Using Genomic Databases for Sequence-Based Biological Discovery

Andreas D Baxevanis

The inherent potential underlying the sequence data produced by the International Human Genome Sequencing Consortium and other systematic sequencing projects is, obviously, tremendous. As such, it becomes increasingly important that all biologists have the ability to navigate through and cull important information from key publicly available databases. The continued rapid rise in available sequence information, particularly as model organism data is generated at breakneck speed, also underscores the necessity for all biologists to learn how to effectively make their way through the expanding "sequence information space." This review discusses some of the more commonly used tools for sequence discovery; tools have been developed for the effective and efficient mining of sequence information. These include LocusLink, which provides a gene-centric view of sequence-based information, as well as the 3 major genome browsers: the National Center for Biotechnology Information Map Viewer, the University of California Santa Cruz Genome Browser, and the European Bioinformatics Institute's Ensembl system. An overview of the types of information available through each of these front-ends is given, as well as information on tutorials and other documentation intended to increase the reader's familiarity with these tools.

INTRODUCTION

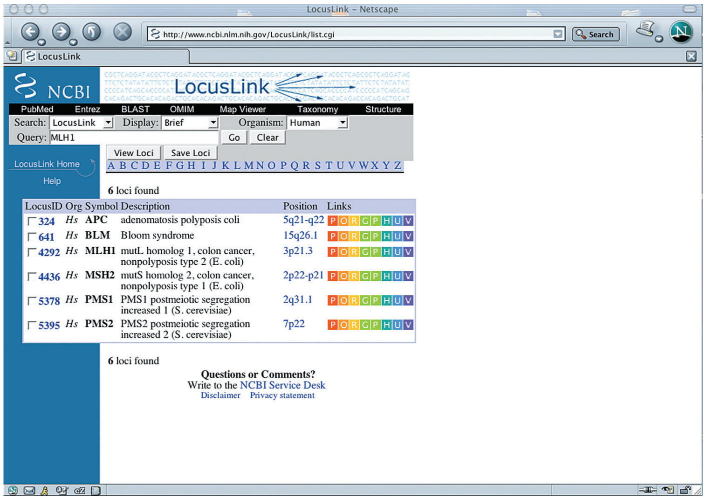
In April 2003, the scientific community celebrated the achievement of the Human Genome Project's major goal: completion of a high-accuracy sequence of the human genome. The significance of attaining this goal, which many have compared with landing a man on the moon, cannot be underestimated. This milestone firmly marks the entrance of modern biology into the genomic era (and *not* the post-genomic era, as many have stated), changing the way in which biological and clinical research will be conducted in the future. The intelligent use of sequence data from human and model organisms, along with technological innovations fostered by the Human Genome Project, will lead to significant advances in our understanding of diseases that have a genetic basis and, more importantly, in how health care is delivered from this point forward.

The completion of human genome sequencing has provided the biological community an opportunity to look forward and begin to think about how to use genomic approaches in a way that will lead to tangible health benefits. To that end, the National Human Genome Research Institute led a 2-year process involving hundreds of scientists and members of the public in more than a dozen workshops and individual consultations. The result of this process has led to the publication of a document entitled *A Vision for the Future of Genomics Research* (1). This "vision document" sets forth a number of "grand challenges" organized around 3 major themes: genomics to biology, genomics to health, and genomics to society. These grand challenges are intended to provide ambitious, interdisciplinary research goals for the scientific community that will eventually translate the promise of the Human Genome Project into improved human health.

As part of this vision, 6 critical "cross-cutting elements" were identified as being relevant to all 3 of the thematic areas. One of these areas is computational biology, an area whose importance will continue to increase as more and more sequence data becomes available, as data sets continue to get larger and larger, and as the complexity of both the data and the kinds of questions being addressed become more sophisticated. The focus on computational biology (or, as it is more often called, bioinformatics) underscores that both laboratory- and computationally-based approaches will be necessary to do cutting-edge research in the future. In the same way that investigators are trained in basic biochemistry and molecular biology techniques, a basic understanding of bioinformatic techniques as part of the biologist's arsenal will be absolutely indispensable in the future.

The database that most biologists are familiar with is GenBank (2), an annotated collection of all publicly available DNA and protein sequences maintained by the National Center for Biotechnology Information (NCBI) at the National Institutes of Health. At the time of this writing, GenBank contained 35.6 billion nucleotide bases, representing 29.8 million sequences in more than 119000 species (2). Whereas the inherent value of these data cannot be understated, the sheer magnitude of data presents a conundrum to the inexperienced user, not just because of the size of the "sequence information space," but because the information space continues to get larger, growing at an exponential pace. GenBank's size doubles once every 12 to 14 mo; this translates to 45 new sequences being deposited every minute and 7 new structures becoming available every day. This exponential growth rate is expected to continue well into the future, particularly because of the September 2002 announcement of "high priority" model organisms earmarked for

Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health,
Bethesda, Maryland, USA.



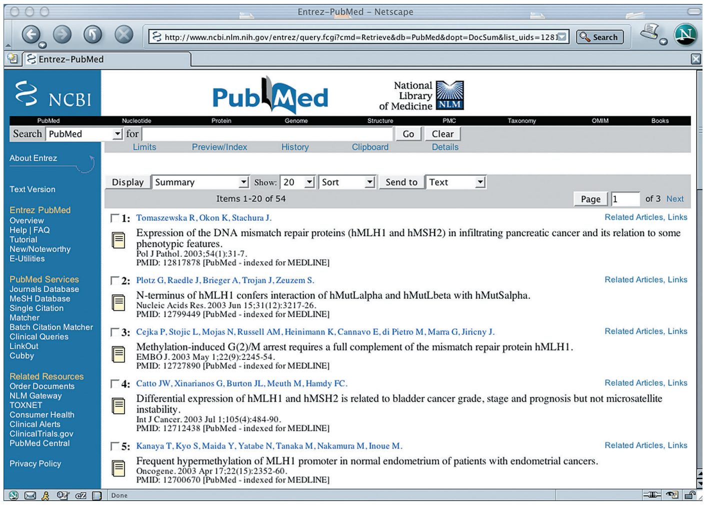
Search: LocusLink Display: Brief Organism: Human
Query: MLH1 Go Clear

6 loci found

LocusID	Org	Symbol	Description	Position	Links
□ 324	*Hs*	**APC**	adenomatosis polyposis coli	5q21-q22	P O R G P H U V
□ 641	*Hs*	**BLM**	Bloom syndrome	15q26.1	P O R G P H U V
□ 4292	*Hs*	**MLH1**	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)	3p21.3	P O R G P H U V
□ 4436	*Hs*	**MSH2**	mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)	2p22-p21	P O R G P H U V
□ 5378	*Hs*	**PMS1**	PMS1 postmeiotic segregation increased 1 (S. cerevisiae)	2q31.1	P O R G P H U V
□ 5395	*Hs*	**PMS2**	PMS2 postmeiotic segregation increased 2 (S. cerevisiae)	7p22	P O R G P H U V

6 loci found

Figure 1. Results of a LocusLink query, using "MLH1" as the search term. The report returns information on the MLH1 gene, as well as for related genes. A brief description is given for each found locus, as well as its chromosomal location. The colored alphabet blocks found to the right of each entry are explained in detail in the text.



Search PubMed for

Display Summary Show: 20 Sort Send to Text

Items 1-20 of 54 Page 1 of 3 Next

□ 1: Tomaszewska R, Okon K, Stachura J. Related Articles, Links
Expression of the DNA mismatch repair proteins (hMLH1 and hMSH2) in infiltrating pancreatic cancer and its relation to some phenotypic features.
Pol J Pathol. 2003;54(1):31-7.
PMID: 12817878 [PubMed - indexed for MEDLINE]

□ 2: Plotz G, Raedle J, Brieger A, Trojan J, Zeuzem S. Related Articles, Links
N-terminus of hMLH1 confers interaction of hMutLalpha and hMutLbeta with hMutSalpha.
Nucleic Acids Res. 2003 Jun 15;31(12):3217-26.
PMID: 12799449 [PubMed - indexed for MEDLINE]

□ 3: Cejka P, Stojic L, Mojas N, Russell AM, Heinimann K, Cannavo E, di Pietro M, Marra G, Jiricny J. Related Articles, Links
Methylation-induced G(2)/M arrest requires a full complement of the mismatch repair protein hMLH1.
EMBO J. 2003 May 1;22(9):2245-54.
PMID: 12727890 [PubMed - indexed for MEDLINE]

□ 4: Catto JW, Xinarianos G, Burton JL, Meuth M, Hamdy FC. Related Articles, Links
Differential expression of hMLH1 and hMSH2 is related to bladder cancer grade, stage and prognosis but not microsatellite instability.
Int J Cancer. 2003 Jul 1;105(4):484-90.
PMID: 12712438 [PubMed - indexed for MEDLINE]

□ 5: Kanaya T, Kyo S, Maida Y, Yatabe N, Tanaka M, Nakamura M, Inoue M. Related Articles, Links
Frequent hypermethylation of MLH1 promoter in normal endometrium of patients with endometrial cancers.
Oncogene. 2003 Apr 17;22(15):2352-60.
PMID: 12700670 [PubMed - indexed for MEDLINE]

Figure 2. A PubMed window showing papers on MLH1. Each entry gives the names of the authors, the title of the paper, and the citation information. The abstract of each paper can be found by clicking on the hyperlinked list of authors.

sequencing. The continued, rapid rise in available sequence information underscores the necessity for all biologists to learn how to effectively make their way through this sequence space. GenBank (or any other biological database, for that matter) serves little purpose unless the data can be easily searched and entries retrieved in a usable, meaningful format. Otherwise, sequencing efforts will have had no useful end, because the biological community as a whole would not be able to make use of the information hidden within these millions of bases and amino acids. Much effort has gone into making these data available to the biological community, and several of the most highly used interfaces resulting from these efforts are the focus of this review.

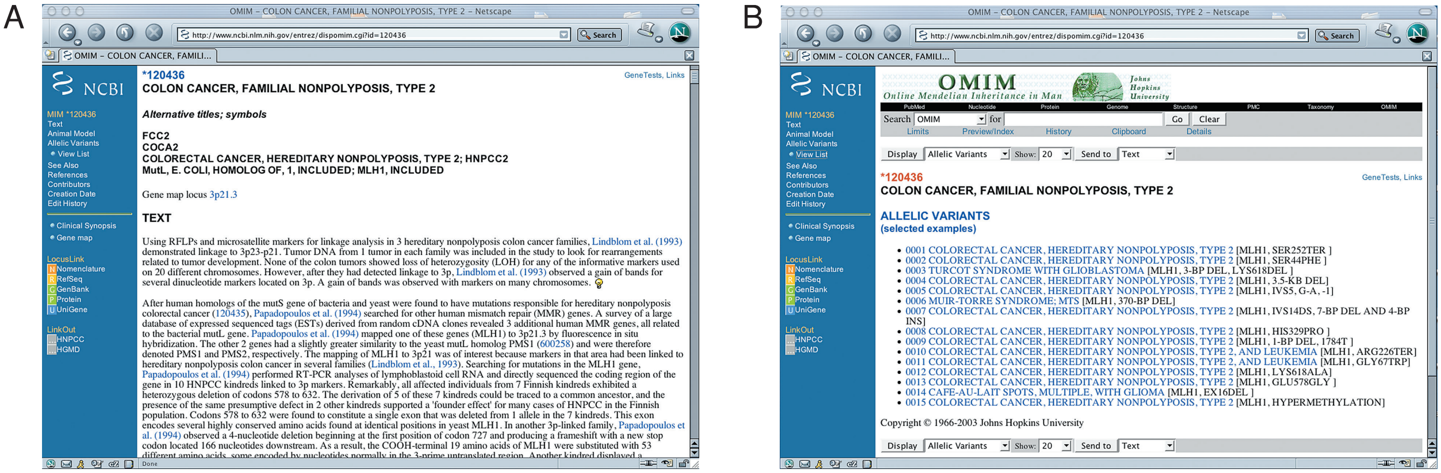
GENE-CENTRIC INFORMATION RETRIEVAL

One of the most commonly used interfaces for retrieving biological data is LocusLink, which allows access to a series of component databases in a gene-centric fashion (3). The kinds of information that can be retrieved using LocusLink include nucleotide and protein sequences, bibliographic information from PubMed, information on homologous genes, and single nucleotide polymorphism (SNP) data. LocusLink queries are very easy to perform, requiring only the name of the gene of interest or the gene symbol. As an example, we will search for information on the gene MLH1. From the LocusLink home page (http://www.ncbi.nlm.nih.gov/LocusLink), the user would simply type "MLH1" into the query box, select "Human" as the organism, then click *Go*. The results of the query are shown in Figure 1. Note that MLH1 is the 3rd entry in the returned list; the other entries returned are also related to MLH1 and may be of interest to the user. Clicking on the LocusID number at the left of the entry (here, *4292*) returns some basic information about the gene, giving the official gene symbol and name, as well as a brief summary of the function of the gene.

The colored alphabet blocks that appear to the right of each entry are hyperlinked and allow the user to find all available information on the particular gene of interest. Clicking on the 1st *P* (in red) in the row marked MLH1 takes the user to PubMed, the National Library of Medicine's collection of citations from the primary literature (Figure 2). The list is arranged in reverse chronological order (most recent 1st) and clicking on the hyperlinked author lists returns the abstract for that particular paper. Also notice the *Related Articles, Links* hyperlink to the right of each author list; clicking on these links returns a list of all of the papers that are similar in subject matter to the original paper.

Returning to the original hit list, clicking the *O* block takes the user to Online Mendelian Inheritance in Man (OMIM), an often-overlooked but incredibly valuable resource for anyone studying genetic or genomic disorders (4). OMIM provides concise textual information from the published literature on most human conditions having a genetic basis, as well as pictures illustrating the condition or disorder (where appropriate) and full citation information (Figure 3A). One of the most important features of OMIM is that for each gene, a list of allelic variants is available (see Figure 3B). Clicking on any of the items in the list returns information on how specific mutations (shown in the square brackets at the end of each entry) were identified and the effect of those specific mutations in patients. In this case, a number of allelic variants have been identified in the MLH1 gene leading to hereditary nonpolyposis colon cancer, but also related to some other syndromes as well.

The next 3 alphabet blocks would take the user to actual sequence information for that gene. The *R* stands for RefSeq, and clicking on the *R* would take the user to the "reference sequence" for that entry. The RefSeq project at NCBI is geared toward reducing redundancy in the public databases, with the goal of representing each molecule in the central dogma (DNA, mRNA, or protein) by 1 and only 1 sequence. Often times, a user will do a query and get back a long list of sequences, all representing the same biological entity, and it is often unclear which entry should be used; by using the curated RefSeq entry, the user can be assured that they are using the most accurate sequence information available. The *G*



*120436

COLON CANCER, FAMILIAL NONPOLYPOSIS, TYPE 2

Alternative titles; symbols

FCC2
COCA2
COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2; HNPCC2
MutL, E. COLI, HOMOLOG OF, 1, INCLUDED; MLH1, INCLUDED

Gene map locus 3p21.3

TEXT

Using RFLPs and microsatellite markers for linkage analysis in 3 hereditary nonpolyposis colon cancer families, Lindblom et al. (1993) demonstrated linkage to 3p23-p21. Tumor DNA from 1 tumor in each family was included in the study to look for rearrangements related to tumor development. None of the colon tumors showed loss of heterozygosity (LOH) for any of the informative markers used on 20 different chromosomes. However, after they had detected linkage to 3p, Lindblom et al. (1993) observed a gain of bands for several dinucleotide markers located on 3p. A gain of bands was observed with markers on many chromosomes.

After human homologs of the mutS gene of bacteria and yeast were found to have mutations responsible for hereditary nonpolyposis colorectal cancer (120435), Papadopoulos et al. (1994) searched for other human mismatch repair (MMR) genes. A survey of a large database of expressed sequenced tags (ESTs) derived from random cDNA clones revealed 3 additional human MMR genes, all related to the bacterial mutL gene. Papadopoulos et al. (1994) mapped one of these genes (MLH1) to 3p21.3 by fluorescence in situ hybridization. The other 2 genes had a slightly greater similarity to the yeast mutL homolog PMS1 (600258) and were therefore denoted PMS1 and PMS2, respectively. The mapping of MLH1 to 3p21 was of interest because markers in that area had been linked to hereditary nonpolyposis colon cancer in several families (Lindblom et al., 1993). Searching for mutations in the MLH1 gene, Papadopoulos et al. (1994) performed RT-PCR analyses of lymphoblastoid cell RNA and directly sequenced the coding region of the gene in 10 HNPCC kindreds linked to 3p markers. Remarkably, all affected individuals from 7 Finnish kindreds exhibited a heterozygous deletion of codons 578 to 632. The derivation of 5 of these 7 kindreds could be traced to a common ancestor, and the presence of the same presumptive defect in 2 other kindreds supported a 'founder effect' for many cases of HNPCC in the Finnish population. Codons 578 to 632 were found to constitute a single exon that was deleted from 1 allele in the 7 kindreds. This exon encodes several highly conserved amino acids found at identical positions in yeast MLH1. In another 3p-linked family, Papadopoulos et al. (1994) observed a 4-nucleotide deletion beginning at the first position of codon 727 and producing a frameshift with a new stop codon located 166 nucleotides downstream. As a result, the COOH-terminal 19 amino acids of MLH1 were substituted with 53 different amino acids, some encoded by nucleotides normally in the 3-prime untranslated region. Another kindred displayed a

*120436

COLON CANCER, FAMILIAL NONPOLYPOSIS, TYPE 2

ALLELIC VARIANTS
(selected examples)

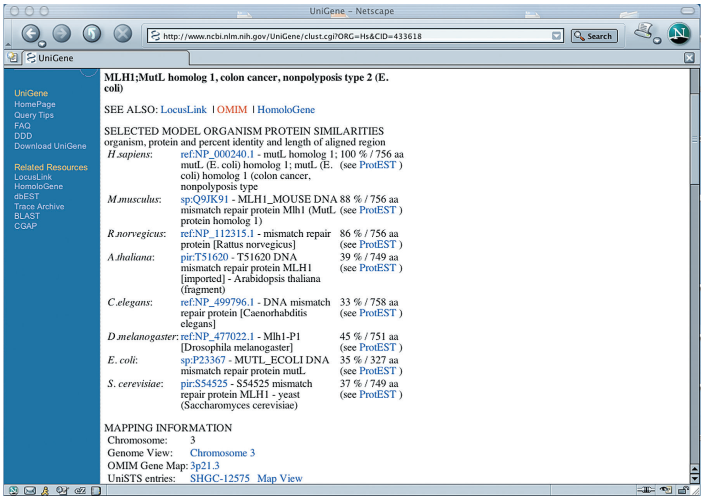
- 0001 COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2 [MLH1, SER252TER]
- 0002 COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2 [MLH1, SER44PHE]
- 0003 TURCOT SYNDROME WITH GLIOBLASTOMA [MLH1, 3-BP DEL, LYS618DEL]
- 0004 COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2 [MLH1, 3.5-KB DEL]
- 0005 COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2 [MLH1, IVS5, G-A, -1]
- 0006 MUIR-TORRE SYNDROME; MTS [MLH1, 370-BP DEL]
- 0007 COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2 [MLH1, IVS14DS, 7-BP DEL AND 4-BP INS]
- 0008 COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2 [MLH1, HIS329PRO]
- 0009 COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2 [MLH1, 1-BP DEL, 1784T]
- 0010 COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2, AND LEUKEMIA [MLH1, ARG226TER]
- 0011 COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2, AND LEUKEMIA [MLH1, GLY67TRP]
- 0012 COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2 [MLH1, LYS618ALA]
- 0013 COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2 [MLH1, GLU578GLY]
- 0014 CAFE-AU-LAIT SPOTS, MULTIPLE, WITH GLIOMA [MLH1, EX16DEL]
- 0015 COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2 [MLH1, HYPERMETHYLATION]

Copyright © 1966-2003 Johns Hopkins University

Figure 3. Information from Online Mendelian Inheritance in Man (OMIM) for the MLH1 gene. (A) OMIM entries begin with general information on the gene; the section marked *Text* provides an "executive summary" of relevant information on the gene and is curated on a regular basis by experts in the field. (B) For each gene where information is available, a list of allelic variants can be obtained; clicking on any of the entries provides more detailed information on that particular variant. See text for details.

and *P* are for GenBank and Protein, respectively, and will return *all* nucleotide and protein entries available for the gene of interest.

The remaining 3 alphabet blocks take users to more specialized resources, and for purposes of this discussion, we will take them slightly out of order, beginning with the *U*. The *U* stands for UniGene, a long-standing effort at NCBI to collapse expressed sequence tags (ESTs) (5), mRNA sequences, and coding sequences into discrete clusters. All of the sequences represented in a particular UniGene cluster share similarity at their 5′ end, and the presence of at least 1 EST in each cluster means that, by definition, the cluster represents the transcription product of a distinct gene. An example of a UniGene entry page, found by clicking on the *U* resulting from the original query, can be found in Figure 4. The page begins with information on selected model organism protein similarities, showing the best match of this UniGene cluster to sequences in other organisms. Scrolling down, users find a section on mapping information; here, the gene of interest is on chromosome 3, at 3p21.3. Users can click on the *Genome View* link in this section to jump directly to the NCBI Map Viewer and look at the genomic context of the gene; the maps will be discussed in greater detail below. Following the mapping information is a rather extensive section listing what tissues the gene has been shown to be expressed in, along with a link to any available information gathered through SAGE mapping (6). Finally, the UniGene entry ends with a list of the actual mRNA and EST sequences that have been brought together to comprise this cluster.



MLH1;MutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)

SEE ALSO: LocusLink | OMIM | HomoloGene

SELECTED MODEL ORGANISM PROTEIN SIMILARITIES
organism, protein and percent identity and length of aligned region

H.sapiens:	ref:NP_000240.1 - mutL homolog 1; mutL (E. coli) homolog 1; mutL (E. coli) homolog 1 (colon cancer, nonpolyposis type	100 % / 756 aa (see ProtEST)
M.musculus:	sp:Q9JK91 - MLH1_MOUSE DNA mismatch repair protein Mlh1 (MutL protein homolog 1)	88 % / 756 aa (see ProtEST)
R.norvegicus:	ref:NP_112315.1 - mismatch repair protein [Rattus norvegicus]	86 % / 756 aa (see ProtEST)
A.thaliana:	pir:T51620 - T51620 DNA mismatch repair protein MLH1 [imported] - Arabidopsis thaliana (fragment)	39 % / 749 aa (see ProtEST)
C.elegans:	ref:NP_499796.1 - DNA mismatch repair protein [Caenorhabditis elegans]	33 % / 758 aa (see ProtEST)
D.melanogaster:	ref:NP_477022.1 - Mlh1-P1 [Drosophila melanogaster]	45 % / 751 aa (see ProtEST)
E. coli:	sp:P23367 - MUTL_ECOLI DNA mismatch repair protein mutL	35 % / 327 aa (see ProtEST)
S. cerevisiae:	pir:S54525 - S54525 mismatch repair protein MLH1 - yeast (Saccharomyces cerevisiae)	37 % / 749 aa (see ProtEST)

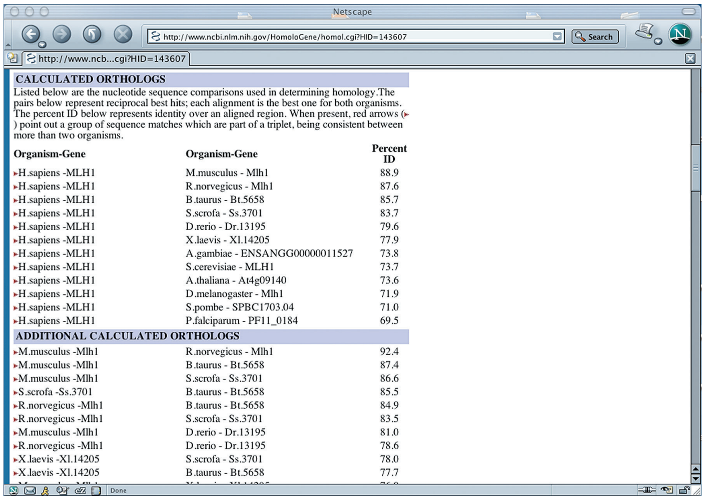
MAPPING INFORMATION

Chromosome:	3
Genome View:	Chromosome 3
OMIM Gene Map:	3p21.3
UniSTS entries:	SHGC-12575 Map View

Figure 4. A UniGene display showing information on the MLH1 cluster. Each UniGene entry contains information on protein similarities, mapping data, expression information, and links to the mRNA and EST sequences comprising the cluster. See text for details.

Moving back in the list to the *H*, users can access a service at NCBI called HomoloGene. HomoloGene includes curated and calculated orthologs and homologs for genes from human, mouse, rat, and zebrafish. The curated orthologs come from the Mouse Genome Database (7), the Zebrafish Information Resource (8), and published reports in the literature. Calculated homologs and orthologs are derived from direct nucleotide sequence comparisons between all UniGene clusters from each pair of organisms. Here, the functional definition of homolog is "the best match between a UniGene cluster in one organism and a cluster in a second organism," providing a representation of *similarity*. The functional definition of ortholog is when 2 sequences in different organisms are best matches *to one another*—a reciprocal best match—inferring that these sequences are direct descendants of a sequence in a common ancestor and most likely share common characteristics such as domain structure and biological function. Figure 5 shows the section of a HomoloGene entry for MLH1 where calculated orthologs are presented, giving the user a quick way of viewing and collecting sequence information on a given gene represented in a variety of organisms.

The final remaining alphabet block is the *V*, for variation. Clicking on the *V* brings the user to dbSNP, the database of human single nucleotide polymorphisms maintained by NCBI. The most important part of this page is the section shown in Figure 6, beginning with the header *Gene Model*. Immediately above the



Netscape

http://www.ncbi.nlm.nih.gov/HomoloGene/homol.cgi?HID=143607

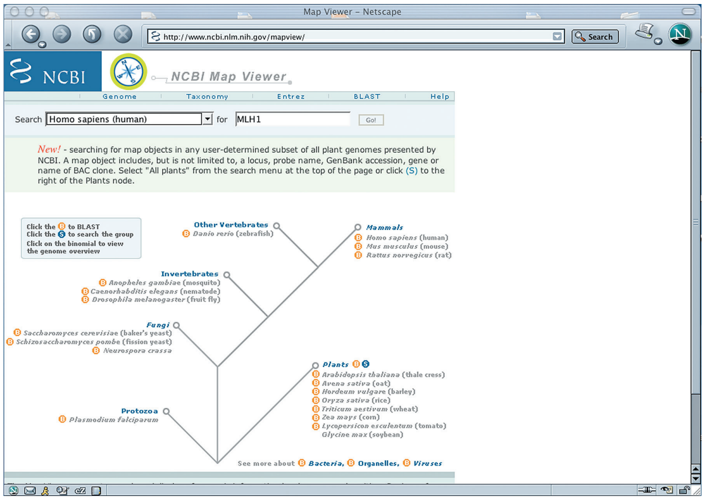
CALCULATED ORTHOLOGS

Listed below are the nucleotide sequence comparisons used in determining homology.The pairs below represent reciprocal best hits; each alignment is the best one for both organisms. The percent ID below represents identity over an aligned region. When present, red arrows (▸) point out a group of sequence matches which are part of a triplet, being consistent between more than two organisms.

Organism-Gene	Organism-Gene	Percent ID
▸H.sapiens -MLH1	M.musculus - Mlh1	88.9
▸H.sapiens -MLH1	R.norvegicus - Mlh1	87.6
▸H.sapiens -MLH1	B.taurus - Bt.5658	85.7
▸H.sapiens -MLH1	S.scrofa - Ss.3701	83.7
▸H.sapiens -MLH1	D.rerio - Dr.13195	79.6
▸H.sapiens -MLH1	X.laevis - Xl.14205	77.9
▸H.sapiens -MLH1	A.gambiae - ENSANGG00000011527	73.8
▸H.sapiens -MLH1	S.cerevisiae - MLH1	73.7
▸H.sapiens -MLH1	A.thaliana - At4g09140	73.6
▸H.sapiens -MLH1	D.melanogaster - Mlh1	71.9
▸H.sapiens -MLH1	S.pombe - SPBC1703.04	71.0
▸H.sapiens -MLH1	P.falciparum - PF11_0184	69.5
ADDITIONAL CALCULATED ORTHOLOGS		
▸M.musculus -Mlh1	R.norvegicus - Mlh1	92.4
▸M.musculus -Mlh1	B.taurus - Bt.5658	87.4
▸M.musculus -Mlh1	S.scrofa - Ss.3701	86.6
▸S.scrofa -Ss.3701	B.taurus - Bt.5658	85.5
▸R.norvegicus -Mlh1	B.taurus - Bt.5658	84.9
▸R.norvegicus -Mlh1	S.scrofa - Ss.3701	83.5
▸M.musculus -Mlh1	D.rerio - Dr.13195	81.0
▸R.norvegicus -Mlh1	D.rerio - Dr.13195	78.6
▸X.laevis -Xl.14205	S.scrofa - Ss.3701	78.0
▸X.laevis -Xl.14205	B.taurus - Bt.5658	77.7

Done

Figure 5. The HomoloGene entry for MLH1. Here, both calculated and curated orthologs to the MLH1 gene in a number of organisms are shown, as well as the percent identity between human MLH1 and its counterpart in other organisms.



Map Viewer – Netscape

http://www.ncbi.nlm.nih.gov/mapview/

Search Homo sapiens (human) for MLH1 Go!

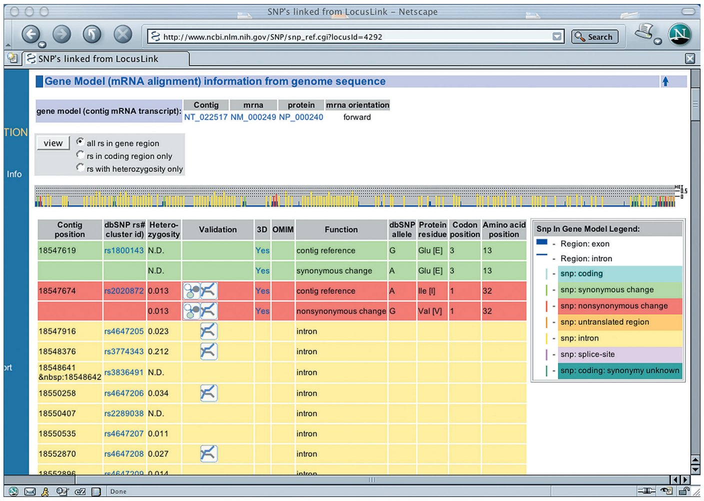
New! - searching for map objects in any user-determined subset of all plant genomes presented by NCBI. A map object includes, but is not limited to, a locus, probe name, GenBank accession, gene or name of BAC clone. Select "All plants" from the search menu at the top of the page or click (S) to the right of the Plants node.

Click the B to BLAST
Click the S to search the group
Click on the binomial to view the genome overview

Other Vertebrates
B Danio rerio (zebrafish)
Mammals
B Homo sapiens (human)
B Mus musculus (mouse)
B Rattus norvegicus (rat)
Invertebrates
B Anopheles gambiae (mosquito)
B Caenorhabditis elegans (nematode)
B Drosophila melanogaster (fruit fly)
Fungi
B Saccharomyces cerevisiae (baker's yeast)
B Schizosaccharomyces pombe (fission yeast)
B Neurospora crassa
Plants B S
B Arabidopsis thaliana (thale cress)
B Avena sativa (oat)
B Hordeum vulgare (barley)
B Oryza sativa (rice)
B Triticum aestivum (wheat)
B Zea mays (corn)
B Lycopersicon esculentum (tomato)
Glycine max (soybean)
Protozoa
B Plasmodium falciparum
See more about B Bacteria, B Organelles, B Viruses

Figure 7. The NCBI Map Viewer home page. From this page, users can select any of the available organisms for which map information is available and perform targeted queries (by gene, location, or any of a number of other criteria). Information on constructing queries can be found by following the *Help* hyperlink in the upper right.

table is a gray bar with a series of vertical yellow, red, green, and blue bars. This gray bar represents the MLH1 gene, and a key for the color-coded features is shown immediately below the gray bar to the right. Briefly, exons are shown as thicker blue bands and introns are shown as thinner blue bands. The most important vertical marks are the ones in green and red; the green marks indicate where there is an SNP that leads to a synonymous change at the protein level, whereas the red marks indicate where there is a SNP that results in a nonsynonymous change at the protein level. More information on the individual SNPs can be found by looking at the table that follows immediately below the gray bar. In this case, the 1st entry in the table shows a synonymous change (G → A, resulting in no change in the glutamine residue at position 13), whereas the 2nd entry shows a nonsynonymous change (A → G, resulting in a change of an isoleucine to a valine at position 32). This information is extremely useful in a wide variety of studies, ranging from the more intelligent design of site-directed mutagenesis experiments to three-dimensional structural studies aimed at deciphering the net effect of a mutation on the structure (and function) of a protein.



SNP's linked from LocusLink – Netscape

http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?locusId=4292

Gene Model (mRNA alignment) information from genome sequence

gene model (contig mRNA transcript):	Contig	mrna	protein	mrna orientation
	NT_022517	NM_000249	NP_000240	forward

view ● all rs in gene region ○ rs in coding region only ○ rs with heterozygosity only

Contig position	dbSNP rs# cluster id)	Hetero-zygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
18547619	rs1800143	N.D.		Yes		contig reference	G	Glu [E]	3	13
		N.D.		Yes		synonymous change	A	Glu [E]	3	13
18547674	rs2020872	0.013		Yes		contig reference	A	Ile [I]	1	32
		0.013		Yes		nonsynonymous change	G	Val [V]	1	32
18547916	rs4647205	0.023				intron				
18548376	rs3774343	0.212				intron				
18548641 :18548642	rs3836491	N.D.				intron				
18550258	rs4647206	0.034				intron				
18550407	rs2289038	N.D.				intron				
18550535	rs4647207	0.011				intron				
18552870	rs4647208	0.027				intron				
18552896	rs4647209	0.014				intron				

Snp In Gene Model Legend:
▬ - Region: exon
— - Region: intron
- snp: coding
- snp: synonymous change
- snp: nonsynonymous change
- snp: untranslated region
- snp: intron
- snp: splice-site
- snp: coding: synonymy unknown

Done

Figure 6. Single nucleotide polymorphism (SNP) data for human MLH1. The upper portion of the figure presents the gene model (position of introns and exons) and a graphical overview of where the various known SNPs for this gene are located. The table provides more detailed information about each characterized SNP. See text for details.

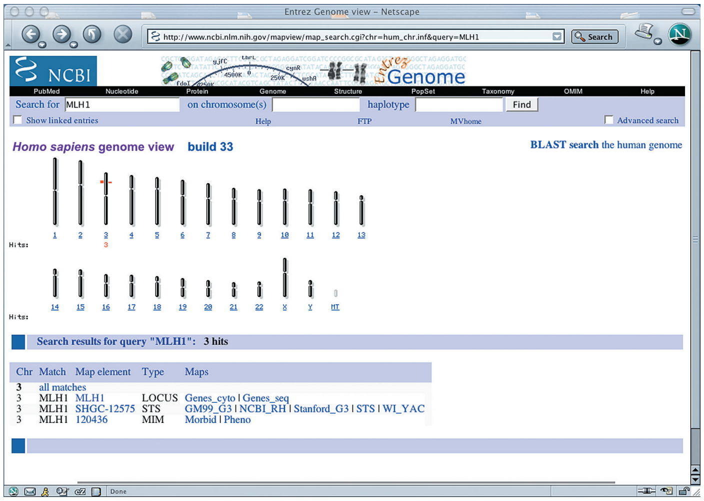
Whereas all of these component resources have been illustrated in the context of a LocusLink search, each one of these databases can be queried individually, and the URLs for all of these resources are provided in Table 1.

GENOME BROWSERS

LocusLink obviously provides a very easy-to-use, gene-centric view of the "sequence information space," but what if a scientist is more interested in seeing the gene of interest in context, particularly now that human genome sequencing is complete? A number of portals called *genome browsers* have been developed that allow users to access genomic data and, more importantly, view annotations that have been made on the underlying sequence data.

As alluded to above, NCBI has its own Map Viewer, a tool through which experimentally verified genes, predicted genes, genomic markers, physical maps, genetic maps, and sequence variation data can be viewed. Currently, the Map Viewer can be used to view the genomes of 19 organisms (Figure 7), with the number increasing as the genomes of more and more model organisms are sequenced. The Map Viewer is integrated into other NCBI tools, allowing one to link between the Map Viewer, LocusLink, and the main integrated information retrieval system at NCBI, Entrez.

Continuing with the MLH1 example, to find the genomic context of the MLH1 gene, one would simply change the pull-down menu marked *Search* on the Map viewer home page (http://www.ncbi.nlm.nih.gov/mapview/; see Figure 5) to *Human*, type "MLH1" in the text box, then press *Go*. The resulting screen is



Search for MLH1 on chromosome(s) haplotype

□ Show linked entries □ Advanced search

Homo sapiens genome view build 33

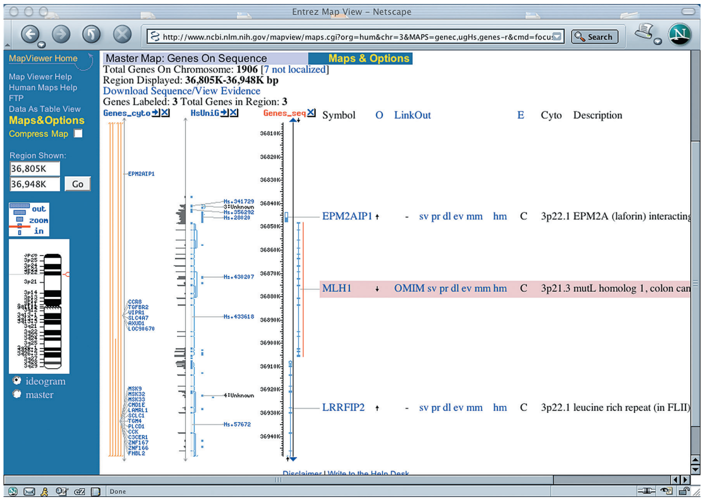
Hits: 3

Search results for query "MLH1": 3 hits

Chr	Match	Map element	Type	Maps
3	all matches			
3	MLH1	MLH1	LOCUS	Genes_cyto | Genes_seq
3	MLH1	SHGC-12575	STS	GM99_G3 | NCBI_RH | Stanford_G3 | STS | WI_YAC
3	MLH1	120436	MIM	Morbid | Pheno

Figure 8. The results of an NCBI Map Viewer search, using "human" as the organism and "MLH1" as the query. The query returned 3 matches, one of which is to the MLH1 locus. See text for details.

shown in Figure 8. Notice that there are several red bars marking chromosome 3; below the chromosome, the number 3 appears, indicating 3 hits. Below the pictogram is a list of all the matches found for MLH1: the 1st corresponds to the gene itself ("locus"), an STS marker is next, followed by an entry from OMIM ("MIM"). Clicking on the hyperlinked *MLH1* in the "locus" line returns the map shown in Figure 9, which is the default map display. The header above the actual map gives some overview information about the map itself: there are 1906 genes on chromosome 3, and the region that is currently being displayed (the "sequence coordinates") are from 36805K to 36948K. These numbers are also indicated in the blue bar to the left of the maps; the region displayed is also indicated by the red tick mark next to the ideogram in the blue sidebar, relative to the known cytogenetic banding patterns on chromosome 3.



Master Map: Genes On Sequence — Maps & Options

Total Genes On Chromosome: **1906** [7 not localized]

Region Displayed: **36,805K-36,948K bp**

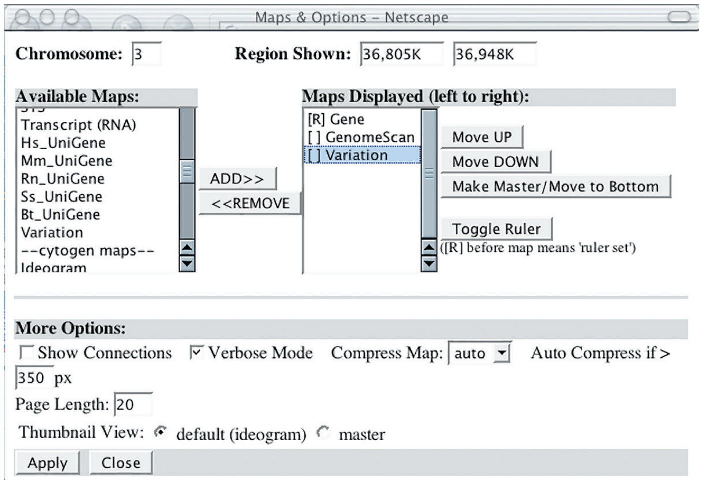
Download Sequence/View Evidence

Genes Labeled: **3** Total Genes in Region: **3**

Region Shown: 36,805K – 36,948K

Symbol		O	LinkOut	E	Cyto	Description
EPM2AIP1	↑	-	sv pr dl ev mm hm	C	3p22.1	EPM2A (laforin) interacting
MLH1	↓	OMIM	sv pr dl ev mm hm	C	3p21.3	mutL homolog 1, colon can
LRRFIP2	↑	-	sv pr dl ev mm hm	C	3p22.1	leucine rich repeat (in FLII)

Figure 9. The default map view. Three maps are displayed: the cytogenetic gene map, the UniGene cluster map, and the "Genes_seq" map (known and putative genes that have been placed as a result of alignments of mRNAs to individual contigs). See text for details.



Chromosome: 3 Region Shown: 36,805K 36,948K

Available Maps:
STS
Transcript (RNA)
Hs_UniGene
Mm_UniGene
Rn_UniGene
Ss_UniGene
Bt_UniGene
Variation
--cytogen maps--
Ideogram

Maps Displayed (left to right):
[R] Gene
[ ] GenomeScan
[ ] Variation

([R] before map means 'ruler set')

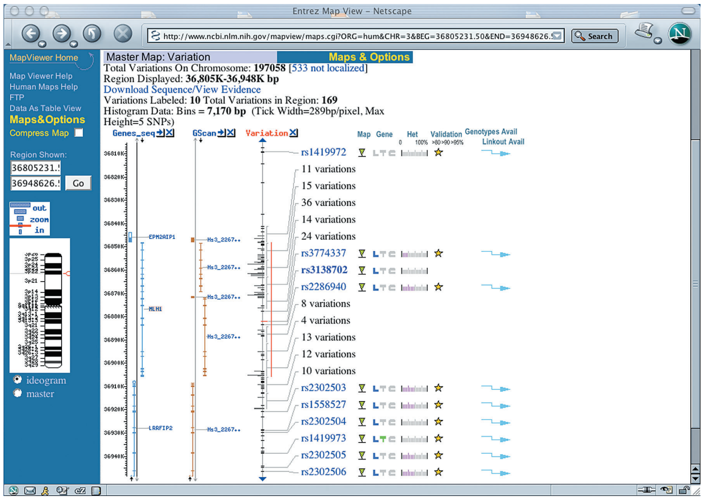
More Options:
□ Show Connections ☑ Verbose Mode Compress Map: auto Auto Compress if > 350 px
Page Length: 20
Thumbnail View: ● default (ideogram) ○ master

Figure 10. The Maps & Options window. This control panel is used to select from the available maps and change the order in which the maps are displayed.

Three maps are shown in the main window, to the left, as long, vertical bars. The map marked *Genes_cyto* (for "genes-cytogenetic") shows the cytogenetic locations of genes as reported in LocusLink. Twenty genes, in addition to MLH1, have been cytogenetically mapped to this region of chromosome 3. The next map, marked *HsUniG* (for "Human UniGene") shows the positions of UniGene clusters (described above); put another way, mRNA and EST sequences that comprise a UniGene cluster map to this region. On the left side of this particular map are gray bars that form what appears to be a histogram. These bars are intended to illustrate the density of aligned mRNAs and ESTs in this region.

The thick blue lines to the right of this map are intended to illustrate exons. The final, right-most map is labeled *Genes_seq* (for "gene sequence"). The map occupying the right-most position in any view is called the "master map," and the information appearing to the right of all the maps pertains to that master map. Three genes are plotted on the master map in this particular view: an EPM2A-interacting protein 1, then the MLH1 gene, which was the basis of the query (highlighted in red), and finally a leucine-rich repeat interacting protein 2 (LRRFIP2). For each gene, an indication of the gene's structure is given by the blue line running along the right side of the map, with exons being represented as thick blue bars and introns being represented as the thinner, intervening blue lines. Finally, note the arrow immediately to the right of each gene name; this arrow represents the direction of transcription for each gene.

Whereas the default map view is useful to gain a sense of what a particular genomic region looks like, there are additional maps available that may shed more light on the biological properties of a particular area of interest. To change the maps that are shown in any particular view, the user can click on the link marked *Maps & Options* that appears at the top of any Map Viewer page (see Figure 9, top). Clicking on the link brings up the Maps and Options window, allowing the user to now customize the view (Figure 10). For purposes of this example, we will remove the cytogenetic and UniGene map; to do so, the user would highlight each of these in the *Maps Displayed* list on the



Entrez Map View – Netscape

http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?ORG=hum&CHR=3&BEG=36805231.50&END=36948626.5

Master Map: Variation

Total Variations On Chromosome: **197058** [533 not localized]

Region Displayed: **36,805K-36,948K bp**

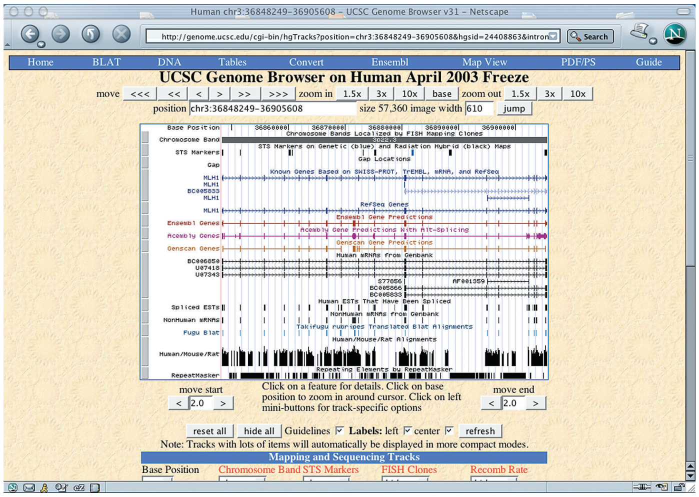
Download Sequence/View Evidence

Variations Labeled: **10** Total Variations in Region: **169**

Histogram Data: Bins = **7,170 bp** (Tick Width=289bp/pixel, Max Height=5 SNPs)

Region Shown: 36805231.5 36948626.5

Figure 11. A new map view for the MLH1 gene, using the options shown in Figure 10. Notice that the Variation map showing all known SNPs is now the "master map." See text for details.



Human chr3:36848249-36905608 – UCSC Genome Browser v31 – Netscape

http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr3:36848249-36905608&hgsid=24408863&intron

UCSC Genome Browser on Human April 2003 Freeze

position chr3:36848249-36905608 size 57,360 image width 610

Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options

Guidelines ☑ **Labels:** left ☑ center ☑

Note: Tracks with lots of items will automatically be displayed in more compact modes.

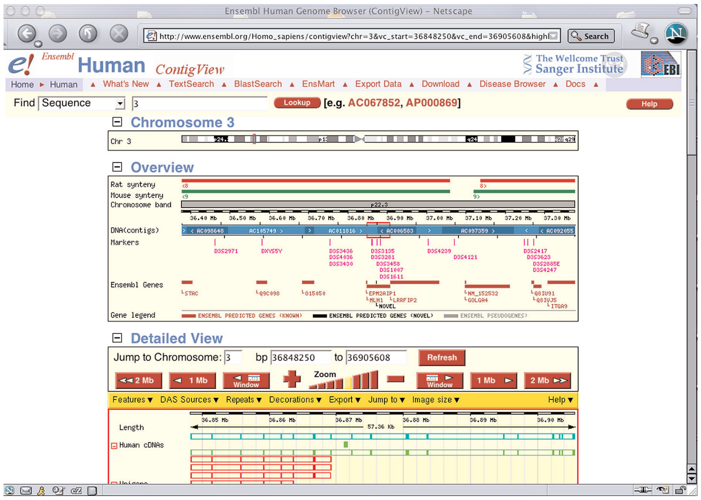
Figure 12. The UCSC Genome Browser. The region shown is for the human MLH1 gene. The overall organization is as a series of "tracks" that go from left to right, as opposed to the NCBI maps that go from top to bottom. The appearance of each track can be customized so that information appears at various densities; individual tracks can be selected or de-selected using toggles that appear below the graphic. Controls at the top of the window can be used to either zoom in or out or navigate 5′ or 3′ of the featured area.

right, then click <<*Remove*. Suppose that we wished to add the GenomeScan map (to see where all predicted genes are located) and the Variation map (to see the position of all known SNPs). To do so, select each of these from the list on the left, then click *Add*>>. When done, the window should be similar to that shown in Figure 10. Once these selections are made, the user would click the *Apply* button, and then the *Close* button. This will then recast the map as shown in Figure 11. Looking first at the 2 maps to the left, we see that the Genes_seq map and the Gscan (GenomeScan) map are not the same. The reason for this is that the GenomeScan map gives the results of a gene prediction algorithm (9), and the Genes_seq map's annotation is based on known and putative genes that have been placed as a result of alignments of mRNAs to individual contigs. Note that the MLH1 gene is marked in red (as before, the center-most gene on the map).

The master map (the right-most map) is now the Variation map, giving a different display than before. As with the UniGene map in Figure 10, the gray bars shown to the left of the Variation map indicate the density of SNPs at any given position. Some positions are simply marked with the number of variations (for example, "11 variations"), indicating that the map is too dense to display information on each individual SNP; simply zoom out to get more information at those positions (see below). In this view, numerous SNPs can be seen, each marked with an "rs" number. Clicking on that rs number would bring the user to the dbSNP page for that particular SNP, which is similar in appearance (but not identical) to the Variation page shown in the LocusLink example above (see Figure 5). Moving across from the rs number is a series of columns of interest. The column labeled *Map* indicates whether a particular SNP has been mapped to the genome. If the SNP has been mapped to a single position, a single green down-arrow would be shown (as in Figure 11); if the SNP has been mapped to multiple positions, a double down-arrow would be shown. The column labeled *Gene* indicates whether the SNP of interest is associated with a particular genomic feature. In each row of the Gene column, notice that there is an L, T, and C either "lit up" or "grayed out." If the L ("locus," blue) is lit, as it is for most of the SNPs in Figure 11, that indicates that the SNP lies within 2 kb of the 5′ end of a gene or within 500 bases of the 3′ end of a gene. If the T ("transcript," green) is lit, the marker overlaps with a known mRNA. Finally, if the C ("coding," orange) is lit, part or all of the SNP marker position overlaps with the coding region of a gene. The columns that follow provide additional information about the quality of the SNP marker, and more information on each of these can be found by clicking on the blue column headers.

In addition to changing the maps shown in any given view, the user can navigate by clicking anywhere on the ideogram on the left, or zoom in and out by clicking on the "out-zoom-in" picture above the ideogram. There are also short, gray bars at the top and bottom of each map that allow the user to "scroll up" or "scroll down," moving to the next genomic segment.

This particular example only scratches the surface of what can be done with the NCBI Map Viewer. To further complicate matters, there are 2 other genome browsers that have found widespread usage and should be in the arsenal of every molecular biologist. These browsers take slightly different approaches to visualizing genomic data, and users may prefer using 1 browser to another to answer a particular biological question. The 1st of these, the UCSC Genome Browser (10), is based on the concept of "tracks," where each track represents a particular type of annotation; this roughly corresponds to NCBI's maps. The annotation tracks available through UCSC include known genes, predicted genes, EST alignments, and cross-species homologies, to name a few. One of the strengths of the UCSC browser lies in its ease of navigation and the ability for individual users to display their own custom annotation tracks on a map, allowing them to correlate their own experimental data to publicly available data. An


Ensembl Human Genome Browser (ContigView) – Netscape

http://www.ensembl.org/Homo_sapiens/contigview?chr=3&vc_start=36848250&vc_end=36905608&highl

Find Sequence 3 [e.g. AC067852, AP000869]

Jump to Chromosome: 3 bp 36848250 to 36905608
Search
Ensembl
Human
ContigView
The Wellcome Trust Sanger Institute
EBI
Home ▸ Human ▴ What's New ▴ TextSearch ▴ BlastSearch ▴ EnsMart ▴ Export Data ▴ Download ▴ Disease Browser ▴ Docs ▴
Lookup
Help
Chromosome 3
Chr 3
p24.3
p13
q24
q28 q29
Overview
Rat synteny
Mouse synteny
Chromosome band
p22.3
36.40 Mb 36.50 Mb 36.60 Mb 36.70 Mb 36.80 Mb 36.90 Mb 37.00 Mb 37.10 Mb 37.20 Mb 37.30 Mb
DNA(contigs)
AC098648 AC105749 AC011816 AC006583 AC097359 AC092055
Markers
D3S2971 DXYS5Y D3S3436 D3S4036 D3S3430 D3S3135 D3S3281 D3S3458 D3S1007 D3S1611 D3S4239 D3S4121 D3S2417 D3S3623 D3S2885E D3S4247
Ensembl Genes
STAC Q9C098 O15050 EPM2AIP1 MLH1 LRRFIP2 NOVEL NM_152532 GOLGA4 Q8IU91 Q8IVJ5 ITGA9
Gene legend
ENSEMBL PREDICTED GENES (KNOWN) ENSEMBL PREDICTED GENES (NOVEL) ENSEMBL PSEUDOGENES)
Detailed View
Refresh
◂◂ 2 Mb ◂ 1 Mb ◂ Window + Zoom − Window ▸ 1 Mb ▸ 2 Mb ▸▸
Features ▾ DAS Sources ▾ Repeats ▾ Decorations ▾ Export ▾ Jump to ▾ Image size ▾ Help ▾
Length
36.85 Mb 36.86 Mb 36.87 Mb 36.88 Mb 36.89 Mb 36.90 Mb
57.36 Kb
Human cDNAs
Unigene

Figure 13. The Ensembl browser. The region shown is for the human MLH1 gene. The page begins with a chromosomal view, and then moves through various levels of detail. Controls in the section marked *Detailed View* can be used to either zoom in or out or navigate 5′ or 3′ of the featured area.

example of the UCSC display is given in Figure 12, showing the region around the MLH1 gene. Unlike the NCBI maps that run from top to bottom, the UCSC tracks run from left to right, allowing for more data to be displayed in a more intuitive fashion.

Finally, the Ensembl browser, developed by the Wellcome Trust Sanger Institute and EMBL's European Bioinformatics Institute, contains comprehensive genome annotation resulting from gene predictions and experimental data for 9 species (11). One of Ensembl's features is the ability to easily perform comparative analyses between species through the availability of DNA-DNA alignments, orthologous protein information, and large-scale synteny information. Ensembl's "ContigView" around the MLH1 gene is shown in Figure 13. As the user moves down the Web page, more detailed information is provided, moving from the chromosomal level in the 1st panel to an overview of the region of interest in the 2nd panel to a detailed view in the 3rd panel; scrolling to the bottom of the page (not shown) would eventually bring the user to the base pair level of detail.

BEYOND THE BASICS

While space obviously precludes an in-depth treatment of any of the 3 major genome browsers, a number of useful guides and papers have been published to help biologists make intelligent use of these powerful tools. Recently, National Human Genome Research Institute published *A User's Guide to the Human Genome* (12). The majority of the guide is devoted to a series of worked examples, providing an overview of the types of data available, details on how these data can be browsed, and step-by-step instructions and strategies for using many of the most commonly used tools for sequence-based discovery. In addition, each browser's Web site provides instructional information intended to assist the novice user in using the browsers to their best advantage. The URLs for these resources are given in Table 1.

Obviously, the range of publicly available data goes well beyond just the types of data discussed in this review. Because major public sequence databases like GenBank need to be able to store data in a generalized fashion, they often do not contain more specialized types of information that would be of interest to specific groups within the biological community. Many smaller, specialized databases have emerged to fill this gap, often developed and curated by biologists "in the trenches" to address the needs of their fellow investigators. These databases, which contain information ranging from strain crosses to gene expression data, provide a valuable supplement to the major sequence repositories, and the reader is encouraged to make intelligent use of both types of databases in their searches. An annotated list of such databases can be found in the yearly database issue of *Nucleic Acids Research* (13).

Table 1. Web-based resources for sequence analysis

Major Public Sequence Databases	
GenBank	http://www.ncbi.nlm.nih.gov
EMBL	http://www.ebi.ac.uk/embl/index.html
DNA Databank of Japan (DDBJ)	http://www.ddbj.nig.ac.jp
Gene-Centric Information Retrieval	
LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink
PubMed	http://www.ncbi.nlm.nih.gov/Entrez
OMIM	http://www.ncbi.nlm.nih.gov/Omim
UniGene	http://www.ncbi.nlm.nih.gov/UniGene
HomoloGene	http://www.ncbi.nlm.nih.gov/HomoloGene
dbSNP	http://www.ncbi.nlm.nih.gov/SNP
Genome Browsers	
Ensembl	http://www.ensembl.org
NCBI Map Viewer	http://www.ncbi.nlm.nih.gov/cgibin/Entrez/map_search
UCSC Genome Browser	http://genome.ucsc.edu
Educational Resources	
A User's Guide to the Human Genome	http://www.nature.com/genomics
Current Topics in Genome Analysis	http://www.genome.gov/COURSE2003

As is undoubtedly apparent by this point, there is no substitute for actually placing one's hands on the keyboard to learn how to effectively search and use genomic sequence data. Readers are strongly encouraged to take advantage of the resources presented in this review, grow in confidence and capability by working with the available tools, and begin to apply bioinformatic methods and strategies toward advancing their own research interests.

Address correspondence and reprint requests to *Andreas D Baxevanis, Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Building 50, Room 5222, Bethesda, Maryland 20852. Phone: 301-496-8570; fax: 301-480-2634; e-mail: andy@nhgri.nih.gov.*

Submitted August 27, 2003; accepted for publication October 1, 2003.

REFERENCES

1. Collins FS, Green ED, Guttmacher AE, Guyer MS. (2003) A vision for the future of genomics research. *Nature* 422:835-47.
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. (2003) GenBank. *Nucleic Acids Res.* 31:23-7.
3. Baxevanis AD. Information retrieval from biological databases. In: *Bioinformatics: a practical guide to the analysis of genes and proteins*. 2nd edition. Baxevanis AD and Ouellette BFF (eds.) John Wiley and Sons, New York, pp. 155-85.
4. Hamosh A et al. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 30:52-5.
5. Wolfsberg TG, Landsman D. Expressed sequence tags. In: *Bioinformatics: a practical guide to the analysis of genes and proteins*. 2nd edition. Baxevanis AD and Ouellette BFF (eds.) John Wiley and Sons, New York, pp. 283-302.
6. Velculescu VE, Vogelstein B, Kinzler KW. (2000) Analyzing uncharted transcriptomes with SAGE. *Trends Genet.* 16:423-5.
7. Blake JA et al. (2003) MGD: the Mouse Genome Database. *Nucleic Acids Res.* 31:193-5.
8. Sprague J et al. (2003) The Zebrafish Information Network (ZFIN): the zebrafish model organism database. *Nucleic Acids Res.* 31:241-3.
9. Yeh RF, Lim LP, Burge CB. (2001) Computational inference of homologous gene structures in the human genome. *Genome Res.* 11:803-16.
10. Karolchik D et al. (2003) The UCSC Genome Browser database. *Nucleic Acids Res.* 31:51-4.
11. Clamp M et al. (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* 31:38-42.
12. Wolfsberg TG, Wetterstrand KA, Guyer MS, Collins FS, Baxevanis AD. (2002) A user's guide to the human genome. *Nat. Genet.*, vol. 32 supplement.
13. Baxevanis AD. (2003) The Molecular Biology Database Collection: 2003 update. *Nucleic Acids Res.* 31:1-12.