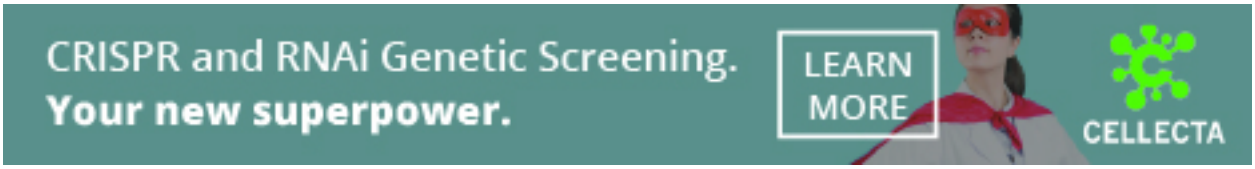
Human chromosome-specific cDNA libraries: new tools for gene identification and genome annotation.

R G Del Mastro, L Wang, A D Simmons, et al.

Genome Res. 1995 5: 185-194
Access the most recent version at doi:10.1101/gr.5.2.185



RESEARCH

Human Chromosome-specific cDNA Libraries: New Tools for Gene Identification and Genome Annotation

Richard G. Del Mastro,[1,2] Luping Wang,[1,2] Andrew D. Simmons,[1] Teresa D. Gallardo,[1] Gregory A. Clines,[1] Jennifer A. Ashley,[1] Cynthia J. Hilliard,[3] John J. Wasmuth,[3] John D. McPherson,[3] and Michael Lovett[1,4]

[1]Department of Biochemistry and the McDermott Center for Human Growth and Development, The University of Texas Southwestern Medical Center, Dallas, Texas 75235-8591; [3]Department of Biological Chemistry and the Human Genome Center, University of California, Irvine, California 92717

To date, only a small percentage of human genes have been cloned and mapped. To facilitate more rapid gene mapping and disease gene isolation, chromosome 5-specific cDNA libraries have been constructed from five sources. DNA sequencing and regional mapping of 205 unique cDNAs indicates that 25 are from known chromosome 5 genes and 138 are from new chromosome 5 genes (a frequency of 79.5%). Sequence complexity estimates indicate that each library contains ~20% of the ~5000 genes that are believed to reside on chromosome 5. This study more than doubles the number of genes mapped to chromosome 5 and describes an important new tool for disease gene isolation.

A detailed map of expressed sequences within the human genome would provide an indispensable resource for isolating disease genes, and would answer fundamental questions relating to gene number, density, and evolution. Unfortunately, this biological annotation of the physical map has been slow in occurring; only 5213 human "genes" have been regionally localized (Bowcock et al. 1995). Of these, only 311 gene sequences have been localized to human chromosome 5 and only 110 of these have been sublocalized within this chromosome. Currently, two strategies are used for mapping large numbers of genes: (1) positional cloning approaches (The Huntington's Disease Collaborative Research Group 1993; Miki et al. 1994; Lefebvre et al. 1995; Roy et al. 1995; Thompson et al. 1995) which use cloned genomic regions to isolate coding regions in a time-consuming "bottom up" approach that results in detailed transcription maps over small (< 1 Mb) regions of the 3000-Mb human genome; and (2) random cDNA sequencing approaches which yield DNA sequence information [Expressed Sequence Tags (eSTs)] (Adams et al. 1991, 1992, 1993a,b; Khan et al. 1991; Wilcox et al. 1991; Okubo et al. 1992). However, while random sequencing does aid gene discovery, it provides no mapping information per se. To provide a useful resource for disease gene isolation eSTs must be mapped in an expensive "top down" approach using genome-wide reagents (Cox et al. 1990; Green and Olson 1990; Khan et al. 1991; Wilcox et al. 1991; Cohen et al. 1993; Polymeropoulos et al. 1993). Both approaches result in gene maps (and eST data bases) that will be incomplete if only a limited spectrum of tissue types are sampled as cDNA libraries. In this report, we describe the use of an entire human chromosome to select cDNAs from complex cDNA pools (Lovett et al. 1991; Parimoo et al. 1991; Morgan et al. 1992; Lovett 1994a,b); in effect a positional cloning project on a chromosomal scale. Clones from these chromosome-specific cDNA libraries have been sequenced and mapped within existing chromosome-specific reagents, obviating the need to conduct genome-wide mapping, and considerably speeding up the analysis. Unlike conventional cDNA libraries, this new type of chromosome-specific reagent

[2]These authors contributed equally to this work.
[4]Corresponding author.
E-MAIL Lovett@ryburn.swmed.edu; FAX (214) 648-1666.

5:185–194 ©1995 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/95 $5.00

contains cDNAs that occur usually at very low levels, now at relatively high levels [quasi-normalization (Weissman 1987)]. They can be constructed from many different tissue types and have wider applications beyond random sequencing strategies; for example, as physical mapping reagents in hybridization-based analyses of genomic contigs.

RESULTS

Direct Selection

Our starting source for human chromosome 5 genomic DNA was a chromosome 5-specific cosmid library constructed by the Los Alamos National Laboratories, Los Alamos, NM (Longmire et al. 1993). This arrayed library of 24,768 cosmid clones represents ~5 chromosome 5 equivalents and a DNA sequence complexity of ~174 Mb. It contains at least one copy of all of the > 50 chromosome 5 loci that we have tested to date. (Warrington et al. 1991, Saltman et al. 1993), and a small (< 5%) contamination with hamster genomic DNA. Cosmid DNAs from all 24,768 Los Alamos National Laboratories chromosome 5 cosmids were prepared, biotinylated, and used en masse in a set of five separate but parallel direct cDNA selection experiments (Lovett 1994). The cDNA sources used were derived from human placenta, fetal brain, thymus, activated T-cells, and HeLa cells (Lovett 1994b). In all cases these were uncloned, PCR-amplifiable pools of cDNAs (to maximize the complexity of cDNAs sampled), derived from cytoplasmic RNAs (to avoid genomic DNA contamination and hnRNA artifacts). Chromosome 5 cosmids and the cDNA pools were hybridized in solution (Lovett 1994a,b) and the cosmids plus cognate cDNAs captured on streptavidin-coated paramagnetic beads. After washing, the cDNAs were eluted, PCR amplified, and recycled through an additional round of selection. To assess enrichment qualitatively, equal amounts of cDNA from the starting cDNA pool, primary selected cDNAs, and secondary selected cDNAs were analyzed by Southern blotting and hybridization with a panel of eight chromosome 5 genes previously described (Wang et al. 1989; Warrington et al. 1991). Figure 1A shows four examples of these controls (ANX6, SPARC, CD74, and COUP). All of these genes show enrichments in the selected material. However, the COUP enrichment appears to peak in the primary material and then slightly decrease in abundance (a pattern that is characteristic of the occurrence of abundance normalization) in contrast with the other cDNAs which continue to increase in the secondary selection (Weissman 1987; Lovett at al. 1991; Parimoo et al. 1991; Morgan et al. 1992; Lovett 1994a,b).

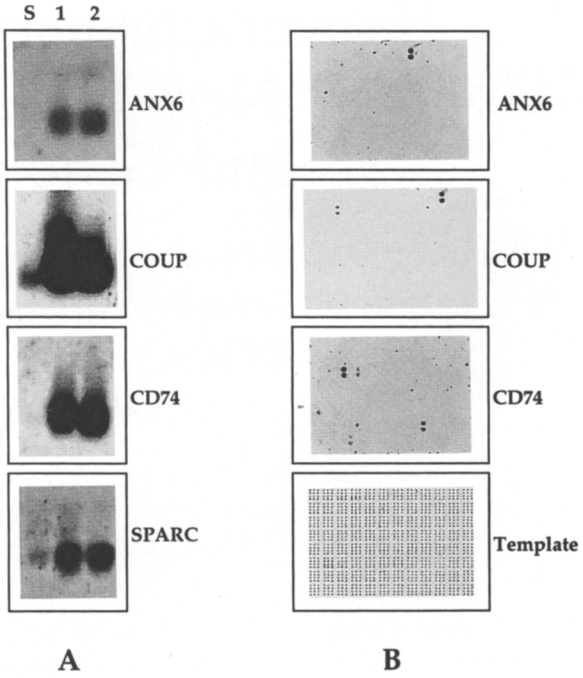


Figure 1 Examples of qualitative and quantitative measures of enrichment of specific chromosome 5 genes. (*A*) Four examples of autoradiographs from Southern blots hybridized with the designated cDNA probes. In each case the tracks contained 1 μg of the starting HeLa cDNA (lane *S*), the primary selected cDNAs (lane *1*), and the secondary selected cDNAs (lane *2*). (*B*) The result of hybridizing the same cDNA probes to a sixfold density filter array of cDNAs (576 clones of the 4608 total are shown here) arrayed in the template pattern shown at the bottom. Each clone was spotted twice to yield a duplicate positive signal when hybridized with radiolabeled probes. Examples of positive signals within this array are shown for ANX6 (one hit), COUP (two hits), and CD74 (four hits).

Quantitation of Enrichment

The secondary selected cDNAs were then cloned into a plasmid vector and 4608 clones were randomly picked from each tissue source, for a total of five selected cDNA libraries. These clones were formatted into high-density, duplicate clone ar-

rays on filters and were hybridized with full-length control reporter genes to quantitate the number of clones present in each library of 4608 clones. Examples of the detected positives within the HeLa array are shown in Figure 1B. To quantitate the abundance of the control cDNAs prior to enrichment, the starting HeLa cDNA was cloned in a bacteriophage vector and 500,000 recombinant plaques were screened by hybridization for the presence of the control genes. These data and data from other selected libraries are summarized in Table 1. Not all of the control genes are expressed in all of the starting tissues. However, in all cases but one (SPARC in the HeLa library), where a control could be detected by PCR in the starting cDNA, it was also present at least once in the array of 4608 clones. As shown in Figure 1A, the SPARC gene is enriched, but is not present in the 4608 picked clones, indicating that the HeLa library has a sequence complexity of > 4608. SPARC was detected twice in a separate screen of 8000 phage clones derived from the same material (Table 1). Substantial enrichments have been obtained in these selections; ~540-fold for the CD74 gene and ~40-fold for the RPL7 cDNA which starts at a relatively high abundance and is actually selected by a pseudogene (see below). Some level of abundance normalization (Weissman 1987; Lovett at al. 1991; Parimoo et al. 1991; Morgan et al. 1992; Lovett 1994a,b) is also evidenced in these figures, with the range of abundance of the controls spanning 21-fold in the HeLa selection (data not shown). The exception to these relatively uniform abundance numbers is the placental selection, in which some individual controls are as high as 2.0% and others are as low as 0.02%. However, the placental selected library was cloned after only one round of selection and it is the second round of selection that is designed to accomplish abundance normalization (Weissman 1987; Lovett at al. 1991; Parimoo et al. 1991; Morgan et al. 1992; Lovett 1994a,b). On average, the frequency of known genes in the selected material is 5/4608 indicat-

Table 1. Summarized enrichment data on cDNA Libraries

Gene[a]	HeLa starting[b]	HeLa selected (%)[c]	Fold enrichment[d]	Selected (%)[e]		
				fetal brain	placenta[f]	T cell
ANX6	<1 in 5×10^5	6 (0.13)	>600	0.20	1.0	0.20
GM2A	<1 in 5×10^5	7 (0.15)	>700	0.10	0.35	0.02
CSF1R	<1 in 5×10^5	1 (0.02)	>100	—	—	—
IRF1	<1 in 5×10^5	1 (0.02)	>100	—	—	0.19
CD74	4 in 5×10^5	20 (0.43)	540	—	2.0	—
SPARC		0 (0.02)[g]	—	0.02	0.15	—
COUP	<1 in 5×10^5	3 (0.06)	>300	—	—	—
FLT4	—	—	—	—	0.02	—
RPL7	10 in 5×10^5	4 (0.08)	40	—	0.17	—

(—) A particular gene was either not expressed in a starting source or was not tested.

[a]Eight cDNA hybridization probes to chromosome 5 genes and one probe to a chromosome 5 pseudogene are listed: annexin 6 (ANX6), GM2 ganglioside activator (GM2A), colony stimulating factor 1 receptor (CSF1R), interferon regulatory factor 1 (IRF1), the minor histocompatibility antigen CD74, the gene for osteonectin (SPARC), the human homolog of the chicken ovalbumin upstream regulatory protein (COUP), the FMS-like tyrosine kinase 4 gene (FLT4), and the ribosomal protein L7 gene (RPL7). The last of these genes is not located on chromosome 5. However, a processed RPL7 pseudogene is located close to CSF1R on chromosome 5.

[b]The number of duplicate positive signals in a screen of a library of 500,000 recombinant phage plaques, derived from HeLa cell cDNA prior to selection, when screened with the various control cDNAs.

[c]The number of duplicate signals detected per 4608 arrayed clones. Numbers in parentheses are percentages overall in all other columns.

[d]Fold enrichments are rounded to the nearest hundred when a gene was undetectable in the starting source and to the nearest ten in the two cases where positives were found in the starting source.

[e]Average insert lengths for the HeLa library were 250 and 500 bp for the placental, fetal brain, thymus, and activated T-cell libraries.

[f]The placental selected library was cloned after the first round of direct selection in contrast with all others that were cloned after two rounds of selection.

[g]A screen of 4608 arrayed clones detected no positive signals; however, a screen of 8000 phage clones derived from the same selected material detected 2 positive clones.

ing that the overall frequency of a given gene is ~1/1000. This suggests that each of the cDNA libraries described herein contains ~20% of all chromosome 5 genes (assuming 5000 chromosome 5 genes).

Validation of Chromosome 5-selected cDNA Libraries by Sequence Analysis

To further validate these libraries, the DNA sequences of 261 randomly picked HeLa-selected clones were derived and analyzed. These data are summarized in Tables 2 and 3. Of the cDNAs sequenced, 25 (9.6%) were 100% homologous to known human chromosome 5 genes (see Table 3 for examples). Among these were cDNAs derived from very low abundance transcripts such as those encoding the transcription factors *IRF1* and *EGR1*. A small percentage of sequenced clones (1.5%) represented a definite background clone (rRNA) and an additional 19.9% contained either CA repeats, Alu repeats, L1 repeats, O-family repeats, or vector sequences. All of these repeats are relatively trivial to remove from any further analysis of these libraries. However, a pre-screen for these was not conducted prior to sequencing. The 10.3% frequency of Alu repeat-containing clones is not significantly different than their frequency in conventional cDNA libraries (Crampton et al. 1981). We have chosen conservatively

Table 2. Summarized DNA Sequencing and Chromosomal Mapping Data

Sequence class[a]	Number[b]	Percentage	Number on chromosome 5[d]	Percentage of total on chromosome 5[e]	Percentage of nonrepetitives on chromosome 5[f]
Genes already known to be located on chromosome 5	25	9.6	25	9.6	12.2
Known genes not already localized to chromosome 5	15*	5.8	9	3.4	4.4
Previously described but unmapped ESTs	23*	8.8	18	6.9	8.8
Novel sequences	142*	54.4	111	42.5	54.1
Alu repeat containing	27	10.3	N.D.	N.D.	N.D.
L1 repeat containing	8	3.0	N.D.	N.D.	
CA repeat containing	15	5.8	N.D.	N.D.	N.D.
O repeat containing	1	0.4	N.D.	N.D.	N.D.
Vector containing	1	0.4	N.D.	N.D.	N.D.
rRNA	4	1.5	0	0	0
Totals	261	100	163	58.8	79.5

(N.D.) Not determined. Mapped primer sets have been deposited with the Genome Data Base (GDB), and specific loci can be found by access to that resource.
[a]Various classes of DNA sequences identified by searches using the BLAST-X, BLAST-N, and FASTA programs for all sequenced DNAs.
[b]Number of sequences in a. PCR primers were designed and synthesized to all sequence classes marked with an asterisk.
[c]The percentage of the total number of sequences.
[d]Number of sequences that were localized to chromosome 5 by a PCR assay of the hybrid panel described in the text and shown in Fig. 2.
[e]The percentage of cDNAs that mapped to chromosome 5.
[f]Percentage values adjusted for removal of repeat-containing clones, which can be readily achieved by screening with radiolabeled repeat probes prior to picking for sequence analysis.

Table 3. Representative examples of sequence similarities and mapping data for chromosome 5 genes

cDNA clone name	Length (bp)	BLAST-N/BLAST-X	Percent homology	Accession no.	Localization on chromosome 5
3G7	273	GP36b glycoprotein	100	U10362	5
1B1	272	DAP-1	100	X76105	5p15.2
1D8	161	49-kD protein	100	L22009	5q32–qter
3A11	282	TGF-β-induced gene H3*	100	M77349	5q31
2-4	274	protocadherin 42	94	L11370	5q32–q33
1B3	161	interferon regulatory factor 1*	100	X14454	5q23–q31
3F2	217	GM2 activator protein*	96	X61094	5q32–q33
3E5	146	putative novel receptor kinase (GPRKG)*	96	U00686	5
4B11	200	calphobindin II (ANX 6)*	100	D00510	5q32–q34
3F1	218	c-fms proto-oncogene for CSF-1*	100	X14720	5q33.3–q34
1H12	161	early growth response protein 1*	100	X52541	5q23–q31
3G11	185	NIB1948 cDNA 3′ end	93	T16862	5q32
1H5	160	partial cDNA sequence; clone c-24g06	79	Z44585	5q32
1G3	336	cDNA clone f11009 5′ end	84	T19261	5q
1F3	179	3′-directed *Mbo*I cDNA; clone pm2938	100	D20052	5q32
3G2	253	cDNA clone c08013 5′ end	87	T19194	5q33.2–qter
3C10	273	EST92610	100	T35864	5q
3B12	248	cDNA clone HIBBF39 5′ end	100	T08503	5q23
3C9	174	novel/Connexin 40	66	P33725	5q
3B5	200	novel/bone morphogenetic protein	73	L35278	5q
2-6	209	integral membrane protein E16	90	M80244	5q14
2-10	227	novel			5q33.2–qter
3B7	189	novel			5p13
4C11	175	novel			5q15–p21
4B7	136	novel			5p11–p13
1D10	249	novel			5p15
3H6	207	novel			5p11–p13
3D2	232	novel			5p15
1-2	260	novel			5q33.2–qter
3G5	179	novel			5p11–p12

Thirty clones are shown with their respective lengths in bp, homologies as detected by Blast-N, or, where BLAST-N showed no significant homologies, a (/) indicates a BLAST-X homology. Seven previously known chromosome 5 genes that were sequenced as part of our random sample are shown by an asterisk (*). Localizations within chromosome 5 are also shown. 5q indicates that the cDNA yielded an ambiguous sublocalization within the deletion hybrids but is clearly on the long arm.

to classify these clones as background clones. However, most of these sequences (with the exception of rDNA clones) are comprised partly of repetitive and partly of novel sequence. It is possible that some of these may be bona fide chromosome 5 sequences, but we have not attempted to map them. A total of 23 sequenced cDNAs were homologous to eSTs present in the publicly accessible data bases and a further 157 were either novel by homology searches ($n = 142$), were 100% homologous to known genes that had not yet been chromosomally localized, or were homologous to a known gene ($n = 15$ for the latter two cases).

Mapping of cDNAs to Chromosome 5

Oligonucleotide primers were designed and synthesized for a total of 180 cDNAs shown in Table 2 and were initially scored by PCR on a mapping panel consisting of human genomic DNA, hamster DNA, HHW105 [a chromosome 5 monochromosomal hybrid in a hamster background (Gilliam et al. 1989)], A15 [a chromosome 15

monochromosomal hybrid in a hamster background (McDaniel and Schultz 1992)], HeLa cDNA, and a negative control. Of the 180 cDNAs tested, 138 cDNAs mapped specifically to chromosome 5 (see Fig. 2 and Table 2) Of these, 9 were unlocalized genes described previously (see Fig. 2 and Table 3 for examples), 18 were eSTs that had not been chromosomally assigned previously, and 111 were novel sequences. In many cases, the single copy nature of the cDNAs was confirmed by whole genome Southern blotting, in addition to PCR assays and DNA sequence comparisons (data not shown). Thus, of the 261 sequenced cDNAs a total of 163 (58.8%) were located on human chromosome 5. When all repeat-containing clones are discounted, 79.5% (163/205) of tested cDNAs were specifically located on chromosome 5. Of the remaining 42 cDNAs, 18 were nonspecific background from other chromosomes, 16 yielded PCR products from HeLa cDNA and not from human genomic DNA indicating that they might cross an intron, and 8 yielded a product on hamster DNA that comigrated with the human product. This latter class presumably represents cDNAs that are highly conserved between human and hamster. Some of these may be genuine chromosome 5 loci. Alternatively, these may be human cDNAs that were spuriously enriched by the contaminating hamster genomic sequences present in the Los Alamos National Laboratory chromosome 5 cosmid library.

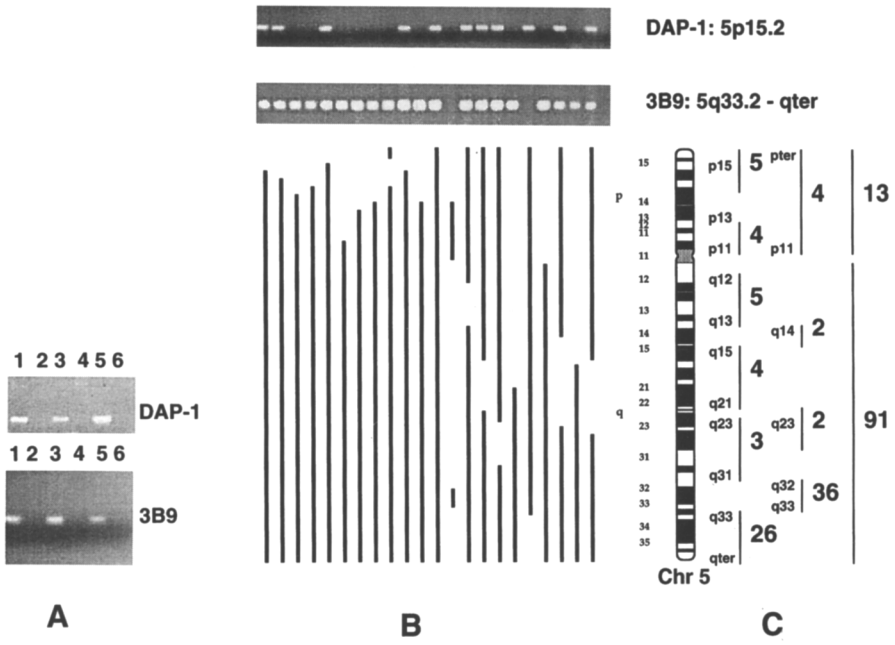


Figure 2 Chromosomal and subchromosomal mapping of cDNAs. (*A*) Two ethidium bromide-stained agarose gels of PCR products derived using primers for clone 1B1 (100% homologous to DAP-1; see Table 3) and 3B9 a cDNA with no significant sequence homologies. The DNAs used in each PCR were human DNA (lane *1*), CHO hamster DNA (lane *2*), HHW105 a monochromosomal 5 hybrid DNA (lane *3*), A15 a monochromosomal 15 hybrid DNA (lane *4*), secondary selected cDNAs (lane *5*), and primers alone (lane *6*). (*B*) Results obtained with the same two sets of primers on 21 deletion hybrids of chromosome 5 plus HHW105 which contains an intact chromosome 5. Bars indicate the areas of the chromosome that are present in each hybrid and they are aligned with the ideogram at *right*. Agarose gel electrophoresis of PCR products of DAP-1 and 3B9 are shown at *top* and are aligned with the various deletion hybrids. DAP-1 localizes to 5p15.2 and 3B9 localizes to 5q33.2-qter. (*C*) Summarizes the number of 104 cDNAs that could be binned to the various locations shown by the bars at the right side of the ideogram. In some cases only a 5q or 5p location was determined. A total of 91 reside on the q arm and 13 reside on the p arm.

Localization of cDNAs within Chromosome 5

A total of 104 sequenced cDNAs were regionally mapped within chromosome 5 using a panel of 21 somatic cell hybrids constructed from deletions of human chromosome 5 (Brant et al. 1993; McPherson et al. 1994). DNAs from these hybrids were used in a series of PCRs to regionally localize the cDNAs. Figure 2 shows an ideogram of chromosome 5 and the various deleted chromosomal regions in the hybrids, as well as two representative sets of PCRs. This figure also summarizes the regional localization of the 104 cDNAs. Thirteen are located on 5p and 91 are located on 5q. The apparent ratio of lengths of 5p:5q is 1:3, and the ratio of 5p:5q genes found here (1:7) may reflect a difference in gene distribution between the arms. A larger than expected number of cDNAs are located in the 5q32-qter region (62/104). This may be the result of a bias in the cosmid source

or in the HeLa cDNAs. However, fluorescence in situ hybridization (FISH) experiments conducted with random cosmids from the library do not reveal any overt bias in the cosmid source (data not shown). It is interesting to note that this is the same region to which a large proportion of CpG islands map within chromosome 5 (Craig and Bickmore 1994).

DISCUSSION

To place these observations in perspective, to date, only 110 expressed sequences have been regionally localized within human chromosome 5 (Bowcock et al. 1995). This represents between 2% and 5% of the total number of genes estimated to reside on this chromosome (Bishop et al. 1974; Fields et al. 1994). Chromosome 5 comprises ~5% of the human genome and thus, randomly picking and sequencing cDNAs from a normalized (but not chromosome-specific) cDNA library (Ko 1990; Patanjali et al. 1991; Soares et al. 1994) should on average reveal a known chromosome 5 gene with a frequency of 0.25% (5%×5%), and an entirely new chromosome 5 gene with a frequency of 4.75% (5%–0.25%). The frequency we observed in this study of 9.6% overall for known chromosome 5 genes, and 79.5% overall for new chromosome 5 genes, thus represent ~40-fold and ~15-fold enrichments, respectively, over the predicted random distribution of cDNAs in a perfectly normalized library.

It is important to appreciate three limitations that are inherent in this approach, and which are shared by the random eST approach. The first limitation is one of completion. A map based upon cDNAs will inevitably miss some genes that exhibit spatially or temporally restricted patterns of expression. This is exacerbated when only a few tissue sources are used for cDNA sequencing. Structural approaches, such as that described by J.A. Trofatter, K.R. Long, J.R. Murrell, C.J. Stotler, J.F. Gusella, and A.J. Buckler (in prep.) do not suffer from this expression-based limitation. For selected libraries this can be ameliorated to some extent by using cDNA from several tissue types at once, as described here. However, even when many tissue sources are used, direct selection, because it is hybridization and PCR dependent, is unlikely to select every gene across a large genomic region (Weissman 1987; Lovett et al. 1991; Parimoo et al. 1991; Morgan et al. 1992; Lovett 1994a). The second limitation involves mapping errors. Direct selection will select related members of gene families and cDNAs homologous to pseudogenes (Weissman 1987; Lovett at al. 1991; Parimoo et al. 1991; Morgan et al. 1992; Lovett 1994a). The RPL7 cDNA used as a control in this study is relevant in this regard. The RPL7 structural gene is not located on chromosome 5. However, an RPL7 processed pseudogene is located on 5q32–q33 close to the CSF1R locus on chromosome 5 (G. Clines and M. Lovett, unpubl.), and this selects its homologous cDNA (see Table 1). This type of mapping error is usually, but not always, eliminated when chromosomal localization is conducted. This type of error also will occur in genome-wide eST mapping. The third limitation concerns the length of selected cDNAs. Because of the length bias inherent in PCR-based schemes, selected cDNAs are an average of 500 bp and represent fragments of the entire cDNA (see footnote to Table 1). Placing these within the physical map of the genome adds considerably to the biological information content of the map, but ideally, these cDNAs would be near full length and unequivocally mark single transcription units.

These limitations are offset to a large degree by the high levels of enrichment and sequence complexity afforded by selected libraries. The mapping of 138 new cDNAs described in this study more than doubles the number of expressed sequences localized within this chromosome. As was mentioned above, the HeLa selected library has a sequence complexity of > 4608 clones and has a reasonably flat cDNA abundance profile indicative of some normalization having occurred. Taken together, these observations argue that this library and the other four that we have constructed, will provide an in-depth and efficient source for the large majority of the cDNAs encoded by human chromosome 5. Therefore, it should be a valuable resource for isolating genes that represent chromosome 5 disease loci. In this study we utilized deletion hybrid mapping panels to regionally localize new genes. The next logical step in the high throughput and detailed mapping of chromosome 5-specific cDNAs, will be the use of chromosome 5 radiation-reduced hybrids [RH (Warrington et al. 1991, 1992; Saltman et al. 1993)] and parallel analysis of chromosome 5-specific cosmids and yeast artificial chromosomes (YACs). This should represent an efficient and cost effective means of rapidly annotating the cosmid, YAC, and RH maps of this chromo-

some with cDNAs derived from five (or more) tissue sources at once. The derivation of such maps will be an essential step in speeding up the isolation of human genetic disease loci.

METHODS

Isolation of Cosmid DNA and Construction of cDNA Pools

The two hundred and fifty-eight 96-well microtiter plates of clones from the Los Alamos National Laboratories chromosome 5-specific cosmid library (Longmire et al. 1993) were stamped onto LB agar containing kanamycin (10 mg/ml), and grown overnight at 37°C. The colonies were scraped off the agar and the cosmid DNA isolated by the alkaline lysis method (Sambrook et al. 1989). The five starting cDNA pools were constructed as described (Lovett et al. 1991; Warrington et al. 1992), using cytoplasmic polyadenylated RNAs (Sambrook et al. 1989). Thymus, placenta, and HeLa cell cDNAs were ligated to oligonucleotides 1 and 2; and fetal brain and activated T-cell cDNAs were ligated to oligonucleotides 3 and 4 (Lovett et al. 1991; Lovett et al. 1994b).

Direct Selection

Direct selection was performed as described (Lovett 1994b) to a Cot½ of 120, except that in the first round of selection, 0.1 μg of cDNA and 1.0 μg of biotin-labeled cosmid DNA (from the entire cosmid library) were used. cDNAs were blocked with total human DNA (1 μg) and cosmid vector DNA (1 μg). Secondary selections were conducted using 0.1 μg of primary selected cDNAs and 1 μg of total cosmid library DNA. An annexin 6 cDNA clone was used as a control to monitor enrichment during the two rounds of selection. Further assessment of enrichment in the two rounds of selection was performed (Lovett et al. 1991; Lovett 1994b) using previously described genes (Table 1; Warrington et al. 1991; Warrington et al. 1992; Saltman et al. 1993).

Cloning, Arraying, and Gridding of the Selected Product

The primary selected products from placenta and the secondary selected products from the other four tissue sources, were cloned into the UDG vector pAMP10 (GIBCO–BRL), in accordance with the manufacturer's recommendations. Prior to cloning, the secondary selected material was PCR amplified using modified primers of oligonucleotides 1 and 3. The modified primer sequences were Oligonucleoide 1-CUA; 5′-CUACUACUACUACTGAGCGGAATTCGTGAGACC-3′ and Oligonucleotide 3-CUA; 5′-CUACUACUACUACTCGAGAATTCTGGATCCTC-3′.

The cloned secondary selected material, from each tissue source, was transformed into MAX Efficiency DH5a Competent Cells (GIBCO–BRL) as recommended by the manufacturer. Clones (sample size 4608) were picked from each transformed tissue source, and arrayed into forty-eight 96-well microtiter plates. Each selected cDNA library was stamped, in duplicate, in a high-density format onto Hybond N+ nylon membrane (Amersham) using a Biomek 1000 Automated Laboratory Workstation (Beckman). The bacteria were grown overnight at 37°C, and the membranes processed as recommended by the manufacturer.

Quantitation of Enrichment

To quantitate the abundance of known chromosome 5 genes within the 4608 clones, gel-purified PCR fragments of known genes, radiolabeled with ^{32}P were individually hybridized at 65°C to the high-density filters of the selected cDNA libraries as described (Simmons et al. 1995). The filters were washed three times in buffer (0.1× SSPE, 0.1% SDS) at 65°C, and were autoradiographed or phosphorimaged (Molecular Dynamics).

Sequence Analysis

HeLa cDNA clones were sequenced using a dye primer terminator cycle sequencing kit (Applied Biosystems), and the data collected by the ABI 373A automated fluorescence sequencer (Applied Biosystems). All sequences were analyzed using the BLAST-N, BLAST-X, and FASTA programs (Altschul et al. 1990).

Mapping Analysis

PCR analysis was used to map the HeLa cDNAs using a panel of mapping DNAs described in the text and figure legends. Oligonucleotide primer sets were designed from the cDNA sequences and were used under the following PCR conditions: 30s at 94°C, 30s at 59°C, and 30s at 72°C for 30 cycles. PCR reactions were performed in a Perkin-Elmer 9600; each 25-μl reaction contained 100 ng DNA, 10 μM each of primer, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.001% gelatin, 200 mM each dNTPs, and 1 unit of *Taq* DNA polymerase (Perkin-Elmer). Primer sets that mapped to chromosome 5, were localized regionally using a panel of 21 somatic cell hybrids under the same PCR conditions as described above. All sequences mapping to chromosome 5 have been submitted as PCR primer pairs to Genome Database, Baltimore (GDB) along with regional assignments (accession nos. G00-624-854 and G00-625-915). The arrayed chromosome 5-specific cDNA libraries described here are available to investigators interested in human chromosome 5 on a cost recovery basis, and subject to participation in a data sharing consortium.

Ambiguities in mapping were scored when one or more hybrids produced negative results that resulted in an uninterpretable regional assignment. In several cases this still allowed assignment to a more general region. We interpret these ambiguities as resulting from as yet uncharacterized microdeletions within some hybrids. Radiation hybrid mapping will eventually resolve these assignments.

ACKNOWLEDGMENTS

We are grateful to Dr. Anne Bowcock for helpful comments on the manuscript. This work was supported by

grant HG00834 from the National Institutes of Health to J.J.W.; and grants HG00368, HG00734, and HG00882 from the National Institutes of Health to M.L.

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

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, and R.F. Moreno. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252:** 1651–1656.

Adams, M.D., M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelley, T.R. Utterback, J.W. Nagle, C. Fields, and J.C. Venter. 1992. Sequence identification of 2,375 human brain genes. *Nature* **355:** 632–634.

Adams, M.D., A.R. Kerlavage, C. Fields, and J.C. Venter. 1993a. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nature Genet.* **4:** 256–267.

Adams, M.D., M.B. Soares, A.R. Kerlavage, C. Fields, and J.C. Venter. 1993b. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nature Genet.* **4:** 373–380.

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

Bishop, J.O., J.G. Morton, M. Rosbash, and M. Richardson. 1974. Three abundance classes in HeLa cell messenger RNA. *Nature* **250:** 199–204.

Bowcock, A.M., M.A. Chipperfield, P. Ceverha, A. Minter-Morrison, B. Bakker, and C.J. Porter. 1995. Report of the DNA committee. Genome Database Baltimore, Maryland. *Cytogenet. Cell Genet.* (in press).

Brant, S.R., M. Bernstein, J.J. Wasmuth, E.W. Taylor, J.D. McPherson, X. Li, S. Walker, J. Pouyssegur, M. Donowitz, C.M. Tse et al. 1993. Physical and genetic mapping of a human apical epithelial Na^+/H^+ exchanger (NHE3) isoform to chromosome 5p15.3. *Genomics* **15:** 668–672.

Cohen, D., I. Chumakov, and J.A. Weissenbach. 1993. First-generation physical map of the human genome. *Nature* **366:** 698–701.

Cox, D.R., M. Burmeister, E.R. Price, S. Kim, and R.M. Myers. 1990. Radiation hybrid mapping: A somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250:** 245–250.

Craig, J.M. and W.A. Bickmore. 1994. The distribution of CpG islands in mammalian chromosomes. *Nature Genet.* **7:** 376–384.

Crampton, J.M., K.E. Davies, and T.F. Knapp. 1981. The occurrence of families of repetitive sequences in a library of cloned cDNA from human lymphocytes. *Nucleic Acids Res.* **9:** 3821–3834.

Fields, C., M.D. Adams, O. White, and J.C. Venter 1994. How many genes in the human genome? *Nature Genet.* **7:** 345–346.

Gilliam, T.C., N.B. Freimer, C.A. Kaufmann, P.P. Powchik, A.S. Bassett, U. Bengtsson, and J.J. Wasmuth. 1989. Deletion mapping of DNA markers to a region of chromosome 5 that cosegregates with schizophrenia. *Genomics* **5:** 940–944.

Green, E.D. and M.V. Olson. 1990. Systematic screening of yeast artificial chromosome libraries by use of the polymerase chain reaction. *Proc. Natl. Acad. Sci.* **87:** 1213–1217.

Khan, A.S., A.S. Wilcox, J.A. Hopkins, and J.M. Sikela. 1991. Efficient double stranded sequencing of cDNA clones containing long poly(A) tails using anchored poly(dT) primers. *Nucleic Acids Res.* **19:** 1715.

Khan, A.S., A.S.Wilcox, M.H. Polymeropoulos, J.A. Hopkins, T.J. Stevens, M. Robinson, A.K. Orpana, and J.M. Sikela. 1992. Single pass sequencing and physical and genetic mapping of human brain cDNA. *Nature Genet.* **2:** 180–185.

Ko, M.S. 1990. An equalized cDNA library by the reassociation of short double-stranded cDNAs. *Nucleic Acids Res.* **18:** 5705–5711.

Lefebvre, S., L. Burglen, S. Reboullet, O. Clermont, P. Burlet, L. Viollet, B. Benichou, C. Cruaud, P. Millasseau, M. Zeviani et al. 1995. Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* **80:** 155–165.

Longmire, J.L., N.C. Brown, L.J. Meincke, M.L. Campbell, K.L. Albright, J.J. Fawcett, E.W. Campbell, R.K. Moyzis, C.E. Hildebrand, G.A. Evans et al. 1993. Construction and characterization of partial digest DNA libraries made from flow-sorted human chromosome 16. *GATA* **10:** 69–76.

Lovett, M. 1994a. Fishing for complements: Finding genes by direct selection. *Trends Genet.* **10:** 352–357.

Lovett, M. 1994b. Direct selection of cDNAs using genomic contigs. In *Current protocols in human genetics* (ed. J. Seidman), pp. 6.3.1. Wiley Interscience, New York.

Lovett, M., J. Kere, and L.M. Hinton. 1991. Direct selection: A method for the isolation of cDNAs encoded by large genomic regions. *Proc. Natl. Acad. Sci.* **88:** 9628–9632.

McDaniel, L.D. and R.A. Schultz. 1992. Elevated sister chromatid exchange phenotype of Bloom syndrome cells is complemented by human chromosome 15. *Proc. Natl. Acad. Sci.* **89:** 7968–7972.

McPherson, J.D., R.A. Morton, C.M. Ewing, J.J. Wasmuth, J. Overhauser, A. Nagafuchi, S. Tsukita, and W.B. Isaacs. 1994. Assignment of the human alpha-catenin gene (CTNNA1) to chromosome 5q21-q22. *Genomics* **19:** 188–190.

Miki, Y., J. Swensen, D. Shattuck-Eidens, P.A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L.M. Bennett, W. Ding et al. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266:** 66–71.

Morgan J.G., G.M. Dolganov, S.E. Robbins, L.M. Hinton, and M. Lovett. 1992. The selective isolation of novel cDNAs encoded by the regions surrounding the human interleukin 4 and 5 genes. *Nucleic Acids Res.* **20:** 5173–5179.

Okubo, K., N. Hori, R. Matoba, T. Niiyama, A. Fukushima, Y. Kojima, and K. Matsubara. 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet.* **2:** 173–179.

Parimoo, S., S.R. Patanjali, H. Shulka, D.D. Chaplin, and S.M. Weissman. 1991. DNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc. Natl. Acad Sci.* **88:** 9623–9627.

Patanjali, S.R., S. Parimoo, and S.M. Weissman. 1991. Construction of a uniform-abundance (normalized) cDNA library. *Proc. Natl. Acad. Sci.* **88:** 1943–1947.

Polymeropoulos, M.H., H. Xiao, J.M. Sikela, M. Adams, J.C. Venter, and C.R. Merrill. 1993. Chromosomal distribution of 320 genes from a brain cDNA library. *Nature Genet.* **4:** 381–386.

Roy, N., M.S. Mahadevan, M. McLean, G. Shutler, Z. Yaraghi, R. Farahani, S. Baird, A. Besner-Johnston, C. Lefebvre, X. Kang et al. 1995. The gene for neuronal apoptosis inhibitory protein is partially deleted in individuals with spinal muscular atrophy. *Cell* **80:** 167–178.

Saltman D.L., G.M. Dolganov, J.A. Warrington, J.J. Wasmuth, and M. Lovett. 1993. A physical map of 15 loci on human chromosome 5q23-q33 by two-color fluorescence in situ hybridization. *Genomics* **16:** 726–732.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. *Molecular cloning: A laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Simmons, A.D., S.A. Goodart, T.D. Gallardo, J. Overhauser, and M. Lovett. 1995. Five novel genes from the cri-du-chat critical region isolated by direct selection. *Hum. Mol. Genet.* **4:** 295–302.

Soares, M.B., M.F. Bonaldo, P. Jelene, L. Su, L. Lawton, and A. Efstratiadis. 1994. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci.* **91:** 9228–9232.

The Huntington's Disease Collaborative Research Group. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72:** 971–983.

Thompson, T.G., C.J. DiDonato, L.R. Simard, S.E. Ingraham, A.H. Burghes, T.O. Crawford, C. Rochette, J.R. Mendell, and J.J. Wasmuth. 1995. A novel cDNA detects homozygous microdeletions in greater than 50% of type I spinal muscular atrophy patients. *Nature Genet.* **9:** 56–62.

Wang, H., S.Y. Tsai, R.G. Cook, W.G. Beattie, M.J. Tsai, and B.W. O'Malley. 1989. COUP transcription factor is a member of the steroid receptor superfamily. *Nature* **340:** 163–165.

Warrington, J.A., L.V. Hall, L.M. Hinton, J.N. Miller, J.J. Wasmuth, and M. Lovett. 1991. Radiation hybrid map of 13 loci on the long arm of chromosome 5. *Genomics* **11:** 701–708.

Warrington, J.A., S.K. Bailey, E. Armstrong, O. Aprelikova, K. Alitalo, G.M. Dolganov, A.S. Wilcox, J.M. Sikela, S.F. Wolfe, M. Lovett et. al. 1992. A radiation hybrid map of 18 growth factor, growth factor receptor, hormone receptor, or neurotransmitter receptor genes on the distal region. *Genomics* **13:** 803–808.

Weissman, S.M. 1987. Molecular genetic techniques for mapping the human genome. *Mol. Biol. Med.* **4:** 133–143.

Wilcox, A.S., A.S. Khan, J.A. Hopkins, and J.M. Sikela. 1991. Use of 3'-untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: Implications for an expression map of the genome. *Nucleic Acids Res.* **19:** 1837–1843.

Received July 27, 1995; accepted August 1, 1995.