Complex genomes at the NCBI

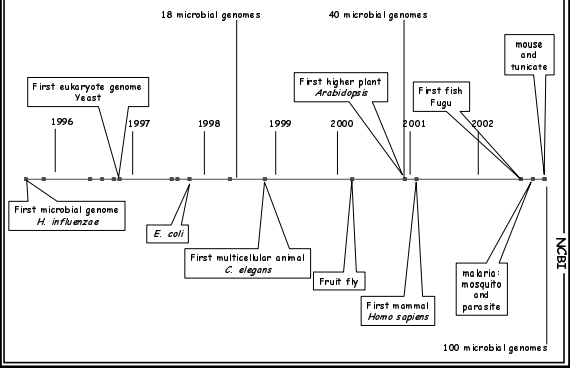
O'Reilly Bioinformatics Technology Conference 2003

Peter Cooper
National Center for Biotechnology Information

NCBI Booths: 25-26

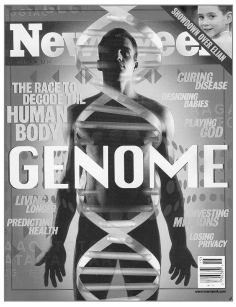
NCBI

The Genome Sequencing Era



NCBI

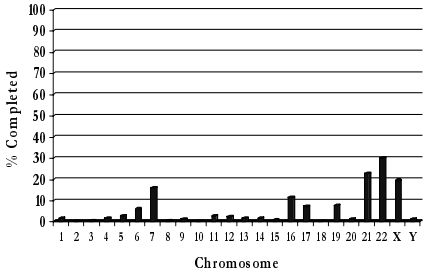
Complex Eukaryotic Genomes



NCBI

Human Genome Sequencing Progress

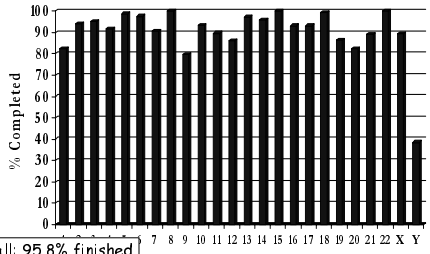
June 1998



NCBI

Human Genome Sequencing Progress

Dec. 2002



Overall: 95.8% finished
3.0% draft
98.8%

NCBI

Complex Genomes Jan. 2003

- Chordates
  - Human
  - Mouse
  - Rat
  - Pufferfish
  - Sea squirt (Ciona)
- Arthropods
  - *D. melanogaster*
  - *D. simulans*
  - *A. gambiae*

- Higher plants
  - *Arabidopsis*
  - Rice
- Fungi
  - *Aspergillus terreus*

NCBI

Coming soon …

- In progress
  - purple sea urchin
  - zebrafish
- NHGRI's Priority Organisms
  - Chicken
  - Cow
  - Dog
  - Chimpanzee
  - Honeybee
  - *Tetrahymena*
  - *Oxytichia*
  - Several fungi
- Over 100 bacterial genomes in progress

NCBI

Talk Outline

- About NCBI
- The Parts
  - Sequences
  - Maps
  - Domains
  - Structures
  - Variations
  - Literature
  - Genes
- Integrated Access
  - Entrez
  - Genome Resources
- Tour with MLH1
- Live Web Searches

NCBI

The National Center for Biotechnology Information (NCBI)

- **Created as a part of NLM in 1988**
  - **Establish public databases**
  - **Research in computational biology**
  - **Develop software tools for sequence analysis**
  - **Disseminate biomedical information**
- **Tools: BLAST(1990), Entrez (1992)**
- **GenBank (1992)**
- **Free MEDLINE (PubMed, 1997)**
- **Human genome (2001)**

NCBI

Sequences

```
103621 aataaaatac acactcacgc acacacatat gtactatata tatatatata tatatatata
103681 tatatatata tatatatata tatatttaaa gagtgtgtgt gtgtgtgtgt gtgtgtgtgt
103741 gtgtgtgagt gtgtgtgtgt gtgtgtgtgt gtgtgtgtgt gtctagatag agagagagag
103801 agagagagag agaataaata aaaaaaaaaa agaaagagag agagagagtg tgtgcgtgtg
103861 tgtgtgtgtg tgtgtgtgtg tatatagaga tatacatata gacagagaga gagagtgtgt
103921 gtatatatat atatatatat atatatatat agannnnnnn nnnnnnnnnn nnnnnnnnnn
103981 nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn
104041 nnnnnnnnnn nnnagtaaca gtacaagtat gatgtgatgt agttagtaaa gtaagcagtc
104101 agacgacatc atattgtttg tttattctaa atatcttttt gagggggcac ttgattaagc
104161 agaatggaaa gaaacatggc ttacgattta tatagactag gataacgttt ttctctcacg
104221 cttgagccac agatctgaca gattatagca gctttaacac acatttcaaa gttgtgtgtc
104281 ataaaaacac tctgtaattc actcaaaatg ctattgatac ccattttcgg atcacaaatt
104341 accggttata aaagctgta                    tct accggtagac gctggtcact
104401 atggctccct ccatgcagc                    taa ttgttaggag tgggggtgtg
104461 aacctacact ggtcttatg                    tac catggcatca atcggtttaa
104521 ttcgatatct cggtgcatct cagtgcattg aagatgcttt ccatacataa ttagattttt
104581 tcttcacaac ataacaattt ttttattcaa atctatttat atatatatat ccagggatgc
104641 catctgggcc cgtagctttt cggatgttta cgcgacaatc aatagatcta gttactttca
104701 ctacagaaac agagagtgtg ctggcattta caatatcggc cgtgggagcg tgcagtgtgt
104761 cacgctgtcc tgggcttcaa agcaggcatg gaacgagttg agctcgtccg gcagagatgc
104821 aacaatgttt acagtggcgg gtttcttact tttaaagtct gtgatttgat taggtctttg
104881 ccacatgctc cttgcgttat tggtgttaaa ttgaccctct accttgtctc tgtattgacg
104941 ttttgatgct ttgacagttt ttccaaggct gtaatttgct cgtttttgtt cctcagcatt
105001 cctggattta taggtggaag ctcgcactga tcgggttgtt caaacttcac tattgatcca
105061 tggtttttgg ttggggtaga cacggacagt ttttgttggc accacgtcgt ctatacattt
105121 tctgatgaag cacgtaactg tgtctgagta cacttcgatg tcatcatcag aagctacctg
```

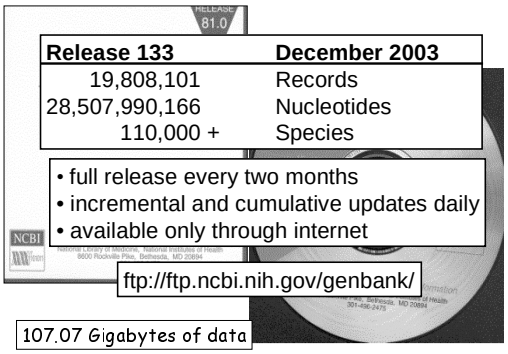
NCBI

Molecular Databases

- Primary Databases
  - Original submissions by experimentalists
  - Database staff organize but don't add additional information
    - **Example: GenBank**
- Derivative Databases
  - Human curated
    - compilation and correction of data
    - **Example: SWISS-PROT, NCBI RefSeq mRNA**
  - Computationally Derived
    - **Example: UniGene**
  - Combinations
    - **Example: NCBI Genome Assembly**

NCBI

GenBank: NCBI's Primary Sequence Database



Release 133	December 2003
19,808,101	Records
28,507,990,166	Nucleotides
110,000 +	Species

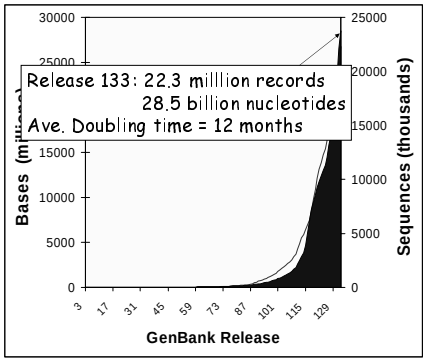
- full release every two months
- incremental and cumulative updates daily
- available only through internet

ftp://ftp.ncbi.nih.gov/genbank/

107.07 Gigabytes of data

NCBI

The Growth of GenBank



NCBI

Traditional GenBank Divisions

- Direct Submissions (Sequin and BankIt)
- Accurate (99.99%)
- Well characterized

BCT	Bacterial and Archeal
INV	Invertebrate
MAM	Mammalian (ex. ROD and PRI)
PHG	Phage
PLN	Plant and Fungal
PRI	Primate
ROD	Rodent
SYN	Synthetic (cloning vectors)
VRL	Viral
VRT	Other Vertebrate

Entrez limit: **gbdiv_XXX[Properties]**

NCBI

Traditional GenBank Record

```
LOCUS       HSHMLHI                 2503 bp    mRNA    linear   PRI 31-MAR-1994
DEFINITION  Human DNA mismatch repair (hmlh1) mRNA, complete cds.
ACCESSION   U07418
VERSION     U07418.1  GI:466461
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 2503)
  AUTHORS   Papadopoulos,N., Nicolaides,N.C., Wei,Y.F., Ruben,S.M.,
            Carter,K.C., Rosen,C.A., Haseltine,W.H., Fleischmann,R.D.,
            Fraser,C.M., Adams,M.D., Venter,J.C., Hamilton,S.R., Petersen,G.M.,
            Watson,P., Lynch,H.T., Peltomaki,P., Mecklin,J.P., Chapelle,A.D.,
            Kinzler,K.W. and Vogelstein,B.
  TITLE     Mutation of a mutL homolog in hereditary colon cancer
  JOURNAL   Science 263 (5153), 1625-1629 (1994)
  MEDLINE   94174309
   PUBMED   8128251
REFERENCE   2  (bases 1 to 2503)
  AUTHORS   Wei,Y.F.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-MAR-1994) Ying-Fei Wei, Molecular Biology, Human
            Genome Sciences, Inc., 9620 Medical Center Drive, Rockville, MD
            20850, USA
```

NCBI

Traditional GenBank Record

```
LOCUS       HSHMLHI                 2503 bp    mRNA    linear   PRI 31-MAR-1994
DEFINITION  Human DNA mismatch repair (hmlh1) mRNA, complete cds.
ACCESSION   U07418
VERSION     U07418.1  GI:466461
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 2503)
  AUTHORS   Papadopoulos,N., Nicolaides,N.C., Wei,Y.F., Ruben,S.M.,
            Carter,K.C., Rosen,C.A., Haseltine,W.H., Fleischmann,R.D.,
            Fraser,C.M., Adams,M.D., Venter,J.C., Hamilton,S.R., Petersen,G.M.,
            Watson,P., Lynch,H.T., Peltomaki,P., Mecklin,J.P., Chapelle,A.D.,
            Kinzler,K.W. and Vogelstein,B.
  TITLE     Mutation of a mutL homolog in hereditary colon cancer
  JOURNAL   Science 263 (5153), 1625-1629 (1994)
  MEDLINE   94174309
   PUBMED   8128251
REFERENCE   2  (bases 1 to 2503)
  AUTHORS   Wei,Y.F.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-MAR-1994) Ying-Fei Wei, Molecular Biology, Human
            Genome Sciences, Inc., 9620 Medical Center Drive, Rockville, MD
            20850, USA
FEATURES             Location/Qualifiers
     source          1..2503
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="3"
                     /map="p21"
                     /tissue_type="gall bladder"
                     /dev_stage="adult"
     gene            1..2503
                     /gene="hmlh1"
     CDS             42..2312
                     /gene="hmlh1"
                     /function="DNA mismatch repair"
                     /note="human homolog of E. coli mutL gene product,
                     Swiss-Prot Accession Number P23367"
                     /codon_start=1
                     /product="MSH2 protein"
                     /protein_id="AAA17374.1"
                     /db_xref="GI:466467"
                     /translation="MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKS
                     TSIQVIVKEGGLKLIQIQDNGTGIRKEDLDIVCERFTTSKLQSFEDLASISTYGFRGE
                     ALASISHVAHVTITTKTADGKCAYRASYSDGKLKAPPKPCAGNQGTQIYVEDLFYNIA
                     TRRKALKNPSEEYGKILEVVGRYSVHNAGISFSVKKQGETVADVRTLPNASTVDNIRS
                     VFGNAVSRELIEIGCEDETLAFKMNGYISNANYSVKKCIFLLFINHRLVESTSLKKAI
                     ETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLNEESILERVQQHIESKL
                     LGSNSSRMYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQK
                     LDAFLQPLSKPLESQPQATVTEDKTDTSSGRARQQDEEMLELPAPAEVAAKNQSLEGD
                     TTKGTSEMSEKKGPTSSNPRERHRESDSDVEMVEDDSRKEMTAACTPRRRIINLTSVLS
                     LQEEINSQQHEVLESEMLHNHSFVGCVNPQWALAQHQTKLYLLNTTKLSEELFYQILIY
                     DFANFGVLKLSEPAPLFDLAMLALDSPESGWTEEDGPKEGLAEYIVEFLKKKAEMLAD
                     YFSLEIDEKGNLIGLPLIINYVPPLESLPIFILRLATEVNWDEEKECFESLSRKECAM
                     FYSIRKQYISEESTLSGQQSEVPGSIPNSKWTYESIVYKASSSILPRNFTEDGNI
                     IGLANLPDLYKVFERC"
```

well annotated

NCBI

Traditional GenBank Record

```
LOCUS       HSHMLHI                 2503 bp    mRNA    linear   PRI 31-MAR-1994
DEFINITION  Human DNA mismatch repair (hmlh1) mRNA, complete cds.
ACCESSION   U07418
VERSION     U07418.1  GI:466461
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 2503)
  AUTHORS   Papadopoulos,N., Nicolaides,N.C., Wei,Y.F., Ruben,S.M.,
            Carter,K.C., Rosen,C.A., Haseltine,W.H., Fleischmann,R.D.,
            Fraser,C.M., Adams,M.D., Venter,J.C., Hamilton,S.R., Petersen,G.M.,
            Watson,P., Lynch,H.T., Peltomaki,P., Mecklin,J.P., Chapelle,A.D.,
            Kinzler,K.W. and Vogelstein,B.
  TITLE     Mutation of a mutL homolog in hereditary colon cancer
  JOURNAL   Science 263 (5153), 1625-1629 (1994)
  MEDLINE   94174309
   PUBMED   8128251
REFERENCE   2  (bases 1 to 2503)
  AUTHORS   Wei,Y.F.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-MAR-1994) Ying-Fei Wei, Molecular Biology, Human
            Genome Sciences, Inc., 9620 Medical Center Drive, Rockville, MD
            20850, USA
FEATURES             Location/Qualifiers
     source          1..2503
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="3"
                     /map="p21"
                     /tissue_type="gall bladder"
                     /dev_stage="adult"
     gene            1..2503
                     /gene="hmlh1"
     CDS             42..2312
                     /gene="hmlh1"
                     /function="DNA mismatch repair"
                     /note="human homolog of E. coli mutL gene product,
                     Swiss-Prot Accession Number P23367"
                     /codon_start=1
                     /product="MSH2 protein"
                     /protein_id="AAA17374.1"
                     /db_xref="GI:466467"
                     /translation="MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKS
                     TSIQVIVKEGGLKLIQIQDNGTGIRKEDLDIVCERFTTSKLQSFEDLASISTYGFRGE
                     ALASISHVAHVTITTKTADGKCAYRASYSDGKLKAPPKPCAGNQGTQIYVEDLFYNIA
                     TRRKALKNPSEEYGKILEVVGRYSVHNAGISFSVKKQGETVADVRTLPNASTVDNIRS
                     VFGNAVSRELIEIGCEDETLAFKMNGYISNANYSVKKCIFLLFINHRLVESTSLKKAI
                     ETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLNEESILERVQQHIESKL
                     LGSNSSRMYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQK
                     LDAFLQPLSKPLESQPQATVTEDKTDTSSGRARQQDEEMLELPAPAEVAAKNQSLEGD
                     TTKGTSEMSEKKGPTSSNPRERHRESDSDVEMVEDDSRKEMTAACTPRRRIINLTSVLS
                     LQEEINSQQHEVLESEMLHNHSFVGCVNPQWALAQHQTKLYLLNTTKLSEELFYQILIY
                     DFANFGVLKLSEPAPLFDLAMLALDSPESGWTEEDGPKEGLAEYIVEFLKKKAEMLAD
                     YFSLEIDEKGNLIGLPLIINYVPPLESLPIFILRLATEVNWDEEKECFESLSRKECAM
                     FYSIRKQYISEESTLSGQQSEVPGSIPNSKWTYESIVYKASSSILPRNFTEDGNI
                     IGLANLPDLYKVFERC"
```

well annotated

NCBI

Traditional GenBank Record

```
LOCUS       HSHMLHI                 2503 bp    mRNA    linear   PRI 31-MAR-1994
DEFINITION  Human DNA mismatch repair (hmlh1) mRNA, complete cds.
ACCESSION   U07418
VERSION     U07418.1  GI:466461
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 2503)
  AUTHORS   Papadopoulos,N., Nicolaides,N.C., Wei,Y.F., Ruben,S.M.,
            Carter,K.C., Rosen,C.A., Haseltine,W.H., Fleischmann,R.D.,
            Fraser,C.M., Adams,M.D., Venter,J.C., Hamilton,S.R., Petersen,G.M.,
            Watson,P., Lynch,H.T., Peltomaki,P., Mecklin,J.P., Chapelle,A.D.,
            Kinzler,K.W. and Vogelstein,B.
  TITLE     Mutation of a mutL homolog in hereditary colon cancer
  JOURNAL   Science 263 (5153), 1625-1629 (1994)
  MEDLINE   94174309
   PUBMED   8128251
REFERENCE   2  (bases 1 to 2503)
  AUTHORS   Wei,Y.F.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-MAR-1994) Ying-Fei Wei, Molecular Biology, Human
            Genome Sciences, Inc., 9620 Medical Center Drive, Rockville, MD
            20850, USA
FEATURES             Location/Qualifiers
     source          1..2503
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="3"
                     /map="p21"
                     /tissue_type="gall bladder"
                     /dev_stage="adult"
     gene            1..2503
                     /gene="hmlh1"
     CDS             42..2312
                     /gene="hmlh1"
                     /function="DNA mismatch repair"
                     /note="human homolog of E. coli mutL gene product,
                     Swiss-Prot Accession Number P23367"
                     /codon_start=1
                     /product="MSH2 protein"
                     /protein_id="AAA17374.1"
                     /db_xref="GI:466467"
                     /translation="MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKS
                     TSIQVIVKEGGLKLIQIQDNGTGIRKEDLDIVCERFTTSKLQSFEDLASISTYGFRGE
                     ALASISHVAHVTITTKTADGKCAYRASYSDGKLKAPPKPCAGNQGTQIYVEDLFYNIA
                     TRRKALKNPSEEYGKILEVVGRYSVHNAGISFSVKKQGETVADVRTLPNASTVDNIRS
                     VFGNAVSRELIEIGCEDETLAFKMNGYISNANYSVKKCIFLLFINHRLVESTSLKKAI
                     ETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLNEESILERVQQHIESKL
                     LGSNSSRMYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQK
                     LDAFLQPLSKPLESQPQATVTEDKTDTSSGRARQQDEEMLELPAPAEVAAKNQSLEGD
                     TTKGTSEMSEKKGPTSSNPRERHRESDSDVEMVEDDSRKEMTAACTPRRRIINLTSVLS
                     LQEEINSQQHEVLESEMLHNHSFVGCVNPQWALAQHQTKLYLLNTTKLSEELFYQILIY
                     DFANFGVLKLSEPAPLFDLAMLALDSPESGWTEEDGPKEGLAEYIVEFLKKKAEMLAD
                     YFSLEIDEKGNLIGLPLIINYVPPLESLPIFILRLATEVNWDEEKECFESLSRKECAM
                     FYSIRKQYISEESTLSGQQSEVPGSIPNSKWTYESIVYKASSSILPRNFTEDGNI
                     IGLANLPDLYKVFERC"
BASE COUNT     723 a    539 c    599 g    642 t
ORIGIN
        1 gttgaacatc tagacgttcc ctgggtcctc ctggcgccaa aatgtcgttc gtggcaggg
       61 ttattcggcg gctgaacgag acagtggtga accgcatcgc gcgcggggaa gttatccagg
      121 gccacgctaa tgtcatcaaa gagatgattg agaatcgtct agatcgcaaa tcccacgatg
      181 ttccaagtgat tgttaaagag ggaggcctga agttgattca gatccacagac aatggcacag
      241 gyatcaggaa agaagatctg gatatctgat gtgaaagctt cactactatg aaactgcagt
      301 ccttgaagga tttagccagt atttctacct atcgcttttg agctgaagcc ttggcccagca
      361 taagccatgt ggctcatgtt actatcacaa cgaaaacagc tgatgcaaag tgtgcatacc
      421 gagcaagtta ctcagatgga aaactgaaag ccccctcctaa accatgtgcc ggcaatcaag
      481 gyaccccagat cacggtgggg gatcttttct acaaatatgc cacgcgggga aaagctttaa
      541 aaaatccaag tgaagaatat gggaaattt tgggagttgt tgcgcgtat tcagtacacc
      601 tgtcccatgc ctcaccctg gatcatatcc gctcgtcttc tgtgaatgtc gtcgtcgtcg
      661 taccaatgc ctcaccctg gatcatatcc gctcgtcttc tgtgaatgct gtcagtcgtc
      721 aactgatagg aattggatgt gaggataaaa ccctatcctt caaaatgaat ggttacatat
      781 ccaatgcaaa ctactcagtg agaagtgcca tcttcttact ctcatcaac catcgctcgg
      841 tagaatcaac ttccttgaga aaagccatag aaacatgtga tgcagcatat ttgccccaaa
      901 acacacaccc attcctgtac ctcagttatg aaatcagtcc cagaatgtg gatgttaatg
      961 tgcaccccac aaagcatgaa gttcaattcc tgcaagagga gagcatcctg gagcggtgg
     1021 agcagcacat cgaggagaag ctccctggct ccaattcctc caggatgtac ttcaccccag
     1081 cttgctacc aggactgctc gcccctctg gggagatggt taaatccaca acaagtctga
     1141 cctcgtcttc tactctcgga agtagtgata aggtctatgc cccaccagatg gttcgtacag
     1201 atccccggga acagaagctt gatgcatttc tgcagcctct gagcaaaccc ctgtccagtc
     1261 agccccaggc cattgtcacc gaggatcaga cagatatttc tagtgcgcg gctaggcagt
     1321 agatgaggga gatgcttgaa ctcccacgccc ctcgtcgagt gctcgccaaa aatcaagact
     1381 gtcaccaccc cccaccaccc cccaccaccc cccaccaccc cccaccaccc cccaccaccc
     1441 gcaaccccag aaagagacat cggagaatct ctgatgtgga aatgctggaa gatgatctcc
     1501 gaaggaaaat gatcgagctt tgtaccccc gggaaaggat cattaacctc acttgtgttt
     1561 tgagtctcca ggaagaatt aatgagagg gacatcgagt tctccggaa atgttgtaca
     1621 accactcctt cgtgggtgtg gtgaatcctc atgtgggcctt gcccacgcat caaaccaagt
     1681 tatacctctt caaacccacc aagtctatg agagactgtt gccacagcat cgtgaatctg
     1741 atcttgccaa tttgtgtgtt ctcaggttac cggagccagc acccgcttcc gacctgccaa
     1801 tgcttgcctt agatagtcca gagagtggct gtaccagagga agatgtgccc aaagaaggag
     1861 ttgctgatca cattgttgag tttctgaaga agaaggctga gatgcttgca gactatttct
     1921 ctttgaaact tgatgaggaa gggaactga ttggattacc cctctgatt gacaactatg
     1981 tgcccccttt ggaggagctg cctatcttca ttcttcgact agccactgag gtgaactcgg
     2041 acgaagaaaa ggatgtttt gaaagcttca gtaaagaatg cgctatgttc tattcatcct
     2101 gaagcagtta cattatcgag gagtcgaccc cctcaagccca gcagatgaga gtgcctgcct
     2161 ccaccaccgc cccaccaccc cccaccaccc cccaccaccc cccaccaccc cccaccaccc
     2221 acattcgcc tcctaaacat ttcacagaag atcgaaatct ctcgcagctt gcccacctgt
     2281 ctgatctata caaagtcttt gagaggtgtt aaatatggtt atttatgcac tgtgggatgt
     2341 gtctctcttt ctcfcgattc cgatacaaag tgttgtatca aagtgtgata tacaaagtgt
     2401 accaacataa gtgttggtag cacttaagac ttatacttgc cttctgatag tattcctta
     2461 tacacagtgg attgatatta aataaataga tgtgtcttaa cat
```

well annotated

the sequence is the data

NCBI

Bulk GenBank Divisions

- Batch Submission and htg (email and ftp)
- Inaccurate
- Poorly Characterized

EST	Expressed Sequence Tag
STS	Sequence Tagged Site
GSS	Genome Survey Sequence
HTG	High Throughput Genomic

NCBI

GenBank Bulk Sequence

```
LOCUS       AZ211804                 408 bp    DNA     linear   GSS 31-AUG-2000
DEFINITION  SP_0153_B2_F09_T7A Strongylocentrotus purpuratus, purple sea urchin
            , sperm genomic BAC library Strongylocentrotus purpuratus genomic
            clone Plate=153 Col=18 Row=L, DNA sequence.
ACCESSION   AZ211804
VERSION     AZ211804.1  GI:8427953
KEYWORDS    GSS.
SOURCE      Strongylocentrotus purpuratus
  ORGANISM  Strongylocentrotus purpuratus
            Eukaryota; Metazoa; Echinodermata; Eleutherozoa; Echinozoa;
            Echinoidea; Euechinoidea; Echinacea; Echinoida;
            Strongylocentrotidae; Strongylocentrotus.
REFERENCE   1  (bases 1 to 408)
  AUTHORS   Cameron,R.A., Mahairas,G., Rast,J.P., Martinez,P., Biondi,T.R.,
            Swartzell,S., Wallace,J.C., Poustka,A.J., Livingston,B.T., Wray
            ,G.A., Ettensohn,C.A., Lehrach,H., Britten,R.J, Davidson,E.H. and
            Hood,L.
  TITLE     A sea urchin genome project: Sequence scan, virtual map, and
            additional resources
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 97(17), 9514-9518 (2000)
  MEDLINE   20402566
  PUBMED    10920195
COMMENT     Contact: Cameron, RA, Davidson, EH, Hood, L
            Division of Biology 156-29
            California Institute of Technology
            Pasadena California 91125, USA
            Tel: (626) 395-8421
            Fax: (626) 793-3047
            Email: acameron@caltech.edu
            Plate: 153  row: L  column: 18
            Seq primer: T7
            Class: BAC end#
            High quality sequence stop: 408.
```

produced in quantity

NCBI

GenBank Bulk Sequence

```
LOCUS       AZ211804                 408 bp    DNA     linear   GSS 31-AUG-2000
DEFINITION  SP_0153_B2_F09_T7A Strongylocentrotus purpuratus, purple sea urchin
            , sperm genomic BAC library Strongylocentrotus purpuratus genomic
            clone Plate=153 Col=18 Row=L, DNA sequence.
ACCESSION   AZ211804
VERSION     AZ211804.1  GI:8427953
KEYWORDS    GSS.
SOURCE      Strongylocentrotus purpuratus
  ORGANISM  Strongylocentrotus purpuratus
            Eukaryota; Metazoa; Echinodermata; Eleutherozoa; Echinozoa;
            Echinoidea; Euechinoidea; Echinacea; Echinoida;
            Strongylocentrotidae; Strongylocentrotus.
REFERENCE   1  (bases 1 to 408)
  AUTHORS   Cameron,R.A., Mahairas,G., Rast,J.P., Martinez,P., Biondi,T.R.,
            Swartzell,S., Wallace,J.C., Poustka,A.J., Livingston,B.T., Wray
  TITLE
  JOURNAL
  MEDLINE
  PUBMED
COMMENT
            High quality sequence stop: 408.
FEATURES             Location/Qualifiers
     source          1..408
                     /organism="Strongylocentrotus purpuratus"
                     /db_xref="taxon:7668"
                     /clone="Plate=153 Col=18 Row=L"
                     /clone_lib="Strongylocentrotus purpuratus, purple sea
                     urchin, sperm genomic BAC library"
                     /note="Organ: sperm; Vector: BACe3.6; BAC Clones in E-Coli
                     DH10B"
                     96 a    150 c     54 g    108 t
ORIGIN
        1 tcccccatatg caccacgccc gtttttcgcg gtcattcaag ccccacatca tgggcctaca
       61 caactggccc tgcatacccg agcagaatca ttttgctgct cgctacaact agcccacacc
      121 cacagcccct ctgcgcgctc aaaccagcca ggctccccac atattattaa ttttttagt
      181 tatactcctt tcattgacag tatactatta actatcaggc aatattgcgc atgctctgtc
      241 ctcaacatac agctcccccag ccattctggc cccccggccc cgggcaccca acaacttggc
      301 caacaattct atccttcatg cctctcttgc tccacaacaa ccatatccta actggtcttt
      361 tacaaacggc tatcatttct aatttcccat ctgtcctccc tccaacac
//
```

low accuracy

poorly characterized

NCBI

EST Division: Expressed Sequence Tags

gbdiv_est[Properties]

```
>IMAGE:275615 5' mRNA sequence
GACAGCATTCGGGCCGAGATGTCTCGCTCCGTGGCCTTAGCTGTGCTCGCGCTACTCTCTCTTTCTGGCC
TGGAGGTATCCAGCGTACTCCAAAGATTCAGGTTTACTCACGTCATCCAGCAGAGAATGGAAAGTCAAAT
TTCCTGAATTGCTATGTGTCTGGGTTTCATCCATCCGACATTGAAGTTGACTTACTGAAGAATGGAGAGA
GAATTGAAAAAGTGGAGCATTCAGACTTGTCTTTCAGCAAGGACTGGTCTTTCTATCTCTTGTACTACAC
TGAATTCACCCCCACTGAAAAAGATGAGTATGCCTGCCGTGTTGAACCATGTNGACTTTGTCACAGNCCC
AAGTTNAGTTTAAGTGGGNATCGAGACATGTAAGGCAGGCATCATGGGAGGTTTTGAAGNATGCCGCNTT
TTGGATTGGGATGAATTCCAAATTTCTGGTTTGCTTGNTTTTTTAATATTGGATATGCTTTTG
```

```
>IMAGE:275615 3', mRNA sequence
NNTCAAGTTTTATGATTTATTTAACTTGTGGAACAAAAATAAACCAGATTAACCACAACCATGCCTTACT
TTATCAAATGTATAAGANGTAAATATGAATCTTATATGACAAAATGTTTCATTCATTATAACAAATTTCC
AATAATCCTGTCAATNATATTTCTAAATTTTCCCCCAAATTCTAAGCAGAGTATGTAAATTGGAAGTTAA
CTTATGCACGCTTAACTATCTTAACAAGCTTTGAGTGCAAGAGATTGANGAGTTCAAATCTGACCAAGAT
GTTGATGTTGGATAAGAGAATTCTCTGCTCCCCACCTCTANGTTGCCAGCCCTC
```

make cDNA library

80-100,000 unique cDNA clones in library

NCBI

What is UniGene?

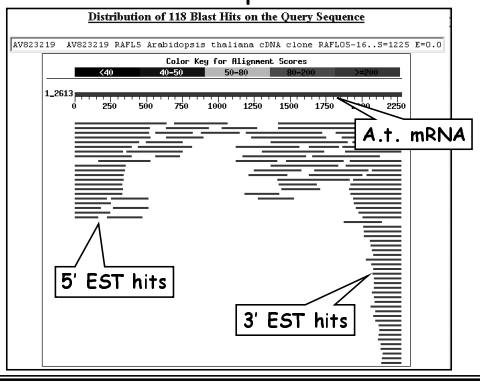
A gene-oriented view of sequence entries

- MegaBlast based automated sequence clustering
- Nonredundant set of gene oriented clusters
- Each cluster a unique gene
- Information on tissue types and map locations
- Includes well-characterized genes and novel ESTs
- Useful for gene discovery and selection of mapping reagents

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

NCBI

EST hits A.t. serine protease mRNA



NCBI

Arabidopsis UniGene Statistics

UniGene Build 14
Dec. 5th, 2002

```
    52,966     mRNAs + gene CDSs
    87,249     EST, 3'reads
    42,444     EST, 5'reads
+   34,790     EST, other/unknown
----------
   217,479    total sequences in clusters

Final Number of Clusters (sets)
===============================
27,159   sets total

25,976   [illegible] one known gene
17,758   [illegible] one EST
16,566   sets contain both genes and ESTs
```

115,000,000 bp
25,498 expected genes
5% uncharacterized transcripts

NCBI

Hs UniGene Statistics

UniGene Build 158
Dec. 5, 2002

```
   115,648     mRNAs + gene CDSs
 1,436,481     EST, 3'reads
 1,779,347     EST, 5'reads
+  634,745     EST, other/unknown
----------
 3,966,221    total sequences in clusters

Final Number of Clusters (sets)
===============================
 128,826 sets total
```

3,000,000,000 base pairs
30 K expected genes
80% uncharacterized transcripts

```
 33,111   s                    one known mRNA
123,884   s                    one EST
 28,169   sets contain both mRNAs and ESTs
```

NCBI

UniGene Collections: Animals

		Sequences	Clusters
Homo sapiens	human	3,966,221	128,826
Mus musculus	mouse	3,348,641	94,528
Rattus norvegicus	rat	345,498	63,430
Bos taurus	cow	130,603	10,675
Sus scrofa	pig	60,253	14,315
Gallus gallus	chicken	118,031	4,993
Danio rerio	zebrafish	218,881	15,968
Xenopus laevis	frog	173,697	19,451
Ciona intestinalis	tunicate	464,757	14,254
Drosophila melanogaster	fruit fly	277,372	14,835
Anopheles gambiae	mosquito	43,713	2,581
Caenorhabditis elegans	nematode	176,719	20,386

Jan. 31, 2003

NCBI

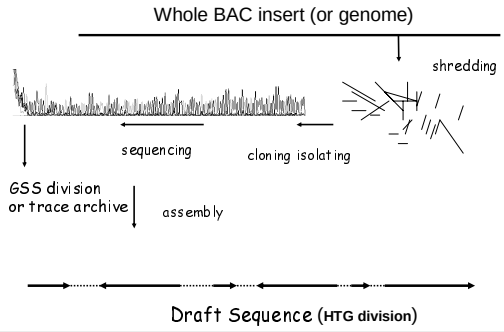
UniGene Collections: Plants

		Sequences	Clusters
Arabidopsis thaliana	thale cress	220,191	27,248
Glycine max	soybean	178,849	8,815
Medicago truncatula	barrel medic	83,918	5,823
Lycopersicon esculentum	tomato	60,091	3,740
Triticum aestivum	wheat	305,454	20,458
Hordeum vulgare	barley	163,844	7,941
Zea mays	maize (corn)	153,669	12,628
Oryzia sativa	rice	83,435	16,886
Chlamydamonas reinhardti	alga	83,548	6,549

Jan. 31, 2003

NCBI

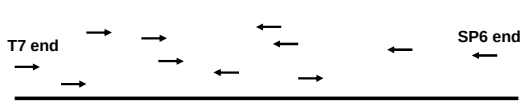
Genome Sequencing - HTG, GSS,(WGS)



NCBI

GSS Division: Genome Survey Sequences

- Genomic equivalent of ESTs
- BAC and other first pass surveys
- BAC end sequences
- Whole Genome Shotgun (some)
- RAPIDS and other anonymous loci

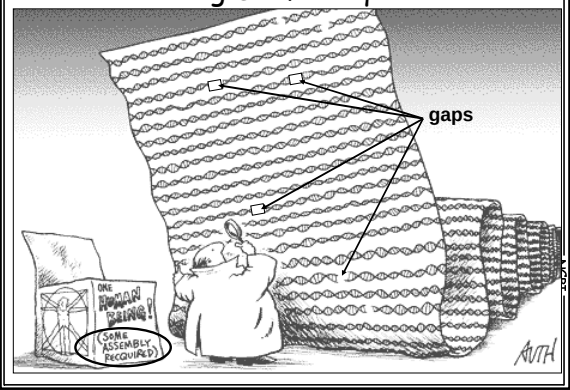


Genomic Clone (BAC)

`gbdiv_gss[Properties]`

NCBI

Working Draft Sequence



NCBI

HTG Division: High Throughput Genome

Zebrafish PAC Clone

phase 1 → ← → → ← HTG
Acc = AC109580.1

phase 2 → → → HTG
Acc =AC109580.11

phase 3 → VRT
Acc = AC109580.14

`gbdiv_htg[Properties]`

40,000 to > 350,000 bp

NCBI

HTG Division: High Throughput Genome

```
LOCUS       AC109580                 2463 bp    DNA     linear   HTG 05-FEB-2002
DEFINITION  Danio rerio clone busml-183j13 strain AB, WORKING DRAFT SEQUENCE, 1
            unordered piece.
ACCESSION   AC109580
VERSION     AC109580.1  GI:18497096
KEYWORDS    HTG; HTGS_PHASE1; HTGS_DRAFT.
SOURCE      Danio rerio (zebrafish)
```

```
LOCUS       AC109580                97770 bp    DNA     linear   HTG 26-OCT-2002
DEFINITION  Danio rerio clone busml-183j13 strain AB, WORKING DRAFT SEQUENCE, 4
            ordered pieces.
ACCESSION   AC109580
VERSION     AC109580.11  GI:24414284
KEYWORDS    HTG; HTGS_PHASE2; HTGS_DRAFT.
SOURCE      Danio rerio (zebrafish)
```

```
LOCUS       AC109580                94348 bp    DNA     linear   VRT 19-NOV-2002
DEFINITION  Danio rerio clone busml-183j13 strain AB, complete sequence.
ACCESSION   AC109580
VERSION     AC109580.14  GI:24635954
KEYWORDS    HTG.
SOURCE      Danio rerio (zebrafish)
```

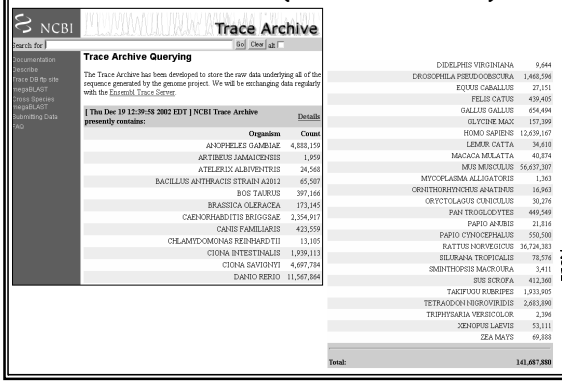
40,000 to > 350,000 bp

NCBI

Whole Genome Shotgun

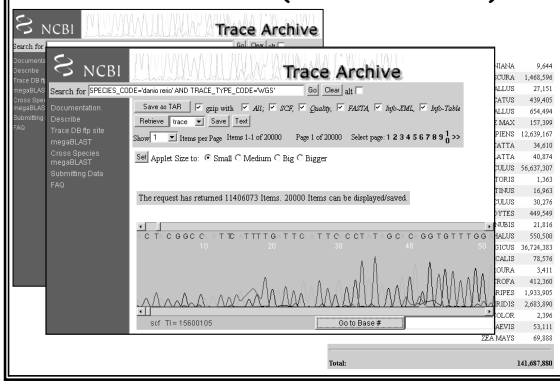
NCBI

Trace Archive (not in GenBank)



NCBI

Trace Archive (not in GenBank)



NCBI

Whole Genome Shotgun in GenBank

`wgs_master[Properties]`

1: AABT00000000
Aspergillus terreus, whole genome shotgun sequencing project
gi|27262064|gb|AABT00000000.1|AABT01000000[27262064]

2: AABU00000000
Drosophila melanogaster strain y; cn bw
sequencing project
gi|27232097|gb|AABU00000000.1|AAB

3: AABS00000000
Ciona intestinalis, whole genome shotg
gi|26556995|gb|AABS00000000.1|AAB

4: AABR00000000
Rattus norvegicus strain BN/SsNHsd/M
sequencing project
gi|25762278|gb|AABR00000000.1|AAB

5: AABL00000000
Plasmodium yoelii yoelii, whole genom
gi|23491527|gb|AABL00000000.1|AAB

6: CAAB00000000
Takifugu rubripes, whole genome shotg
gi|22418063|emb|CAAB00000000.1|CA

7: CAAC00000000
Caenorhabditis briggsae, whole genome shotgun sequencing project
gi|22417469|emb|CAAC00000000.1|CAAC01000000[22417469]

8: AAAB00000000
Anopheles gambiae str. PEST whole genome shotgun sequencing project
gi|19683831|gb|AAAB00000000.1|AAAB01000000[19683831]

9: AAAD00000000
Mus musculus chromosome 16, whole genome shotgun sequencing project
gi|21281728|gb|AAAD00000000.1|AAAD01000000[21281728]

10: CAAA00000000
Mus musculus, whole genome shotgun sequencing project
gi|20800445|emb|CAAA00000000.1|CAAA01000000[20800445]

11: AAAA00000000
Oryza sativa (indica cultivar-group) cultivar 93-11 whole genome shotgun sequencing project
gi|19924305|gb|AAAA00000000.1|AAAA01000000[19924305]

NCBI

Whole Genome Shotgun in GenBank

```
LOCUS       AABS01002501            3216 bp    DNA     linear   INV 12-DEC-2002
DEFINITION  Ciona intestinalis cionaS501, whole genome
ACCESSION   AABS01002501 AABS01000000
VERSION     AABS01002501.1  GI:26556991
KEYWORDS    WGS.
SOURCE      Ciona intestinalis
  ORGANISM  Ciona intestinalis
            Eukaryota; Metazoa; Chordata; Urochordata;
            Phlebobranchia; Cionidae; Ciona.
REFERENCE   1  (bases 1 to 3216)
  AUTHORS   Dehal,P., Satou,Y., Campbell,R.K., Chapman,
            Tomaso,A., Davidson,B., Di Gregorio,A., Gel
            Harafuji,N., Hastings,K.E., Ho,I., Hotta,K.
            Kawashima,T., Lemaire,P., Martinez,D., Meik
            Necula,S., Nonaka,M., Putnam,N., Rash,S., S
            Terry,A., Yamada,L., Wang,H.G., Awazu,S., A
            Branno,M., Chin-Bow,S., DeSantis,R., Doyle,
            Keys,D.N., Saga,S., Sawyamshi,H., Hino,K., IM
            Kano,S., Kobayashi,K., Kobayashi,M., Lee,B.
            Manohar,C., Matassi,G., Medina,M., Mochizuk
            Morishita,T., Miura,S., Nakayama,A., Nishis
            Ohta,F., Oishi,K., Rigoutsos,I., Sano,M., S
            Shoguchi,E., Shin-I,T., Spagnuolo,A., Stain
            Tassy,O., Takatori,N., Tokuoka,M., Yagi,K.
            Zhang,C., Hyatt,P.D., Larimer,F., Detter,C.
            Glavina,T., Hawkins,T., Richardson,P., Lucas
            Levine,M., Satch,N. and Rokhsar,D.S.
  TITLE     The Draft Genome of Ciona intestinalis: Inm
            Vertebrate Origins
  JOURNAL   Science 298 (5601), 2157-2167 (2002)
   PUBMED   12481130
REFERENCE   2  (bases 1 to 3216)
  AUTHORS   .
  CONSRTM   DOE Joint Genome Institute
  TITLE     Direct Submission
  JOURNAL   Submitted (26-NOV-2002) Production Sequenci
            Genome Institute, 2800 Mitchell Drive, Wal
```

```
  JOURNAL   Submitted (26-NOV-2002) Production Sequencing Facility, DOE Joint
            Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA
COMMENT     This record contains the initial draft assembly of the Ciona
            intestinalis genome as described in Dehal et al. Science 2002. The
            sequence spans 114.7 million base pairs of an estimated 153-159
            million base pair genome, with the unassembled sequence comprising
            predominantly repetitive sequence, as described in Dehal et al. 99%
            of known genes are recovered in this draft sequence; we estimate
            one gene-disrupting misassembly per 75 genes (500 kilobases).
            Strings of N's in this record represent gaps between contigs, and
            the length of each string corresponds to the estimated size of the
            gap based on spanning clones. Gaps with an expected size less than
            50 kp are represented by a string of 50 N's. In some cases small
            gaps may represent unresolved overlaps from the contigs flanking
            the gap. The sequence is a mosaic of two haplotypes found in a
            single individual selected for sequencing from Half Moon Bay
            California. While no quality scores are available for this
            assembly, we estimate the indel rate at better than one per 16
            kilobases (Dehal et al.). Sequences near gaps are likely to be
            lower quality. This draft sequence will be updated as additional
            data are incorporated into the assembly.
FEATURES             Location/Qualifiers
     source          1..3216
                     /organism="Ciona intestinalis"
                     /db_xref="taxon:7719"
BASE COUNT     1087 a    662 c    694 g    723 t    50 others
ORIGIN
        1 ttctggctgt gcccatgttt acagtaacca gttttacact ccccatgaat cacctgatgtc
       61 atatcctgtc tacatgacac acaggggtgt gaggtcatcg ctgtgtgtaca gctatatatt
      121 gtgtgtcatac aacatgccac atagatatat agtcatgtcta tgtacacata acaggggccc
      181 gggtatcgaa ttaatggctg gatgcatggc tgactcggac ggcaacacaca tccatccatta
      241 tccactcagac gtcgatgctg agcaacctgg gaaaggaggg atccatctcc atgtcaagctt
      301 gttaaactta tigaatataa atcgcttigt tctactcaag actactcgca cacctcagca
      361 gccactctca aaatccagat aaaaacgtaa atactcttca tgtgagtgtg ggcgtgtaac
      421 tacatctacac ttatatacatt ttaaatacct ttgacactt acactatata gacgtcgtac
      481 ttatatactta actgccacaa taaaggtggt tgactgtggg atccaaatcg tgtttaactta
      541 tggcttttatt tacaaatagc tgacatcata acatagtcat gataaaacat ttactattag
      601 cattataagt aactacgaca tgtacacagt gactgatccc cccgaaagcaa aattatctaa
      661 caatcaaata tgcgtgtgag gaggggttgg tatcactcag gctcgtcatca gagtcttcag
      721 gttggcttca cctgagatta aacaagctga ggggggatgcc cccatcatga gaacccatcaa
      781 tgatcaaagt atgtgatcat aggcgtcgat ttttcacgt tgtaaaccaa tctcagggag
      841 gcgaaatana tgtcagtgta tgcttaacta cgacgcacgt gaaggctaaa ccacgacaaa
      901 cgtgggcgca aaactacgac gcacgtgaag gctaaactac gaacgcacgtg aaggctaaac
      961 tatgacgcac gtgaaggcta agctacgacg cacgtgaagg ctaaactacg acgcacgtga
     1021 aggctaaaac acgacgcacg tgaaggctaa actacgacgc acgtgaaggc taaactacga
     1081 cgcacgtgaa ggctaaacta cgacgcacgt gaaggctaaa acacgacaaa cgtgaaggct
     1141 aaactacgac gcacgtgaag gctaaactac gaacgcacgtg aaggctaaaa cacgacgcac
     1201 gtgaaggcta aactacgacg cacgtgaagg ctaaactacn nnnnnnnnnn nnnnnnnnnn
     1261 nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn gatatgaggg tctatcatcg aaagaaatct
```

NCBI

STS Division : Sequence Tagged Sites

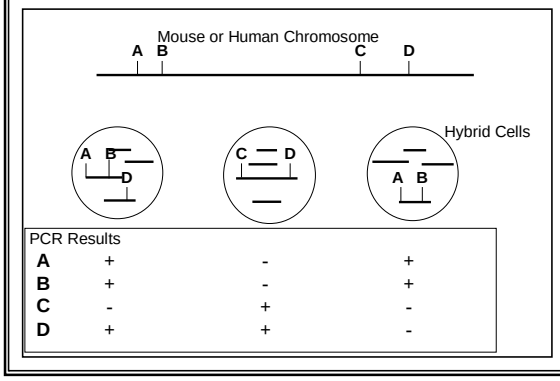
- Segment of gene, EST , mRNA or genomic DNA of known position (microsatellite)
- PCR with STS primers gives one product per genome
- Basis of Radiation Hybrid Mapping
  - UniGene
  - Genome Assembly
- Related resource: Electronic PCR

http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi

gbdiv_sts[Properties]

NCBI

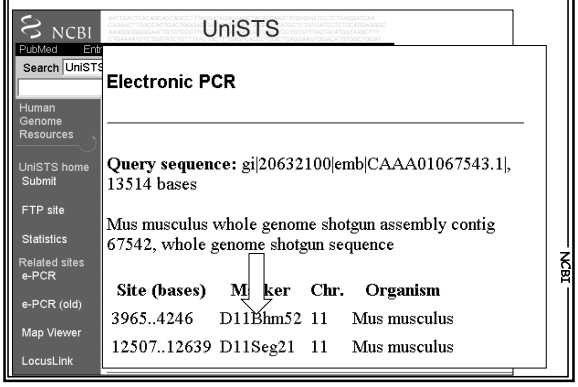
RH mapping using STSs



PCR Results			
A	+	-	+
B	+	-	+
C	-	+	-
D	+	+	-

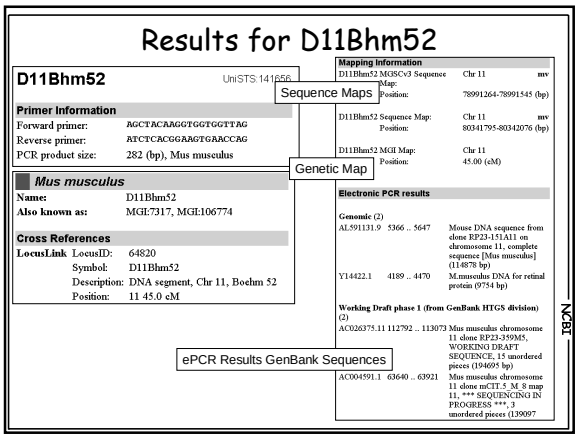
NCBI

Electronic PCR



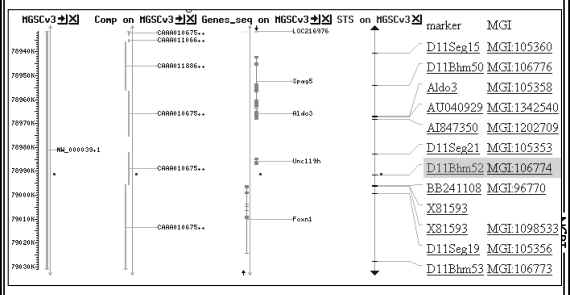
NCBI

Results for D11Bhm52



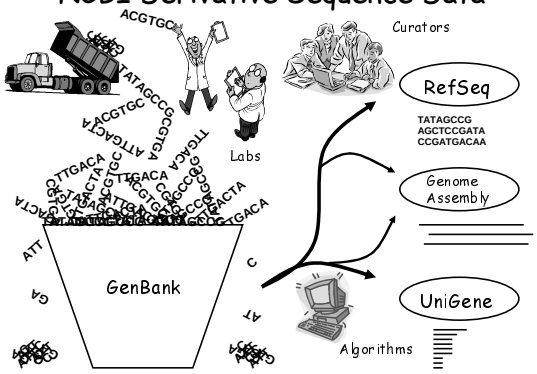
NCBI

Preview of Maps: Markers on MGSCv3



NCBI

NCBI Derivative Sequence Data



NCBI

RefSeq: NCBI's Derivative Sequence Database

- **Curated transcripts and proteins**
  - reviewed
  - human, mouse, rat, cow, fruit fly, zebrafish, arabidopsis *C.elegans*

  Model transcripts and proteins
- **Assembled Genomic Regions (contigs)**
  - draft human genome
  - mouse genome
- **Chromosome records**
  - microbial
  - organelle

`srcdb_refseq[Properties]`

NCBI

RefSeq Benefits

- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current sequence data and biology
- data validation
- format consistency
- distinct accession series
- stewardship by NCBI staff and collaborators

NCBI

RefSeq Accession Numbers

mRNAs and Proteins

NM_123456	Curated mRNA
NP_123456	Curated Protein
NR_123456	Curated non-coding RNA
XM_123456	Predicted mRNA
XP_123456	Predicted Protein
XR_123456	Predicted non-coding RNA

Gene Records

NG_123456	Reference Genomic Sequence

Assemblies

NT_123456	Contig
NW_123456	WGS Supercontig
NC_123455	Chromosome

NCBI

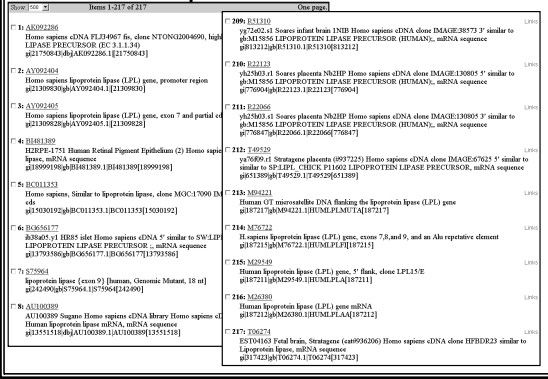
Useful Entrez Limits: RefSeqs

Query Limit	Accession Prefixes	Description
`srcdb_refseq_reviewed[prop]`	NC_, NG_, NM_, NP_	Reviewed
`srcdb_refseq_provisional[prop]`	NC_, NM_, NP_	Provisional
`srcdb_refseq_model[prop]`	XM_, XP_, XR_	Model

Reviewed Transcripts: `srcdb_refseq_reviewed[prop] AND biomol_mrna[prop]`

NCBI

GenBank Sequences: Human Lipoprotein Lipase



NCBI

Curated RefSeq Records: NM_, NP_

```
LOCUS       Lpl                     3969 bp    mRNA    linear   ROD 07-JAN-2002
DEFINITION  Mus musculus lipoprotein lipase (Lpl), mRNA.
ACCESSION   NM_008509
VERSION     NM_008509.1  GI:6678709
KEYWORDS    .
SOURCE      Mus musculus.
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 3969)
  AUTHORS   Kirchgessner,T.G., Svenson,K.L., Lusis,A.J. and Schotz,M.C.
  TITLE     The sequence of cDNA encoding lipoprotein lipase. A member of a
            lipase gene family
  JOURNAL   J. Biol. Chem. 262 (18), 8463-8466 (1987)
  MEDLINE   87250454
   PUBMED   3597382
REFERENCE   2  (bases 1 to 3969)
  AUTHORS   Zechner,R., Newman,T.C., Steiner,E. and Breslow,J.L.
  TITLE     The structure of the mouse lipoprotein lipase gene: a B1 repetitive
            element is inserted into the 3' untranslated region of the mRNA
  JOURNAL   Genomics 11 (1), 62-76 (1991)
  MEDLINE   92112227
   PUBMED   1765386
REFERENCE   3  (bases 1 to 3969)
  AUTHORS   Hua,X.X., Enerback,S., Hudson,J., Youkhana,K. and Gimble,J.M.
  TITLE     Cloning and characterization of the promoter of the murine
            lipoprotein lipase-encoding gene: structural and functional
            analysis
  JOURNAL   Gene 107 (2), 247-258 (1991)
  MEDLINE   92084117
   PUBMED   1748295
REFERENCE   4  (bases 1 to 3969)
```

NCBI

Curated RefSeq Records: NM_, NP_

```
LOCUS       Lpl                     3969 bp    mRNA    linear   ROD 07-JAN-2002
DEFINITION  Mus musculus lipoprotein lipase (Lpl), mRNA.
ACCESSION   NM_008509
VERSION     NM_008509.1  GI:6678709
KEYWORDS    .
SOURCE      Mus musculus.
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 3969)
  AUTHORS   Kirchgessner,T.G., Svenson,K.L., Lusis,A.J. and Schotz,M.C.
  TITLE     The sequence of cDNA encoding lipoprotein lipase. A member of a
            lipase gene family
  JOURNAL   J. Biol. Chem. 262 (18), 8463-8466 (1987)
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from M60847.1.
REFERENCE   2  (bases 1 to 3969)
  AUTHORS   Zechner,R., Newman,T.C., Steiner,E. and Breslow,J.L.
  TITLE     The structure of the mouse lipoprotein lipase gene: a B1 repetitive
            element is inserted into the 3' untranslated region of the mRNA
  JOURNAL   Genomics 11 (1), 62-76 (1991)
  MEDLINE   92112227
   PUBMED   1765386
REFERENCE   3  (bases 1 to 3969)
  AUTHORS   Hua,X.X., Enerback,S., Hudson,J., Youkhana,K. and Gimble,J.M.
  TITLE     Cloning and characterization of the promoter of the murine
            lipoprotein lipase-encoding gene: structural and functional
            analysis
  JOURNAL   Gene 107 (2), 247-258 (1991)
  MEDLINE   92084117
   PUBMED   1748295
REFERENCE   4  (bases 1 to 3969)
```

NCBI

Curated RefSeq Records: NM_, NP_

```
LOCUS       Lpl                     3969 bp    mRNA    linear   ROD 07-JAN-2002
DEFINITION  Mus musculus lipoprotein lipase (Lpl), mRNA.
ACCESSION   NM_008509
VERSION     NM_008509.1  GI:6678709
KEYWORDS    .
SOURCE      Mus musculus.
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 3969)
  AUTHORS   Kirchgessner,T.G., Svenson,K.L., Lusis,A.J. and Schotz,M.C.
  TITLE     The sequence of cDNA encoding lipoprotein lipase. A member of a
            lipase gene family
  JOURNAL   J. Biol. Chem. 262 (18), 8463-8466 (1987)
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from M60847.1.
REFERENCE   2  (bases 1 to 3969)
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from M15856.1.
            Summary: LPL encodes lipoprotein lipase, which is expressed in
            heart, muscle, and adipose tissue. LPL functions as a homodimer,
            and has the dual functions of triglyceride hydrolase and
            ligand/bridging factor for receptor-mediated lipoprotein uptake.
            Severe mutations that cause LPL deficiency result in type I
            hyperlipoproteinemia, while less extreme mutations in LPL are
            linked to many disorders of lipoprotein metabolism.
  JOURNAL   Gene 107 (2), 247-258 (1991)
  MEDLINE   92084117
   PUBMED   1748295
REFERENCE   4  (bases 1 to 3969)
```

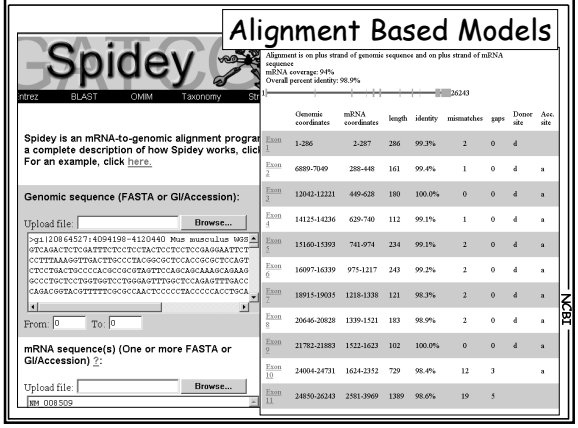
NCBI

Curated RefSeq Records: NM_, NP_

```
LOCUS       Lpl                     3969 bp    mRNA    linear   ROD 07-JAN-2002
DEFINITION  Mus musculus lipoprotein lipase (Lpl), mRNA.
ACCESSION   NM_008509
VERSION     NM_008509.1  GI:6678709
KEYWORDS    .
SOURCE      Mus musculus.
     CDS            200..1624
                    /gene="Lpl"
                    /EC_number="3.1.1.34"
                    /codon_start=1
                    /product="lipoprotein lipase"
                    /protein_id="NP_032535.1"
                    /db_xref="GI:6678710"
                    /db_xref="LocusID:16956"
                    /db_xref="MGD:96820"
                    /translation="MESKALLLVVLGVWLQSLTAFRGGVAAADAGRDFSDIESKFALR
                    TPEDTAEDTCHLIPGLADSVSNCHFNHSSKTFVVIHGWTVTGMYESWVPKLVAALYKR
                    EPDSNVIVVDWLYRAQQHYPVSAGYTKLVGNDVARFINWMEEEFKYPLDNVHLLGYSL
                    GAHAAGVAGSLTNKKVNRITGLDPAGPNFEYAEAPSRLSPDDADFVDVLHTFTRGSPG
                    RSIGIQKPVGHVDIYPNGGTFQPGCNIGEAIRVIAERGLGDVDQLVKCSHERSIHLFI
                    DSLLNEENPSKAYRCNSKEAFEKGLCLSCRKNRCNNLGYEINKVRAKRSSKMYLKTRS
                    QMPYKVFHYQVKIHFSGTENGKQHNQAFEISLYGTVAESENIPFTLPEVSTNKTYSFL
                    IYTEVDIGELLMMKLKWMSDSYFSWFDWWSSPSFVIERIEVKAGETQKKVIFCAREKV
            Severe mutations that cause LPL deficiency result in type I
            hyperlipoproteinemia, while less extreme mutations in LPL are
            linked to many disorders of lipoprotein metabolism.
  JOURNAL   Gene 107 (2), 247-258 (1991)
  MEDLINE   92084117
   PUBMED   1748295
REFERENCE   4  (bases 1 to 3969)
```

NCBI

Alignment Based Models



NCBI

Alignment Based Models

Exon 7: 18915-19035 (genomic); 1218-1338 (mRNA)

XM_134193

```
18915     TTCCCTACAGTGTTCCATTACCAAGTCAAGATTCACTTTTCTGGGACTGA
          ||||||||||||||||||||||||||||||||||||||||||||||||||
1218      TGTTCCATTACCAAGTCAAGATTCACTTTTCTGGGACTGA
          V  F  H  Y  Q  V  K  I  H  F  S  G  T  E
```

NM_008509

```
18955     GGATGGCAAGCAACACAACCAGGCCTTCGAGATTTCTCTGTACGGCACAG
          | |||||||||||||||||||||||||||||| ||||||||||||||||||
1258      GAATGGCAAGCAACACAACCAGGCCTTCGAAATTTCTCTGTACGGCACAG
            N  G  K  Q  H  N  Q  A  F  E  I  S  L  Y  G  T
```

AA change

```
19005     TGGCCGAGAGCGAGAACATTCCCTTCACCCTGTGAGTATCT
          |||||||||||||||||||||||||||||||||
1308      TGGCCGAGAGCGAGAACATTCCCTTCACCCT
          V  A  E  S  E  N  I  P  F  T  L
```

NCBI

Alignment Generated Transcripts: XM_, XP_

```
LOCUS       Lpl                     3862 bp    mRNA    linear   ROD 16-MAY-2002
DEFINITION  Mus musculus lipoprotein lipase (Lpl), mRNA.
ACCESSION   XM_134193
VERSION     XM_134193.1  GI:20864361
KEYWORDS    .
SOURCE      Mus musculus.
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 3862)
  AUTHORS   NCBI Annotation Project.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-MAY-2002) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
COMMENT     GENOME ANNOTATION REFSEQ:  This model reference sequence was
            predicted from NCBI contig NW_000244 by automated computational
            analysis using gene prediction method: BLAST, supported by mRNA
            evidence.
            Also see:
                Documentation of NCBI's Annotation Process

FEATURES             Location/Qualifiers
     source          1..3862
                     /organism="Mus musculus"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="8"
     gene            1..3862
                     /gene="Lpl"
                     /db_xref="LocusID:16956"
                     /db_xref="MGD:96820"
     CDS             199..1623
                     /gene="Lpl"
                     /codon_start=1
                     /product="lipoprotein lipase"
                     /protein_id="XP_134193.1"
                     /db_xref="GI:20864362"
```

NCBI

RefSeq Contig: NT_, NW_

```
LOCUS       NT_007933            64386157 bp    DNA     linear   CON 01-AUG-2002
DEFINITION  Homo sapiens chromosome 7 reference genomic contig.
ACCESSION   NT_007933
VERSION     NT_007933.10  GI:22050620
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chor
            Mammalia; Eutheria; Prim
REFERENCE   1  (bases 1 to 64386157)
  AUTHORS   NCBI Annotation Project.
  TITLE     Direct Submission
  JOURNAL   Submitted (31-JUL-2002)
            Information, NIH, Bethe
COMMENT     GENOME ANNOTATION REFSEQ
            genomic sequence data.
            sequence.
            On Aug 1, 2002 this sequ
            COMPLETENESS: not full
FEATURES             Location/Qualif
     source          1..64386157
                     /organism="Hom
                     /db_xref="taxon
                     /chromosome="7"
     gene            230..24397
                     /gene="WBSCR16"
                     /note="DKFZp434
                     /db_xref="Locus
     mRNA            join(230..298,3
                     10510..10683,11
                     /gene="WBSCR16"
                     /note="Derived
                     gene prediction
                     evidence inclu
                     /transcript_id=
                     /db_xref="GI:23
                     /db_xref="Locus
     CDS             join(240..298,3
```

```
LOCUS       NW_000272            27755676 bp    DNA     linear   CON 04-JUN-2002
DEFINITION  Mus musculus WGS supercontig Ma6_WIFeb01_97.
ACCESSION   NW_000272
VERSION     NW_000272.1  GI:20915425
KEYWORDS    .
SOURCE      house mouse.
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 27755676)
  AUTHORS   NCBI Annotation Project.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-MAY-2002) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
COMMENT     GENOME ANNOTATION REFSEQ:  NCBI contigs are derived from assembled
            genomic sequence data. They may include both draft and finished
            sequence.
FEATURES             Location/Qualifiers
     source          1..27755676
                     /organism="Mus musculus"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="6"
     gene            6728..13494
                     /gene="LOC243722"
                     /db_xref="InterimID:243722"
     mRNA            join(6728..6872,7967..8218,11029..11098,11202..11335,
                     11368..11503,11641..11766,12089..12225,12970..13494)
                     /gene="LOC243722"
                     /product="similar to nuclear matrix transcription factor"
                     /note="Derived by automated computational analysis using
                     gene prediction method: GenomeScan."
                     /transcript_id="XM_145115.1"
                     /db_xref="GI:20915285"
                     /db_xref="InterimID:243722"
     CDS             join(6786..6872,7967..8218,11029..11098,11202..11335,
                     11368..11503,11641..11766,12089..12225,12970..13494)
                     /gene="LOC243722"
                     /codon_start=1
```

NCBI

RefSeq Chromosomes: NC_

Organelles

Views:

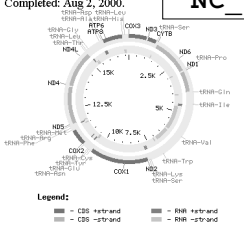
Protein View

Coding Regions

RNA Genes

Loligo bleekeri mitochondrion, complete genome

Accession: NC_002507
Total Bases Sequenced: 17211 bp
Completed: Aug 2, 2000.



Mitochondrion
Organism: Loligo bleekeri
Genetic Code: 5
Lineage: Eukaryota; Metazoa; Mollusca; Cephalopoda; Coleoidea; Teuthoida; Myopsida; Loliginidae; Loligo.

NCBI

RefSeq Chromosomes: NC_

Organelle

Views:

Protein Vie

Coding Re

RNA Gene

```
LOCUS       NC_002507            17211 bp    DNA     circular INV 02-AUG-2000
DEFINITION  Loligo bleekeri mitochondrion, complete genome.
ACCESSION   NC_002507
VERSION     NC_002507.1  GI:979062
KEYWORDS    .
SOURCE      mitochondrion Loligo b
  ORGANISM  Loligo bleekeri
            Eukaryota; Metazoa; Mo
            Myopsida; Loliginidae; Loligo.
REFERENCE   1  (bases 1 to 7998; 17061 to 17211)
  AUTHORS   Tomita,K., Ueda,T. and Watanabe,K.
  TITLE     7-Methylguanosine at the anticodon wobble position of squid
            mitochondrial tRNA(Ser)GCU: molecular basis for assignment of
            AGA/AGG codons as serine in invertebrate mitochondria
  JOURNAL   Biochim. Biophys. Acta 1399 (1), 78-82 (1998)
  MEDLINE   98390274
   PUBMED   9714749
REFERENCE   2  (bases 1 to 6; 7978 to 17211)
  AUTHORS   Sasuga,J., Yokobori,S., Kaifu,M., Ueda,T., Nishikawa,K. and
            Watanabe,K.
  TITLE     Gene contents and organization of a mitochondrial DNA segment of
            the squid Loligo bleekeri
  JOURNAL   J. Mol. Evol. 48 (6), 692-702 (1999)
  MEDLINE   99246380
   PUBMED   10229573
REFERENCE   3  (bases 1 to 17211)
  AUTHORS   Tomita,K., Yokobori Si,S., Oshima,T., Ueda,T. and Watanabe,K.
  TITLE     The Cephalopod Loligo bleekeri Mitochondrial Genome: Multiplied
            Noncoding Regions and Transposition of tRNA Genes
  JOURNAL   J. Mol. Evol. 54 (4), 486-500 (2002)
  MEDLINE   21953360
   PUBMED   11956687
REFERENCE   4  (bases 1 to 17211)
  AUTHORS   Yokobori,S.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-JUL-1999) Department of Molecular Biology, Tokyo
```

NCBI

Third Party Annotation (TPA) Database

- Annotations of existing GenBank sequences
- Allows for community annotation of genomes
- Direct submissions
  - BankIt

`tpa[Properties]`

NCBI

TPAs in Entrez

Items 1-111 of 111 — One page

1: BK000627 — Links
TPA: Homo sapiens equilibrative nucleoside transporter 4 (ENT4) gene, complete cds
gi|25418479|tpg|BK000627.1|[25418479]

2: BK000646 — Links
TPA: Homo sapiens keratin 5b (K5B) gene, complete cds
gi|25140705|tpg|BK000646.1|[25140705]

3: BK000688 — Links
TPA: Homo sapiens synovial sarcoma X breakpoint 8 protein (SSX8) mRNA, complete cds
gi|25140703|tpg|BK000688.1|[25140703]

4: BK000689 — Links
TPA: Homo sapiens synovial sarcoma X breakpoint 9 protein (SSX9) mRNA, complete cds
gi|25140701|tpg|BK000689.1|[25140701]

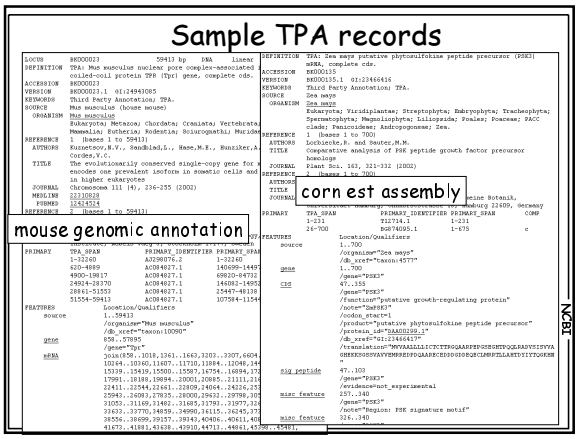
5: BK000686 — Links
TPA: Homo sapiens synovial sarcoma X breakpoint 6 protein (SSX6) mRNA, complete cds
gi|25140699|tpg|BK000686.1|[25140699]

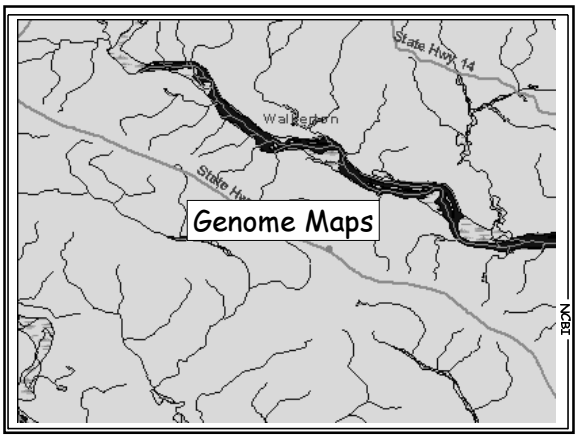
6: BK000687 — Links
TPA: Homo sapiens synovial sarcoma X breakpoint 7 protein (SSX7) mRNA, complete cds
gi|25140697|tpg|BK000687.1|[25140697]

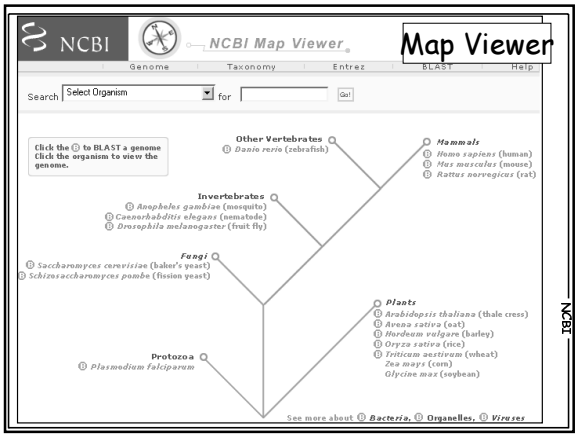
7: BK000442 — Links
TPA: Drosophila melanogaster FMRFamide receptor (FMRFaR) mRNA, complete cds
gi|25140695|tpg|BK000442.1|[25140695]

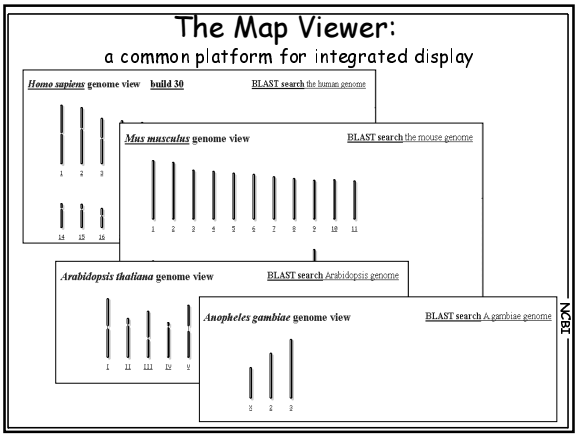
- Genome annotations
  - genes
  - transcripts
- EST assemblies

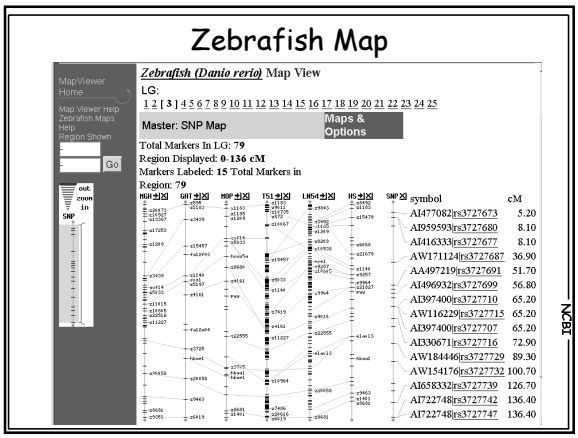
NCBI

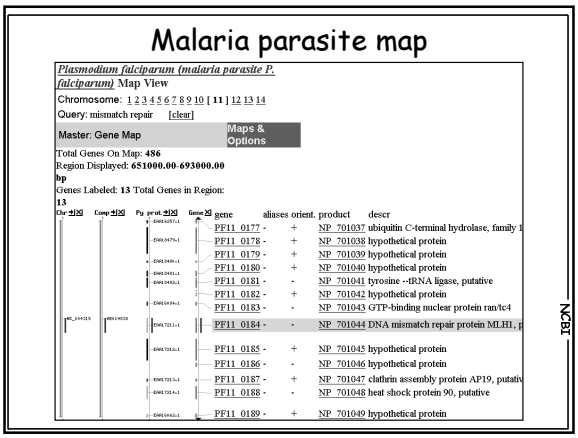

Sample TPA records
mouse genomic annotation
corn est assembly
NCBI


Genome Maps
NCBI

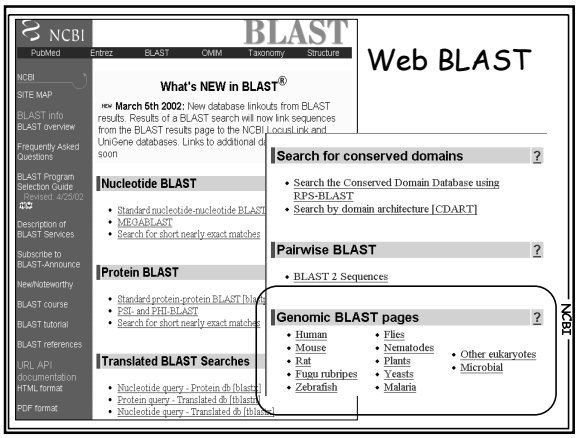

NCBI
NCBI Map Viewer
Map Viewer
Genome
Taxonomy
Entrez
BLAST
Help
Search
Select Organism
for
Go!
Click the to BLAST a genome
Click the organism to view the genome.
Other Vertebrates
Danio rerio (zebrafish)
Mammals
Homo sapiens (human)
Mus musculus (mouse)
Rattus norvegicus (rat)
Invertebrates
Anopheles gambiae (mosquito)
Caenorhabditis elegans (nematode)
Drosophila melanogaster (fruit fly)
Fungi
Saccharomyces cerevisiae (baker's yeast)
Schizosaccharomyces pombe (fission yeast)
Plants
Arabidopsis thaliana (thale cress)
Avena sativa (oat)
Hordeum vulgare (barley)
Oryza sativa (rice)
Triticum aestivum (wheat)
Zea mays (corn)
Glycine max (soybean)
Protozoa
Plasmodium falciparum
See more about Bacteria, Organelles, Viruses
NCBI


The Map Viewer:
a common platform for integrated display
Homo sapiens genome view
build 30
BLAST search the human genome
Mus musculus genome view
BLAST search the mouse genome
Arabidopsis thaliana genome view
BLAST search Arabidopsis genome
Anopheles gambiae genome view
BLAST search A.gambiae genome
NCBI


Zebrafish Map
Zebrafish (Danio rerio) Map View
LG:
Master: SNP Map
Maps & Options
Total Markers In LG: 79
Region Displayed: 0-136 cM
Markers Labeled: 15 Total Markers in Region: 79
symbol
cM
NCBI

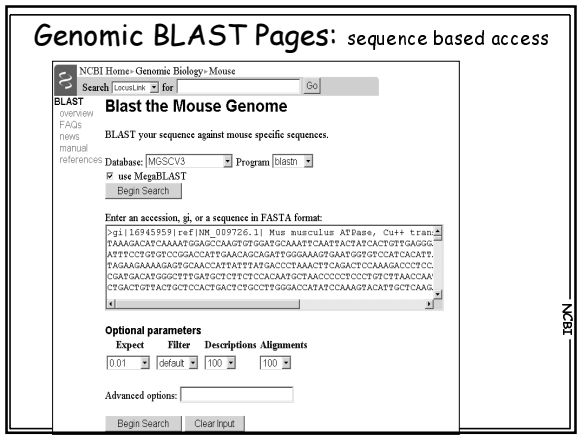

Malaria parasite map
Plasmodium falciparum (malaria parasite P. falciparum) Map View
Chromosome:
Query: mismatch repair
Master: Gene Map
Maps & Options
Total Genes On Map: 486
Region Displayed: 651000.00-693000.00 bp
Genes Labeled: 13 Total Genes in Region: 13
gene
aliases
orient.
product
descr
NCBI

Web BLAST



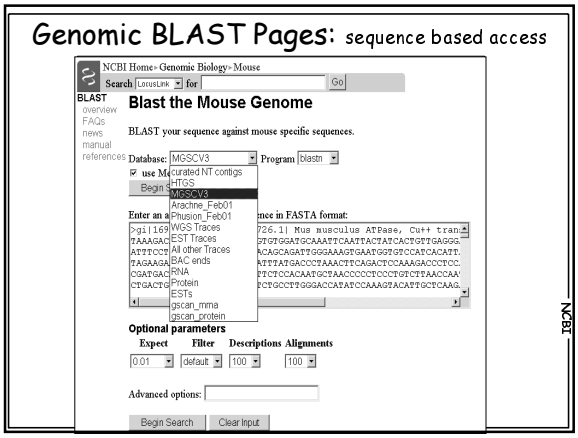
NCBI

Genomic BLAST Pages: sequence based access



NCBI

Genomic BLAST Pages: sequence based access



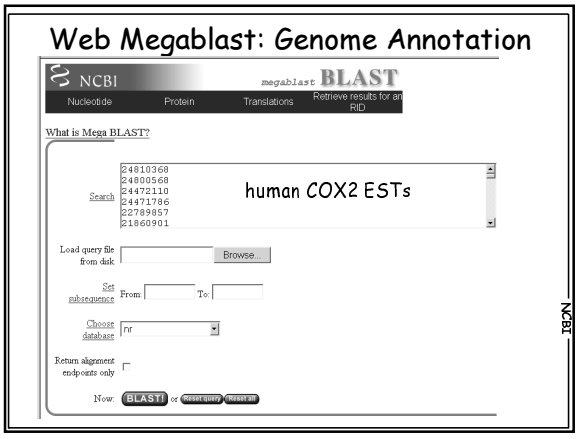
NCBI

Megablast: NCBI's Genome Annotator

- Long alignments between similar DNA sequences
- Concatenation of query sequences
- Faster than blastn
- Discontinuous Megblast allows cross-species comparisons

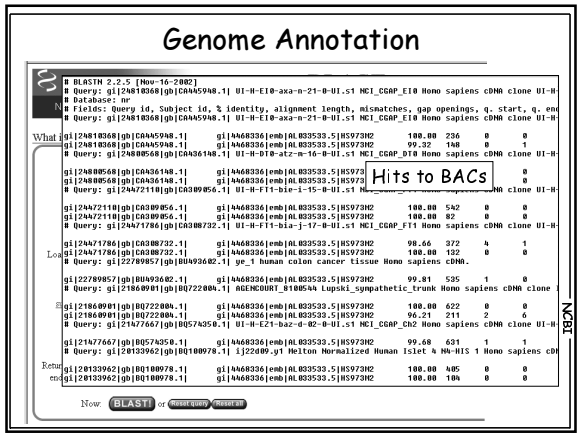
NCBI

Web Megablast: Genome Annotation



NCBI

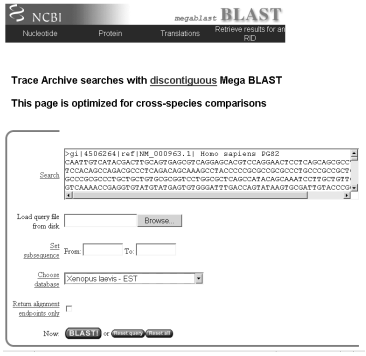
Genome Annotation



NCBI

Discontiguous MegaBlast:

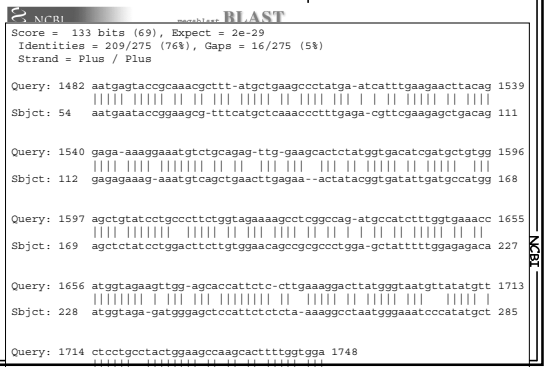
Human vs. Xenopus



NCBI

Discontiguous MegaBlast:

Human vs. Xenopus



NCBI

NCBI

3D Domains vs. Conserved Domains

3D Domains

- Protein folds are divided into compact, independent structural elements based on breakpoints between elements of secondary structure
- Derived by computation only
- Purely 3D objects that can be discontinuous in sequence

Conserved Domains

- Based on protein multiple sequence alignments created from a combination of sequence and structure data
- Curated by hand and computer (Pfam, SMART, NCBI)
- Conserved sequence elements that perform a common function

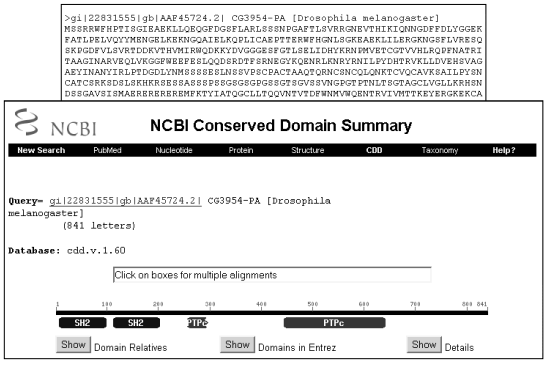
NCBI

Conserved Domain Database

- PSSMs created from
  - SMART
  - PFAM
  - LOAD
  - Clusters of Orthologous Groups (COGs)
  - CD: NCBI Curated Domains (structure and sequence)
- Searchable using RPS BLAST
- Domain Architecture Browser (CDART)

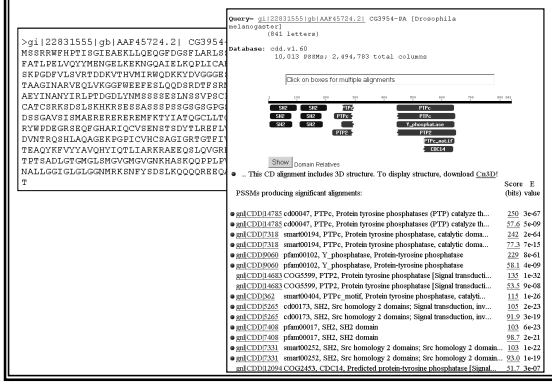
NCBI

Functional Domains



NCBI

Functional Domains: Details



NCBI

Domains in Entrez

NCBI CDs

1: cd00173 — SH2: Src homology 2 domains; Signal transduction, involved in recognition of phosphorylated tyrosine (pTyr). SH2 domains typically bind pTyr-containing ligands via two surface pockets, a pTyr and hydrophobic binding pocket, allowing proteins with SH2 domains to localize to tyrosine phosphorylated sites.

2: cd00047 — PTPc: Protein tyrosine phosphatases (PTP) catalyze the dephosphorylation of phosphotyrosine peptides; they regulate phosphotyrosine levels in signal transduction pathways. The depth of the active site cleft renders the enzyme specific for phosphorylated Tyr (pTyr) residues, instead of pSer or pThr. This family has a distinctive active site signature motif, HCSAGxGRxG. Characterized as either transmembrane, receptor-like or non-transmembrane (soluble) PTPs. Receptor-like PTP domains tend to occur in two copies in the cytoplasmic region of the transmembrane proteins, only one copy may be active.

3: smart00404 — PTPc_motif: Protein tyrosine phosphatase, catalytic domain motif;

4: smart00252 — SH2: Src homology 2 domains; Src homology 2 domains bind phosphotyrosine-containing polypeptides via 2 surface pockets. Specificity is provided via interaction with residues that are distinct from the phosphotyrosine. Only a single occurrence of a SH2 domain has been found in S. cerevisiae.

5: smart00194 — PTPc: Protein tyrosine phosphatase, catalytic domain;

6: pfam00102 — Y_phosphatase: Protein-tyrosine phosphatase.

7: pfam00017 — SH2: SH2 domain.

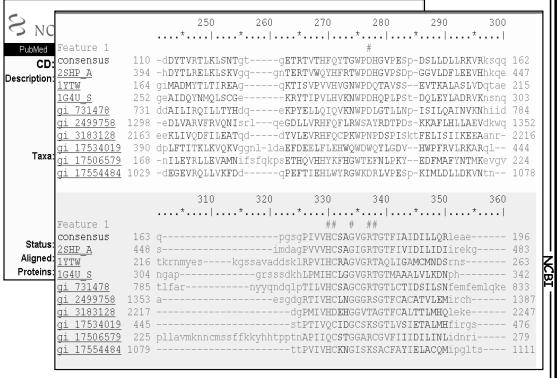
8: COG5599 — PTP2: Protein tyrosine phosphatase [Signal transduction mechanisms]

9: COG2453 — CDC14: Predicted protein-tyrosine phosphatase [Signal transduction mechanisms]

COGS

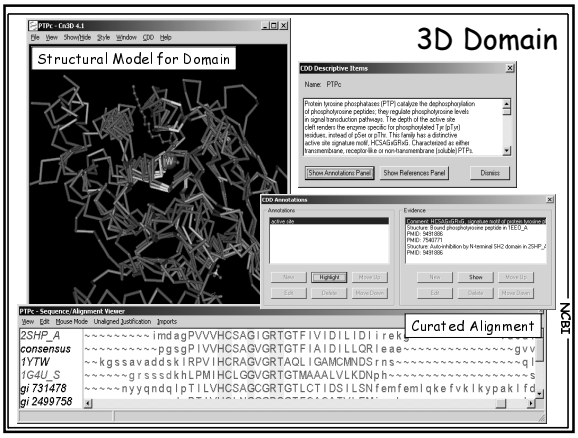
NCBI

CDD Curated Domain



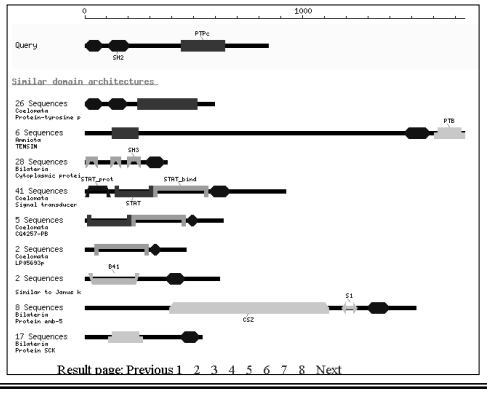
NCBI

3D Domain



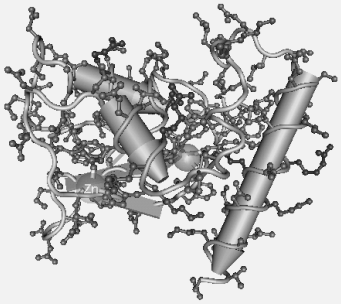
NCBI

CDART related domains



NCBI

Structures



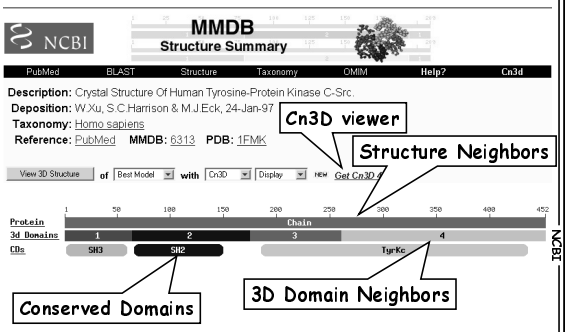
NCBI

MMDB: Molecular Modeling Data Base

- Derived from experimentally determined PDB records
- Value added to PDB records including:
  - Addition of explicit chemical graph information
  - Validation
  - Inclusion of Taxonomy, Citation, and other information
  - Conversion to ASN.1 data description language
- Structure neighbors determined by Vector Alignment Search Tool (VAST)

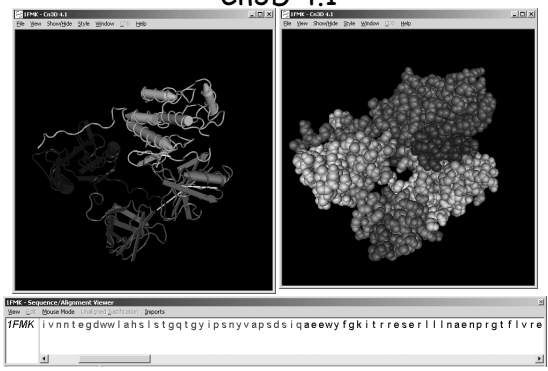
NCBI

Structure Summary



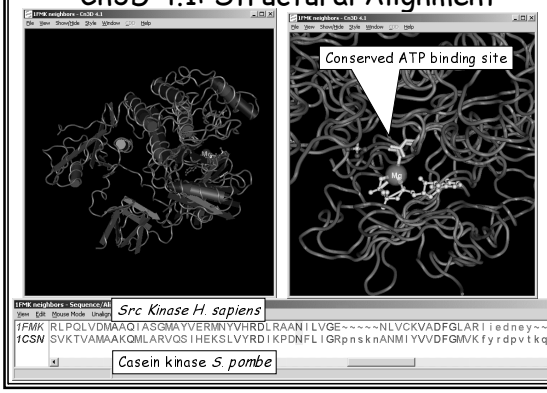
NCBI

Cn3D 4.1



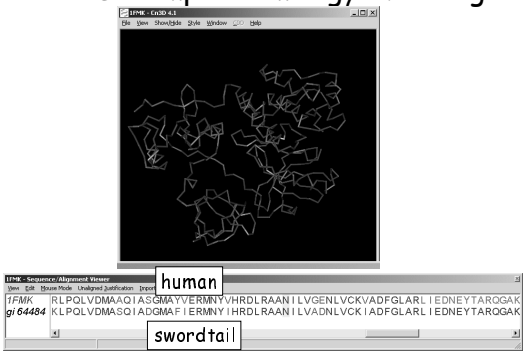
NCBI

Cn3D 4.1: Structural Alignment



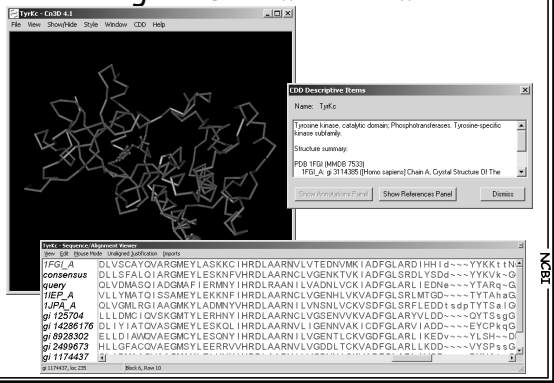
NCBI

Cn3D: Simple Homology Modeling



NCBI

Using Cn3D to model domains



NCBI

Variations

NCBI

dbSNP Data

- SNPs, Microsatellites, indels
- Clinical and Neutral Polymorphisms
- Separate submission pathway
- Resource for annotating genomes

NCBI

dbSNP Release 109

BUILD 109 STATISTICS:

Organism	Number of Submissions (ss#'s)	Number of RefSNP Clusters (rs#'s) (# validated)	Number of (ss#'s) with frequency	Number of (ss#'s) with genotype
human	4892258	3078968 (515340)	196054	32101
African malaria mosq	430291	429587 (0)		
mouse	115466	111235 (79747)		79752
zebrafish	2031	2025 (2025)		2031
chickens	1210	1204 (0)		
nematode	1065	1065 (0)		
soybean	281	278 (278)		281
thale cress	184	184 (184)		184
plasmodium	203	4 (0)		
Collared flycatcher	37	37 (37)	37	
European pied flycat	28	20 (20)	28	
lowland gorilla	4	4 (0)		
chimpanzee	2	2 (2)	2	
pig	2	2 (0)		
rat	1	1 (0)		

RefSNPs mapping onto sequence: Prion Protein

PRNP NM_000311

Contig position	dbSNP rs# (cluster id)	Heterozygosity	3D	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
4605494	rs1894281	N.D.		untranslated region				
4605720	rs1894282	N.D.		untranslated region				
4605785	rs1894283	N.D.		untranslated region				
4606037	rs1894284	N.D.		untranslated region				
4606385	rs1931969	N.D.		untranslated region				
4611080	rs2025114	N.D.		untranslated region				
4611090	rs2025113	N.D.		untranslated region				
4611091	rs2025112	N.D.		untranslated region				
4611252	rs2756270	N.D.		untranslated region				
4611391 4611392	rs3054138	N.D.		untranslated region				
4612045	rs2756269	N.D.		untranslated region				
4613317	rs2756268	N.D.		untranslated region				
4618308	rs1799990	N.D.	Yes	coding reference	A	Met [M]	1	129
		N.D.	Yes	nonsynonymous change	G	Val [V]	1	129
4618578	rs1800014	N.D.	Yes	coding reference	G	Glu [E]	1	219
		N.D.	Yes	nonsynonymous change	A	Lys [K]	1	219
4620175	rs14521	N.D.		untranslated region				

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

NCBI

Validated RefSNP

Submitter records for this RefSNP Cluster

The submission **ss2420046** has the longest flanking sequence of all cluster members and was used to instantiate sequence for **rs1799990** during BLAST analysis for the current build.

NCBI Assay ID	Handle|Submitter ID	Validation Status	Entry Date	Update Date	Build Added	Molecule Type	Sequence Orientation	Observed Alleles	Ancestral Allele	Success Rate
ss2420046	HGBASE|SNP000002608		11/07/00	01/29/01	89	Genomic	forward	A/G		unknown

Fasta sequence (Legend)

>gnl|dbSNP|rs1799990|allelePos=26|totalLen=51|ta

AGTGGTGGGG GGCCTTGGCG GCTAC
R
TGCTGGGAAG TGCCATGAGC AGGCC

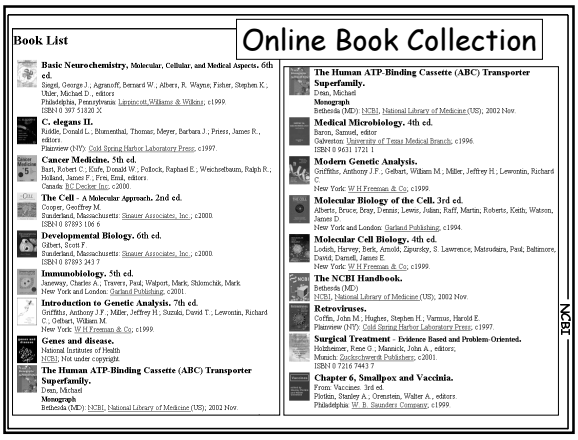
- validated by multiple, independent submissions to the refSNP cluster
- validated by frequency or genotype data
- validated by submitter confirmation
- all alleles have been observed in at least two chromosomes apiece

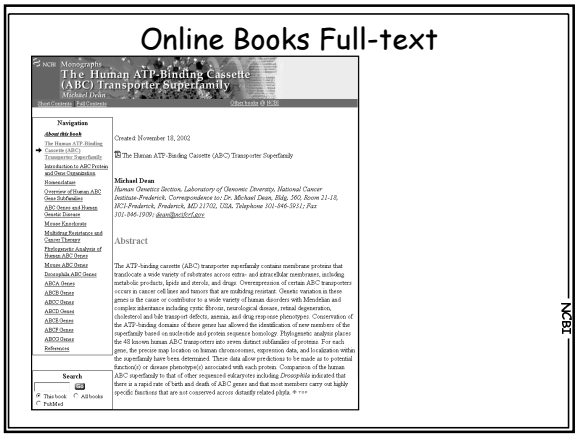
NCBI

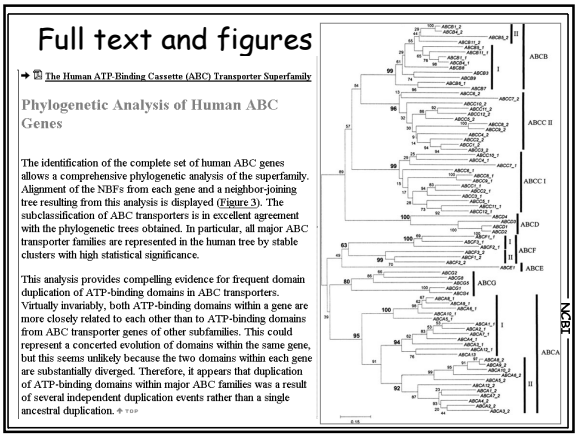
Literature

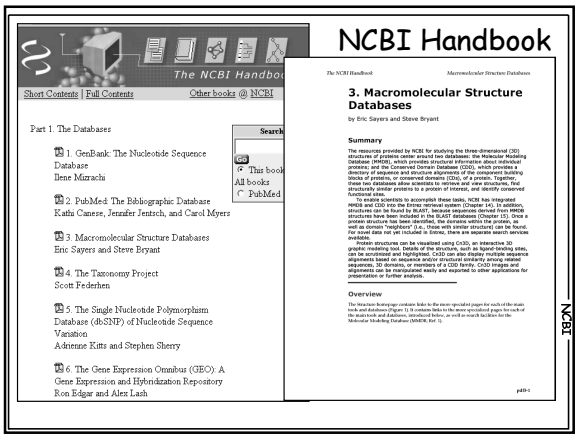
NCBI

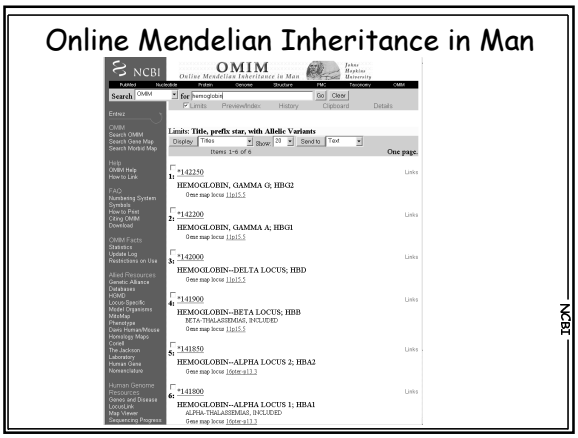
PubMed
Database resources of the National Center for Biotechnology Information: 2002 update
NCBI


Online Book Collection
Book List
NCBI

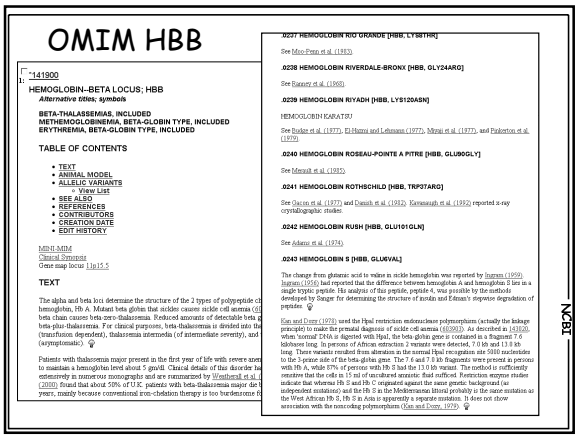

Online Books Full-text
NCBI


Full text and figures
Phylogenetic Analysis of Human ABC Genes
NCBI

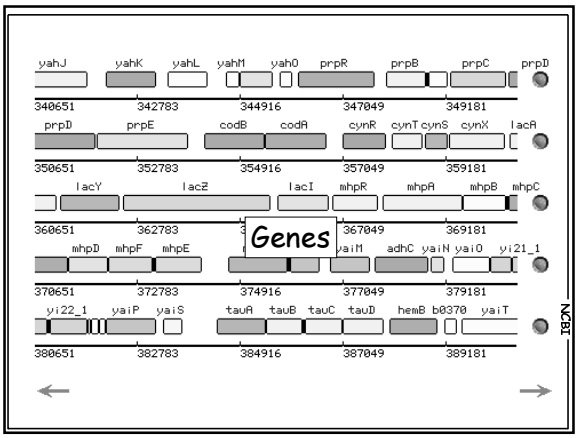

NCBI Handbook
3. Macromolecular Structure Databases
NCBI


Online Mendelian Inheritance in Man
OMIM
NCBI

OMIM HBB



NCBI



NCBI

LocusLink: Gene Based Access

A single query interface to …

- Sequences
  - RefSeqs
  - GenBank
- Maps – the Human Genome Map
  - RH
  - Cytogenetic
  - Assembled Genomic Sequence
- Genome annotations
- Entrez links

NCBI

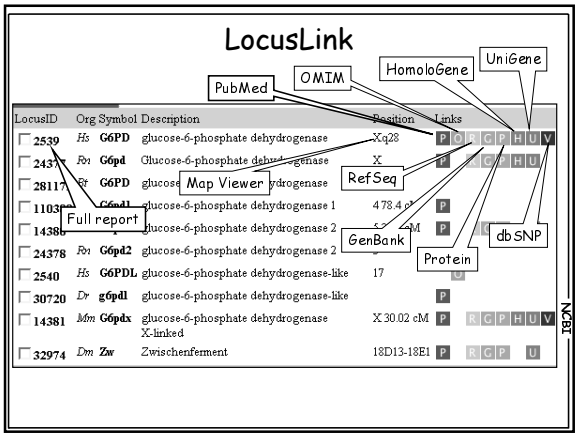
LocusLink: Selected Higher Genomes

Species		Loci
• Human	*Hs*	32,312
• Mouse	*Mm*	65,528
• Rat	*Rn*	36,521
• Cow	*Bt*	2,053
• Zebrafish	*Dr*	1,765
• C. elegans	*Ce*	22,207
• Fruit Fly	*Dm*	27,516

January, 29 2003

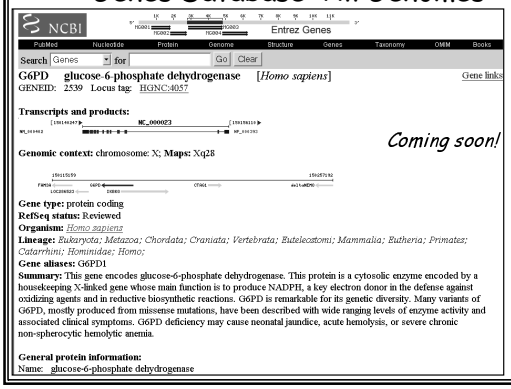
NCBI

LocusLink



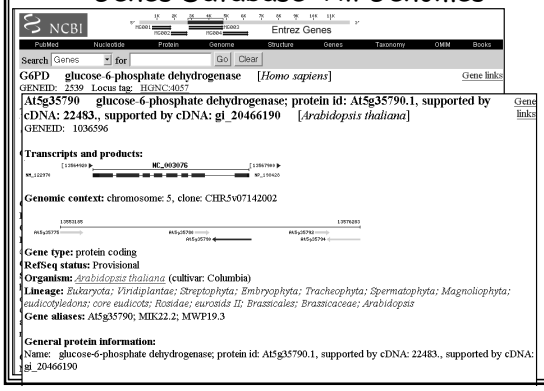
NCBI

Genes Database: All Genomes

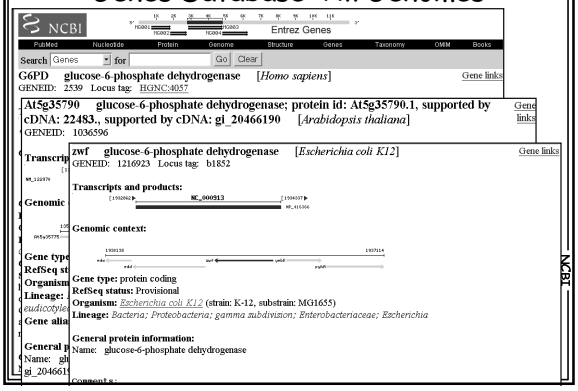


NCBI

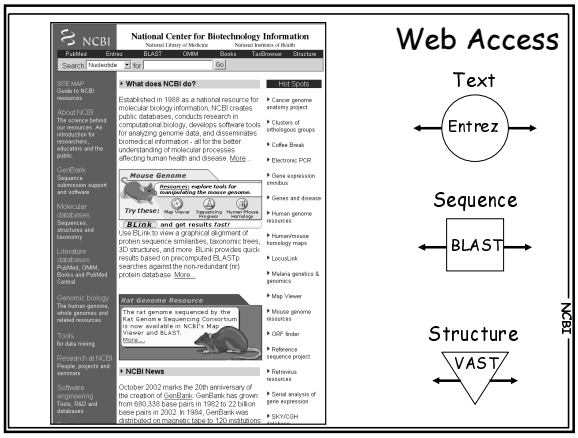
Genes Database: All Genomes



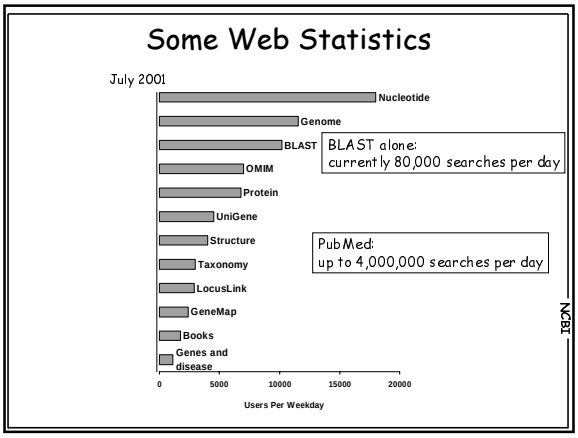
Genes Database: All Genomes



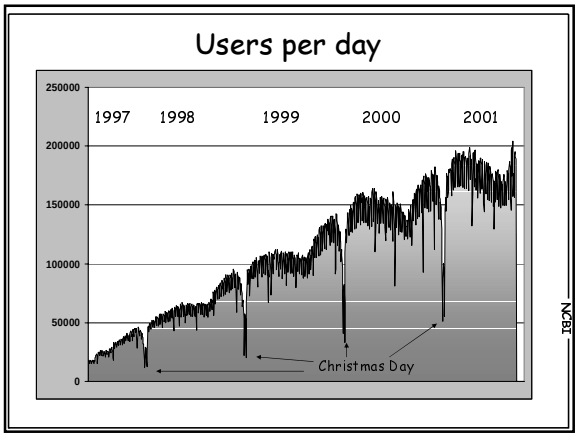
Web Access



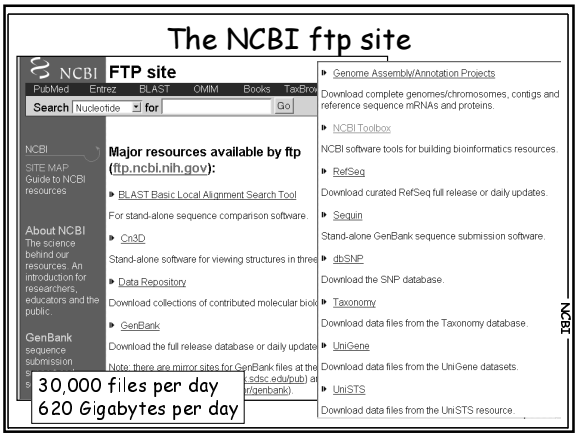
Some Web Statistics

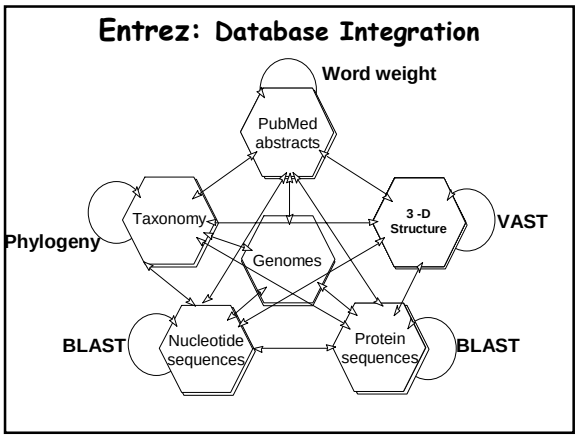


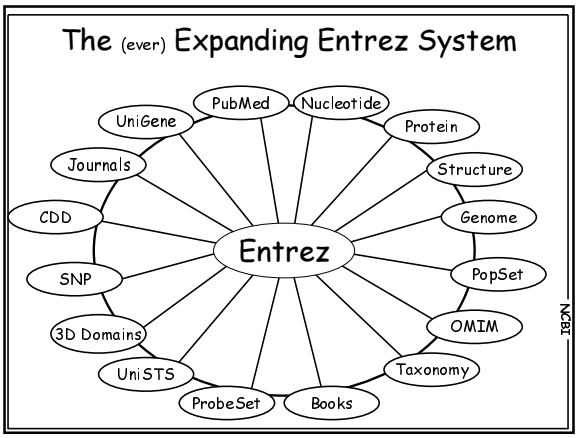
Users per day

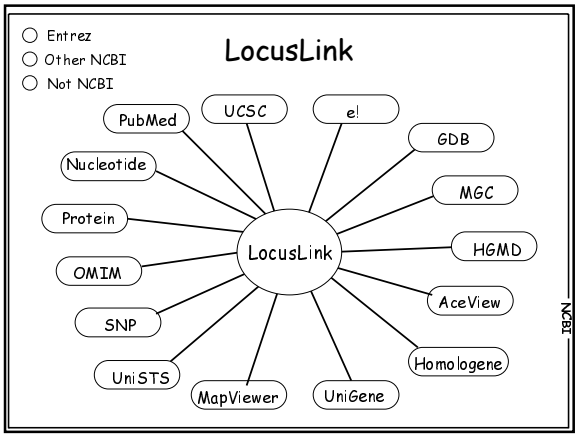


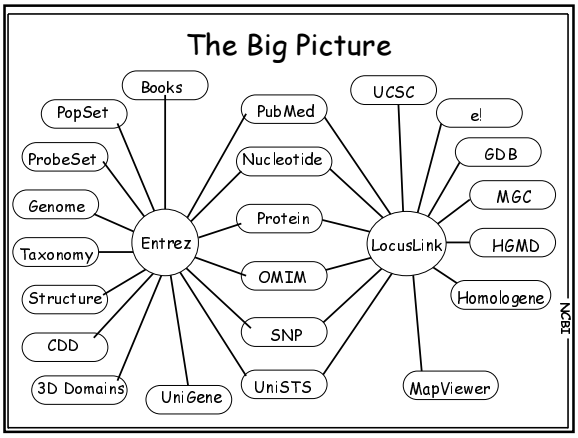
The NCBI ftp site

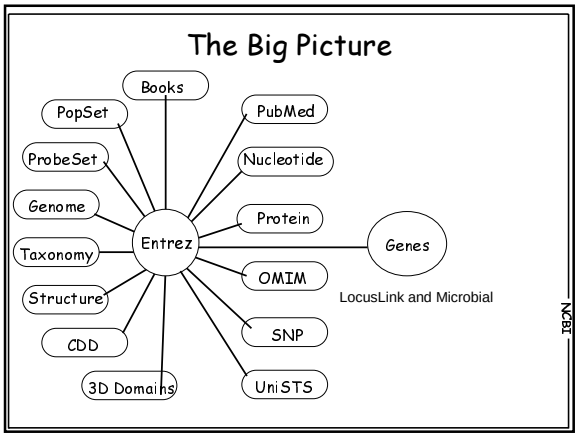


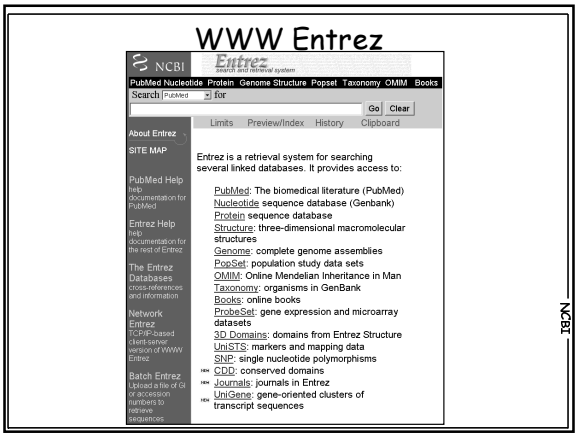

Entrez: Database Integration
Word weight
PubMed abstracts
Taxonomy
3 -D Structure
VAST
Phylogeny
Genomes
BLAST
Nucleotide sequences
Protein sequences
BLAST


The (ever) Expanding Entrez System
PubMed
Nucleotide
UniGene
Protein
Journals
Structure
CDD
Genome
Entrez
SNP
PopSet
3D Domains
OMIM
UniSTS
Taxonomy
ProbeSet
Books
NCBI


Entrez
Other NCBI
Not NCBI
LocusLink
PubMed
UCSC
e!
GDB
Nucleotide
MGC
Protein
LocusLink
HGMD
OMIM
AceView
SNP
Homologene
UniSTS
MapViewer
UniGene
NCBI


The Big Picture
Books
UCSC
PopSet
PubMed
e!
ProbeSet
Nucleotide
GDB
Genome
Protein
MGC
Taxonomy
Entrez
LocusLink
HGMD
OMIM
Structure
Homologene
SNP
CDD
3D Domains
UniGene
UniSTS
MapViewer
NCBI


The Big Picture
Books
PopSet
PubMed
ProbeSet
Nucleotide
Genome
Protein
Entrez
Genes
Taxonomy
OMIM
LocusLink and Microbial
Structure
SNP
CDD
3D Domains
UniSTS
NCBI

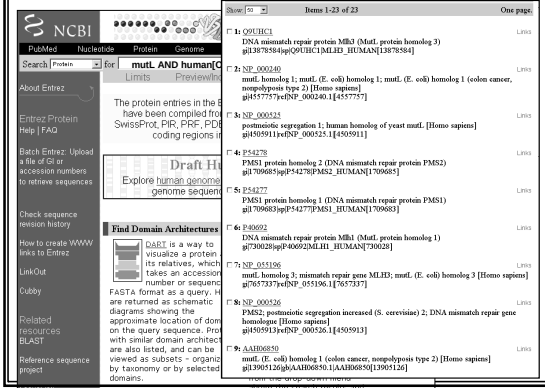

WWW Entrez
NCBI
Entrez
PubMed Nucleotide Protein Genome Structure Popset Taxonomy OMIM Books
Search PubMed for
Go Clear
Limits Preview/Index History Clipboard
About Entrez
SITE MAP
PubMed Help
Entrez Help
The Entrez Databases
Network Entrez
Batch Entrez
Entrez is a retrieval system for searching several linked databases. It provides access to:
PubMed: The biomedical literature (PubMed)
Nucleotide sequence database (Genbank)
Protein sequence database
Structure: three-dimensional macromolecular structures
Genome: complete genome assemblies
PopSet: population study data sets
OMIM: Online Mendelian Inheritance in Man
Taxonomy: organisms in GenBank
Books: online books
ProbeSet: gene expression and microarray datasets
3D Domains: domains from Entrez Structure
UniSTS: markers and mapping data
SNP: single nucleotide polymorphisms
CDD: conserved domains
Journals: journals in Entrez
UniGene: gene-oriented clusters of transcript sequences
NCBI

Entrez Databases

PubMed	Biomedical literature
Books	Online textbooks
Nucleotide	GenBank, EMBL, DDBJ, RefSeq, PDB
Protein	[GenBank, EMBL, DDBJ], RefSeq, SWISS-PROT, PIR, PRF, PDB
Genome	Complete genomes
Taxonomy	Organisms in NCBI sequence databases
Structure	MMDB: experimental 3D structures
Domains	CDD: conserved protein domains
3D Domains	Compact 3D protein domains in MMDB
OMIM	Online Mendelian Inheritance in Man
SNP	Single nucleotide polymorphisms (and more!)
UniSTS	Sequence Tagged Site markers
ProbeSet	Gene expression and microarray datasets
PopSet	Population study datasets
UniGene	Nonredundant gene oriented clusters of expressed sequences

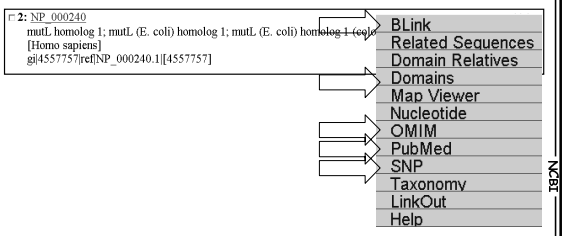
NCBI

Human DNA Mismatch Repair Protein



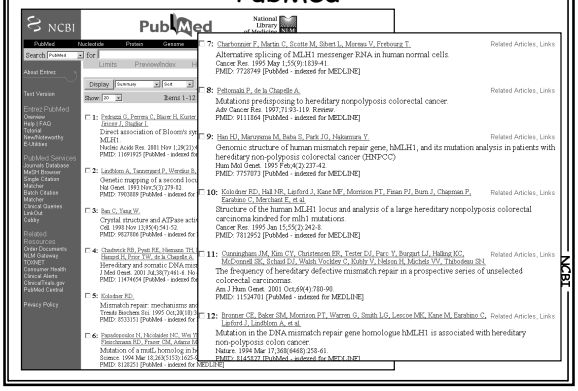
NCBI

Links and Neighbors Human mutL



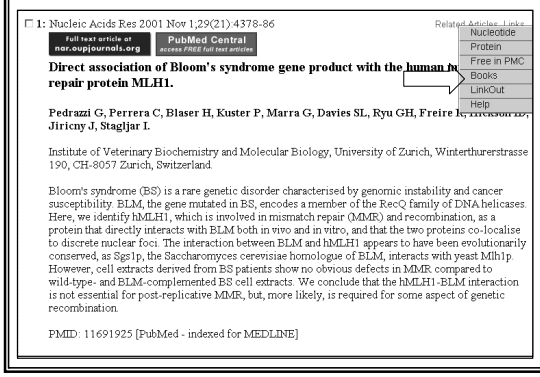
NCBI

PubMed



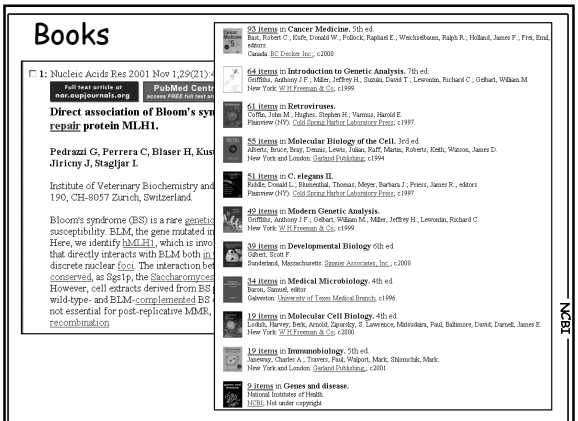
NCBI

Abstract: Full Text, Books



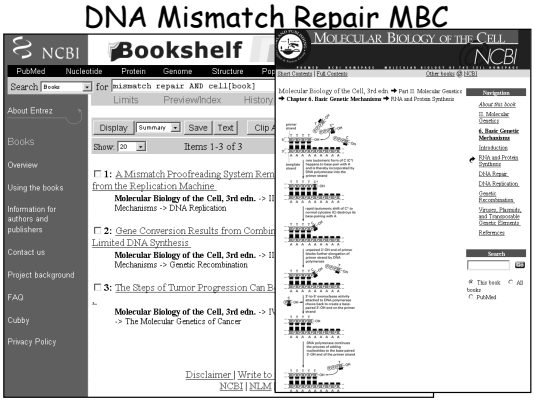
NCBI

Books



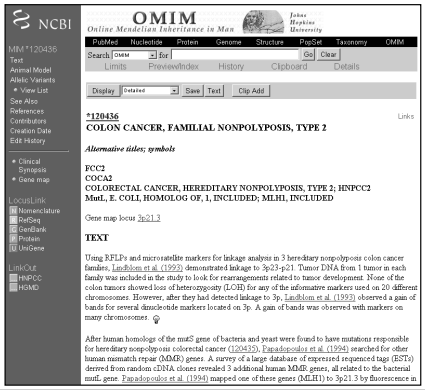
NCBI

DNA Mismatch Repair MBC



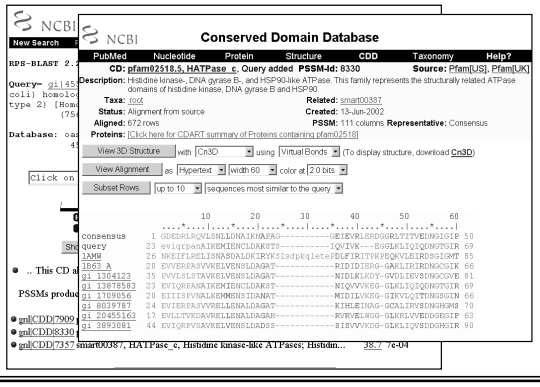
NCBI

OMIM



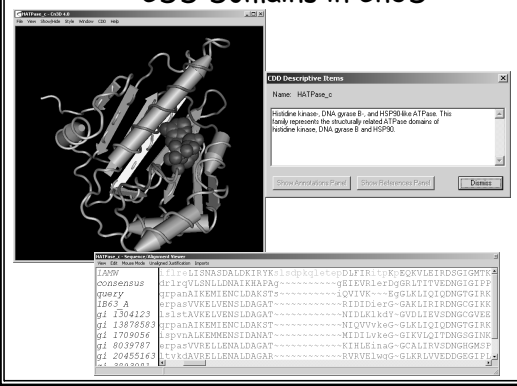
NCBI

Conserved Domains



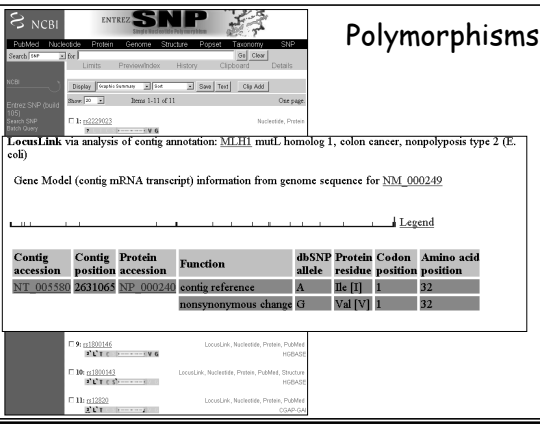
NCBI

CDD Domains in Cn3D



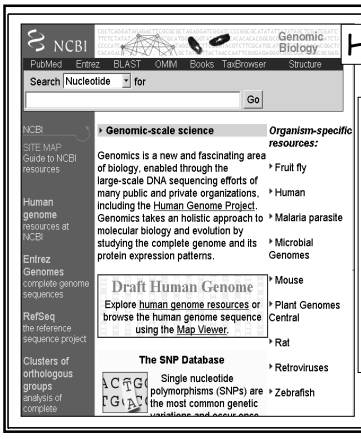
NCBI

Polymorphisms



NCBI

Higher Genomes

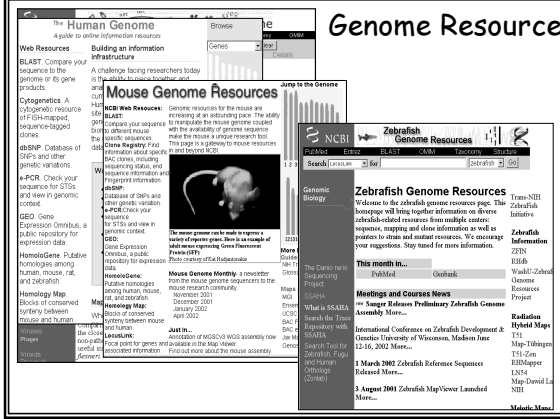


Complete (or nearly)

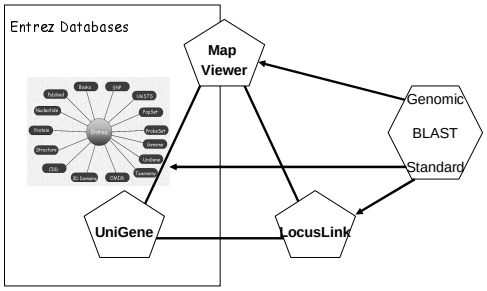
- Human
- Mouse
- Fruit Fly
- Malaria Mosquito
- Fugu
- *C. elegans*
- *P. falciparum*
- *Arabidopsis*
- Rice

NCBI

Genome Resources

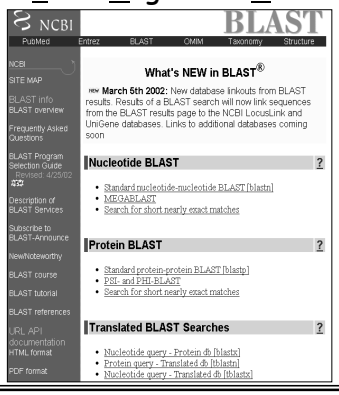


Higher Genomes Access

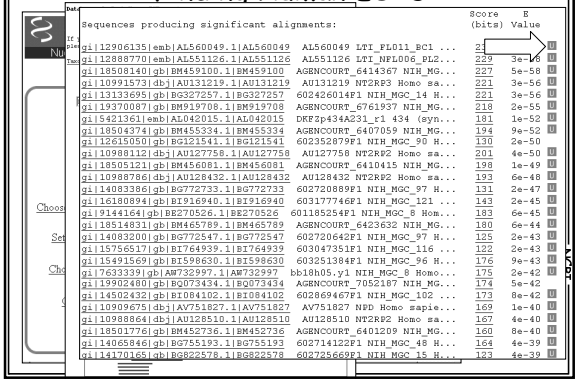


Genome Resources with MLH1

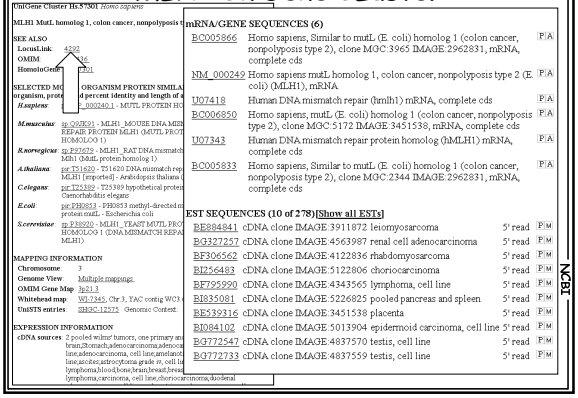
Basic Local Alignment Search Tool

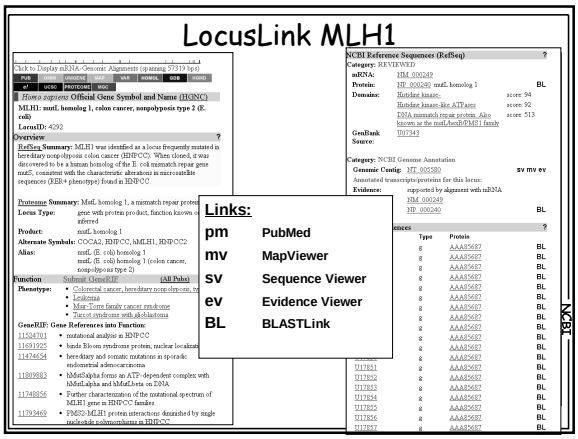


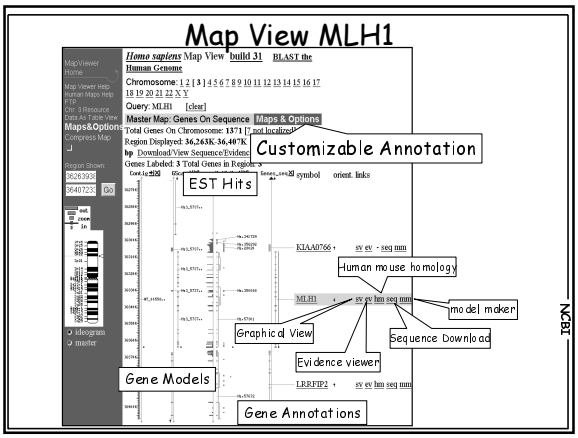
Finding Human ESTs

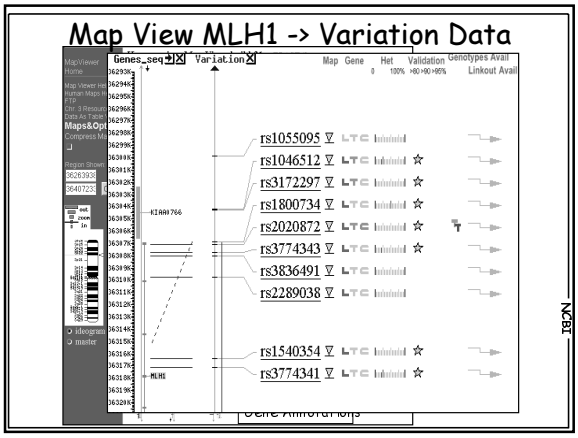


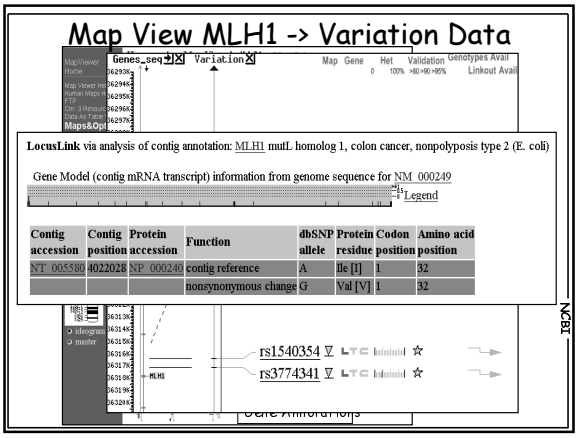
MLH1 UniGene CLuster

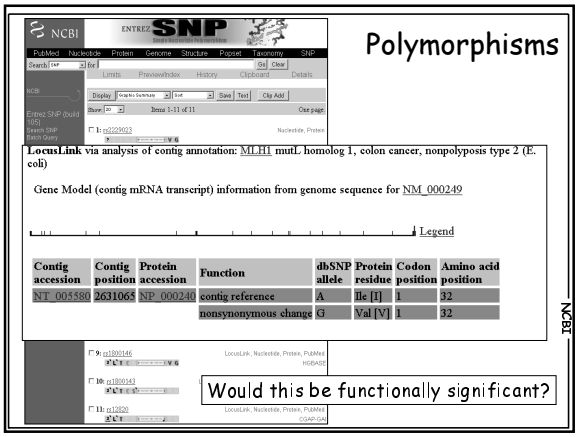


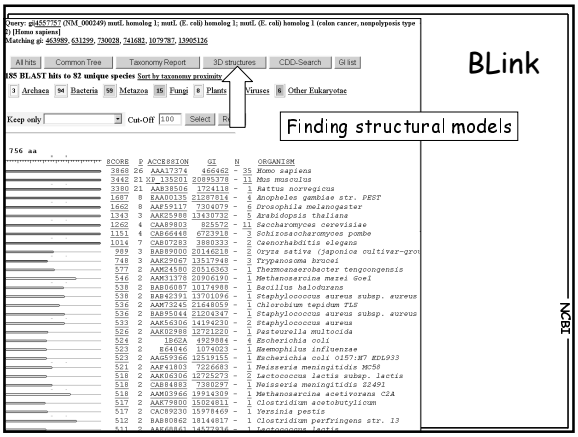

LocusLink MLH1
Links:
pm PubMed
mv MapViewer
sv Sequence Viewer
ev Evidence Viewer
BL BLASTLink
NCBI


Map View MLH1
Customizable Annotation
EST Hits
Human mouse homology
model maker
Graphical View
Sequence Download
Evidence viewer
Gene Models
Gene Annotations
NCBI

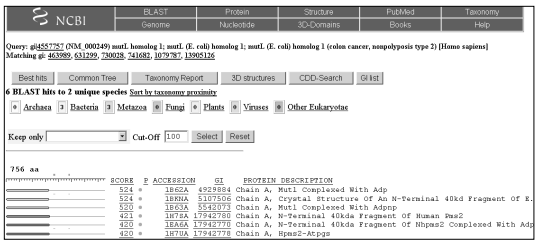

Map View MLH1 -> Variation Data
rs1055095
rs1046512
rs3172297
rs1800734
rs2020872
rs3774343
rs3836491
rs2289038
rs1540354
rs3774341
NCBI


Map View MLH1 -> Variation Data
LocusLink via analysis of contig annotation: MLH1 mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
Gene Model (contig mRNA transcript) information from genome sequence for NM_000249
Legend
Contig accession | Contig position | Protein accession | Function | dbSNP allele | Protein residue | Codon position | Amino acid position
NT_005580 | 4022028 | NP_000240 | contig reference | A | Ile [I] | 1 | 32
nonsynonymous change | G | Val [V] | 1 | 32
rs1540354
rs3774341
NCBI


Polymorphisms
LocusLink via analysis of contig annotation: MLH1 mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
Gene Model (contig mRNA transcript) information from genome sequence for NM_000249
Legend
Contig accession | Contig position | Protein accession | Function | dbSNP allele | Protein residue | Codon position | Amino acid position
NT_005580 | 2631065 | NP_000240 | contig reference | A | Ile [I] | 1 | 32
nonsynonymous change | G | Val [V] | 1 | 32
Would this be functionally significant?
NCBI

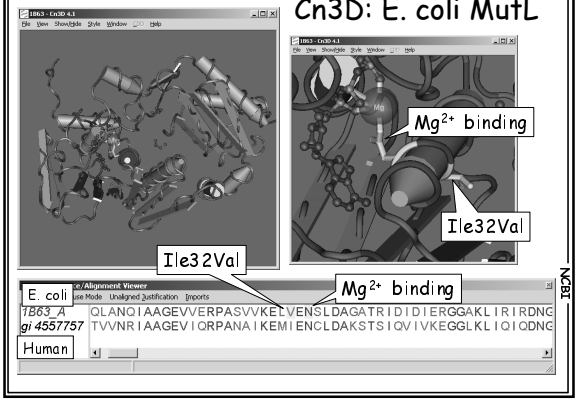

BLink
Finding structural models
NCBI

BLink: Finding Modeling Template



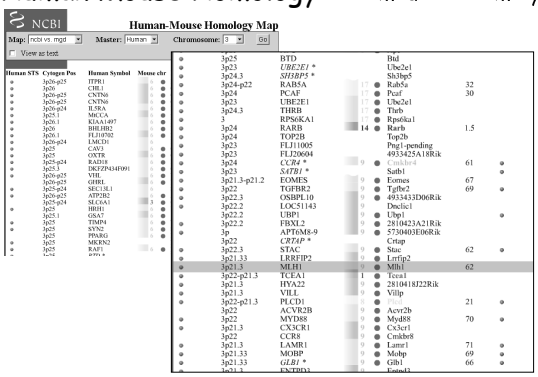
NCBI

Cn3D: E. coli MutL



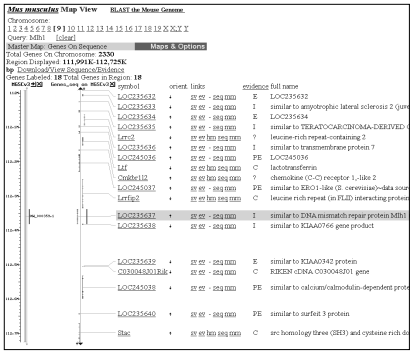
NCBI

Human Mouse Homology: the mouse assembly



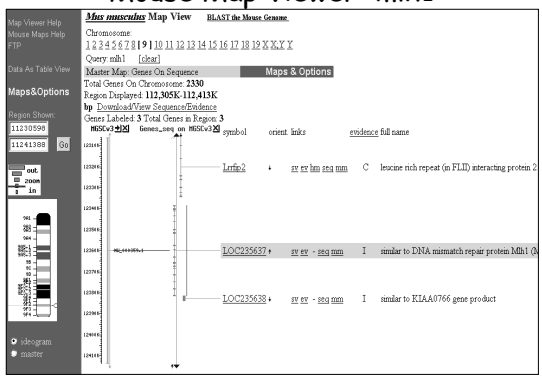
NCBI

The Mouse Assembly



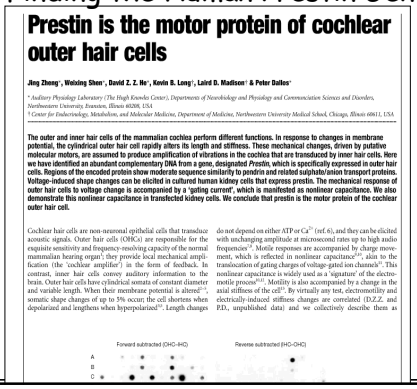
NCBI

Mouse Map Viewer: mlh1



NCBI

Finding the Human Prestin Gene



Prestin is the motor protein of cochlear outer hair cells

Jing Zheng[1], Weixing Shen[1], David Z. Z. He[1], Kevin B. Long[2], Laird D. Madison[2] & Peter Dallos[1]

[1] Auditory Physiology Laboratory (The Hugh Knowles Center), Departments of Neurobiology and Physiology and Communication Sciences and Disorders, Northwestern University, Evanston, Illinois 60208, USA
[2] Center for Endocrinology, Metabolism, and Molecular Medicine, Department of Medicine, Northwestern University Medical School, Chicago, Illinois 60611, USA

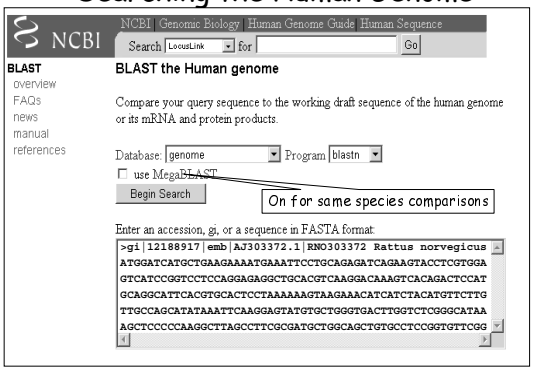
The outer and inner hair cells of the mammalian cochlea perform different functions. In response to changes in membrane potential, the cylindrical outer hair cell rapidly alters its length and stiffness. These mechanical changes, driven by putative molecular motors, are assumed to produce amplification of vibrations in the cochlea that are transduced by inner hair cells. Here we have identified an abundant complementary DNA from a gene, designated *Prestin*, which is specifically expressed in outer hair cells. Regions of the encoded protein show moderate sequence similarity to pendrin and related sulphate/anion transport proteins. Voltage-induced shape changes can be elicited in cultured human kidney cells that express prestin. The mechanical response of outer hair cells to voltage change is accompanied by a 'gating current', which is manifested as nonlinear capacitance. We also demonstrate this nonlinear capacitance in transfected kidney cells. We conclude that prestin is the motor protein of the cochlear outer hair cell.

Cochlear hair cells are non-neuronal epithelial cells that transduce acoustic signals. Outer hair cells (OHCs) are responsible for the exquisite sensitivity and frequency-resolving capacity of the normal mammalian hearing organ[1]; they provide local mechanical amplification (the 'cochlear amplifier') in the form of feedback. In contrast, inner hair cells convey auditory information to the brain. Outer hair cells have cylindrical somatic of constant diameter and variable length. When their membrane potential is altered[2–5], somatic shape changes of up to 5% occur; the cell shortens when depolarized and lengthens when hyperpolarized[6]. Length changes do not depend on either ATP or Ca^{2+} (ref. 6), and they can be elicited with unchanging amplitude at microsecond rates up to high audio frequencies[7,8]. Motile responses are accompanied by charge movement, which is reflected in nonlinear capacitance[2,9], akin to the translocation of gating charges of voltage-gated ion channels[10]. This nonlinear capacitance is widely used as a 'signature' of the electromotile process[2,9]. Motility is also accompanied by a change in the axial stiffness of the cell[11]. By virtually any test, electromotility and electrically-induced stiffness changes are correlated (D.Z.Z. and P.D., unpublished data) and we collectively describe them as

Forward subtracted (OHC-IHC)
Reverse subtracted (IHC-OHC)

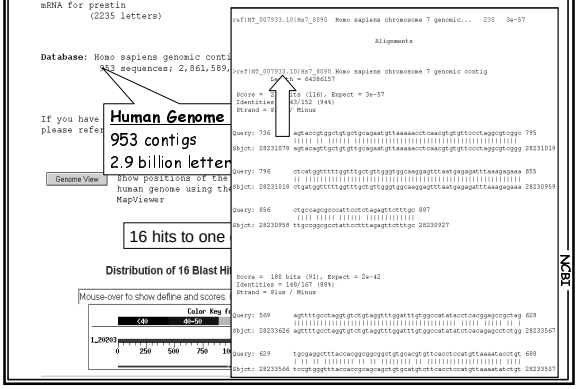
NCBI

Searching the Human Genome



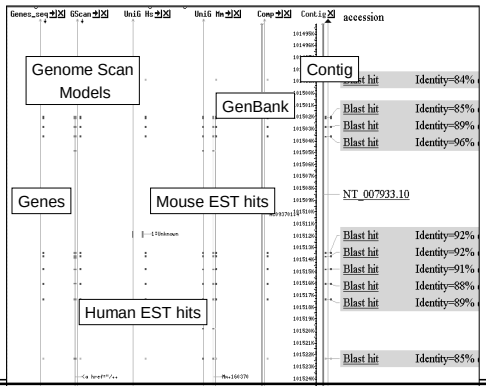
NCBI

BLAST Results



NCBI

Genomic Context of BLAST Hits



NCBI

Service Addresses

•*General Help*	`info@ncbi.nlm.nih.gov`
•*BLAST*	`blast-help@ncbi.nlm.nih.gov`

NCBI

Live Web Demos

- Finding a gene by …
  - Name (Albumin)
  - Database ID (AI863603)
  - BLAST mouse NPHP4

NCBI