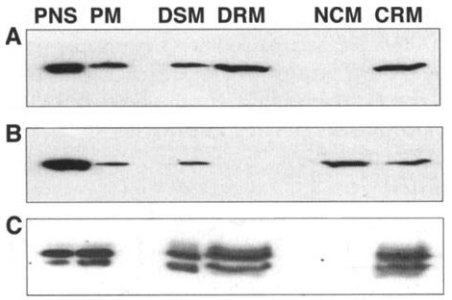

Fig. 5. Fractionation of MDCK plasma membrane reveals differential partitioning of acyl- and prenyl-modified YFP. MyrPalm-mYFP L221K and mYFP-GerGer F223R stably transfected MDCK cells were selected with 200 ng/ml G418 (Gibco). Fractionations were performed as described (*5*, *7*) with the addition of a Percoll gradient to separate plasma membrane (PM) from intracellular membranes in a postnuclear supernatant (PNS) (*22*, *23*) before carbonate or detergent fractionation. One-half of the PM was fractionated with detergent (*5*) and one-half with carbonate (*7*, *24*). Equal volume fractions were collected of detergent-soluble membrane (DSM), detergent-resistant membrane (DRM), noncaveolar membrane (NCM), and caveolae-rich membrane (CRM). (**A**) Western blot with antibody against GFP (Covance) of MyrPalm-mYFP fractions shows enrichment in detergent-insoluble and caveolae-rich fractions. (**B**) Blot by means of antibody against GFP of mYFP-GerGer fractions shows enrichment in detergent-soluble and noncaveolar membranes. (**C**) Immunoblot of endogenous caveolin by antibody against caveolin (Transduction Labs), with partitioning identical to that of the MyrPalm-mYFP.

FRET experiments on live cells. MyrPalm-mYFP (Fig. 5A) partitioned primarily into a detergent-resistant membrane fraction (DRM) and almost exclusively with low-density caveolae-rich membranes (CRM), consistent with the partitioning of endogenous caveolin (Fig. 5C). Conversely, mYFP-GerGer was relatively excluded from DRM and CRM (Fig. 5B) (*7*). These experiments further confirm that lipid modifications alone are sufficient to confer specific sublocalization into or outside of lipid rafts within the plasma membrane.

Our data examining fluorescent proteins attached to the cytosolic side of the plasma membrane complement previous studies using FRET between dye-labeled antibodies, toxins, or ligands to look for clustering of proteins anchored by glycosylphosphatidylinositol (GPI) linkages to the extracellular leaflet of the plasma membrane (*14*, *15*, *20*). Until now, all *Aequorea*-derived GFPs and mutants of any color have contained the hydrophobic patch of Ala^{206}, Leu^{221}, and Phe^{223} responsible for dimerization of the beta barrels. Our new mutations to positively charged residues should prevent dimerization of all colors and are advisable whenever assessing intermolecular interactions of pairs of GFP fusion proteins.

These mutants allowed direct determination in living cells that acylated proteins associate in a manner predicted by models for clustering of proteins into the liquid-ordered phase (*5*). They cluster with each other in lipid rafts and with full-length caveolin-1, a marker for caveolae, suggesting similar lipid compositions and environments in the two structures. Raft disruption with MβCD not only disaggregates acylated fluorescent proteins but also alters numerous signaling events (*15*, *16*, *21*), demonstrating the importance of these domains for cellular function.

References and Notes

1. K. Simons, D. Toomre, *Nature Rev. Mol. Cell. Biol.* **1**, 31 (2000).
2. E. Ikonen, *Curr. Opin. Cell. Biol.* **13**, 470 (2001).
3. C. C. Garner, J. Nash, R. L. Huganir, *Trends Cell. Biol.* **10**, 274 (2000).
4. M. Colledge, J. Scott, *Trends Cell. Biol.* **9**, 216 (1999).
5. K. A. Melkonian, A. G. Ostermeyer, J. Z. Chen, M. G. Roth, D. A. Brown, *J. Biol. Chem.* **274**, 3910 (1999).
6. P. S. Pyenta, D. Holowkà, B. Baird, *Biophys. J.* **80**, 2120 (2001).
7. K. S. Song *et al.*, *J. Biol. Chem.* **271**, 9690 (1996).
8. J. E. Schnitzer, D. P. McIntosh, A. M. Dvorak, J. Liu, P. Oh, *Science* **269**, 1435 (1995).
9. R. Y. Tsien, *Annu. Rev. Biochem.* **67**, 509 (1998)
10. A. Miyawaki, R. Y. Tsien, *Methods Enzymol.* **327**, 472 (2000).
11. M. D. Resh, *Biochim. Biophys. Acta* **1451**, 1 (1999).
12. P. J. Casey, *Science* **268**, 221 (1995).
13. CFP and YFP were enhanced versions ECFP and YFP 10C Q69K (*9*). The signal for geranylgeranylation (GerGer) was a COOH-terminal CLLL from Rho; PalmPalm, an NH_2-terminal MLCCMRRTKQ from Gap43; MyrPalm, an NH_2-terminal MGCIKSKRKDNLNDDE from Lyn kinase. Full-length cDNA of bovine caveolin-1 was fused to the NH_2-terminus of CFP and YFP. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
14. A. K. Kenworthy, M. Edidin, *J. Cell Biol.* **142**, 69 (1998).
15. Supplementary methods and references are available on *Science* Online at www.sciencemag.org/cgi/content/full/296/5569/913/DC1.
16. L. J. Pike, J. M. Miller, *J. Biol. Chem.* **273**, 22298 (1998).
17. T. M. Laue, W. F. Stafford, *Annu. Rev. Biophys. Biomol. Struct.* **28**, 75 (1999).
18. G. N. Phillips, in *Green Fluorescent Protein: Properties, Applications, and Protocols*. M. Chalfie and S. Kain, Eds. (Wiley-Liss, New York, 1998), pp. 77.
19. F. Yang, L. G. Moss, G. N. Phillips Jr., *Nature Biotechnol.* **14**, 1246 (1996).
20. R. Varma, S. Mayor, *Nature* **394**, 798 (1998).
21. S. Parpal, M. Karlsson, H. Thorn, P. Stralfors, *J. Biol. Chem.* **276**, 9670 (2001).
22. E. J. Smart, Y. S. Ying, C. Mineo, R. G. Anderson, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 10104 (1995).
23. R. D. Lasley, P. Narayan, A. Uittenbogaard, E. J. Smart, *J. Biol. Chem.* **275**, 4417 (2000).
24. V. O. Rybin, X. Xu, M. P. Lisanti, S. F. Steinberg, *J. Biol. Chem.* **275**, 41447 (2000).
25. We thank L. Burns, H. Purkey, and M. Petrassi for help with analytical ultracentrifugation and S. Adams, G. Walkup, R. Wachter, and J. Henley for discussion and suggestions. Supported by NIH grants NS27177 (R.T.), DK54441 (A.N.), 2T32 GM07752 (J.V.), and HHMI.

29 November 2001; accepted 25 February 2002

Large-Scale Transcriptional Activity in Chromosomes 21 and 22

Philipp Kapranov,[1] Simon E. Cawley,[1] Jorg Drenkow,[1] Stefan Bekiranov,[1] Robert L. Strausberg,[2] Stephen P. A. Fodor,[1] Thomas R. Gingeras[1]*

The sequences of the human chromosomes 21 and 22 indicate that there are approximately 770 well-characterized and predicted genes. In this study, empirically derived maps identifying active areas of RNA transcription on these chromosomes have been constructed with the use of cytosolic polyadenylated RNA obtained from 11 human cell lines. Oligonucleotide arrays containing probes spaced on average every 35 base pairs along these chromosomes were used. When compared with the sequence annotations available for these chromosomes, it is noted that as much as an order of magnitude more of the genomic sequence is transcribed than accounted for by the predicted and characterized exons.

Transcriptionally active regions of the human genome have been mapped by a combination of the alignment of cDNA sequences to genomic sequences and the annotation of genome sequences to predict coding regions (*1*–*4*). The goal of this study was to develop an empirical map of the transcriptionally active regions of the human genome at the nucleotide level and to relate this map to the sequence annotations derived from the two general approaches listed above. Previous reports (*5*) have relied on the annotations of the human genome to guide the authors in choosing which genomic regions to evaluate to determine the transcription profile of a cell type (exon arrays). In contrast, we have

[1]Affymetrix, Santa Clara, CA 95051, USA. [2]National Cancer Institute, Bethesda, MD 20892, USA.

*To whom correspondence should be addressed. E-mail: tom_gingeras@affymetrix.com

used an empirical approach to create a collection of transcript maps using uniformly spaced oligonucleotide probes (25-nucleotide oligomers) that interrogate either every base or on average every 35 base pairs (bp) of the sequences of human chromosomes 21 and 22 in a systematic fashion. The advantages of this approach include: (i) the identification of new regions of transcription not yet observed by previous experimentation or sequence analysis, (ii) the detection of RNA transcripts that have little or no coding capacity, and (iii) the identification of alternative RNA isoforms of previously annotated genes.

Labeled double-stranded cDNAs made from cytoplasmic polyadenylated [poly $(A)^+$] RNA from 11 different tumor and fetal cell lines [S1 (*6*)] were hybridized to high-density oligonucleotide arrays of two types. The first array design interrogated 362,901 contiguous nucleotides of DiGeorge syndrome minimal critical region (DGCR) (*7*–*9*) on chromosome 22 using a perfect complement (PM) and mismatch (MM) complement oligonucleotide probe for each base (DGCR array). The second design interrogated approximately 35 million nonrepetitive base pairs of chromosomes 21 and 22 (Chrom 21_22 arrays) using 1,011,768 probe pairs synthesized on a three-array set [S2 (*6*)].

Determination of whether or not a probe pair detected RNA target was made using a range of threshold values for the ratio (R) of PM to MM measurements and for the difference (D) of PM – MM values [S3 (*6*)]. Because of the overlap of interrogating probes used in the design of the DGCR array, it was possible to join positive probes separated at most by a certain distance (maxgap) and to then reject resultant regions whose length was less than a particular threshold (minrun), thus building maps with contiguous runs (contigs) of RNA [S4 (*6*)]. Contigs for the Chrom 21_22 array data were not constructed because of the distance between the probes used in this design. By fixing the R and D thresholds for each experiment, it was possible to calculate (i) false positive, specificity, and sensitivity rates on the basis of spiked bacterial RNA transcripts containing specific deletions and (ii) human sensitivity (only for the DGCR array) based on exon sequences detected by reverse transcriptase–polymerase chain reaction (RT-PCR) [S1 and table S1 (*6*)].

Chromosomes 21 and 22 have at least 225 and 545 genes, respectively, that are either well characterized or predicted. Of these, approximately 127 and 247 are "known genes" (*10*, *11*) that contain 1430 and 3134 exons on chromosomes 21 and 22, respectively (*12*). Figure 1 provides an overview of the previously identified and array-predicted transcription activity on chromosomes 21 and 22. By dividing the genome sequences of chromosomes 21 and 22 into 57-kb increments (average gene length on chromosome 21) (*11*), a total of 1220 gene-sized loci can be created across both chromosomes. Given that the average distance between each interrogating probe pair is 35 bp, the positive probe and exon densities (*13*) for each locus are plotted and can be compared. The correlation between the exon and positive probe densities demonstrated a significant relationship over the majority of the lengths of both chromosome sequences. Of the 1,011,768 probe pairs that interrogate approximately 35,000,000 nonrepetitive bp of both chromosomes, 26,516 (2.6%) probe pairs are located within the 4,564 annotated exons of well-characterized genes. Of these annotation-focused probes, 69.8 and 40.7% detect RNA transcript

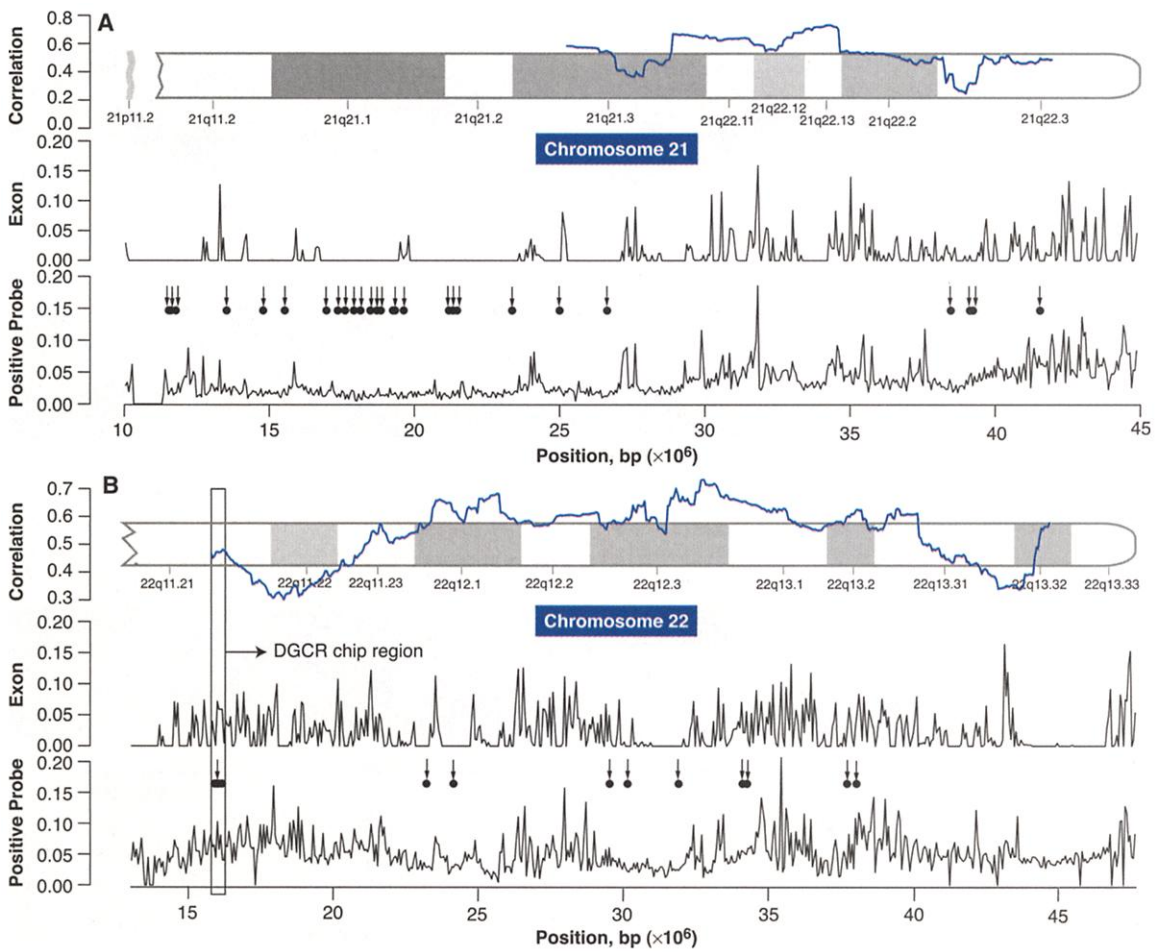

Fig. 1. Correlation of the positive probe and exon density maps [5% false positive (FP) rate] [S5 (*6*)] for chromosomes 21 (**A**) and 22 (**B**). For each map, the lowest graph depicts the positive probe density present in 57-kb bins (average genomic size for genes on chromosome 21 (*11*). Above this plot is the density of nucleotides located within exons present in each bin. The graph overlaying a cartoon of each chromosome is the local correlation coefficient of the exon density and the positive probe density calculated over a 5.7-Mb window. A correlation coefficient is not calculated in regions where the percentage of positive exon density falls below 25% over the 5.7-Mb window. Thus, the chromosome 21 region near the centromere that is relatively sparse in exon annotations is not analyzed for correlation with positive probe density given the relative lack of variation in the exon density. Above the positive probe density maps are the experimentally verified regions (downward arrows). The DGCR region of chromosome 22 is boxed in (B). High-resolution maps of the DGCR are shown in Fig. 2.

in at least 1 or 5 of 11 cell lines, respectively (Table 1). The percentage of the overall positive probes detected was 34.8 and 9.6% in 1 or 5 of 11 cell lines, respectively. This indicates that 94 and 88% of the probes detecting transcripts are located outside annotated exons in 1 or 5 of 11 cell lines, respectively. Approximately, 50% of these positive probes are located <300 bp distant from the nearest annotated exon. This is reflected in the close correlation between the exon and positive probe densities (Fig. 1).

A high-resolution map at 1-bp resolution that describes the locations of annotated exonic sequences and the array-based detected transcriptionally active regions has been developed [S3 and S4 (*6*)], and four examples of the annotated regions from DGCR are depicted in Fig. 2. The formation of contigs for this map allowed us to lower the estimated false positive rate for each of the 11 cell lines [S5 and S6 (*6*)]. Similar to that observed with the lower resolution maps of chromosomes 21 and 22, most of the detected transcripts (59.4 to 65.9%) are located away from the annotated exonic and expressed sequence tag (EST) sequences (Table 2).

At this resolution, for example, alternative transcripts in the DGCR6 locus were observed (Fig. 2A). Exons 1 and 5 appear to have RNA transcripts longer than previously represented. Additionally, there is evidence for transcriptional activity within intron 3. RT-PCR analysis and recent studies support these data (*14*).

Similar alterations could be made to the annotations of three other regions of DGCR (Fig. 2, B through D). RT-PCR analysis and subsequent sequencing of the transcripts in intron 5 revealed an extended version of DiGeorge syndrome gene E (DGS-E), as well as transcripts 5′ to this gene. Additional limited RT-PCR analysis provided confirmatory evidence for the presence of these and other transcripts within the DGCR2 locus (Fig. 2B). Similarly, novel transcripts have been observed and confirmed in intron 1 of DGCR5 (Fig. 2D) and in the vicinity of the highly expressed SCL25A1 gene (Fig. 2C). Thus, these maps have not only been useful in estimating the overall fraction of the human genome that is transcribed but they have also served as a guide for directing further biochemical and molecular efforts to isolate novel transcripts. High-resolution maps for the entire sequence of the DGCR and the nonrepeat sequences of chromosomes 21 and 22 are also available (*6*).

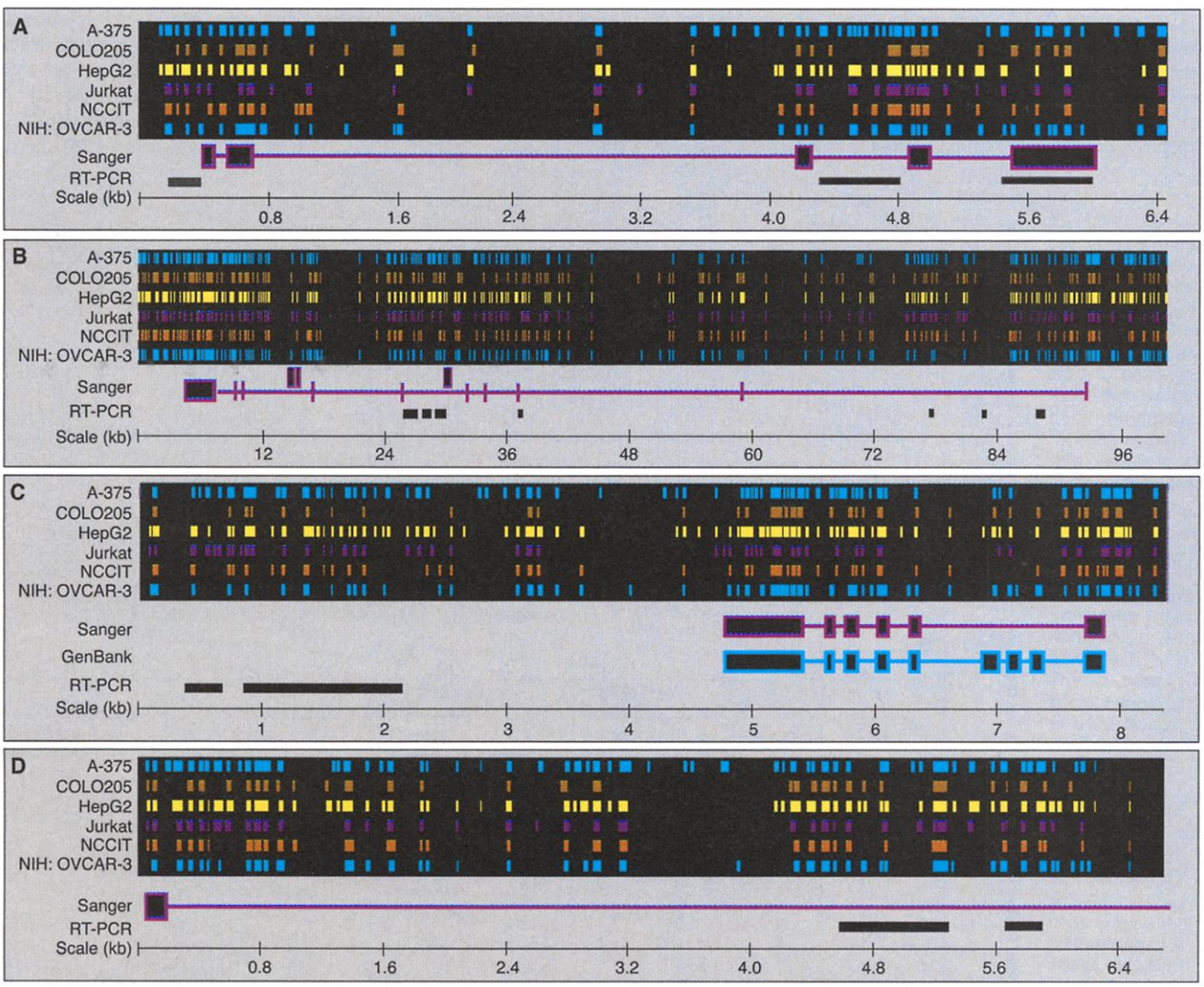

Fig. 2. High-resolution maps of four regions within DGCR of chromosome 22 (22q11.2). For each map, the contigs predicted by the DGCR array for 6 of the 11 cell lines analyzed is presented. Below the array map are cartoons derived from the Sanger hand-curated map of this region (*19*) or from GenBank. Selected regions suggested by the array map were further analyzed with the use of RT-PCR. The sequenced products from these analyses are mapped below the Sanger and ESTs maps (**A**) DGCR6 gene region [Golden Path (GP) sequence 15,833,950–15,840,390], (**B**) DGCR2 region (GP sequence 15,959,850–16,057,850), (**C**) SLC25A1 gene and flanking region (GP sequence 16,098,590–16,107,090), and (**D**) DGCR5 exon1 region (GP sequence 15,898,300–15,905,040).

A total of 63 individual, array-predicted transcription loci on chromosomes 21 and 22 distant from annotated exons and ESTs were selected. Our expectations were that these "annotation or transcript poor" loci would represent the predictions with the highest probability of being identified as false positives by the array, and, thus, successful identification of transcribed loci in such regions would serve as strong supportive evidence for the array-guided maps. Furthermore, any transcribed regions identified distant from known exons or ESTs are more likely to represent novel transcripts. PCR primers were designed at or near positive probes or contigs detected by the arrays. Two approaches were used for the verification: RT-PCR on cytosolic poly $(A)^+$ RNA and PCR on cDNA libraries prepared from cytosolic poly $(A)^+$ RNA of HepG2, NIH:OVCAR-3, and PC3 cell lines (*15*). Positive RT-PCR or PCR products were identified in 44 out of 63 loci [table S2 (*6*)]. Northern hybridization experiments were also conducted using poly $(A)^+$ RNA from 7 of the 11 cell lines and cloned RT-PCR products from 16 loci as probes. Seven of 16 loci yielded detectable RNA transcripts ranging in size from 0.6 to 10 kb, of which three loci were characterized by multiple transcripts of distinct or indistinct size [for examples see fig. S3 (*6*)].

The RT-PCR and sequence analyses of the cytosolic poly $(A)^+$ RNA samples and cDNA libraries indicated that at least 44 of 63 loci predicted by the array experiments to be sites of novel transcripts were transcribed. The filter-based hybridization experiments and large number of cycles required to detect RT-PCR or PCR products strongly suggest that the observed novel RNAs are present at low copy number per cell, providing some explanation as to why these transcripts have not previously been observed.

This survey of cytoplasmic poly $(A)^+$ RNA obtained from 11 developmentally diverse cell lines indicated that there may be as much as an order of magnitude greater sites of transcription of mature RNA transported into the cytoplasm than can be accounted for by the current annotation of the sequence of the human genome. However, it is important to note that the number of transcriptionally active sites does not necessarily correspond to an equal number of genes. The maps obtained in this study do not provide a description of the structural relations among the detected regions of transcription (i.e., the start and termination of each exon in a transcript). The detection of two adjoining regions of positive probes does not necessarily require that both probe sets are detecting the same RNA molecule. In fact, because of our use of labeled double-stranded cDNAs as targets in these experiments, the data presented here does not distinguish the strandedness of the detected RNAs. In some cases, detected expressed sequences cannot necessarily be guaranteed to originate from the transcription of chromosomes 21 and 22. Expression from paralogs in other genomic loci could lead to the same observed signals. However, the central conclusion to be drawn from these data is that a significantly larger proportion of the genome is transcribed and specifically transported as mature cytoplasmic poly $(A)^+$ RNA than previously considered.

Why were these sequences not observed previously? The answer to this question may reside in three observations. First, many of these detected transcripts may be structurally connected to other partially characterized RNAs (i.e., ESTs). Second, our estimates of the relative abundance of the majority of detected transcripts is very low, based on the required levels of RT-PCR and PCR amplification required to detect them in the RNA and cDNA libraries and the time of exposure needed to detect those observed by Northern hybridization [fig. S3 (*6*)]. Lastly, the sequences for many cDNA clones are discarded as possible heteronuclear RNA contaminants because of the low coding potential found.

The answer to the question of what function these transcripts might have requires additional large-scale characterization of the observed novel transcripts. Interestingly, noncoding RNAs are emerging as a rapidly expanding functional class of transcripts important for splicing, nucleolar and ribosomal structures, telomeric sequence addition, transport and insertion of proteins into membranes, down-regulation of translation, and chromosomal inactivation (*16*, *17*). Though for many of these noncoding RNA groups hundreds of new members are being identified in mammalian cells (*18*), the function of these identified transcripts must await additional characterization and perhaps reveal a hidden transcriptome.

Table 1. Proportion of chromosomes 21 and 22 transcribed. Analyzed were 1,011,768 probe pairs, of which 26,516 query exons as annotated in the known mRNA databases such as RefSeqs, Sanger hand-curated and GenBank. ESTs not included as part of expressed portion of genome. (*12*). Numbers in parentheses indicate percent positive probe pairs.

Cell lines	Positive probes overall	Positive probes in exons
1 of 11	268,466 (26.5%)	17,924 (67.6%)
5 of 11	98,231 (9.7%)	10,903 (41.1%)

Table 2. Proportion of DGCR (22q11.2) transcribed. The values are calculated on the basis of 213,009 probe pairs interrogating nonrepetitive (NR) bases of which 61,842 probe pairs are located within annotated expressed regions of the DGCR. The target false positive (FP) rate is that of each individual cell line [S5 and S6 (6)]. Expr. bases refers to the databases mentioned in Table 1 plus all the GenBank mRNAs and ESTs mapping to this region.

Cell lines	FP	Positive NR bases	Positive NR expr. bases	Positive NR nonexpr. bases
1 of 11	3%	50,885 (23.9%)	17,421 (34.2%)	33,464 (65.8%)
	5%	63,908 (30.0%)	21,788 (34.1%)	42,120 (65.9%)
5 of 11	3%	11,623 (5.5%)	4,724 (40.6%)	6,899 (59.4%)
	5%	20,097 (9.7%)	7,477 (37.2%)	12,620 (62.8%)

References and Notes

1. www.ncbi.nlm.nih.gov/LocusLink/
2. G. M. Rubin *et al.*, *Science* **287**, 2012 (2000).
3. H. Caron *et al.*, *Science* **291**, 1289 (2001).
4. F. A. Wright *et al.*, *Genome Biol.* **2**, 1 (2001).
5. R. J. Lipshutz *et al.*, *Nature Genet.* **21** (suppl.), 20 (1999); D. D. Shoemaker *et al.*, *Nature* **409**, 922 (2001).
6. Supplementary materials for this manuscript are being hosted at three Internet sites: Affymetrix at www.netaffx.com/transcriptome/; National Cancer Institute at http://cgap.nci.nih.gov/Info/2002.1; and *Science* Online at www.sciencemag.org/cgi/content/full/296/5569/916/DC1.
7. M. L. Budarf *et al.*, *Nature Genet.* **10**, 269 (1995).
8. M. Li *et al.*, *Am. J. Hum. Genet.* **55**, A10 (1994).
9. W. Gong *et al.*, *Hum. Mol. Genet.* **5**, 789 (1996).
10. I. Dunham *et al.*, *Nature* **402**, 489 (1999).
11. M. Hattori *et al.*, *Nature* **405**, 311 (2000).
12. Best in genome alignments of RefSeq, GenBank mRNA annotated as having complete CDS sequences and Sanger annotations have been combined to form the set of "known exons."
13. We calculate the fraction of positive probe pairs as the number of probe pairs defined as positive using R = 1.5 and D = 12Q (where Q is an estimate of chip-wide noise as computed by the Affymetrix GeneChip software) in at least 8 of 11 cell lines divided by the number of interrogating probe pairs, in nonoverlapping 57 kb windows for both chromosomes 21 and 22.
14. L. Edelmann *et al.*, *Genome Res.* **11**, 208 (2001).
15. RT-PCR procedure was carried out using the *C. therm.* Polymerase One-Step RT-PCR System (Roche, Indianapolis, IN). The RT-PCR procedure used 10 to 50 ng of cytosolic poly $(A)^+$ RNA, following the manufacturer's instructions. TaqGold polymerase (Applied Biosystems, Foster City, CA) was used to amplify putative transcripts from plasmid cDNA libraries (200 ng per reaction) following the manufacturer's instructions. At least 40 cycles of amplification were required to see the products. PCR products were cloned in pCR4-TOPO vector (Invitrogen, Carlsbad, CA) or sequenced directly.
16. V. A. Erdmann *et al.*, *Cell. Mol. Life Sci.* **58**, 960 (2001).
17. R. L. Kelley, M. I. Kuroda, *Cell* **103**, 9 (2000).
18. A. Huttenhofer *et al.*, *EMBO J.* **20**, 2943 (2001).
19. www.sanger.ac.uk/HGP/Chr22

20. We thank E. Schell, X.-M. Zhu, and M. Mittmann for design of photolithographic masks; G. Helt for implementation of mapping browser; A. Lau for assistance with figure preparation; M. Budarf for helpful discussions and cell lines; J. Corbeil and D. D. Richman for use of radiographic facilities and donation of two cell lines; L. Hong for cDNA libraries; and A. Lash, J. Thierry-Mieg, D. Thierry-Mieg, L. Wagner, N. Bhat, and L. Grouse for advice and guidance. Support for this work was provided by NCI/SAIC contract 21XS019A and by Affymetrix, Inc.

3 December 2001; accepted 22 March 2002