Published in final edited form as:
Mol Vis. ; 5: 5.

Identifying and mapping novel retinal-expressed ESTs from humans

Kimberly Malone[1], **Melanie M. Sohocki**[1], **Lori S. Sullivan**[1,2], and **Stephen P. Daiger**[1,2]

1*Human Genetics Center, School of Public Health, The University of Texas Health Science Center, Houston, TX*

2*Department of Ophthalmology and Visual Science, The University of Texas Health Science Center, Houston, TX*

Abstract

Purpose—The goal of this study was to develop efficient methods to identify tissue-specific expressed sequence tags (ESTs) and to map their locations in the human genome. Through a combination of database analysis and laboratory investigation, unique retina-specific ESTs were identified and mapped as candidate genes for inherited retinal diseases.

Methods—DNA sequences from retina-specific EST clusters were obtained from the TIGR Human Gene Index Database. Further processing of the EST sequence data was necessary to ensure that each EST cluster represented a novel, non-redundant mapping candidate. Processing involved screening for homologies to known genes and proteins using BLAST, excluding known human gene sequences and repeat sequences, and developing primers for PCR amplification of the gene encoding each cDNA cluster from genomic DNA. The EST clusters were mapped using the GeneBridge 4.0 Radiation Hybrid Mapping Panel with standard PCR conditions.

Results—A total of 83 retinal-expressed EST clusters were examined as potential novel, non-redundant mapping candidates. Fifty-five clusters were mapped successfully and their locations compared to the locations of known retinal disease genes. Fourteen EST clusters localize to candidate regions for inherited retinal diseases.

Conclusions—This pilot study developed methodology for mapping uniquely expressed retinal ESTs and for identifying potential candidate genes for inherited retinal disorders. Despite the overall success, several complicating factors contributed to the high failure rate (33%) for mapping EST-clustered sequences. These include redundancy in the sequence data, widely dispersed sequences, ambiguous nucleotides within the sequences, the possibility of amplifying through introns and the presence of repetitive elements within the sequence. However, the combination of database analysis and laboratory mapping is a powerful method for identification of candidate genes for inherited diseases.

A rapidly growing area of genome research is the analysis of expressed sequence tags (ESTs). ESTs are generated when large numbers of randomly selected cDNA clones from specific tissues are sequenced partially. The resulting collection of ESTs reflects the level and complexity of gene expression in the sampled tissue. ESTs can be used to rapidly identify expressed genes because they are usually unique to the cDNA from which they are derived and they correspond to a specific gene in the genome.

Correspondence to: Stephen P. Daiger, Human Genetics Center, School of Public Health, PO Box 20334, Houston, TX, 77225-0334; Phone: (713) 500-9829; FAX: (713) 500-0900 email: sdaiger@utsph.sph.uth.tmc.edu.

The goal of this project was to identify potential candidate genes for inherited retinopathies using bioinformatic tools to ascertain retina-specific gene sequences and to map these sequences in a human radiation hybrid panel. The sequence data source was the TIGR Human Gene Index Database. Clusters of expressed sequences (ESTs) representing the same transcript are assembled by TIGR based on sequence overlap. ESTs that share one or more stretches of high sequence identity are grouped into a cluster. The set of sequences that forms the cluster is then reduced to a single tentative human consensus sequence (THC). These THCs were used to further analyze retina-specific clusters. It is important to note that some retinopathies are due to mutations in genes expressed in a number of tissues besides the retina. However, by choosing to analyze abudant, retina-specific ESTs only, we performed a survey of the most likely candidates for inherited retinal diseases.

To date, less than half of the more than 100 genes causing inherited retinal diseases have been cloned (RetNet) [1], and it is highly likely that additional disease loci will identified. Therefore, another goal of this project was to develop efficient methods for EST analysis and to uncover any inherent limitations in using ESTs for mapping purposes. We identified novel retina-specific ESTs using bioinformatic tools and subsequently, through laboratory analysis, localized these ESTs to specific chromosomes. Results using a similar approach to map retina/pineal-specific ESTs, distinct from those described here, are reported elsewhere [2].

METHODS

The purpose of this project was to identify retina-specific gene sequences in public databases for subsequent mapping using a radiation hybrid panel. We used the TIGR database to ascertain candidate sequences and compared these with data from the UniGene and GenBank databases. ESTs (expressed sequence tags) are grouped into clusters of overlapping, highly similar sequences, called THCs or “tentative human consensus sequences” in TIGR. The ESTs within a THC are presumed to derive from a single gene. The tissue origin of each EST within a THC, and the number of ESTs per cluster, provide information on the tissue distribution and relative abundance of the gene transcript. This information was used to select candidate sequences for mapping.

Retina-specific ESTs were identified in the TIGR database version 3.3 on July 1, 1998. Duplicate entries and identical clusters with different THC numbers were eliminated manually. Expression information and map locations (if known) of each cluster were acquired by entering the GenBank accession number of at least one STS (sequence tagged site) for each cluster into UniGene. Repeat sequences within THC sequences were masked using RepeatMasker. BLAST homology searches were performed [3] using the NCBI server. Clusters identified by BLAST as representing known genes or containing additional non-retinal ESTs were excluded from the study. Only clusters that contained retinal ESTs exclusively (or retinal and tumor-derived ESTs) and not previously mapped were considered for further analysis.

PCR primers for each cluster were designed using the Primer3 program. Primer pairs were optimized for PCR in human genomic DNA using AmpliTaq Gold polymerase (Perkin-Elmer) with a standard protocol of 35 cycles and an annealing temperature gradient within the Stratagene Robocycler thermocycler [2]. The resulting DNA fragments were separated on standard 2% agarose gels. The sequence of fragments that were not of the expected size was determined by treating an aliquot of the genomic PCR product with shrimp alkaline phosphatase and exonulease (Amersham) followed by manual sequencing with the AmpliCycle™ Sequencing Kit (Perkin-Elmer) using primers end-labeled with ^{32}P. The sequence fragments were separated on 6% Long Ranger™ (FMC Bioproducts) denaturing acrylamide gels. Each cluster that amplified in genomic DNA was localized in the genome using the GeneBridge 4.0 Radiation Hybrid Panel (Research Genetics). The screening results

were submitted to the GeneBridge 4.0 mapping server at the Whitehead Institute using a minimum LOD score of 15 for placement. To obtain chromosomal band identification, the resulting mapping data were compared to the information in the Stanford Radiation Hybrid mapping database and databases at the Whitehead Institute. Primer pairs that successfully identified a specific cluster location were submitted to GenBank for STS accession numbers.

RESULTS

In total, 1,315 EST clusters containing sequences from exclusively retinal cDNA clones were obtained from the TIGR Human Genome Database on July 1, 1998 (Table 1). In random primed libraries, two different ESTs can be derived from non-overlapping segments of the same gene thus causing redundancy in the database. Due to redundancy and multiple entries of the same EST in the TIGR database, 348 ESTs were removed from further consideration. To select for highly expressed transcripts, we chose to evaluate only EST clusters containing three or more independent sequences. This reduced the number of potential candidates to 276 EST clusters. One hundred and forty-nine (54%) of the EST clusters were mapped previously according to the UniGene database; 84 (30%) of the ESTs showed identity to known genes in the GenBank database; and 67 (24%) contained Alu, SINEs, LINEs, or other repeat elements. (These categories overlap.) The remaining 83 ESTs represent novel retinal clusters with no significant match to any sequence in the databases.

Because ESTs are obtained by single-pass sequencing, some sequences contain errors and have ambiguous nucleotides. Due to this problem, we were unable to make suitable primer pairs for 7 clusters. The STS for each of the remaining clusters was optimized in genomic DNA prior to PCR assay in the radiation hybrid panel. Of the remaining 76 mapping candidates, 11 were composed of widely dispersed sequences and mapped to multiple chromosomes in the GeneBridge 4.0 Radiation Hybrid Panel (Table 2). An additional 17 of the mapping candidates failed to amplify in either genomic DNA or in the radiation hybrid panel. The amplified fragment for THC137122 was much larger than the expected size, but sequencing revealed that the fragment included an intron flanked by the coding sequence for this cDNA cluster. Fifty-five of the 83 potential candidate ESTs were successfully mapped and their locations compared to the locations of known retinal disease genes. The EST name, number of cDNAs per cluster, and chromosomal mapping location (including flanking markers) for each cluster are shown in Table 3. Table 3 also lists the GenBank accession number assigned to each unique primer pair used to map these ESTs.

Each THC sequence examined in this study was found in UniGene; similarities to known genes, map locations, and information on tissue expression were obtained. The total number of ESTs per mapped cluster (3 or more) is shown in Table 3, giving a rough indication of abundance. Using TIGR, UniGene, and GenBank, we reduced the number of retina-specific EST clusters by excluding clusters that represented known genes, were mapped previously, or that included transcripts derived from tissues other than the retina. (Clusters with ESTs from tumors or transformed cell lines were not excluded because these ESTs may derive from transcripts expressed subsequent to transformation). Also, the human genome contains large segments of repetitive DNA sequences, such as Alu, SINEs, or LINEs, and these repeat elements can pose a potential problem for EST analysis [4]. To reduce the probability of analyzing THCs containing repetitive elements, we screened each retina-specific THC for repeats using the analysis program RepeatMasker, thus eliminating several additional clusters.

DISCUSSION

Following these procedures, we identified 83 novel retinal EST clusters as potential mapping candidates. PCR primers were designed for each potential candidate and the primer product

was mapped using the GeneBridge 4.0 Radiation Hybrid Panel. By this process, we localized 55 unique retina-specific genes, 14 of which map within the candidate region for an inherited retinopathy (Table 4).

One potential problem in mapping ESTs is the considerable degree of redundancy in the data and the overlap with more completely characterized, traditional GenBank entries that represent functionally cloned mRNAs and genes [5]. In this study, 84 (30%) of the 276 retinal EST clusters could be identified as known genes in the UniGene database and 149 (54%) were mapped previously. After completion of this study, the International RH Mapping Consortium released GeneMap98, the latest expression map of the human genome. Eleven ESTs localized in this study were confirmed by GeneMap 98.

Another problem with using ESTs for mapping is the presence of repetitive DNA sequences. Despite screening of each of the EST clusters for repetitive elements, 11 of the mapping candidates localized to multiple chromosomes in the radiation hybrid panel. Possible explanations are that the retinal sequence is a member of a dispersed gene family or the existence of multiple pseudogenes.

ESTs are generated from single pass sequencing of random cDNA clones and, as a consequence, they may contain inaccurate regions and ambiguous nucleotides. Due to the possibility of incorrect nucleotides occurring within the primer sequence, these sequences may cause difficulties in primer design. The primers may not anneal to the DNA and therefore fail to amplify in a PCR reaction. This could be an explanation for the relatively high failure rate (20%) of EST mapping in this study. Other explanations could be the presence of primer-dimers or other amplification artifacts.

Despite the problems with EST mapping, the 55 EST clusters mapped in this study represent novel, retina-specific genes and potential candidates for inherited retinopathies. Fourteen of these genes fall within the candidate region for a mapped, but not cloned, form of retinal disease. The specific retinal expression of these novel genes distinguish them from the retina/pineal-specific ESTs identified in a related study [2]. This confirms the utility of using ESTs to identify and map novel inherited retinal genes in the human genome.

Acknowledgements

We thank Odessa L. June, Human Genetics Center, the University of Texas Health Science Center, Houston for expert technical assistance. Supported by grants from the Foundation Fighting Blindness and the George Gund Foundation, by grants from the William Stamps Farish Fund and the M. D. Anderson Foundation, by NIH grant EY07142 and NIH-NEI National Institutional Service Award EY07024.

References

1. Daiger SP, Rossiter BF, Greenberg J, Christoffels A, Hide W. Data services and software for identifying genes and mutations causing retinal degeneration. Invest Ophthalmol Vis Sci 1998;39:S295.
2. Sohocki MM, Malone KA, Sullivan LS, Daiger SP. Localization of retina/pineal expressed sequences (ESTs): identification of novel candidate genes for inherited retinal disorders. Genomics. 1999In press
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10. [PubMed: 2231712]
4. Eichler EE. Masquerading repeats: paralogous pitfalls of the human genome. Genome Res 1998;8:758–62. [PubMed: 9724321]
5. Boguski MS, Schuler GD. ESTablishing a human transcript map. Nat Genet 1995;10:369–71. [PubMed: 7670480]

6. Mitchell SJ, McHale DP, Campbell DA, Lench NJ, Mueller RF, Bundey SE, Markham AF. A syndrome of severe mental retardation, spasticity, and tapetoretinal degeneration linked to chromosome 15q24. Am J Hum Genet 1998;62:1070–6. [PubMed: 9545391]

Table 1

Cluster analysis and Mapping Summary

Category	Number	Percent Evaluated
Initial retina-specific clusters in TIGR	1,315	-
Non-redundant retina-specific clusters in TIGR	967	-
Clusters with 3 or more retina-specific ESTs	276	100
Not previously mapped	147	53
Not containing known repetitive elements	83	30
Primers possible	76	28
Amplification successful	59	21
Successfully mapped	55	20

Table 2

EST Clusters that Map to Multiple Chromosomes

THC Cluster Name	Cluster Size
138546	4
158349	3
158912	3
158805	3
160519	5
160461	12
161469	3
164254	4
164269	6
198769	4

Table 3

Mapped retina-specific EST clusters (in chromosomal order)

The LOD scores used to map the retina-specific EST clusters are listed in GenBank.

Lab ID	GenBank accession	THC cluster name	Cluster size	Map Location	Flanking Markers
KAM56	G42367	226080	3	1p36.31	D1S436-WI5273
KAM55	G42361	224384	26	1q12	WI4283-D1S418
KAM37	G42354	164265	3	1q21.1	WI8123-WI497
KAM23	G42345	161757	4	1q22	WI9711-WI2862
KAM20	G44322	160453	3	1q42.2	GCT1E07-D1S204
KAM02	G44318	127587	3	1q44-qter	D1S204-tel
KAM11	G42340	158775	10	2q33.1	D2S117-WI3652
KAM16	G42654	158977	4	2q34	WI1247-WI3694
KAM08	G42338	158328	3	2q36.3-q37.1	D2S331-WI9093
KAM31	G44320	164210	3	2q36.3-q37.1	D2S396-D2S331
KAM19	G42655	160353	4	3p25.3	D3S1263-WI3115
KAM51	G42359	203022	3	3p25.1	D3S1263-WI4179
KAM49	G44321	202944	5	3q11.1-q11.2	D3S2372-CHLC.GATA71D10
KAM47	G42366	199898	3	3q11.2	CHLC.GATA71D10-D3S2372
KAM12	G42341	158800	3	3q27.3-q28	WI1327-WI3287
KAM41	G42659	165112	4	4p16.1-p15.32	WI4048-D4S2925
KAM27	G42348	163404	6	4q13.2	WI6336-WI9200
KAM18	G42343	160307	8	4q13.2	WI6336-WI9200
KAM34	G42352	164244	3	4q27	WI4262-WI3273
KAM48	G42663	200437	3	4q31.22	WI6076-WI10758
KAM30	G42349	164085	8	5q11.2	D5S660-WI7435
KAM35	G42364	164249	3	5q32	D5S500-D5S402
KAM53	G42360	210676	15	5q32	WI5208-WI9844
KAM46	G42662	198781	3	5q35.1-q32.5	tel-WI6737
KAM45	G42661	197126	3	6p21.1	WI5337-D6S273
KAM33	G42351	164228	3	6q21	AFMA084ZE9-D6S468
KAM36	G42353	164251	8	6q21	WI4792-AFMA084ZE9
KAM38	G42365	164306	5	6q21	WI4792-AFMA084ZE9
KAM50	G42358	202984	5	6q21-p22.1	D6S468-WI4792
KAM05	G42363	138357	4	6q25.3-q26	NIB330-AFMA155TD9
KAM10	G42652	158697	3	7p22-p22.3	WI5230-D7S673
KAM43	G42357	195719	3	7p22.3	D7S531-pter
KAM07	G42337	138645	6	7p14.1	WI6721-WI8144
KAM15	G42362	158931	3	7q21.13	WI652-D7S651
KAM40	G42658	164734	5	8pter-8p23.3	CHLC.GCT1802-WI6240
KAM14	G44319	158860	16	9q21.33	D9S197-WI6378
KAM17	G42342	159003	8	10q21.1	D2S546-WI8929
KAM28	G42656	163847	3	10q21.1	D10S207-WI5577
KAM09	G42339	158358	7	10q24.1-q25.1	WI4132-WI1905
KAM44	G42660	197125	4	11q14.21	WI63485-AFM248MC9
KAM25	G42346	163092	3	11q23.2-q24.1	WI7642-WI6355
KAM04	G42651	137122	4	11q34.3	WI6090-D11S1354
KAM52	G42664	208444	3	13q11	WI8572-tel
KAM32	G42350	164222	3	13q33.3-q34	CHLC.GATA21F07-D13S277
KAM06	G44323	138621	4	13q34-qter	D13S277-WI6695
KAM01	G42650	125808	8	13q34	tel-WI5283
KAM13	G42653	158844	4	15q24.3	WI6335-WI7183
KAM03	G42336	135357	4	17q11.1-q11.2	D17S798-WI6890
KAM42	G42356	195679	3	17q21.31	WI5817-D17S797

Lab ID	GenBank accession	THC cluster name	Cluster size	Map Location	Flanking Markers
KAM39	G42355	164707	3	18q21.2	WI3038-CHLC.GATA30B03
KAM24	G42666	161771	3	19qter	D19S220-WI6526
KAM29	G42657	163961	3	19qter	D19S218-WI5264
KAM26	G42347	163385	5	20q13.2	D20S196-WI9189
KAM22	G42344	160923	3	Xq24-q25	WI9952-WI5191

Table 4

EST clusters mapping to locations containing the locus for an inherited retinal disease (in chromosomal order)

EST name(s)	Cluster size	EST location(s)	Disease symbol or type	Disease location	OMIM Number
161757,164265	4,3	1q21-q22	RP18	1q13-q23	601414
158775	7	2q33.1	RP26	2q31-q33	
158800	3	3q27.3-q28	OPA1	3q28-q29	165500
138537	4	6q25.3-q26	RCD1	6q25-q26	180020
138645	6	7p14.1	RP9	7p13-p15	180104
159003,163847	8,3	10q21.1	congenital retinal nonattachment	10q21	221900
197125	3	11q14.21	EVR1,FEVR; VRNI	11q13-q23	133780 193235
38621,125808,16422	4,8,3	13q33.3-qter	STGD2	13q34	153900
158844	4	15q24.3	syndromic retinal degeneration	15q24	[6]
164707	3	18q21.2	CORD1	18q21.1-q21.3	600624