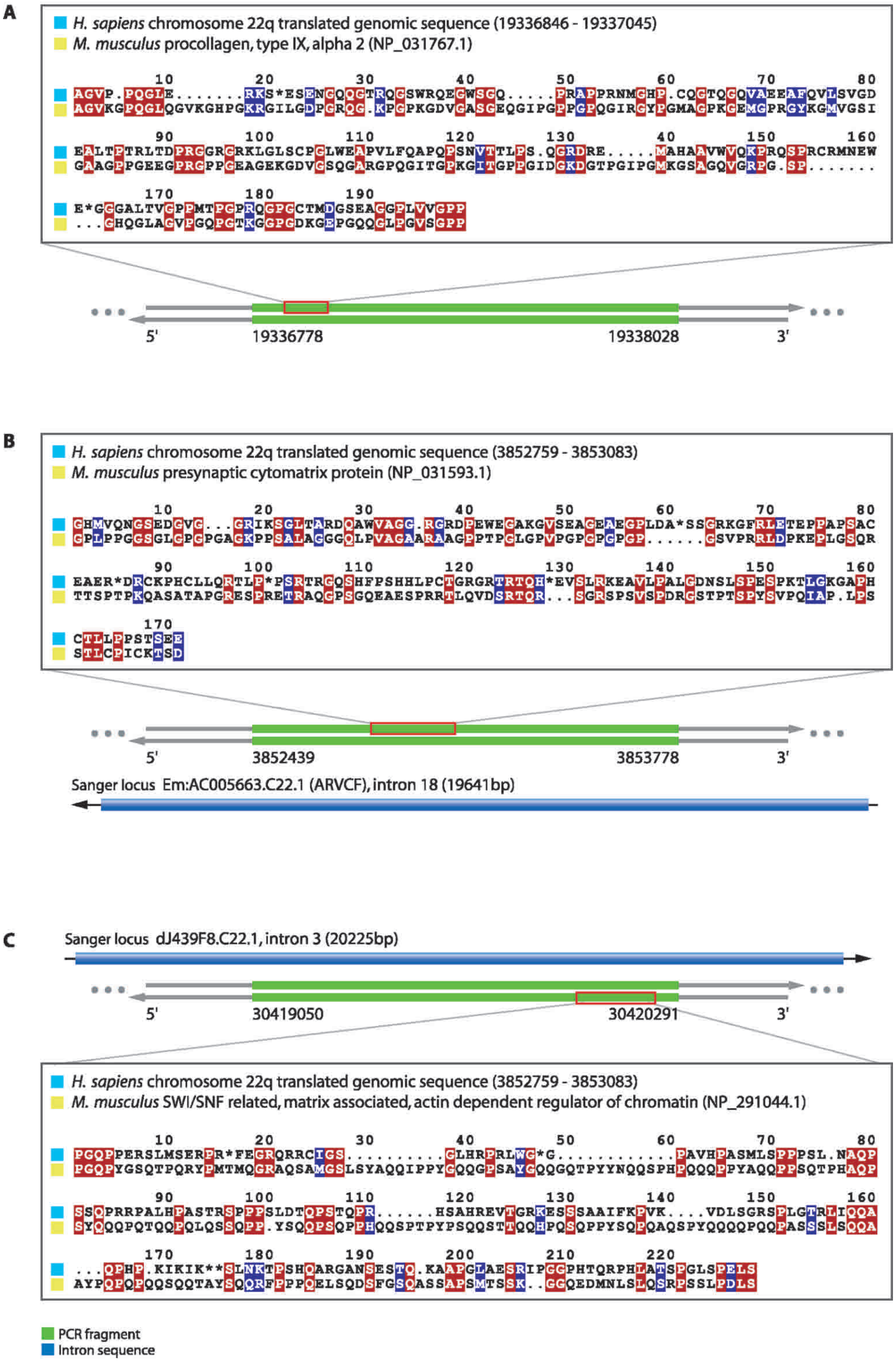
A
H. sapiens chromosome 22q translated genomic sequence (19336846 - 19337045)
M. musculus procollagen, type IX, alpha 2 (NP_031767.1)
10 20 30 40 50 60 70 80
AGVP.PQGLE.......RKS*ESENGQQGTRQGSWRQEGWSGQ.....PRAPPRNMGHP.CQGTQGQVAEEAFQVLSVGD
AGVKGPQGLQGVKGHPGKRGILGDPGRQG.KPGPKGDVGASGEQGIPGPPGPQGIRGYPGMAGPKGEMGPRGYKGMVGSI
90 100 110 120 130 40 150 160
EALTPTRLTDPRGGRGRKLGLSCPGLWEAPVLFQAPQPSNVTTLPS.QGRDRE.....MAHAAVWVQKPRQSPRCRMNEW
GAAGPPGEEGPRGPPGEAGEKGDVGSQGARGPQGITGPKGITGPPGIDGKDGTPGIPGMKGSAGQVGRPG.SP.......
170 180 190
E*GGGALTVGPPMTPGPRQGPGCTMDGSEAGGPLVVGPP
...GHQGLAGVPGQPGTKGGPGDKGEPGQQGLPGVSGPP
5'
19336778
19338028
3'
B
H. sapiens chromosome 22q translated genomic sequence (3852759 - 3853083)
M. musculus presynaptic cytomatrix protein (NP_031593.1)
10 20 30 40 50 60 70 80
GHMVQNGSEDGVG...GRIKSGLTARDQAWVAGG.RGRDPEWEGAKGVSEAGEAEGPLDA*SSGRKGFRLETEPPAPSAC
GPLPPGGSGLGPGPGAGKPPSALAGGGQLPVAGAARAAGPPTPGLGPVPGPGPGPGP......GSVPRRLDPKEPLGSQR
90 100 110 120 130 140 150 160
EAER*DRCKPHCLLQRTLP*PSRTRGQSHFPSHHLPCTGRGRTRTQH*EVSLRKEAVLPALGDNSLSPESPKTLGKGAPH
TTSPTPKQASATAPGRESPRETRAQGPSGQEAESPRRTLQVDSRTQR...SGRSPSVSPDRGSTPTSPYSVPQIAP.LPS
170
CTLLPPSTSEE
STLCPICKTSD
5'
3852439
3853778
3'
Sanger locus Em:AC005663.C22.1 (ARVCF), intron 18 (19641bp)
C
Sanger locus dJ439F8.C22.1, intron 3 (20225bp)
5'
30419050
30420291
3'
H. sapiens chromosome 22q translated genomic sequence (3852759 - 3853083)
M. musculus SWI/SNF related, matrix associated, actin dependent regulator of chromatin (NP_291044.1)
10 20 30 40 50 60 70 80
PGQPPERSLMSERPR*FEGRQRRCIGS..........GLHRPRLWG*G............PAVHPASMLSPPPSL.NAQP
PGQPYGSQTPQRYPMTMQGRAQSAMGSLSYAQQIPPYGQQGPSAYGQQGQTPYYNQQSPHPQQQPPYAQQPPSQTPHAQP
90 100 110 120 130 140 150 160
SSQPRRPALHPASTRSPPPSLDTQPSTQPR......HSAHREVTGRKESSSAAIFKPVK....VDLSGRSPLGTRLIQQA
SYQQQPQTQQPQLQSSQPP.YSQQPSQPPHQQSPTPYPSQQSTTQQHPQSQPPYSQPQAQSPYQQQQPQQPASSSLSQQA
170 180 190 200 210 220
...QPHP.KIKIK**SLNKTPSHQARGANSESTQ.KAAPGLAESRIPGGPHTQRPHLATSPGLSPELS
AYPQPQPQQSQQTAYSQQRFPPPQELSQDSFGSQASSAPSMTSSK..GGQEDMNLSLQSRPSSLPDLS
PCR fragment
Intron sequence

Figure 4. Mouse protein homology within translated PCR fragment sequences. (*A*) Homology match between mouse sequence and a positive microarray fragment whose sequence coordinates lie outside annotated genes. (*B*,*C*) Examples of mouse protein matches to human genomic sequences that are opposite annotated introns. In both cases the homology match is antisense to the intron. NCBI/UCSC sequence coordinates are offset by 13 Mb to approximate the size of the unsequenced p arm.

regions (TARs) are highly conserved. Furthermore, these regions were mapped at a higher resolution using differential hybridization mapping (DHM) with oligonucleotide arrays. By using DHM we verified that many unannotated regions are transcribed; we found a significant fraction of expression is within introns and antisense introns. These studies indicate that a considerable portion of the human genome is transcribed, often in unexpected places.

Many predicted genes are expressed

We found that many known and predicted genes that have been identified previously are expressed in a single tissue type. This includes the majority (60.8%) of all the known genes on the chromosome. In addition, we were also able to detect 40.2% and 35.8% of the related and predicted genes, respectively. Our success rate is similar to that of other studies (de Souza et al. 2000; Shoemaker et al. 2001). However, those studies used the out-of-date annotation accompanying the original Chromosome 22 sequence, which contained many more related and predicted genes. In contrast, our study used the latest Sanger 2.3 annotation, in which many related and predicted genes have now been classified as known (98 and 50, respectively, relative to the initial Sanger Centre data release for Chromosome 22). Nonetheless, our microarray analysis was able to verify the few remaining predicted and related genes, demonstrating the sensitivity of this approach.

Why have the unannotated TARs eluded detection?

In addition to the annotated regions, we found expression of many Chromosome 22 regions that have not been detected previously. There are probably two reasons for this. First, the unannotated TARs may be expressed at low abundance. Only 25% of the 118 hybridizing fragments from unannotated regions detected discrete transcripts using RNA blot analysis. We suggest that those fragments that did not detect mRNAs using RNA blot analysis encode low-abundance transcripts. The second reason we may have found novel TARs is that our approach interrogates most of the unique sequences of the chromosome and is thus more comprehensive than most other methods.

Several hypotheses may explain the biological functions of the novel transcribed regions. It is likely that in many cases these encode low-abundance proteins of new genes. This has the potential to increase, possibly by as much as twofold, the number of human genes above the present estimate of 30,000–35,000 (i.e., to 70,000 total; Ewing and Green 2000). It is also possible that the transcribed regions correspond to previously missed exons of known genes. A third possibility is that they may function as noncoding RNAs (i.e., siRNAs, snoRNAs, hnRNAs, or other small RNAs); in this capacity they might serve in a structural, catalytic, or regulatory capacity. For instance, if the novel coding segments produce antisense transcripts, they might control the levels, export, or translation of genes encoded on the opposite strand. Regardless of their functions, these newly discovered expression regions are clearly an important source of new biological information, as many of them are highly conserved among mammals.

The microarray approach is comprehensive

A variety of other studies have been used to annotate Chromosome 22. SAGE, ESTs, and Orestes have identified a number of coding segments on the chromosome (Saccone et al. 1996; Deloukas et al. 1998; de Souza et al. 2000; Liang et al. 2000; Caron et al. 2001). However, these studies are biased toward detecting the most abundant transcripts, and they are often limited by the short stretches of DNA sequences. The microarray (or "in chipo") approach is more suitable for expression profiling because several different tissue types can be analyzed in parallel to determine tissue-specific abundance. Also, this approach can be used to elucidate other annotation features, whereas the previously mentioned techniques cannot; for example, identification of transcription-factor-binding sites via hybridization of chromatin immunoprecipitated DNA probes.

Two independent microarray studies have also investigated the transcriptional activity of the chromosome. Shoemaker et al. (2001) prepared oligonucleotide probes to represent many predicted exons from Genscan (Burge and Karlin 1997). Although the method was able to detect transcripts for 185 (57%) of the 325 Genscan predicted genes, their study did not examine the majority of nonrepetitive sequence on Chromosome 22. Moreover, the microarray used in that study was printed using ink-jet technology, and was therefore only intended for a single application. Our approach is much more comprehensive and universally applicable to a wide range of experiments.

An independent study by Kapranov et al. (2002) interrogated transcriptional activity using high-density oligonucleotide arrays containing 25-nt oligonucleotides spaced, on average, 10 nt apart to cover most of the nonrepetitive DNA of Chromosomes 21 and 22. The study also found that many unannotated regions of the chromosome are expressed; however, there are a number of differences between that study and ours. First, they did not report that expression is observed from within annotated introns, nor did they assess the degree of homology between expressed sequences and those in other genomes to establish evidence for conserved regions. Second, cDNA probes from different sources were used. Kapranov et al. (2002) used probes from RNA isolated from 11 cancer cell lines, whereas we used placental poly(A)$^+$ RNA. Third, they used double-stranded cDNA probes prepared to the RNA, thus, they could not determine which strand is expressed in the oligonucleotide hybridizations.

Although no microarray is entirely comprehensive, the PCR-based array has several advantages. First, it contains large regions of contiguous sequence information, ensuring that no information is omitted. However, the PCR-based array is of lower resolution, and the exact hybridizing region must be determined by other methods such as DHM. Despite the inherent differences in the two approaches, a thorough comparison of their expressed sequences with ours reveals extensive overlap. Of our 2504 hybridizing fragments, 10% (250) were not detected in the Affymetrix investigation, indicating that the two methodologies are complementary.

Another advantage of the PCR arrays is that they can be prepared in an academic lab and at high throughput. Thus, the approach is easily amenable to serially hybridizing many tissue types to determine tissue-specific transcripts. This array is also a versatile tool for many other purposes such as identifying transcription-factor-binding sites in conjunction with chromatin immunoprecipitations. Ultimately, we envision this approach producing annotation features of all chromosomes on a large scale. These transcription or TAR maps may also serve a comparative evolutionary function as well. Typically, whole genome sequences are compared to find similarities that have been preserved through evolution. Although this is a valid and useful approach, TAR maps may also be compared to find conserved expressed sequences. The latter may be a useful way to determine evolutionary differences for species as well as the evolutionary changes in chromosomes.

Perhaps most importantly, our results have been made available to other investigators in a Web database containing experimental microarray data mapped to genes, pseudogenes, SNPs, and other chromosomal annotation features (available online at http://array.mbb.yale.edu/chr22). This database is a significant step toward an accessible universal resource for all the annotations on Chromosome 22.

Materials and methods

Construction of the human Chromosome 22 array: sequence analysis and primer selection

Chromosome 22q spans 34.5 Mb, of which 45% consists of repetitive elements (e.g., SINES, LINES, retroviral DNA, and low-complexity sequence) identified by the RepeatMasker program (A.F.A. Smit and P. Green, unpubl.). The remaining sequence fragments of sufficient size to facilitate large-scale PCR (≥300 bp) accounted for only 87% of the nonrepetitive DNA; the sizes of many high-complexity fragments fell below this threshold. To improve the sequence coverage, a dynamic programming algorithm was developed (Berman et al. 2002) to recover many of the smaller high-complexity fragments by strategically incorporating short repetitive elements located between them, thereby joining the adjacent fragments into larger contiguous sequences amenable to PCR. This procedure generates an optimal tile path for the masked genomic sequence, simultaneously maximizing (1) the coverage of high-complexity DNA from the target sequence and (2) the number of sequence fragments within a specified size range (in our case, 300 bp–1.4 kb), while minimizing the number of repetitive nucleotides included in the amplified sequences. Following this analysis, the final set of target sequences amounted to 17.4 Mb, or 92% of the nonrepetitive DNA of Chromosome 22. PCR primer pairs were selected using the Primer3 software [written by S. Rozen and H.J. Skaletsky (1996); code available online at http://www-genome.wi.mit.edu/genome_software/other/primer3.html], and were designed to have similar melting temperatures in a 55°C–70°C range, low alignment scores, and preferably a 3′ C or G base for increased binding efficiency. Sequences exceeding 1.4 kb were subdivided prior to the primer design stage, defining the upper bound of amplicon size. To ensure complete interfragment coverage between these adjacent sequences, the 5′ primer sequences for amplicons (2 . . *n*) from subdivided fragments were replaced with the reverse complement of the 3′ primer sequences from the amplicon directly preceding them. The modified primer pairs were examined for inter- and intraoligo alignment, and the 3′ ends of problematic sequences were adjusted to reduce the potential for primer-dimer formation.

Construction of the human Chromosome 22 array: DNA and slide production

PCR reactions were performed using 2× QIAGEN MasterMix, 0.5 µM of each primer, and 65 ng of HeLa genomic DNA as template. Fragments were analyzed by agarose gel electrophoresis, and only those products that migrated as a single band of the predicted size were arrayed. PCR products were precipitated with a 1:1 mixture of ethanol:isopropanol and dried and resuspended in 25 µL of water. The fragments were mixed with an equal volume of DMSO for printing. Slides were printed in house with an SDDC-2 arrayer (ESI-Virtek) on Corning CMT GAPS slides. Arrays were cross-linked, and print quality was confirmed by staining for total DNA with POPO-3 (Molecular Probes).

The quality of the array was analyzed by DNA sequencing and COT hybridization experiments as discussed in the Results.

Hybridizing the placental transcriptome

Using Ambion's amino-allyl cDNA labeling kit, 1.5 µg of poly(A)$^+$ mRNA that had been purified three times with oligo(dT) (Ambion) was reverse-transcribed. The reactions were primed with both oligo(dT) and random decamers in an equimolar mix in the presence of an amino-allyl-modified cytosine. After reverse transcription, the template mRNA was degraded in the presence of NaOH at 70°C. The cDNAs were ethanol-precipitated and resuspended in 0.1 M $NaHCO_3$ to facilitate coupling of the Cy5 mono-amine dye (Amersham) to the amino-allyl functional group. After the coupling reaction, the labeled cDNAs were separated from unincorporated Cy5 mono-amine dye using a Sephadex column provided with the amino-allyl cDNA labeling kit. Labeled probes were then ethanol-precipitated and resuspended in 5× SSC, 25% formamide, and 15 µg of COT1 DNA (Invitrogen) to block. Samples were hybridized at 42°C as described (Hegde et al. 2000).

Determination of positives

Microarrays were scanned with an Axon 4000A scanner, and images were analyzed with GenePix Pro3.0 software. The raw GenePix output was processed as follows to identify positive hybridized fragments: (1) Spots with aberrant morphology, or those with intensities below the threshold of detection were discarded. (2) Within individual experiments, spot pairs (fragments printed in duplicate side by side) were excluded from further analysis if the variation (= $I_1 - I_2/I_1 + I_2$) between them was >3 standard deviations of the error distribution of the data points. (3) The six replicate experiments were normalized with one another to scale the Cy5 intensity spreads to a common range. We calculated a resampled variance for each experiment and scaled the distributions so they had equal variances. Different scale factors were calculated for each block of spots on the slide to correct for intensity variations dependent on slide location (Goryachev et al. 2001; Yang et al. 2002). (4) The final Cy5 intensity for each Chromosome 22 fragment was obtained as the mean for duplicate spots within an experiment and the median value across replicate experiments. We also recorded the number of experiments in which the fragment is hybridized. (5) We counted the number of fragments that hybridized in 1, . . . ,*n* replicate experiments. We only considered those fragments that hybridized in 5 or more replicates. Here, fragments that hybridize in fewer than 5 experiments were considered false positives. In Figure 5 we plot the percentage of fragments that hybridize in 5 or more experiments against different Cy5 intensities. The Cy5 intensity cutoff of 200 for positive hybridized fragments was determined empirically from the plot, on which we observe a sharp rise in the proportion of fragments in 5 or more experiments; at this intensity we identify 2504 positive hybridized fragments with a false-positive rate of 5% (Fig. 5).

RNA blot verification of novel TARs

To verify TARs, a total of 118 Northern blots were analyzed. Northern blots of triple-purified poly(A)$^+$ placental mRNA were purchased from Ambion. Five blots were cut into a total of 50 single-lane strips. Each strip was prehybridized in ULTRAhyb (Ambion) buffer for 2 h and then hybridized using probes prepared from novel TAR PCR products using a Strip-EZ DNA labeling kit (Ambion). Hybridizations were carried out overnight at 42°C. Strips were washed twice in Northern Max (Ambion) high-stringency buffer followed by three washes in Northern Max (Ambion) low-stringency buffer. Single-lane filters were stripped according to the Strip-EZ protocol.

Differential hybridization array

The 60-nt oligonucleotides were purchased from Illumina. They were resuspended in 50% DMSO at 50 µM. Oligonucleotide slides were printed and hybridized as above.

Differential hybridization mapping determination of positives

The oligo-slides were scanned and processed using the same method as for the Chromosome 22 array. To identify positive hybridized oligonucleotides, the final Cy5 signals for oligonucleotide pairs (strand and antistrand) were compared with each other, providing a measure of pairwise differences in hybridization ($= I_{strand} - I_{antistrand}/I_{strand} + I_{antistrand}$). Oligonucleotides that had no detectable signal or that were filtered from the data set were assigned an intensity value of 0. The distribution of the pairwise differences approximated a normal distribution, and a set of 119 outlier pairs was selected as being differentially hybridized ($p < 0.001$). For each pair, the oligonucleotide with the higher Cy5 signal was identified as being positive-hybridized.

Mouse homology comparison

Positive fragments intersecting genes known to be mouse orthologs were identified as follows. A comprehensive set of annotated human genes on Chromosome 22 with established homology to mouse genes was compiled using 5 data sets obtained from the NCBI [National Center for Biotechnology Information, Human/Mouse Homology Maps (May 2002); http://www.ncbi.nlm.nih.gov/Homology]. These consist of human–mouse orthologs identified by homology between the genetic map represented in the Mouse Genome Database (MGD; Blake et al. 2002) and the Whitehead/MRC radiation hybrid map (Hudson et al. 2001) with the NCBI Build 28/UCSC HG10 human genome assembly (UCSC Human Genome Project Working Draft, December 2001 assembly; http://genome.cse.ucsc.edu). Each homologous gene found on Chromosome 22 was cross-referenced with Sanger-annotated genes, and the positive fragments that intersect them were identified. To assess the degree of sequence similarity between the remaining positive microarray fragments and mouse sequences, the fragment sequences were queried against the draft mouse genome (NCBI Mouse Genome Release 27) using BLASTN for nucleotide–nucleotide compari-

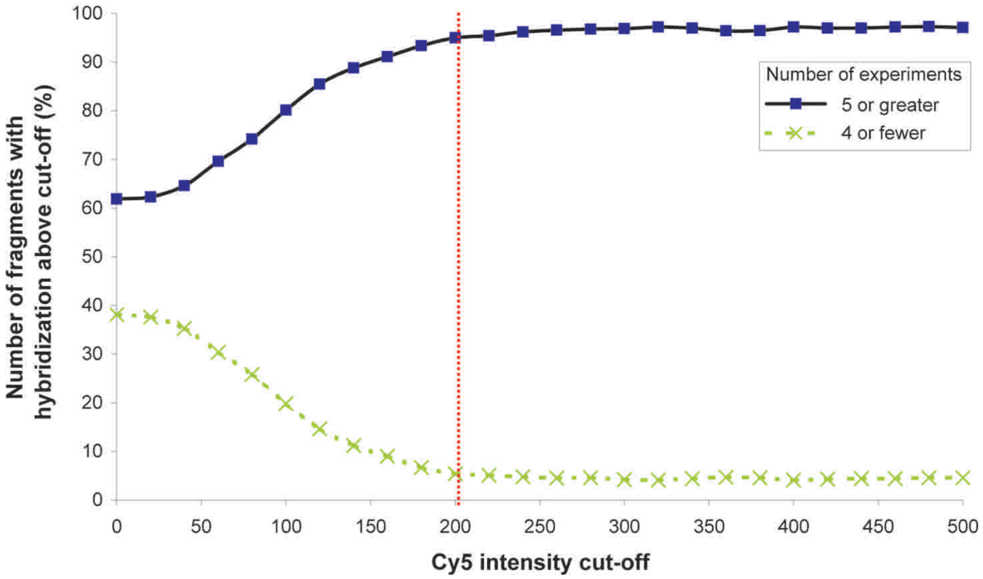

Figure 5. False positives. Determination of intensity cutoff in determining positive hybridizing fragments. There is a clear leveling of consistency after 200 intensity units. The plot demonstrates that fragments with an intensity >200 were present with that intensity or higher in 5 out 6 replicate experiments. Fragments that hybridized in 4 or less of the 6 replicate experiments and with an intensity ≥200 were summed to give a false-positive rate of 5%.

sons (Altschul et al. 1990), and to the RefSeq repository of mouse protein sequences (Pruitt and Maglott 2001) using BLASTX for six-frame translational nucleotide–protein comparisons. In each case a threshold *e*-value of 0.0001 was used to select significant matches, with the additional restriction that only matches exceeding 200 nt were considered significant for the mouse genomic DNA comparison.

Prediction of potential exon sequences

Candidate sequences from hybridizing fragments were searched against the NRDB and Ensembl protein sequence databases using the TBLASTX program with six-frame translation (Altschul et al. 1997). The matches then were filtered for repetitive sequences with the RepeatMasker program. To eliminate overlapping results, homology matches were filtered such that lower-scoring matches that overlapped with a higher-scoring match by >40 nt were discarded. The three gene prediction programs Genscan (Burge and Karlin 1997), GrailEXP (Xu and Uberbacher 1997), and GeneID (Guigo et al. 1992) were also applied to each amplicon sequence. For each resulting set of exon predictions, a nonredundant list was made such that better-scoring predictions were chosen in preference to lower-scoring ones. GrailEXP makes predictions using a large database of ESTs, cDNAs, and mRNAs; these predictions are chosen in preference to any other prediction. The remaining exon predictions were chosen in the following order of preference: (1) Genscan with exon probability ≥ 0.1, (2) GrailEXP, (3) GeneID, (4) Genscan with exon probability < 0.1. Any additional potential exons produced from the homology searches detailed above were also included. The final nonredundant list of exon predictions was then used to derive 60-nt oligonucleotides by selecting unique internal sequences from each predicted exon region using the Primer3 software [written by S. Rozen and H.J. Skaletsky (1996); code available online at http://www-genome.wi.mit.edu/genome_software/other/primer3.html].

Chromosome 22 microarray database

Following the microarray design and construction, a Web-accessible database was developed for chromosome-wide gene annotation and analysis of microarray data generated by the project. The system brings together all of the known and predicted features on Chromosome 22 from many disparate sources, for the purpose of coordinating genomic information with experimental data. Annotated features such as known genes (Dunham et al. 1999; Hubbard et al. 2002), predicted exons (GenomeScan gene predictions contributed by Ru-Fang Yeh and Chris Burge, Massachusetts Institute of Technology), pseudogenes (Harrison et al. 2002), and SNPs (Balasubramanian et al. 2002) are aligned to the positional coordinates of the Chromosome 22 microarray fragments in an automated fashion. Users of the system can upload scanned and quantitated microarray data files, then browse through the results to identify any genes, pseudogenes, or SNPs with which enriched microarray fragments intersect on the chromosome. PCR fragments or features of interest may then be explored in greater detail using a variety of graphic- and text-based views, with relevant links to external resources. Specific genes or chromosomal regions may also be located on the array directly, using search functions that relate their nucleotide positions to the corresponding microarray fragments. Thus, researchers are able to correlate vast amounts of experimental data with existing knowledge in a rapid and intuitive way. At present the database contains ~200 experimental records comprising 3 million individual data points.

Mapping of Affymetrix probes

A recent study constructed a high-density array of 25-nt probes to detect the transcribed sequences on Chromosomes 21 and 22. This study prepared cRNA probes from 11 cell lines that were hybridized to the oligonucleotide microarrays (Kapranov et al. 2002). These 25-nt probes were developed using the original Chromosome 22 contig sequences, corresponding to the initial Sanger Centre data release (Dunham et al. 1999). To relate our transcription data to the results of this study, a procedure was developed to map the positive oligonucleotide sequences to the present assembly of Chromosome 22 on which our microarray was constructed. The original contig sequences were obtained, and each was subdivided into 500-bp fragments. These subsequences were aligned with the present assembly of Chromosome 22q with BLASTN (Altschul et al. 1990), using a long word length of 400 bp to obtain a single optimal match for each fragment. The center positions of the positive oligonucleotides were known relative to the original contig sequences; an offset could therefore be computed for each 25-nt oligonucleotide with the offset shifting its coordinates according to the chromosomal location of the contig fragment on which the oligonucleotide was originally placed. Using this method, short oligonucleotide sequences could be accurately located on the updated chromosome assembly, while avoiding the many spurious homology matches that would result from comparing each 25-bp sequence to the entire chromosome directly.

Acknowledgments

We thank Cindy Jackson and James Hudson for oligonucleotide synthesis, and Cathy Cormier for critical comments on the manuscript. This research was supported by NIH grant P50 HG02357. J.L.R. is supported by an NIH predoctoral training grant; G.E. is supported by NIH postdoctoral fellowship F32 HG02446-01; P.B., R.M., and S.H. are supported in part by NIH grants P50 HG02357, 5 T32 HD07180, and R01 CA77808; and N.M.L. is funded by the Anna Fuller Fund for cancer research.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

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

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Balasubramanian, S., Harrison, P., Hegyi, H., Bertone, P., Luscombe, N., Echols, N., McGarvey, P., Zhang, Z., and Gerstein, M. 2002. SNPs on human chromosomes 21 and 22—Analysis in terms of protein features and pseudogenes. *Pharmacogenomics* **3:** 393–402.

Berman, P., Bertone, P., DasGupta, B., Gerstein, M., Kao, M.-Y., and Snyder, M. 2002. Fast optimal genome tiling with applications to microarray design and homology search. In *Proceedings of the Second International Workshop on Algorithms in Bioinformatics* (eds. R. Guigo and D. Gusfield), Lecture Notes in Computer Science, Vol. 2452, pp. 419–433. Springer, Heidelberg.

Blake, J.A., Richardson, J.E., Bult, C.J., Kadin, J.A., and Eppig, J.T. 2002. The Mouse Genome Database (MGD): The model organism database for the laboratory mouse. *Nucleic Acids Res.* **30:** 113–115.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., et al. 2001. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* **291:** 1289–1292.

Deloukas, P., Schuler, G.D., Gyapay, G., Beasley, E.M., Soderlund, C., Rodriguez-Tome, P., Hui, L., Matise, T.C., McKusick, K.B., Beckmann, J.S., et al. 1998. A physical map of 30,000 human genes. *Science* **282:** 744–746.

de Souza, S.J., Camargo, A.A., Briones, M.R., Costa, F.F., Nagai, M.A., Verjovski-Almeida, S., Zago, M.A., Andrade, L.E., Carrer, H., El-Dorry, H.F., et al. 2000. Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc. Natl. Acad. Sci.* **97:** 12690–12693.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.

Ewing, B. and Green, P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25:** 232–234.

Goryachev, A.B., Macgregor, P.F., and Edwards, A.M. 2001. Unfolding of microarray data. *J. Comput. Biol.* **8:** 443–461.

Guigo, R., Knudsen, S., Drake, N., and Smith, T. 1992. Prediction of gene structure. *J. Mol. Biol.* **226:** 141–157.

Harrison, P.M., Hegyi, H., Balasubramanian, S., Luscombe, N.M., Bertone, P., Echols, N., Johnson, T., and Gerstein, M. 2002. Molecular fossils in the human genome: Identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* **12:** 272–280.

Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Hughes, J.E., Snesrud, E., Lee, N., and Quackenbush, J. 2000. A concise guide to cDNA microarray analysis. *Biotechniques* **29:** 548–556.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. 2002. The Ensembl genome database project. *Nucleic Acids Res.* **30:** 38–41.

Hudson, T.J., Church, D.M., Greenaway, S., Nguyen, H., Cook, A., Steen, R.G., Van Etten, W.J., Castle, A.B., Strivens, M.A., Trickett, P., et al. 2001. A radiation hybrid map of mouse genes. *Nat. Genet.* **29:** 201–205.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P., and Gingeras, T.R. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916–919.

Kumar, A., Harrison, P.M., Cheung, K.H., Lan, N., Echols, N., Bertone, P., Miller, P., Gerstein, M.B., and Snyder, M. 2002. An integrated approach for finding overlooked genes in yeast. *Nat. Biotech.* **20:** 58–63.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., and Quackenbush, J. 2000. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* **25:** 239–240.

Pruitt, K.D. and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29:** 137–140.

Saccone, S., Caccio, S., Kusuda, J., Andreozzi, L., and Bernardi, G. 1996. Identification of the gene-richest bands in human chromosomes. *Gene* **174:** 85–94.

Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G., et al. 2001. Experimental annotation of the human genome using microarray technology. *Nature* **409:** 922–927.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Xu, Y. and Uberbacher, E.C. 1997. Automated gene identification in large-scale genomic sequences. *J. Comput. Biol.* **4:** 325–338.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. 2002. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30:** e15.