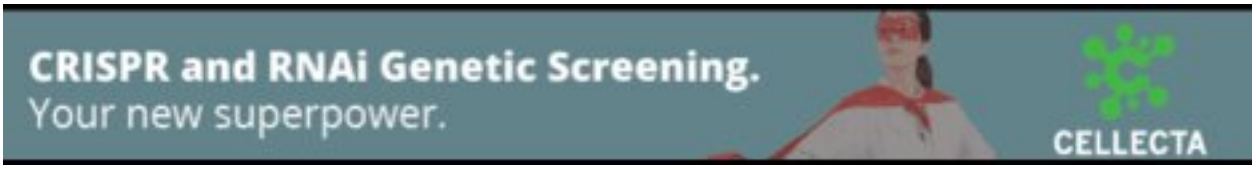
2006 expressed-sequence tags derived from human chromosome 7-enriched cDNA libraries.

J W Touchman, G G Bouffard, L A Weintraub, et al.

Genome Res. 1997 7: 281-292
Access the most recent version at doi:10.1101/gr.7.3.281



RESEARCH

2006 Expressed-Sequence Tags Derived from Human Chromosome 7-Enriched cDNA Libraries

Jeffrey W. Touchman,[1] Gerard G. Bouffard,[1] Lauren A. Weintraub,[1] Jacquelyn R. Idol,[1] Luping Wang,[2] Christiane M. Robbins,[1] Jesse C. Nussbaum,[1] Michael Lovett,[2] and Eric D. Green[1,3]

[1]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892; [2]Department of Otorhinolaryngology and The McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, Texas 75235

The establishment and mapping of gene-specific DNA sequences greatly complement the ongoing efforts to map and sequence all human chromosomes. To facilitate our studies of human chromosome 7, we have generated and analyzed 2006 expressed-sequence tags (ESTs) derived from a collection of direct selection cDNA libraries that are highly enriched for human chromosome 7 gene sequences. Similarity searches indicate that approximately two-thirds of the ESTs are not represented by sequences in the public databases, including those in dbEST. In addition, a large fraction (68%) of the ESTs do not have redundant or overlapping sequences within our collection. Human DNA-specific sequence-tagged sites (STSs) have been developed from 190 of the ESTs. Remarkably, 180 (96%) of these STSs map to chromosome 7, demonstrating the robustness of chromosome enrichment in constructing the direct selection cDNA libraries. Thus far, 140 of these EST-specific STSs have been assigned unequivocally to YAC contigs that are distributed across the chromosome. Together, these studies provide >2000 ESTs highly enriched for chromosome 7 gene sequences, 180 new chromosome 7 STSs corresponding to ESTs, and a definitive demonstration of the ability to enrich for chromosome-specific cDNAs by direct selection. Furthermore, the libraries, sequence data, and mapping information will contribute to the construction of a chromosome 7 transcript map.

[The ESTs and STSs are listed at http://www.nhgri.nih.gov/DIR/GTB/CHR7 and http://www.cshl.org/gr.]

The systematic generation of single-pass partial sequences of cDNA clones (expressed-sequence tags, or ESTs) has proven to be a powerful and useful approach for acquiring large amounts of information about human genes (Sikela and Auffray 1993; Matsubara and Okubo 1993; Adams et al. 1995). However, the value of ESTs is greatly enhanced when they are mapped to specific chromosomal regions or, preferably, to individual clone contigs (Hudson et al. 1995). Such mapping information can become available either by the proactive localization of an EST by physical mapping methods or by its identification during genomic sequencing of an accurately mapped clone. Human ESTs with defined map locations are extremely valuable for isolating disease genes by a positional cloning approach (Ballabio 1993; Collins 1995). Unfortunately, despite the tremendous number of ESTs that have been deposited into the public databases, only a fraction have been mapped to individual chromosomes (Schuler et al. 1996).

Human chromosome 7 contains an estimated 170 Mb of DNA (Trask et al. 1989; Morton 1991), corresponding to ~5% of the human genome. A well-advanced yeast artificial chromosome (YAC)-based physical map has been constructed that provides an average sequence-tagged site (STS) spacing of <80 kb across the entire length of the chromosome (G.G. Bouffard, J.R. Idol, V.V. Braden, L.M. Iyer, A.F. Cunningham, L.A. Weintraub, J.W. Touchman, R.M. Mohr-Tidwell, D.C. Peluso, R.S. Fulton, et al., in prep.). This map has been integrated rigorously with the Genethon genetic map

[3]Corresponding author.
E-MAIL egreen@nhgri.nih.gov; FAX (301) 402-4735.

(Weissenbach et al. 1992; Gyapay et al. 1994; Dib et al. 1996) and a chromosome 7 radiation hybrid (RH) map (E.A. Stewart, R.M. Myers, and D.R. Cox, unpubl.). Together, these maps provide a highly integrated framework for constructing an extensive transcript map, which in turn would benefit efforts to identify genes associated with genetic disorders mapping to chromosome 7, such as Williams–Beuren syndrome (Ewart et al. 1993), cerebral cavernous malformations (Dubovsky et al. 1995; Gunel et al. 1995; Johnson et al. 1995; Marchuk et al. 1995), Saethre–Chotzen syndrome (Lewanda et al. 1994; Rose et al. 1994), two forms of retinitis pigmentosa (Inglehearn et al. 1993, 1994; Jordan et al. 1993; McGuire et al. 1996), syndromic and nonsyndromic forms of hearing loss (Baldwin et al. 1995; Van Camp et al. 1995; Coyle et al. 1996; Sheffield et al. 1996), and several others. To date, only 212 genes (Genome Database, September 1996) have been assigned to chromosome 7, and even fewer to more precise subchromosomal locations. This represents only a small fraction of the estimated 3500–5000 genes predicted to reside on the chromosome (Fields et al. 1994). Although efforts to map ESTs and other cDNA sequences to various physical maps of human chromosomes have been successful [e.g., the study reported by Schuler et al. (1996)], the methodologies and levels of resolution have varied widely (Polymeropoulos et al. 1992, 1993; Hudson et al. 1995; Korenberg et al. 1995; Schuler et al. 1996).

To facilitate the construction of a chromosome 7 transcript map, we have established a set of chromosome 7-enriched cDNA libraries using the technique of direct cDNA selection (Lovett et al. 1991; Del Mastro et al. 1995). Single-pass sequences were generated from large numbers of clones derived from fetal brain, placenta, HeLa cell, and thymus primary cDNA. Here we report the characteristics of these chromosome 7-enriched cDNA libraries, the establishment of 2006 ESTs, and the development of 180 EST-specific STSs.

RESULTS

Direct Selection cDNA Libraries

As a source of human chromosome 7 genomic DNA, we used a cosmid library constructed from flow-sorted chromosome 7 (Lawrence Livermore National Laboratories, Livermore, CA). Analysis of this library revealed that upward of 20% of the clones contained hamster DNA (data not shown), representing contamination acquired during sorting from the monochromosomal human–hamster hybrid cell line. Thus, 9600 random cosmids were hybridized with human- and hamster-specific probes, from which a set of 7707 human-specific clones representing ~1.5 genomic equivalents of chromosome 7 were chosen. Cosmid DNA was isolated from these clones, pooled, and analyzed by fluorescent in situ hybridization (FISH), revealing hybridization to the entire length of chromosome 7 but not to other human chromosomes (B. Wolf-Ledbetter, unpubl.).

Four separate but parallel direct cDNA selection experiments were performed with the pooled chromosome 7 cosmid DNA (essentially as described by Del Mastro et al. 1995), using a mixture of randomly primed and oligo(dT)-primed cDNA from fetal brain, placenta, HeLa cells, and thymus. Importantly, the cDNA from these sources represented uncloned pools and were derived from cytoplasmic RNA (to increase complexity and avoid hnRNA contamination, respectively). For each library, the selected cDNA population was cloned into a plasmid vector and 4608 random clones were individually picked and arrayed.

EST Generation

Single-pass sequences were obtained from randomly selected clones from the fetal brain, placenta, HeLa, and thymus direct selection cDNA libraries. To date, 2172 ESTs have been generated from the four libraries, with the greatest effort focused on the fetal brain library (a total of 1042 sequences established). To assess the various types of background in the libraries, the 2172 ESTs were compared to mitochondrial, yeast genomic, human repetitive, cloning vector, *Escherichia coli* genomic, and structural RNA sequences. The results of these analyses are summarized in Table 1. In general, the relative amount of background sequences in the libraries is low. However, the placenta and HeLa libraries harbor a notable fraction of clones with sequences derived from the cosmid vector (Lawrist 16) as an artifact of the direct selection process. If necessary, the latter clones could be readily identified by hybridization prior to further sequence analysis of these libraries. The identified background sequences were separated from our collection, resulting in a final set of 2006 ESTs. These ESTs have been deposited into dbEST (National Center for Biotechnology Information) and are listed in a World Wide Web-based table at http://www.nhgri.nih.gov/DIR/GTB/CHR7 or http://www.cshl.org/gr, with a representative sample from this table provided in Figure 1. Prior to further analysis, the ESTs were compared to a collection of known human repetitive sequences (Jurka

Table 1. Summary of ESTs Generated from Direct Selection cDNA Libraries

Total ESTs generated[a]	Fetal brain (1042)	Placenta (524)	HeLa (555)	Thymus (51)	**Total (2172)**
Background sequences					
no insert	3	0	2	0	**5**
cosmid vector (Lawrist 16)	0	99	49	1	**149**
mitochondria	1	1	2	0	**4**
yeast	0	0	0	0	**0**
E. coli	1	0	4	0	**5**
structural RNA	0	0	3	0	**3**
Total background sequences	5	100	60	1	**166**
ESTs					
after removal of background	1037	424	495	50	**2006**
with repetitive sequences	196	130	173	21	**520**
completely masked[b]	69	42	92	14	**217**
remaining after masking	968	382	403	36	**1789**
dbEST match[c]	369 (36%)	121 (29%)	131 (27%)	3 (6%)	**624 (31%)**
Exact human match[d]	83	41	30	1	**155 (8%)**
Nonexact human match[e]	57	23	16	1	**97 (5%)**
Nonhuman match[f]	47	17	17	0	**81 (4%)**

[a]Does not include sequence data assessed as poor quality (see Methods for standards).
[b]Following masking of identifiable repeats, remaining sequences with a final length of ≤50 bp were considered completely masked and were excluded from further analysis.
[c]Matches to a dbEST record(s) at $P \leq 1.0e^{-6}$.
[d]A similarity of ≥95% to a human gene using BLASTN against the nonredundant GenBank database was scored as an exact human match.
[e]A similarity of 60%–94% to a human gene using BLASTN against the nonredundant GenBank database was scored as a nonexact human match.
[f]A similarity of ≥60% to a gene from another species using BLASTN against the nonredundant GenBank database was scored as a nonhuman match.

et al. 1992), and identified repetitive elements were "masked" (Claverie 1994). After masking, 217 ESTs (11%) were left with ≤50 bp of sequence and were excluded from further analysis.

Database Comparison

Nucleotide similarity searches of the remaining 1789 ESTs (2006 less the 217 excluded by complete masking; see Table 1) were performed against dbEST using BLASTN (Altschul et al. 1990). Only a minority (31%) of the ESTs matched a corresponding sequence in dbEST at $P \leq 1.0e^{-6}$ (Table 1). The percentage of dbEST matches varied among the libraries, with the fetal brain library having the most matches (36%) and thymus the least (6%). The 3′ ends of transcripts represent the majority of ESTs in the public databases and contain most of the repetitive sequences found in mRNAs. In comparing our ESTs to dbEST, we omitted all portions of sequences containing common human repetitive motifs, which may have slightly reduced the frequency of matches to existing ESTs. Nonetheless, these data indicate that the majority of our ESTs represent either newly identified genes or portions of genes not detected by the larger EST sequencing projects (Adams et al. 1995; Hillier et al. 1996).

The same set of 1789 ESTs was also compared to the nonredundant GenBank database using BLASTN and BLASTX (Altschul et al. 1990). A complete summary of these analyses (including information on the EST names, GenBank accession numbers, and similarity search results) is listed in an electronic table available at http://www.nhgri.nih.gov/DIR/GTB/CHR7 or http://www.cshl.org/gr (see Fig. 1). A number of the ESTs clearly matched known genes (Table 1), including some chromosome 7 genes that are associated with heritable disorders [e.g., 7H15B02 and the mismatch repair gene *PMS2* (Nicolaides et al. 1994), 7B05A02 and the *hERG* gene (Curran et al. 1995), and 7B19A12 and the *GLI3* gene (Vortkamp et al. 1991)]. In addition, some ESTs

EST Name	GenBank Accession	BLASTN (nr)	BLASTX (nr)	BLASTN (dbEST)	STS Name
7B05A02	AA076910	U04270	U04270	W70583	
7B08B10	AA077119	U62293		H38692	sWSS3437
7H01A01	AA077768			D44966	
7H03E04	AA077843	M74587			
7P02B10	AA078345				sWSS3899
7P05A06	AA078319	U41015	U41015	N26293	sWSS3821
7P08A01	AA078632				
7T01H07	AA078534			F10758	

Figure 1 Representative sample of the electronic summary table containing relevant information about the chromosome 7-enriched ESTs. Information about the 2006 chromosome 7-enriched ESTs is summarized in an electronic table that can be accessed through the World Wide Web at http://www.nhgri.nih.gov/DIR/GTB/CHR7 or http://www.cshl.org/gr. For each EST, the following information is provided: EST name [in each case starting with 7 followed by a unique letter corresponding to the library of origin: (B) fetal brain; (P) placenta; (H) HeLa cells; (T) thymus], GenBank accession number, accession number corresponding to the best match (at $P \leq 1.0e^{-6}$) within the nonredundant (nr) GenBank database based on BLASTN comparison, accession number corresponding to the best match (at $P \leq 1.0e^{-6}$) within the nonredundant (nr) GenBank database based on BLASTX comparison, accession number corresponding to the best match (at $P \leq 1.0e^{-6}$) within dbEST based on BLASTN comparison, and corresponding STS name (when applicable). The complete table is sorted by EST name and contains "hot links" to the relevant GenBank records.

demonstrated weak matches with known human genes, whereas others showed strong matches to nonhuman genes (97 and 81 ESTs, respectively; see Table 1). There were only three cases identified where all or part of an EST sequence was identical to an intron of a known gene, suggesting that little hnRNA contamination was present in the starting cDNA sources used for library construction. Arbitrarily, 62 ESTs that identified known genes were examined in greater detail. Of these, 81% represented coding sequence, 17% were identical to (or partially overlapping with) the 3′-untranslated region, and 2% corresponded to the 5′-untranslated region. The under-representation of 5′-untranslated regions in part reflects their frequent absence in the cDNA sequences deposited in GenBank.

Sequence Redundancy

To assess the complexity and depth of the direct selection cDNA libraries, sequence neighboring was performed on the set of 1789 apparently nonrepetitive ESTs. This analysis entails comparing each EST with all others in a pairwise fashion, allowing the sequences to be grouped into clusters (see Methods). Although 69% of the ESTs did not encounter a similar sequence in our collection, the remaining 564 identified one or more sequence neighbors (Fig. 2). These 564 sequences formed 190 unique clusters, with most of these containing only two or three overlapping sequences. Thus, the ESTs reported here form a nonredundant set of 1415 sequence clusters.

STS Generation

A critical issue was the extent to which our ESTs were enriched for chromosome 7. To examine this, 250 unique EST sequences from the fetal brain, placenta, HeLa, and thymus libraries were used to design PCR assays, which in turn were used to test human genomic DNA and appropriate human–hamster hybrid cell lines to determine the chromosomal location of each corresponding EST-specific STS (Green et al. 1991; Green 1993). Of the 250 PCR assays, 180 (72%) uniquely amplified the appropriate product from chromosome 7. Another five (2%) amplified the appropriate product from both human and hamster DNA (i.e., the corresponding STS was conserved in human and hamster and thus could not be assigned unequivocally to chromosome 7). There were 43 (17%) that did not amplify an appropriate product from human genomic DNA. This failure rate is typical of efforts to develop STSs from cDNA sequences, with the associated hazards of unknown intron positions (Hudson et al. 1995; Lamerdin et al. 1995). There were 12 (5%) that appeared to amplify repetitive sequences, with the appropriate product being generated from some or all human chromosome-containing cell lines. Of note, many of these STSs did not appear to be grossly repetitive, being present on only a select subset of human chromosomes. Importantly, only 10 (4%) of the PCR assays uniquely amplified the appropriate product from a human chromosome other than chromosome 7, indicating that the direct cDNA selection was highly effective at enriching for gene sequences from this chromosome. The relevant information about the 180 chromosome 7-specific STSs developed from EST sequences has been deposited into both the STS

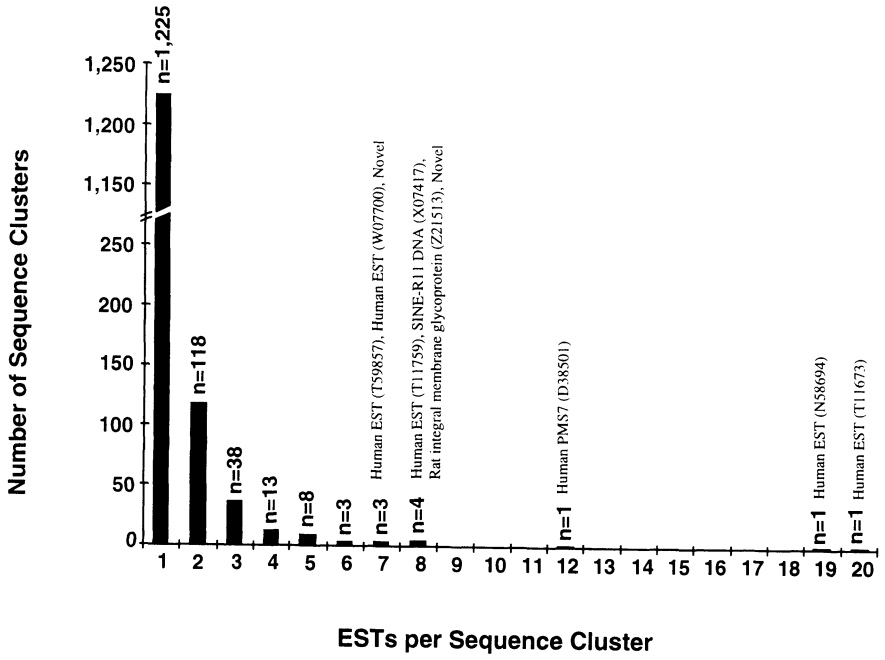


Figure 2 Sequence clusters in chromosome 7-enriched ESTs. The set of 1789 apparently nonrepetitive ESTs (see Table 1) were subjected to sequence neighboring (see Methods), and overlapping sequences were grouped into clusters. The number of unique clusters is indicated as a function of the number of ESTs in each cluster. Thus, 1225 clusters contain only a single EST, whereas 118 clusters contain two ESTs, etc. Each sequence cluster containing seven or more ESTs was compared to the dbEST and nonredundant GenBank databases, and in all but two cases a matching sequence was identified (with the names and accession numbers indicated above the appropriate bar).

Division of GenBank (dbSTS) and the Genome Database (GDB), with a summary table available electronically at http://www.nhgri.nih.gov/DIR/GTB/CHR7 or http://www.cshl.org/gr. A representative sample from this table is provided in Figure 3.

Physical Mapping

The availability of a well-established YAC-based physical map of human chromosome 7 (G.G. Bouffard, J.R. Idol, V.V. Braden, L.M. Iyer, A.F. Cunningham, L.A. Weintraub, J.W. Touchman, R.M. Mohr-Tidwell, D.C. Peluso, R.S. Fulton et al., in prep.) allows our EST-specific STSs to be readily mapped. To date, 140 of the STSs have been localized to individual YAC contigs. Figure 4 shows the rough cytogenetic positions of these mapped ESTs, in each case based on FISH analysis of a YAC containing or in close physical proximity to the EST. These data indicate that our ESTs were derived from all across chromosome 7, with some suggestion of a greater proportion localizing to the light bands.

DISCUSSION

The single-pass partial sequencing of cDNAs to generate ESTs has proven to be a powerful and successful way to assemble a profile of genes expressed in a particular organism, tissue, or cell type (Adams et al. 1991, 1992, 1993a,b, 1995; Khan et al. 1992; McCombie et al. 1992; Okubo et al. 1992; Waterston et al. 1992; Hofte et al. 1993; Takeda et al. 1993; Liew et al. 1994; Sudo et al. 1994; Frigerio et al. 1995; Houlgatte et al. 1995; Lamerdin et al. 1995; Hillier et al. 1996). We have extended this paradigm by coupling direct cDNA selection with EST production to establish a profile of expressed genes from a single human chromosome. More than 2000 ESTs, which are almost entirely derived from human chromosme 7, have been generated from developing brain, placenta, HeLa cells, and thymus. Importantly, the majority of these ESTs represent sequences that were not deposited previously into public databases. Thus far, 180 of these ESTs have been used to develop STSs and have been definitively mapped to chromosome 7, with 140 already sublocalized on the YAC-based physical map of the chromosome to within ~80 kb (on average) of another STS. Given the current state of our chromosome 7 physical map, there is roughly an 86% chance that a chromosome 7 EST will map to a YAC clone containing a Genethon genetic marker (Weissenbach et al. 1992; Gyapay et al. 1994; Dib et al. 1996) or a 97% chance that it will map to a YAC contig containing such a genetic marker (G.G. Bouffard, J.R. Idol, V.V. Braden, L.M. Iyer, A.F. Cunningham, L.A. Weintraub, J.W. Touchman, R.M. Mohr-Tidwell, D.C. Peluso, R.S. Fulton et al., in prep.). The ability to readily place our ESTs on a well-established physical map in a manner that virtually assures their localization relative to a genetic marker is particularly valuable for efforts aimed at identifying disease genes on the chromosome (Ballabio 1993; Collins 1995).

STS Name	EST Name	GenBank Accession	GDB Accession	Genetic Location	Cytogenetic Location
sWSS3437	7B08B10	G30714	GDB:4585101	D7S485, D7S2524	7p13
sWSS3438	7B08E01	G30715	GDB:4585105	D7S2549, D7S663	7q11
sWSS3439	7B09B09	G30716	GDB:4585109		7
sWSS3440	7B09C01	G30717	GDB:4585113	D7S2531, D7S512	7q31
sWSS3445	7B07H08	G30718	GDB:4585117	D7S518, D7S2509	7q22
sWSS3447	7B12C06	G30719	GDB:4585119		7q31
sWSS3448	7B12A05	G30720	GDB:4585123	D7S481, D7S2553	7p22
sWSS3449	7B11H06	G30721	GDB:4585127	D7S1870	7q11.23

Figure 3 Representative sample of the electronic summary table containing relevant information about the chromosome 7 EST-specific STSs. The 180 chromosome 7-specific STSs developed from the EST sequences are summarized in an electronic table that can be accessed through the World Wide Web at http://www.nhgri.nih.gov/DIR/GTB/CHR7 or http://www.cshl.org/gr. For each STS, the following information is provided: STS name (in each case starting with the prefix sWSS followed by a unique number), EST name (see Fig. 1), GenBank accession number for the STS, GDB accession number for the STS, the approximate genetic location of the STS (if known), and the approximate cytogenetic location of the STS (see Fig. 4). The genetic and cytogenetic locations were established by mapping the STSs to individual YAC contigs (G.G. Bouffard, J.R. Idol, V.V. Braden, L.M. Iyer, A.F. Cunningham, L.A. Weintraub, J.W. Touchman, R.M. Mohr-Tidwell, D.C. Peluso, R.S. Fulton et al., in prep.). For the genetic location, the nearest Genethon genetic marker (Weissenbach et al. 1992; Gyapay et al. 1994; Dib et al. 1996), indicated by the assigned D number, is given. When the STS is localized physically between two genetic markers, both are listed (in order from 7p-ter to 7q-ter). When the STS is physically localized adjacent to a single genetic marker (but not with high confidence relative to other genetic markers), only the one marker is listed. For the cytogenetic location, a 7 (without a subchromosomal band assignment) indicates that the STS is not yet assigned to a YAC contig but is known to be derived from chromosome 7. The complete table is sorted by STS name and contains hot links to the corresponding GenBank and GDB records.

In practical terms, the work reported here should benefit the study of chromosome 7 in several important ways. First, the STSs developed from EST sequences will enhance the physical map of the chromosome both by providing additional markers to increase the overall map resolution and by serving as gene-based landmarks for map annotation. Second, the chromosome 7-enriched cDNA libraries and ESTs generated to date will be useful for positional cloning projects involving the chromosome. The libraries themselves can be arrayed at high density and screened by hybridization using large insert clones [e.g., YACs and bacterial artificial chromosomes (BACs)] (Del Mastro et al. 1995; Stickens et al. 1996), whereas the ESTs can provide a rich source of candidate genes for further evaluation (especially after placement on the physical map). Finally, the ESTs will contribute to the annotation of the chromosome 7 genomic sequence. With chromosome 7 slated to be among the first few human chromosomes sequenced, the availability of large numbers of ESTs will be useful both for gene-based annotation and for evaluating the efficacy of the evolving gene prediction programs. In this regard, it is important to emphasize that the majority of our ESTs are novel and not previously existing in the public databases. Thus, the ESTs reported here are complementary to (rather than redundant with) other EST generation efforts.

The targeted generation of ESTs from the 3′ ends of mRNA molecules is widely regarded to be desirable for the subsequent development of STSs (Wilcox et al. 1991; Berry et al. 1995; Boguski and Schuler 1995; Hudson et al. 1995), as the 3′ portions of transcripts (1) contain sequence that is more likely to be contiguous in genomic DNA (i.e., devoid of introns), (2) can provide a single tag for each gene, and (3) tend to be less conserved between gene families and species. However, the generation of ESTs from randomly primed cDNAs is also desirable, as it provides a greater proportion of coding sequence and can potentially be used to assemble multiple ESTs into fuller-length cDNA sequences. By employing a mixture of both randomly primed and oligo(dT)-primed cDNA for constructing our direct selection libraries, we have retained the advantages of each of the above approaches for EST generation. On the one hand, our ESTs have proven to be an excellent source of sequence for the development of STSs. Our success rate in designing PCR assays that amplify genomic DNA (83%) is roughly the same as that achieved by others using 3′ end-derived ESTs (Hudson et al. 1995). On the other hand, our ESTs represent a rich source of novel cDNA sequence, much of which is likely coding, that can be used to form sequence clusters. Of note, the largest EST cluster formed with our collection was nearly 1 kb in length.

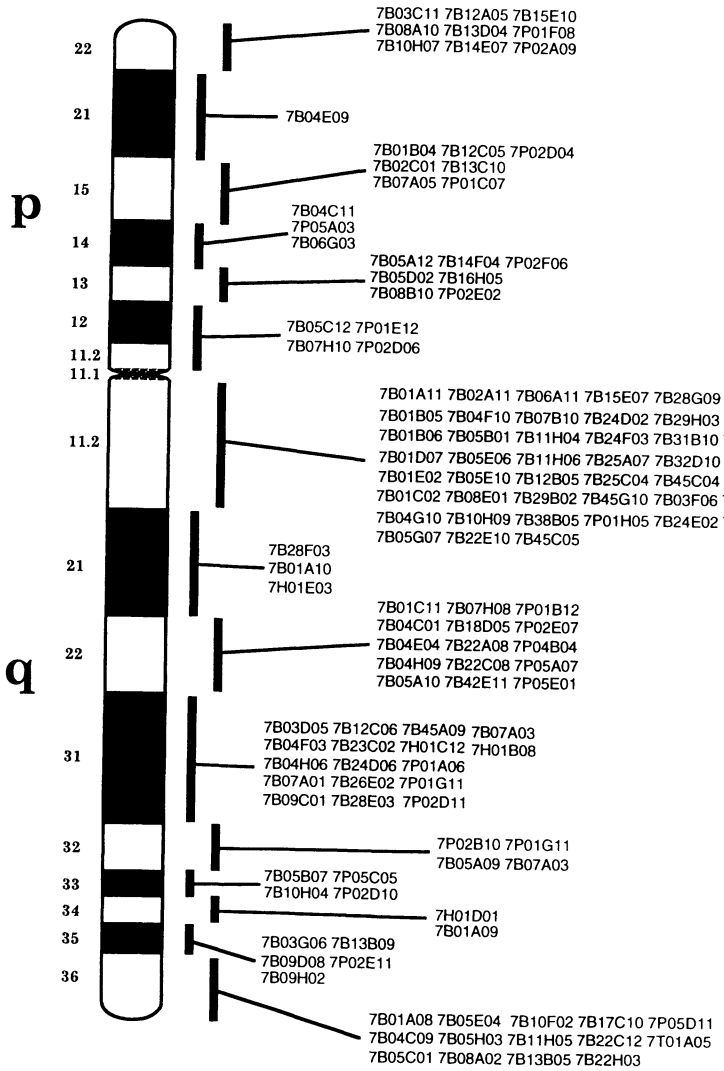


Figure 4 Cytogenetic localization of chromosome 7 ESTs. A total of 140 EST-specific STSs were mapped to individual YAC contigs by PCR-based screening (Green et al. 1995). In each case, either a YAC containing the EST or a YAC in close physical proximity to the EST had already been analyzed by FISH (Green et al. 1994). This information allowed assignment of the 140 ESTs to the indicated cytogenetic positions on chromosome 7. The EST names are listed, with information about the corresponding STSs available at http://www.nhgri.nih.gov/DIR/GTB/CHR7 or http://www.cshl.org/gr.

An important measure of an EST collection is the apparent complexity (i.e., relative uniqueness) of its gene sequences. This is particularly difficult to estimate for libraries derived entirely or partially from randomly primed cDNA, as nonoverlapping portions of the same gene inevitably will be represented. There is evidence that the complexity of the EST collection reported here is high. For example, in those cases where a known human gene was identified by BLASTN analysis ($n = 155$), an average of 1.7 ESTs (with a S.D. of 1.3) was found to match that gene (data not shown). Such a low number likely reflects the "pseudonormalization" inherent in the direct selection process (Del Mastro et al. 1995). By this estimate, one can postulate that upward of 1052 distinct genes may be represented in our EST collection [1789 (see Table 1) divided by 1.7], assuming that there are no significant biases associated with the use of known genes for such a comparison. In a similar fashion, neighboring analysis performed on our ESTs (Fig. 2) revealed that 69% had no sequence neighbors. Among the 1415 sequence clusters formed, the average number of ESTs per cluster was only 1.3 (with a S.D. of 1.1). Thus, although precise assessment of sequence complexity and gene representation would require more rigorous experimental analyses, these data suggest that a highly diverse set of genes are likely represented among our ESTs.

Although the generation and mapping of ESTs cannot by themselves be used to develop a comprehensive transcript map of a human chromosome, the reagents and data reported here can provide an important adjunct to such an effort. In conjunction with genome-wide EST mapping projects (Schuler et al. 1996) and, increasingly, genomic sequencing, our studies should accelerate the establishment of a gene-based map of chromosome 7.

METHODS

Chromosome 7 Genomic DNA

A total of one hundred 96-well microtiter plates of clones from the Lawrence Livermore National Laboratory chromosome 7-specific cosmid library (LL07NC01) were analyzed by hybridization with radiolabeled total human DNA or total hamster DNA. Approximately 20% of the cosmids were found to contain hamster (but not human) DNA (data not shown). All 9600 clones were grown on LB plates containing kanamycin (35 μg/ml), and the hamster DNA-containing colonies were selectively removed. The remaining 7707 human DNA-containing clones were pooled, and cosmid DNA was isolated by standard alkaline lysis fol-

lowed by two rounds of cesium chloride banding (Sambrook et al. 1989).

Direct cDNA Selection

Cytoplasmic RNA was isolated from fetal brain, placenta, HeLa cells, and thymus as described (Clemens 1984). Four chromosome 7-enriched direct selection cDNA libraries were constructed from each of these tissues using the technique described by Del Mastro et al. (1995). Purified DNA from the pooled 7707 human DNA-containing cosmids was biotinylated en masse. In the first round of direct selection, 1 µg of the biotinylated cosmid DNA was hybridized with 0.1 µg of repeat-suppressed cDNA (Lovett 1994; Del Mastro and Lovett 1996) to a C_0t of 200. In the second round of direct selection, 0.1 µg of the biotinylated cosmid DNA was hybridized with 1 µg of repeat-suppressed cDNA selected in the first round to the same C_0t value. Known chromosome 7 genes (aldose reductase for fetal brain and hepatocyte growth factor for placenta, HeLa cells, and thymus) were used to monitor the relative extent of enrichment during the two rounds of direct selection. In all cases, the final enrichment was at least 100-fold (data not shown).

Following the second round of direct selection, each sample was PCR amplified (Del Mastro et al. 1995) and cloned into pAMP10 (Life Technologies, Inc.) using MAX Efficiency DH5α competent cells (Life Technologies, Inc.) according to the manufacturer's instructions. For each library, 4608 clones were selected and arrayed in 96-well formats.

DNA Sequencing

Individual clones were inoculated in 1.2 ml of Terrific Broth (Sambrook et al. 1989) containing 100 µg/ml of ampicillin (in a 2-ml well of a 96-well box) and incubated with agitation at 37°C for 20–24 hr. Plasmid DNA was prepared using a 96-well modified boiling lysis procedure (Advanced Genetics Technologies Corporation), as described by the manufacturer, and suspended in 50 µl of modified TE buffer (10 mM Tris-HCl at pH 7.5; 0.1 mM EDTA). Systematic assessment of the DNA concentration was not performed prior to sequencing; however, when measured, the concentrations were generally between 150 and 250 ng/µl. Fluorescent DNA sequencing reactions were performed with the −21M13 dye primer (Perkin Elmer/Applied Biosystems) by a Catalyst 800 LabStation (Perkin Elmer/Applied Biosystems), and the products were analyzed on an ABI 373A or 377 automated fluorescent sequencer (Perkin Elmer/Applied Biosystems).

Sequence Analysis

ABI sequence trace files were transferred to a SUN SPARC20 workstation (Solaris 2.5 operating system), and the DNA sequence was extracted from each trace using a variation of trace editor program TED (Gleeson and Hillier 1991), which removes low-quality sequence data based on signal-to-noise and shoulder-width ratios (L. Hillier, pers. comm.). Flanking vector sequences were then identified and removed using the program WEP (W. Gish, pers. comm.). A collection of common background sequences that included structural RNAs, human mitochondrial DNA, yeast DNA, *E. coli* DNA, and cloning vector DNA were compared to the entire set of ESTs using BLAST (Altschul et al. 1990), and identified contaminants were eliminated.

EST sequence comparisons were performed by similarity searches against GenBank (October 1996; release 97.0) using BLAST. To reduce matches to common repetitive motifs, ESTs were compared (BLASTN, $P \leq 1.0e^{-5}$) to a collection of known human repetitive sequences (Jurka et al. 1992). Regions of ESTs matching a repeat sequence were masked using XBLAST (Claverie 1994) prior to further analysis. Masked sequences were compared to the nonredundant nucleotide division of GenBank using BLASTN, the nonredundant protein division of GenBank using BLASTX [after filtering low-complexity sequences with SEG and XNU (Wooton and Federhen 1996; Claverie and States 1993)], and the EST division of GenBank (dbEST) using BLASTN. Sequence redundancy of the masked sequences was assessed by neighboring analysis using algorithms derived from those used in Entrez (T. Madden, pers. comm.) followed by assembly of the ESTs with neighbors into clusters using AssemblyLIGN (Eastman Kodak Company).

YAC-Based STS-Content Mapping

STS-specific PCR assays were developed and optimized essentially as described (Green et al. 1991; Green 1993). Oligonucleotide primer pairs suitable for PCR were selected from individual EST sequences using the computer program OSP (Hillier and Green 1991), with the final amplified product size constrained to <150 bp. Each STS is named using the prefix sWSS followed by a unique number. STSs were mapped to a set of YACs highly enriched for chromosome 7 DNA (Green et al. 1995) by PCR-based screening (Green and Olson 1990; Green 1995), allowing their localization to be established within well-defined YAC contigs (Green et al. 1994; G.G. Bouffard, J.R. Idol, V.V. Braden, L.M. Iyer, A.F. Cunningham, L.A. Weintraub, J.W. Touchman, R.M. Mohr-Tidwell, D.C. Peluso, R.S. Fulton et al., in prep.).

ACKNOWLEDGMENTS

We thank Marco Marra, Richard Del Mastro, and Gregory Schuler for thoughtful discussion, Warren Gish, LaDeana Hillier, and Tom Madden for DNA sequence analysis tools, and Valerie Braden, Aimee Cunningham, and Leslie Iyer for technical assistance. We also thank Jane Weissman and Carolyn Tolstoshev for assistance with dbEST and dbSTS submissions, Andy Baxevanis for construction of the electronic tables, and Mark Boguski, Gregory Schuler, Rick Wilson, and Dennis Drayna for critical review of this manuscript. This work was supported in part by National Institutes of Health grant HG00368 (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, R.F. Moreno, A.R. Kerlavage, W.R. McCombie, and J.C. Venter.

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.

Adams, M.D., A.R. Kerlavage, R.D. Fleischmann, R.A. Fuldner, C.J. Bult, N.H. Lee, E.F. Kirkness, K.J. Weinstock, J.D. Gocayne, O. White, G. Sutton, J.A. Blake, R.C. Brandon, M.-W. Chiu, R.A. Clayton, R.T. Cline, M.D. Cotton, J. Earle-Hughes, L.D. Fine, L.M. FitzGerald, W.M. FitzHugh, J.L. Fritchman, N.S.M. Geoghagen, A. Glodek, C.L. Gnehm, M.C. Hanna, E. Hedblom, P.S. Hinkle Jr., J.M. Kelley, K.M. Klimek, J.C. Kelley, L.-I. Liu, S.M. Marmaros, J.M. Merrick, R.F. Moreno-Palanques, L.A. McDonald, D.T. Nguyen, S.M. Pellegrino, C.A. Phillips, S.E. Ryder, J.L. Scott, D.M. Saudek, R. Shirley, K.V. Small, T.A. Spriggs, T.R. Utterback, J.F. Weldman, Y. Li, R. Barthlow, D.P. Bednarik, L. Cao, M.A. Cepeda, T.A. Coleman, E.-J. Collins, D. Dimke, P. Feng, A. Ferrie, C. Fischer, G.A. Hastings, W.-W. He, J.-S. Hu, K.A. Huddleston, J.M. Greene, J. Gruber, P. Hudson, A. Kim, D.L. Kozak, C. Kunsch, H. Ji, H. Li, P.S. Meissner, H. Olsen, L. Raymond, Y.-F. Wei, J. Wing, C. Xu, G.-L. Yu, S.M. Ruben, P.J. Dillon, M.R. Fannon, C.A. Rosen, W.A. Haseltine, C. Fields, C.M. Fraser, and J.C. Venter. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377:** 3–174.

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

Baldwin, C.T., S. Weiss, L.A. Farrer, A.L. De Stefano, R. Adair, B. Franklyn, K.K. Kidd, M. Korostishevsky, and B. Bonne-Tamir. 1995. Linkage of congenital, recessive deafness (DFNB4) to chromosome 7q31 and evidence for genetic heterogeneity in the Middle Eastern Druze population. *Hum. Mol. Genet.* **4:** 1637–1642.

Ballabio, A. 1993. The rise and fall of positional cloning? *Nature Genet.* **3:** 277–279.

Berry, R., T.J. Stevens, N.A.R. Walter, A.S. Wilcox, T. Rubano, J.A. Hopkins, J. Weber, R. Goold, M.B. Soares, and J.M. Sikela. 1995. Gene-based sequence-tagged-sites (STSs) as the basis for a human gene map. *Nature Genet.* **10:** 415–423.

Boguski, M.S. and G.D. Schuler. 1995. ESTablishing a human transcript map. *Nature Genet.* **10:** 369–371.

Claverie, J.-M. and D. States. 1993. Information enhancement methods for large scale sequence analysis. *Comput. Chem.* **17:** 191–201.

Claverie, J.-M. 1994. Large scale sequence analysis. In *Automated DNA sequencing and analysis techniques* (ed. M.D. Adams, C. Fields, and J.C. Venter), pp. 267–279. Academic Press, New York, NY.

Clemens, M.J. 1984. Purification of eukaryotic messenger RNA. In *Transcription and translation: A practical approach* (ed. B.M. Haines and S.Y. Higgins), pp. 211–229. IRL Press, Washington, DC.

Collins, F.S. 1995. Positional cloning moves from perditional to traditional. *Nature Genet.* **9:** 347–350.

Coyle, B., R. Coffey, J.A.L. Armour, E. Gausden, Z. Hochberg, A. Grossman, K. Britton, M. Pembrey, W. Reardon, and R. Trembath. 1996. Pendred syndrome (goitre and sensorineural hearing loss) maps to chromosome 7 in the region containing the nonsyndromic deafness gene DFNB4. *Nature Genet.* **12:** 421–423.

Curran, M.E., I. Splawski, K.W. Timothy, G.M. Vincent, E.D. Green, and M.T. Keating. 1995. A molecular basis for cardiac arrhythmia: *HERG* mutations cause long QT syndrome. *Cell* **80:** 795–803.

Del Mastro, R.G. and M. Lovett. 1996. Isolation of coding sequences from genomic regions using direct selection. In *Methods in molecular biology* (ed. J. Boutwould), Humana Press, Inc., Totowa, NJ.

Del Mastro, R.G., L. Wang, A.D. Simmons, T.D. Gallardo, G.A. Clines, J.A. Ashley, C.J. Hilliard, J.J. Wasmuth, J.D. McPherson, and M. Lovett. 1995. Human chromosome-specific cDNA libraries: new tools for gene identification and genome annotation. *Genome Res.* **5:** 185–194.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette, and J. Weissenbach. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380:** 152–154.

Dubovsky, J., J.M. Zabramski, J. Kurth, R.F. Spetzler, S.S. Rich, H.T. Orr, and J.L. Weber. 1995. A gene responsible for cavernous malformations of the brain maps to chromosome 7q. *Hum. Mol. Genet.* **4:** 453–458.

Ewart, A.K., C.A. Morris, D. Atkinson, W. Jin, K. Sternes, P. Spallone, A.D. Stock, M. Leppert, and M.T. Keating. 1993. Hemizygosity at the elastin locus in a developmental disorder, Williams syndrome. *Nature Genet.* **5:** 11–16.

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

Frigerio, J.-M., P. Berthezene, P. Garrido, E. Ortiz, S. Barthellemy, S. Vasseur, B. Sastre, I. Seleznieff, J.-C. Dagorn, and J.L. Iovanna. 1995. Analysis of 2166 clones from a

human colorectal cancer cDNA library by partial sequencing. *Hum. Mol. Genet.* **4:** 37–43.

Gleeson, T. and L. Hillier. 1991. A trace display and editing program for data from fluorescence based sequencing machines. *Nucleic Acids Res.* **19:** 6481–6483.

Green, E.D. 1993. Physical mapping of human chromosomes: Generation of chromosome-specific sequence-tagged sites. In *Methods in molecular genetics (Vol. 1): Gene and chromosome analysis (Part A)* (ed. K.W. Adolph), pp. 192–210. Academic Press, San Diego, CA.

Green, E.D. 1995. PCR-based screening of yeast artificial chromosome libraries. In *PCR strategies* (ed. M.A. Innis, D.A. Gelfand and J.J. Sninsky), pp. 277–288. Academic Press, Orlando, FL.

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.

Green, E.D., R.M. Mohr, J.R. Idol, M. Jones, J.M. Buckingham, L.L. Deaven, R.K. Moyzis, and M.V. Olson. 1991. Systematic generation of sequence-tagged sites for physical mapping of human chromosomes: Application to the mapping of human chromosome 7 using yeast artificial chromosomes. *Genomics* **11:** 548–564.

Green, E.D., J.R. Idol, R.M. Mohr-Tidwell, V.V. Braden, D.C. Peluso, R.S. Fulton, H.F. Massa, C.L. Magness, A.M. Wilson, J. Kimura, J. Weissenbach, and B.J. Trask. 1994. Integration of physical, genetic and cytogenetic maps of human chromosome 7: Isolation and analysis of yeast artificial chromosome clones for 117 mapped genetic markers. *Hum. Mol. Genet.* **3:** 489–501.

Green, E.D., V.V. Braden, R.S. Fulton, R. Lim, M.S. Ueltzen, D.C. Peluso, R.M. Mohr-Tidwell, J.R. Idol, L.M. Smith, I. Chumakov, D. Le Paslier, D. Cohen, T. Featherstone, and P. Green. 1995. A human chromosome 7 yeast artificial chromosome (YAC) resource: Construction, characterization, and screening. *Genomics* **25:** 170–183.

Gunel, M., I.A. Awad, J. Anson, and R.P. Lifton. 1995. Mapping a gene causing cerebral cavernous malformation to 7q11.2-q21. *Proc. Natl. Acad. Sci.* **92:** 6620–6624.

Gyapay, G., J. Morissette, A. Vignal, C. Dib, C. Fizames, P. Millasseau, S. Marc, G. Bernardi, M. Lathrop, and J. Weissenbach. 1994. The 1993-94 Genethon human genetic linkage map. *Nature Genet.* **7:** 246–249.

Hillier, L. and P. Green. 1991. OSP: A computer program for choosing PCR and DNA sequencing primers. *PCR Methods Applic.* **1:** 124–128.

Hillier, L., G. Lennon, M. Becker, M. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. DuBuque, A. Favello, W. Gish, M. Hawkins, M. Hultman, T. Kucaba, M. Lacy, M. Le, N. Le, E. Mardis, B. Moore, M. Morris, J. Parsons, C. Prange, L. Rifkin, T. Rohlfing, K. Schellenberg, M.B. Soares, F. Tan, J. Thierry-Meg, E. Trevaskis, K. Underwood, P. Wohldman, R. Waterston, R. Wilson, and M. Marra. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6:** 807–828.

Hofte, H., T. Desprez, J. Amselem, H. Chiapello, M. Caboche, A. Moisan, M.F. Jourjon, J.L. Charpenteau, P. Berthomieu, D. Guerrier, J. Giraudat, F. Quigley, F. Thomas, D.Y. Yu, R. Mache, M. Raynal, R. Cooke, F. Grellet, M. Delseny, Y. Parmentier, G. Marcillac, C. Gigot, J. Fleck, G. Phillips, M. Axelos, C. Bardet, D. Tremousaygue, and B. Lescure. 1993. An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from Arabidopsis thaliana. *Plant J.* **4:** 1051–1061.

Houlgatte, R., R. Mariage-Samson, S. Duprat, A. Tessier, S. Bentolila, B. Lamy, and C. Auffray. 1995. The Genexpress Index: A resource for gene discovery and the genic map of the human genome. *Genome Res.* **5:** 272–304.

Hudson, T.J., L.D. Stein, S.S. Gerety, J. Ma, A.B. Castle, J. Silva, D.K. Slonim, R. Baptista, L. Kruglyak, S.-H. Xu, X. Hu, A.M.E. Colbert, C. Rosenberg, M.P. Reeve-Daly, S. Rozen, L. Hui, X. Wu, C. Vestergaard, K.M. Wilson, J.S. Bae, S. Maitra, S. Ganiatsas, C.A. Evans, M.M. DeAngelis, K.A. Ingalls, R.W. Nahf, L.T. Horton Jr., M. Oskin Anderson, A.J. Collymore, W. Ye, V. Kouyoumjian, I.S. Zemsteva, J. Tam, R. Devine, D.F. Courtney, M. Turner Renaud, H. Nguyen, T.J. O'Conner, C. Fizames, S. Faure, G. Gyapay, C. Dib, J. Morissette, J.B. Orlin, B.W. Birren, N. Goodman, J. Weissenbach, T.L. Hawkins, S. Foote, D.C. Page, and E.S. Lander. 1995. An STS-based map of the human genome. *Science* **270:** 1945–1954.

Inglehearn, C.F., S.A. Carter, T.J. Keen, J. Lindsey, A.M. Stephenson, R. Bashir, M. Al-Maghtheh, A.T. Moore, M. Jay, A.C. Bird, and S.S. Bhattacharya. 1993. A new locus for autosomal dominant retinitis pigmentosa on chromosome 7p. *Nature Genet.* **4:** 51–56.

Inglehearn, C.F., T.J. Keen, M. Al-Maghtheh, C.Y. Gregory, M.R. Jay, A.T. Moore, A.C. Bird, and S.S. Bhattacharya. 1994. Further refinement of the location for autosomal dominant retinitis pigmentosa on chromosome 7p (RP9). *Am. J. Hum. Genet.* **54:** 675–680.

Johnson, E.W., L.M. Iyer, S.S. Rich, H.T. Orr, A. Gil-Nagel, J.H. Kurth, J.M. Zabramski, D.A. Marchuk, J. Weissenbach, C.L. Clericuzio, L.E. Davis, B.L. Hart, J.F. Gusella, B.E. Kosofsky, D.N. Louis, L.A. Morrison, E.D. Green, and J.L. Weber. 1995. Refined localization of the cerebral cavernous malformation gene (*CCM1*) to a 4-cM interval of chromosome 7q contained in a well-defined YAC contig. *Genome Res.* **5:** 368–380.

Jordan, S.A., G.J. Farrar, P. Kenna, M.M. Humphries, D.M. Sheils, R. Kumar-Singh, E.M. Sharp, N. Soriano, C. Ayuso, J. Benitez, and P. Humphries. 1993. Localization of an autosomal dominant retinitis pigmentosa gene to chromosome 7q. *Nature Genet.* **4:** 54–59.

Jurka, J., J. Walichiewicz, and A. Milosavljevic. 1992. Prototypic sequences for human repetitive DNA. *J. Mol. Evol.* **35:** 286–291.

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 cDNAs. *Nature Genet.* **2:** 180–185.

Korenberg, J.R., X.-N. Chen, M.D. Adams, and J.C. Venter. 1995. Towards a cDNA map of the human genome. *Genomics* **29:** 364–370.

Lamerdin, J.E., R.S. Athwal, M.S. Kansara, A.K. Sandhu, S.R. Patanjali, S.M. Weissman, and A.V. Carrano. 1995. Chromosomal localization and expressed sequence tag generation of clones from a normalized human adult thymus cDNA library. *Genome Res.* **5:** 359–367.

Lewanda, A.F., E.D. Green, J. Weissenbach, H. Jerald, E. Taylor, M.L. Summar, J.A. Phillips III, M. Cohen, M. Feingold, W. Mouradian, S.K. Clarren, and E.W. Jabs. 1994. Evidence that the Saethre-Chotzen syndrome locus lies between D7S664 and D7S507, by genetic analysis and detection of a microdeletion in a patient. *Am. J. Hum. Genet.* **55:** 1195–1201.

Liew, C.C., D.M. Hwang, Y.W. Fung, C.M. Laurenssen, E. Cuckerman, S. Tsui, and C.Y. Lee. 1994. A catalogue of genes in the cardiovascular system as identified by expressed sequence tags. *Proc. Natl. Acad. Sci.* **91:** 10645–10649.

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

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.

Marchuk, D.A., C.J. Gallione, L.A. Morrison, C.L. Clericuzio, B.L. Hart, B.E. Kosofsky, D.N. Louis, J.F. Gusella, L.E. Davis, and V.L. Prenger. 1995. A locus for cerebral cavernous malformations maps to chromosome 7q in two families. *Genomics* **28:** 311–314.

Matsubara, K. and K. Okubo. 1993. Identification of new genes by systematic analysis of cDNAs and database construction. *Curr. Opin. Biotechnol.* **4:** 672–677.

McCombie, W.R., M.D. Adams, J.M. Kelley, M.G. FitzGerald, T.R. Utterback, M. Khan, M. Dubnick, A.R. Kerlavage, J.C. Venter, and C. Fields. 1992. *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues. *Nature Genet.* **1:** 124–131.

McGuire, R.E., A.A. Jordan, V.V. Braden, G.G. Bouffard, P. Humphries, E.D. Green, and S.P. Daiger. 1996. Mapping the RP10 locus for autosomal dominant retinitis pigmentosa on 7q: Refined genetic positioning and localization within a well-defined YAC contig. *Genome Res.* **6:** 255–266.

Morton, N.E. 1991. Parameters of the human genome. *Proc. Natl. Acad. Sci.* **88:** 7474–7476.

Nicolaides, N.C., N. Papadopoulos, B. Liu, Y. Wei, -F., K.C. Carter, S.M. Ruben, C.A. Rosen, W.A. Haseltine, R.D. Fleischmann, C.M. Fraser, M.D. Adams, J.C. Venter, M.G. Dunlop, S.R. Hamilton, G.M. Peterson, A. De La Chapelle, B. Vogelstein, and K.W. Kinzler. 1994. Mutations of two *PMS* homologues in hereditary nonpolyposis colon cancer. *Nature* **371:** 75–80.

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.

Polymeropoulos, M.H., H. Xiao, A. Glodek, M. Gorski, M.D. Adams, R.F. Moreno, M.G. FitzGerald, J.C. Venter, and C.R. Merril. 1992. Chromosomal assignment of 46 brain cDNAs. *Genomics* **12:** 492–496.

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

Rose, C.S.P., A.A.J. King, D. Summers, R. Palmer, S. Yang, A.O.M. Wilkie, W. Reardon, S. Malcolm, and R.M. Winter. 1994. Localization of the genetic locus for Saethre-Chotzen syndrome to a 6 cM region of chromosome 7 using four cases with apparently balanced translocations at 7p21.2. *Hum. Mol. Genet.* **3:** 1405–1408.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. *Molecular cloning: A laboratory manual*, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schuler, G.D., M.S. Boguski, E.A. Stewart, L.D. Stein, G. Gyapay, K. Rice, R.E. White, P. Rodriguez-Tome, A. Aggarwal, E. Bajorek, S. Bentolila, B.B. Birren, A. Butler, A.B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P.J.R. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, C. Edwards, J.-B. Fan, N. Fang, C. Fizames, C. Garret, L. Green, D. Hadley, M. Harris, P. Harrison, S. Brady, A. Hicks, E. Holloway, L. Hui, S. Hussain, C. Louis-Dit-Sully, J. Ma, A. MacGilvery, C. Mader, A. Maratukulam, T.C. Matise, K.B. McKusick, J. Morissette, A. Mungall, D. Muselet, H.C. Nusbaum, D.C. Page, A. Peck, S. Perkins, M. Piercy, F. Qin, J. Quackenbush, S. Ranby, T. Reif, S. Rozen, C. Sanders, X. She, J. Silva, D.K. Slonim, C. Soderlund, W.-L. Sun, P. Tabar, T. Thangarajah, N. Vega-Czarny, D. Vollrath, S. Voyticky, T. Wilmer, X. Wu, M.D. Adams, C. Auffray, N.A.R. Walter, R. Brandon, A. Dehejia, P.N. Goodfellow, R. Houglatte, J.R. Hudson Jr., S.E. Ide, K.R. Iorio, W.Y. Lee, N. Seki, T. Nagase, K. Ishikawa, N. Nomura, C. Phillips, M.H. Polymeropoulos, M. Sandusky, K. Schmitt, R. Berry, K. Swanson, R. Torres, J.C. Venter, J.M. Sikela, J.S. Beckmann, J. Weissenbach, R.M. Myers, D.R. Cox, M.R. James, D. Bentley, P. Deloukas, E.S. Lander, and T.J. Hudson. 1996. A gene map of the human genome. *Science* **274:** 540–546.

Sheffield, V.C., Z. Kraiem, J.C. Beck, D. Nishimura, E.M. Stone, M. Salameh, O. Sadeh, and B. Glaser. 1996. Pendred syndrome maps to chromosome 7q21-34 and is caused by an intrinsic defect in thyroid iodine organification. *Nature Genet.* **12:** 424–426.

Sikela, J.M. and C. Auffray. 1993. Finding new genes faster than ever. *Nature Genet.* **3:** 189–191.

Stickens, D., G. Clines, D. Burbee, P. Ramos, S. Thomas, D. Hogue, J.T. Hecht, M. Lovett, and G.A. Evans. 1996. The EXT2 multiple exostoses gene defines a family of putative tumor suppressor genes. *Nature Genet.* **14:** 25–32.

Sudo, K., K. Chinen, and Y. Nakamura. 1994. 2058 expressed sequence tags (ESTs) from a human fetal lung cDNA library. *Genomics* **24:** 276–279.

Takeda, J., H. Yano, S. Eng, Y. Zeng, and G.I. Bell. 1993. A molecular inventory of human pancreatic islets: sequence analysis of 1000 cDNA clones. *Hum. Mol. Genet.* **2:** 1793–1798.

Trask, B., G. van den Engh, B. Mayall, and J.W. Gray. 1989. Chromosome heteromorphism quantified by high-resolution bivariate flow karyotyping. *Am. J. Hum. Genet.* **45:** 739–752.

Van Camp, G., P. Coucke, W. Balemans, D. Van Velzen, C. Van de Bilt, L. Van Laer, R.J.H. Smith, K. Fukushima, G.W. Padberg, R.R. Frants, P. Van de Heyning, S.D. Smith, E.H. Huizing, and P.J. Willems. 1995. Localization of a gene for non-syndromic hearing loss (DFNA5) to chromosome 7p15. *Hum. Mol. Genet.* **4:** 2159–2163.

Vortkamp, A., M. Gessler, and K.-H. Grzeschik. 1991. GLI3 zinc-finger gene interrupted by translocations in Greig syndrome families. *Nature* **352:** 539–540.

Waterston, R., C. Martin, M. Craxton, C. Huynh, A. Coulson, L. Hillier, R. Durbin, P. Green, R. Shownkenne, N. Halloran, M. Metzstein, T. Hawkins, R. Wilson, M. Berks, Z. Du, K. Thomas, J. Thierry-Mieg, and J. Sulston. 1992. A survey of expressed genes in *Caenorhabditis elegans*. *Nature Genet.* **1:** 114–123.

Weissenbach, J., G. Gyapay, C. Dib, A. Vignal, J. Morissette, P. Millasseau, G. Vaysseix, and M. Lathrop. 1992. A second-generation linkage map of the human genome. *Nature* **359:** 794–801.

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.

Wooton, J.C. and S. Federhen. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266:** 554–571.

Received November 3, 1996; accepted in revised form January 16, 1997.