Gene finding and Genome annotation

Manfred Zorn

BerkeleyPGA

Bioinformatics Tools for Comparative Analysis

May 14, 2003

What is a Gene?

- Definition: An inheritable trait associated with a region of DNA that codes for a polypeptide chain or specifies an RNA molecule which in turn have an influence on some characteristic phenotype of the organism.

Abstract concept that describes a complex phenomenon

What is Annotation?

- Definition: Extraction, definition, and interpretation of features on the genome sequence derived by integrating computational tools and biological knowledge.

Identifiable features in the sequence

How does an annotation differ from a gene?

- Many annotations describe features that constitute a gene.
- Other annotations may not always directly correspond in this way, e.g., an STS, or sequence overlap

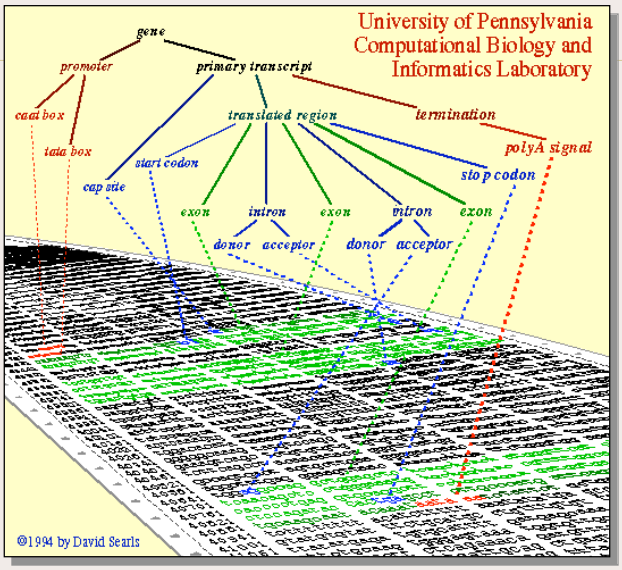
DNA Analysis

- Heuristics
- Statistics
- Artistics

DNA Analysis

- Find the genes
  - Heuristic signals
  - Inherent features
  - Intelligent methods
- Characterize each gene
  - Compare with other genes
  - Find functional components
  - Predict features

What is a Gene?

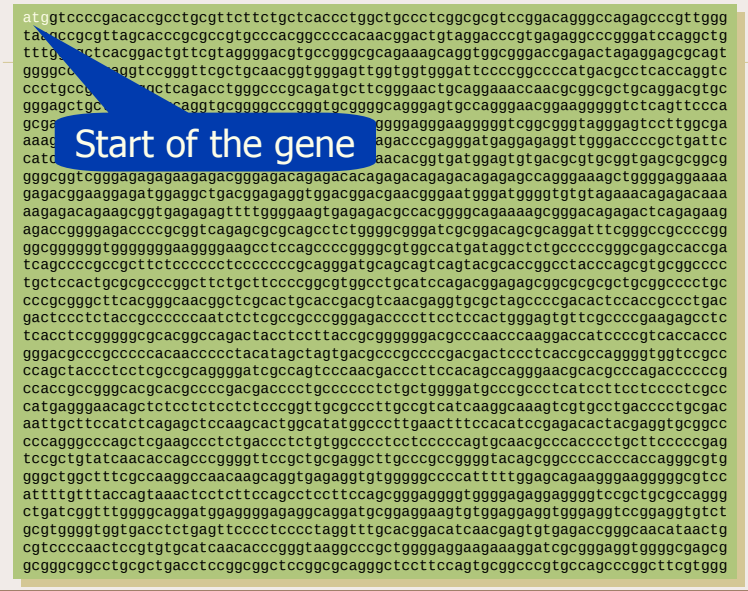


Heuristic Signals

- DNA contains various recognition sites for internal machinery

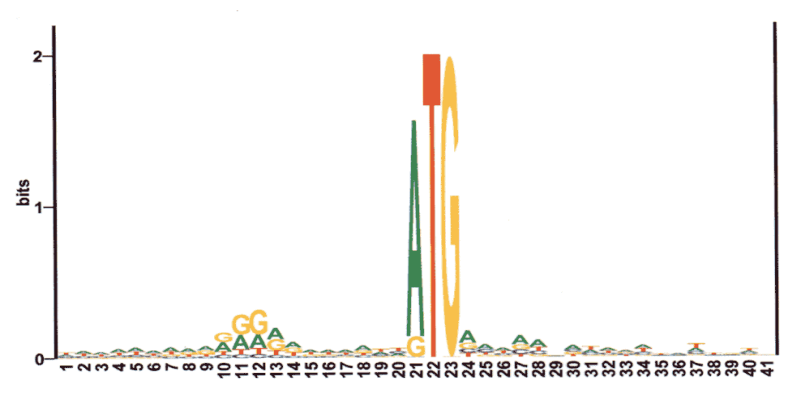
- Promoter signals
- Transcription start signals
- Start Codon
- Exon, Intron boundaries
- Transcription termination signals

Heuristic Signals

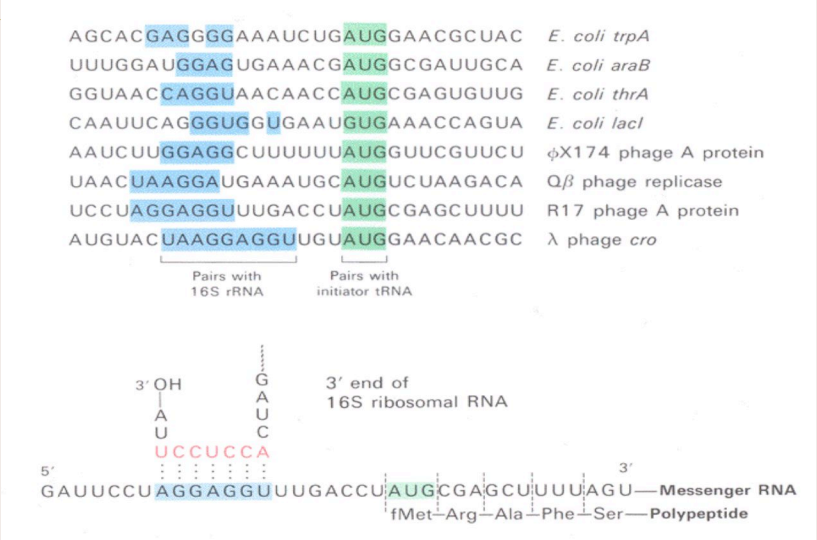


Heuristic Signals

Start Codon



Initiation

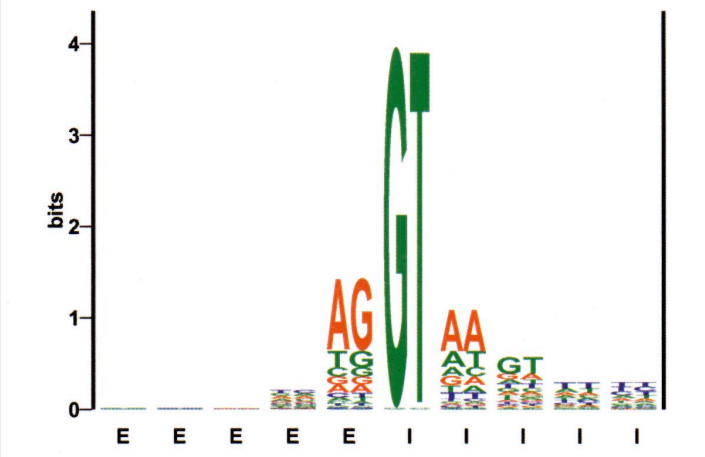


Inherent Features

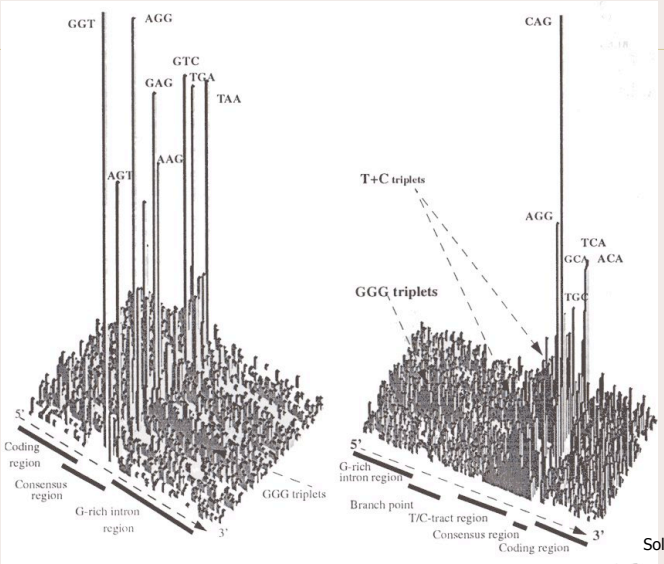
- DNA exhibits certain biases that can be exploited to locate coding regions

- Uneven distribution of bases
- Codon bias
- CpG islands
- In-phase words
- Encoded amino acid sequence
- Imperfect periodicity
- Other global patterns

Donor Splice Site



Inherent Features



Solovyev, 1994

Intelligent Methods

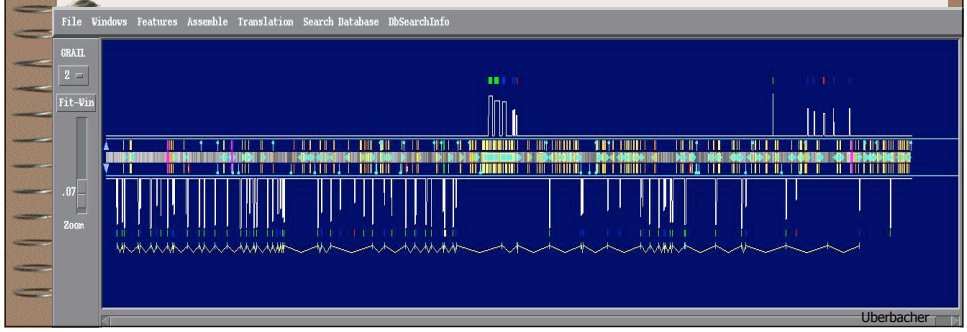
- Pattern recognition methods weigh inputs and predict gene location
  - Content-based methods
  - Site-based methods
  - Comparative methods
- Neural Networks
- Hidden Markov Models
- Stochastic Context-Free Grammar

GRAIL *Uberbacher, Mural*

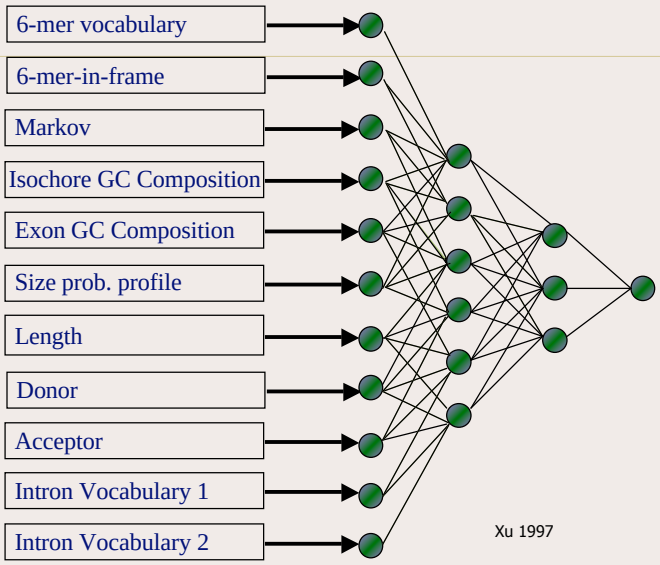
- GRAIL 1
  - Neural network with fixed window length (100 bases)
- GRAIL 1a
  - GRAIL 1 + adjacent information
- GRAIL 2
  - Variable length window, contextual information
- GRAIL-EXP
  - Comparison with partial and complete gene sequences

Analyzing Complex Multi-Gene Regions

- Errors in exon prediction and splice site boundaries
- Gene boundaries uncertain
- Genes can be on both strands

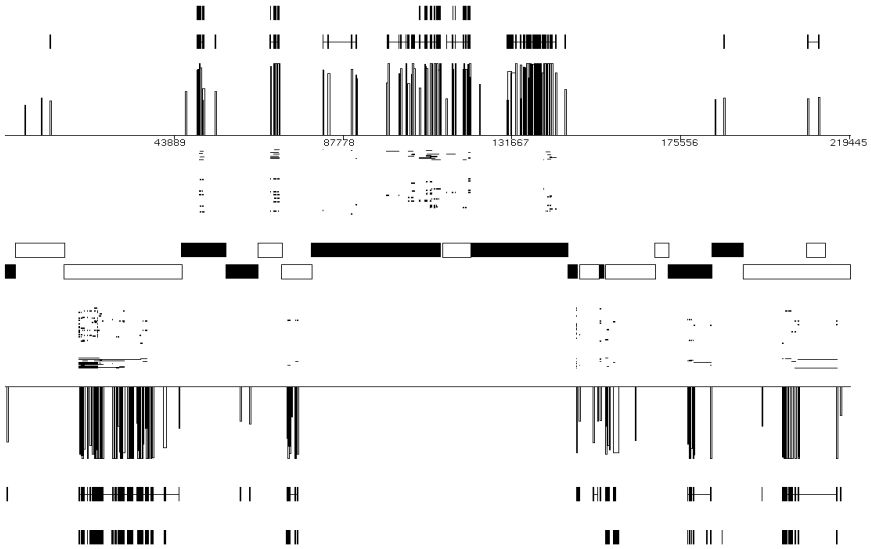


Neural networks



Xu 1997

Grail-EXP



FGENEH/FGENES *Solovyev*

- Looks at several structural features
  - Splice donor/acceptor sites
  - Putative coding regions
  - Intronic regions
- *Linear discriminant analysis* to split exon / non-exon classes
- Dynamic programming to assemble best gene structure

MZEF *Zhang*

- *Quadratic discriminant analysis*
  - Exon length
  - Exon-intron transitions
  - Splice sites
  - Branch sites
  - Exon, strand, frame scores
- Detects internal exons
- No information about gene structure

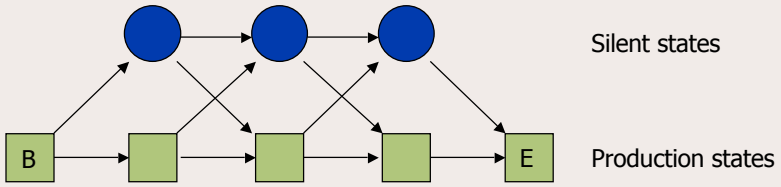
GENSCAN *Burge, Karlin*

- Probabilistic model of sequence composition and gene structure
  - Looks for gene structure descriptions that are consistent with the query sequence to assign probability that sequence stretch is exon, …
  - Best ---> optimal
  - But generates also suboptimal exons

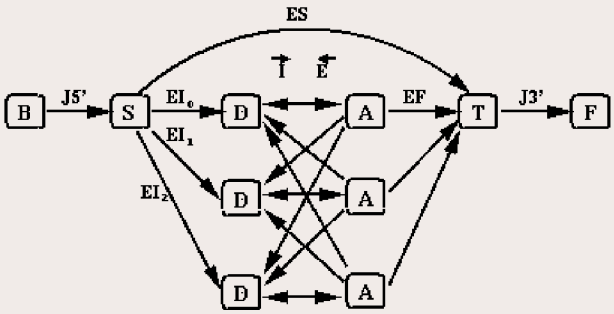
PROCRUSTES *Gelfand*

- Forces sequence into target structure
  - Requires putative gene product
  - Stretches/shortens sequence to fit into model

Hidden Markov Models



GENIE *Kulp, Reese, Haussler*



Strategies

- Select by correlation coefficient
- Select by review paper
- Select by recommendation
- Use them all

Drawbacks

- Most programs are “trained” on existing data
- It’s awfully hard to find new things this way!
    - NTT
    - IPW

Internet Resources

Banbury Cross	http://igs-server.cnrs-mrs.fr/igs/banbury
FGENEH	http://genomic.sanger.ac.uk/gf/gf.shtml
GeneID	http://www1.imim.es/geneid.html
GeneMachine	http://genome.nhgri.nih.gov/genemachine
GENSCAN	http://genes.mit.edu/GENSCAN.html
Genotator	http://www.fruitfly.org/_nomi/genotator/
GRAIL	http://compbio.ornl.gov/tools/index.shtml
GRAIL-EXP	http://compbio.ornl.gov/grailexp
MZEF	http://www.cshl.org/genefinder
PROCRUSTES	http://www-hto.usc.edu/software/procrustes
RepeatMasker	http://ftp.genome.washington.edu/RM/RepeatMasker.html
HMMgene	http://www.cbs.dtu.dk/services/HMMgene
Chapter 10	http://www.wiley.com/legacy/products/subject/life/bioinformatics/chapterlinks.html

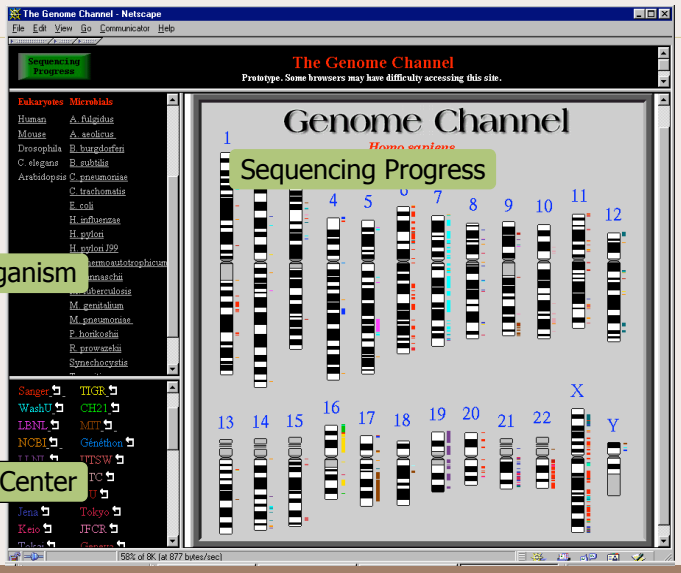
Characterize a Gene

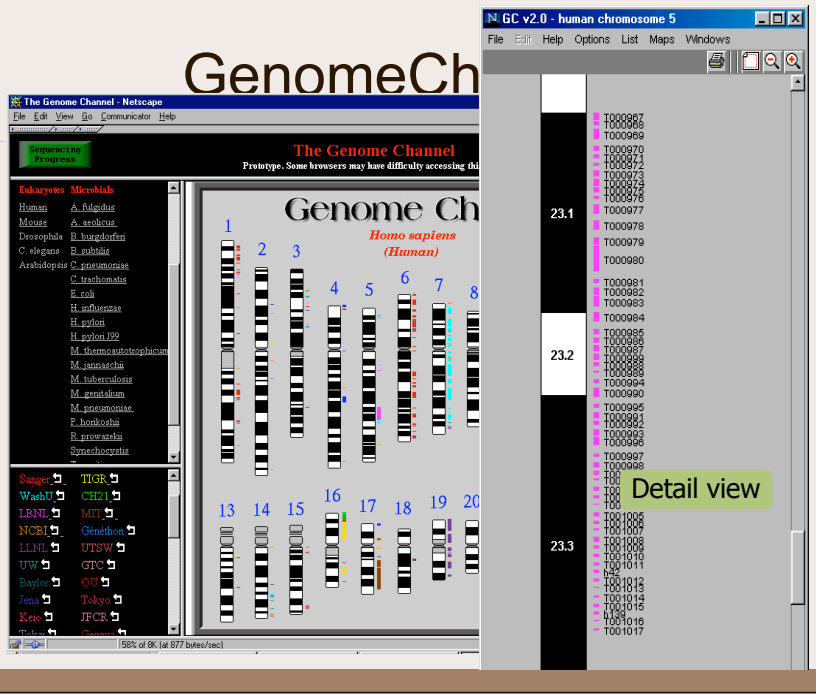
Collect clues for potential function

- Comparison with other known genes, proteins
- Predict secondary structure
- Fold classification

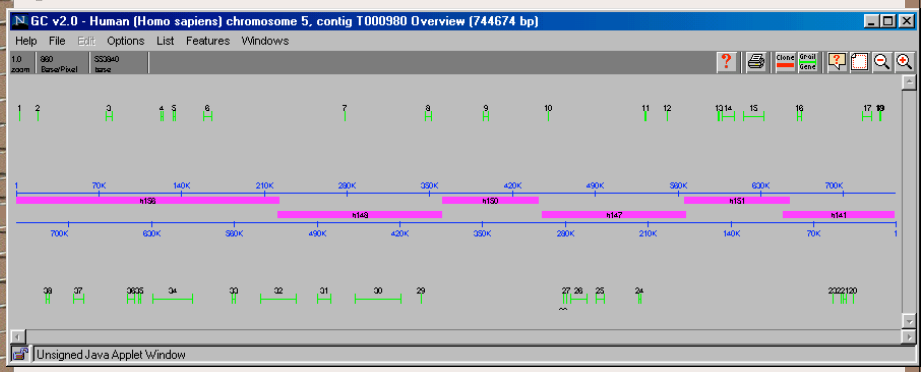
- Gene Expression
- Gene Regulatory Networks
- Phylogenetic comparisons
- Metabolic pathways

GenomeChannel

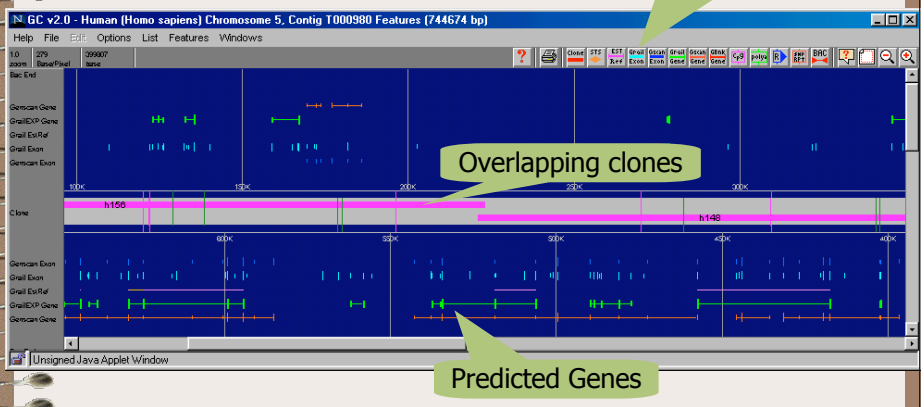




A Contig Overview



Feature Display



Gene Summary Report

Genome Channel Report - Netscape

File Edit View Go Communicator Help

Genome Channel | **Gene Summary Report** | Help

human Chromosome 5, Contig T000980, GrailEXP Gene 32

Links: Protein

```
TYPE            gene
SIZE            31475 bp
ORGANISM*       Homo sapiens
CHROMOSOME*     5
MAP*            5q23.1
GCID            GC05241232 (chr.5.ctg.T000980.gene.grailexp.32)
SIMILARITY      (U52351) neural plakophilin related arm-repeat protei...
FROM_ACC
FROM_NID
SEQ_SOURCE
FEATURES*           Location/Qualifiers
     gene           join(<1..111,12191..12394,28123..28277,28805..28942,
                    31445..31475>)
                    /similarity="(U52351) neural plakophilin related arm-repeat protei..." (blast_score=
                    /evidence="not experimental"
                    /translation=MGTDELDGLLCGEANGKDAESSGCWGKKKKKKKSQDQMFALLF
                    FFRQWDGVGPLPDCAEPPKGIQMLWHPSIVKPYLTLLSECSNPDTLEGAAGALQNL
                    AAGSWKWSVYIRAAVRKEKGLPILVELLRIDNDRVVCAVATALRNMALDVRNKELI
                    GMPVLGPEIKSISKTDRKPCPCGVYIQEKEMFKKQTKQMNHKMELSKGWEDAKAKA
                    D*
     exon           (1..111)
                    /EST=T77214
     exon           (12191..12394)
                    /EST=T77214
     exon           (28123..28277)
     exon           (28805..28942)
     exon           (31445..31475)

BASE COUNT     9381 a     6543 c     6139 g     9412 t
ORIGIN
        1 atgggcacgg acgagctgga cgggctactc tgtggcgagg ccaatggcaa ggatgctgag
       61 agctctgggt gctggggcaa gaagaagaag aaaaagaaat cccaagatca ggtgatgcag
      121 gctgcttgca gctgcatgca attatcccct tcctaattgc aactgtaata tatcctcaat
      181 atatcacaa ccaacccccc atttccatca ccaactaaaa tacctaatcc atctccacaa
```

http://compbio.ornl.gov/cgi-bin/PrtnRpt.pl?human,5,T000980,gene,grailexp,32

BEAUTY - Gene Search Results

BLASTP+BEAUTY Search Results - Netscape

File Edit View Go Communicator Help

Distribution of 29 Blast Hits on the Query Sequence

2822195 (U52351) neural plakophilin related arm-repeat protein ..S= 253 E=6e-67

Color Key for Alignment Scores

<40 | 40-50 | 50-80 | 80-200 | >=200

QUERY 100 200

```
                                                          Score     E
Sequences producing significant alignments:               (bits)  Value
gi|2822195(U52351) neural plakophilin related arm-repeat protei...   253  6e-67
gi|3712673(U96136) delta-catenin [Homo sapiens]                      249  9e-66
gi|2580537(U90331) neural plakophilin related arm-repeat protei...   236  9e-62
gi|1702924|gnl|PID|e259279(X81889) p0071 protein [Homo sapiens]      165  3e-40
gi|1932727(U51269) armadillo repeat protein [Homo sapiens]           109  1e-23
gi|2253589(U52828) delta-catenin [Homo sapiens]                      106  1e-22
gi|3152867(AF062344) p120 catenin isoform 4B [Homo sapiens]           82  4e-15
gi|3152817(AF062319) p120 catenin isoform 2ABC [Homo sapiens]         82  4e-15
gi|3152845(AF062331) p120 catenin isoform 4A [Homo sapiens] >gi...    82  4e-15
```

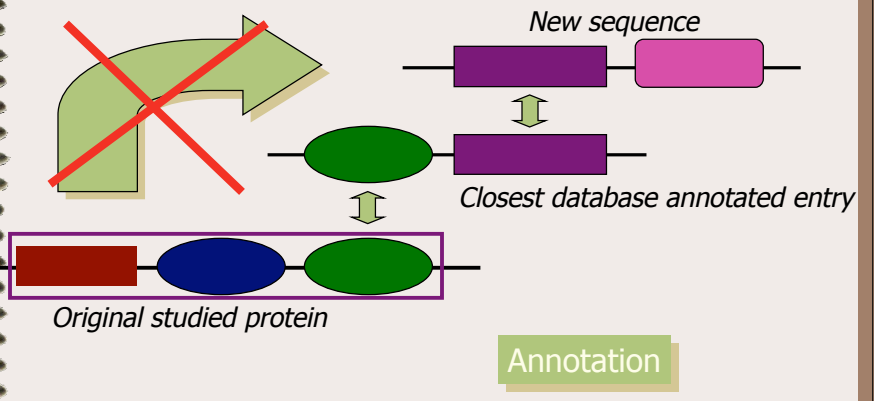
http://grail.lsd.ornl.gov/GC/human/chromosome/5/contig/T000980/gene/grailexp/search/beauty/32.html#2822195

Layers of Information

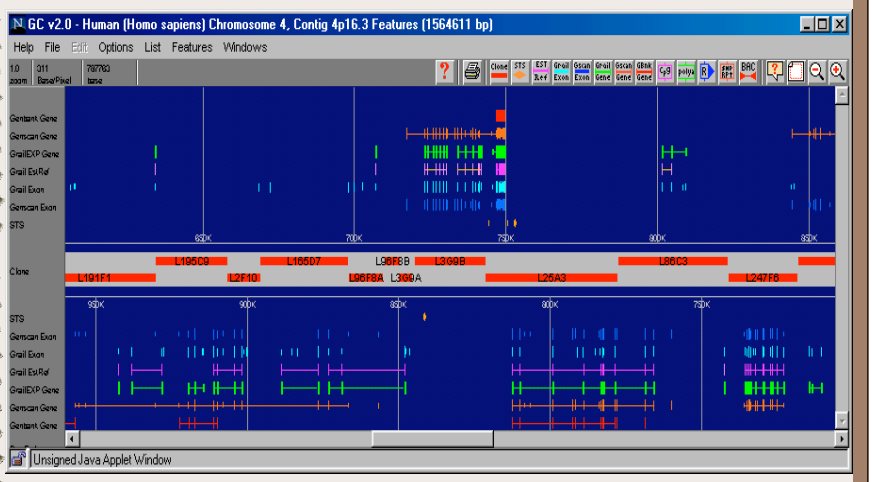
The same base sequence contains many layered instructions!

- Chromosome structure and function
  - Telomers, centromers
- Gene Regulatory information
  - Enhancers, promoters, ...
- Instructions for gene structure
- Instructions for protein
- Instructions for protein post-processing and localization

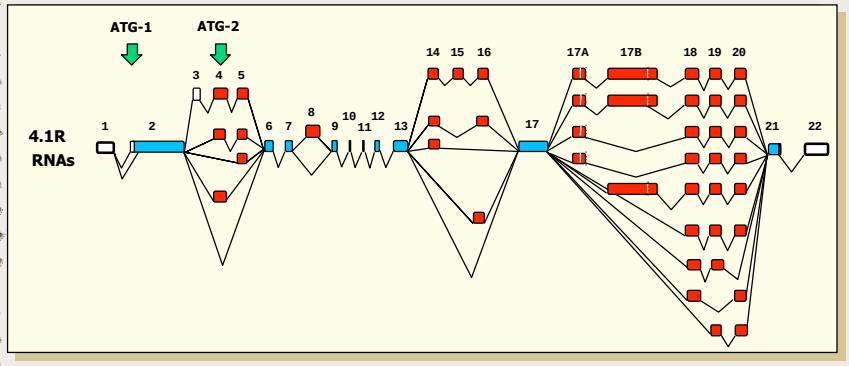
Inherited Annotation Problems in Multi-Domain Proteins



Alternatively Spliced ?



One Gene - Many Proteins



Conboy 1998

As many as 30% of human genes, in particular structural genes, may be alternatively spliced.

Annotated Genomes

Internet Resources

GenomeHub	http://www.nhgri.nih.gov/genome_hub.html
NAR Database Issue	http://nar.oupjournals.org/content/vol29/issue1/

Netscape: NHGRI Genome Hub

Location: http://www.nhgri.nih.gov/genome_hub.html

National Human Genome Research Institute

International Human Genome Sequencing Consortium
Genome Hub

On behalf of the International Human Genome Sequencing Consortium, Genome Hub is a web site that consists of a list of a number of important web sites that contain information about the human genome sequence, other genome sequences, and other relevant genomic information. It is similar to the "Genome Central" sites that can be found on the NCBI and EBI sites. Genome Hub will be maintained and regularly updated with new information.

Use the links below to go to the listed directories containing information pertinent to those areas

- Draft Genome Sequence Assemblies, and other genomic data resources
- Underlying Map Information
- Sequencing Centers of the International Human Genome Sequencing Consortium
- Nucleotide Sequence Databases
- Trace Archives
- SNPs
- cDNAs and ESTs
- Other Organisms
- Other Human Genetics Information
- ELSI Information
- Funding Agencies

top **Draft Genome Sequence Assemblies, and other genomic data resources**

UCSC	contains the assembly of the draft genome sequence that was analyzed in the February 15, 2001 publication
EBI	the Ensembl project provides access to DNA and protein sequences with automatic baseline annotation; allows BLAST analysis
NCBI	in addition to the NCBI assembly of the draft genome sequence, this site provides access to other NCBI resources including BLAST analysis
Oak Ridge National Laboratory	Java viewers for Human genome data.
RIKEN Genomic Sciences Center	gives an overview of the entire human genome structure.

top **Underlying Map Information**

WashU	This page contains links to the fingerprinted clone and accession maps of the human genome
Gene Map 99	Gene map 99 contains data and viewers for radiation hybrid maps.
Marshfield	this page contains comprehensive human genetic linkage maps
Genethon	This site is contains the Genethon genetic linkage maps
RHdb	This is the site for the collaborative Radiation Hybrid mapping

GoldenPath

Netscape: Human Genome Browser Gateway

Location: http://genome.cse.ucsc.edu/goldenPath/octTracks.html

Human Genome Browser

Web tool created by Jim Kent of UC Santa Cruz
7 Oct 2000 draft assembly of the human genome

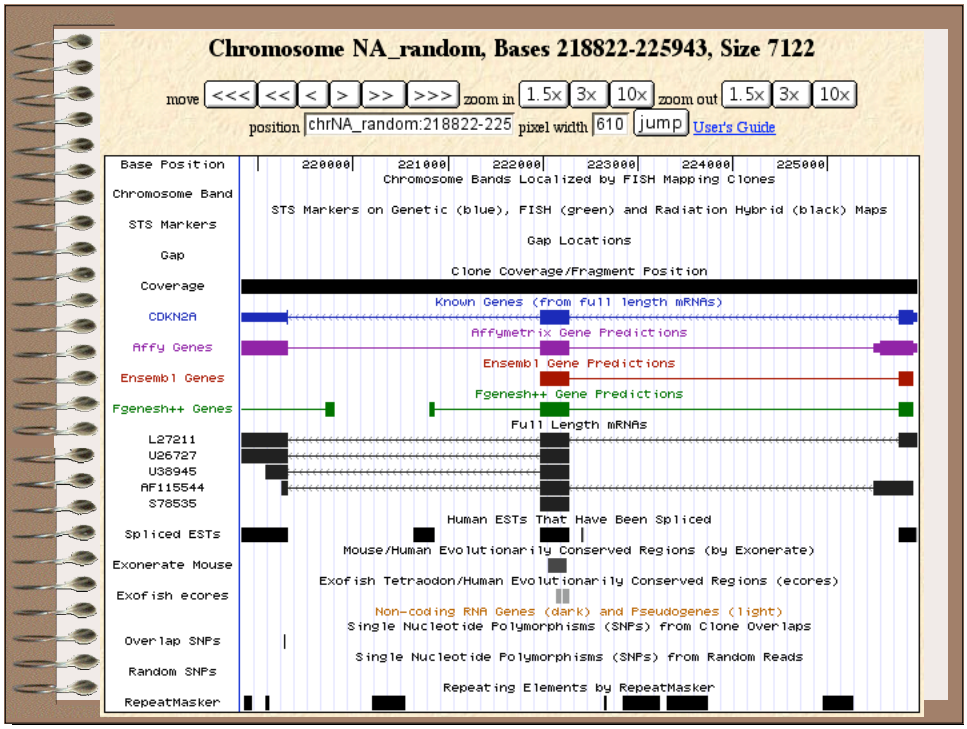
Please enter a position in the genome, set your preferred window width, and press the submit button: (Use BLAT Search to locate a particular sequence in the genome.)

genome position: chr17:45247249-45350955 pixel width 610 Submit

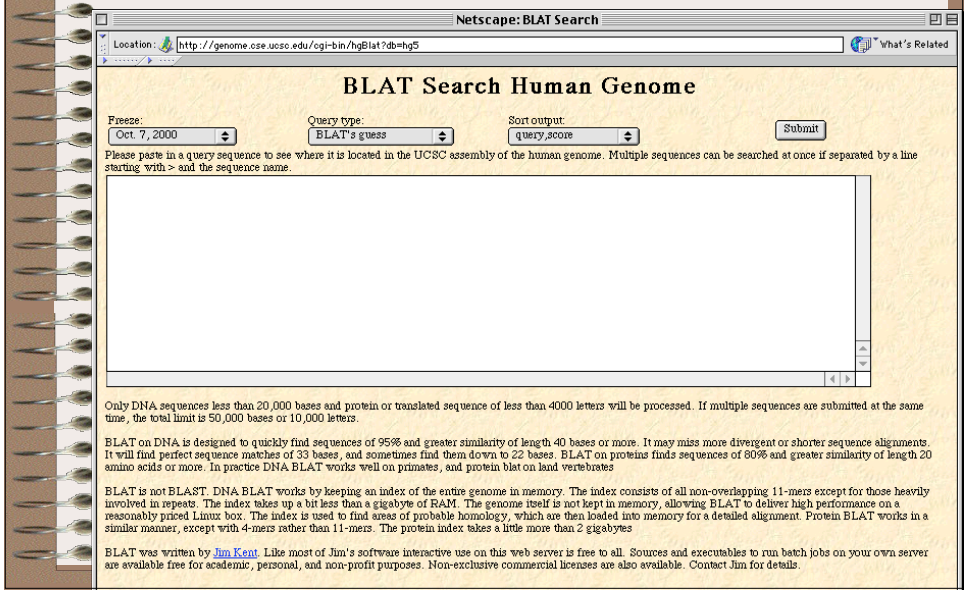
A genome position can be specified by the accession number of a sequenced human genomic clone, an mRNA or EST or STS marker, or a cytological band, a chromosomal coordinate range, or keywords from the Genbank description of an mRNA. See the User Guide for more help.

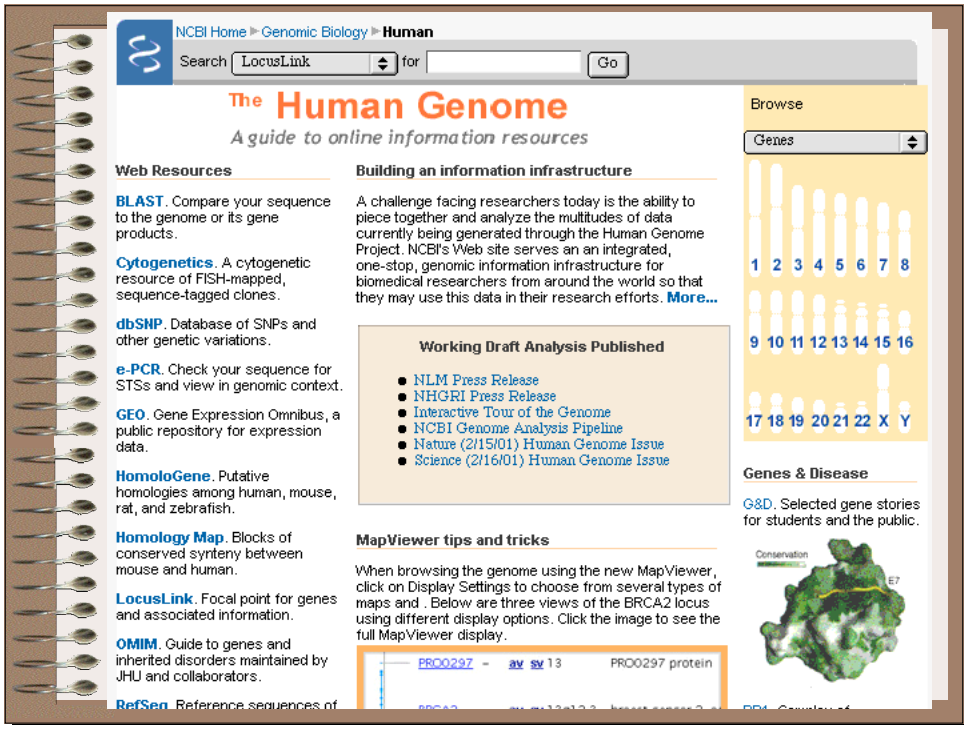
Request:	Genome Browser Response:
chr19	Displays all of chromosome 19
20p13	Displays region for band p13 on chr 20
4q28	Displays band q28 on chr 4, gene region determining red hair color
chr3:1-1000000	Displays first million bases of chr 3, counting from p arm telomere
D16S3046	Displays region around STS marker D16S3046 from the Genethon/Marshfield maps (open "STS Markers" track by clicking to see this marker)
AA205474	Displays region of EST with GenBank acc. AA205474 in BRCA1 cancer gene on chr 17 (open "spliced ESTs" track by clicking to see this EST)
ctg13698	Displays region of the fingerprint clone contig ctg13698 from the Wash. U. map (set "FPC contigs" track to "dense" and refresh to see FPC contigs)
AP001670	Displays region of clone with GenBank accession AP001670 (open "coverage" track by clicking to see this clone)
AF083811	Offers 2 coordinate choices on chr 7 and chr 22 for this mitotic checkpoint mRNA
PRNP	Offers coordinates choices for nested mRNAs for prion gene, chr 20p
pseudogene mRNA	Lists transcribed pseudogenes but not cDNAs
homeobox caudal	Lists mRNAs for caudal homeobox genes
valyl-tRNA	Lists mRNAs for valyl tRNA synthetases but not unsubmitted tRNA
zinc finger	Lists many zinc finger mRNAs
kruppel zinc finger	Lists only kruppel-like zinc fingers
huntington	Lists candidate genes associated with Huntington's disease
zahler	Lists mRNAs deposited by scientist named Zahler
Evans,J.E.	Lists mRNAs deposited by co-author J.E. Evans

Use this last format for entry authors -- even though Genbank searches require Evans JE format, GenBank entries themselves use Evans,J.E. internally.


Chromosome NA_random, Bases 218822-225943, Size 7122
move <<< << < > >> >>> zoom in 1.5x 3x 10x zoom out 1.5x 3x 10x
position chrNA_random:218822-225 pixel width 610 jump User's Guide
Base Position
220000 221000 222000 223000 224000 225000
Chromosome Band
Chromosome Bands Localized by FISH Mapping Clones
STS Markers
STS Markers on Genetic (blue), FISH (green) and Radiation Hybrid (black) Maps
Gap
Gap Locations
Coverage
Clone Coverage/Fragment Position
CDKN2A
Known Genes (from full length mRNAs)
Affy Genes
Affymetrix Gene Predictions
Ensembl Genes
Ensembl Gene Predictions
Fgenesh++ Genes
Fgenesh++ Gene Predictions
L27211
U26727
U38945
AF115544
S78535
Full Length mRNAs
Spliced ESTs
Human ESTs That Have Been Spliced
Exonerate Mouse
Mouse/Human Evolutionarily Conserved Regions (by Exonerate)
Exofish ecores
Exofish Tetraodon/Human Evolutionarily Conserved Regions (ecores)
Non-coding RNA Genes (dark) and Pseudogenes (light)
Overlap SNPs
Single Nucleotide Polymorphisms (SNPs) from Clone Overlaps
Random SNPs
Single Nucleotide Polymorphisms (SNPs) from Random Reads
RepeatMasker
Repeating Elements by RepeatMasker

BLAT Search


Netscape: BLAT Search
Location: http://genome.cse.ucsc.edu/cgi-bin/hgBlat?db=hg5
What's Related
BLAT Search Human Genome
Freeze: Oct. 7, 2000
Query type: BLAT's guess
Sort output: query,score
Submit
Please paste in a query sequence to see where it is located in the UCSC assembly of the human genome. Multiple sequences can be searched at once if separated by a line starting with > and the sequence name.
Only DNA sequences less than 20,000 bases and protein or translated sequence of less than 4000 letters will be processed. If multiple sequences are submitted at the same time, the total limit is 50,000 bases or 10,000 letters.
BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 22 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. In practice DNA BLAT works well on primates, and protein blat on land vertebrates
BLAT is not BLAST. DNA BLAT works by keeping an index of the entire genome in memory. The index consists of all non-overlapping 11-mers except for those heavily involved in repeats. The index takes up a bit less than a gigabyte of RAM. The genome itself is not kept in memory, allowing BLAT to deliver high performance on a reasonably priced Linux box. The index is used to find areas of probable homology, which are then loaded into memory for a detailed alignment. Protein BLAT works in a similar manner, except with 4-mers rather than 11-mers. The protein index takes a little more than 2 gigabytes
BLAT was written by Jim Kent. Like most of Jim's software interactive use on this web server is free to all. Sources and executables to run batch jobs on your own server are available free for academic, personal, and non-profit purposes. Non-exclusive commercial licenses are also available. Contact Jim for details.

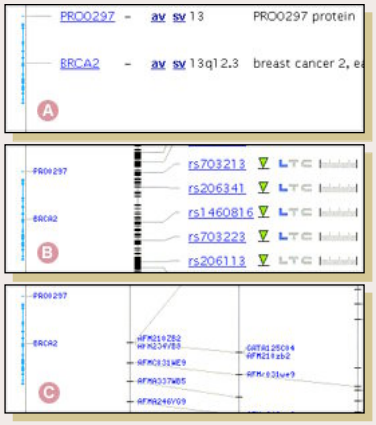


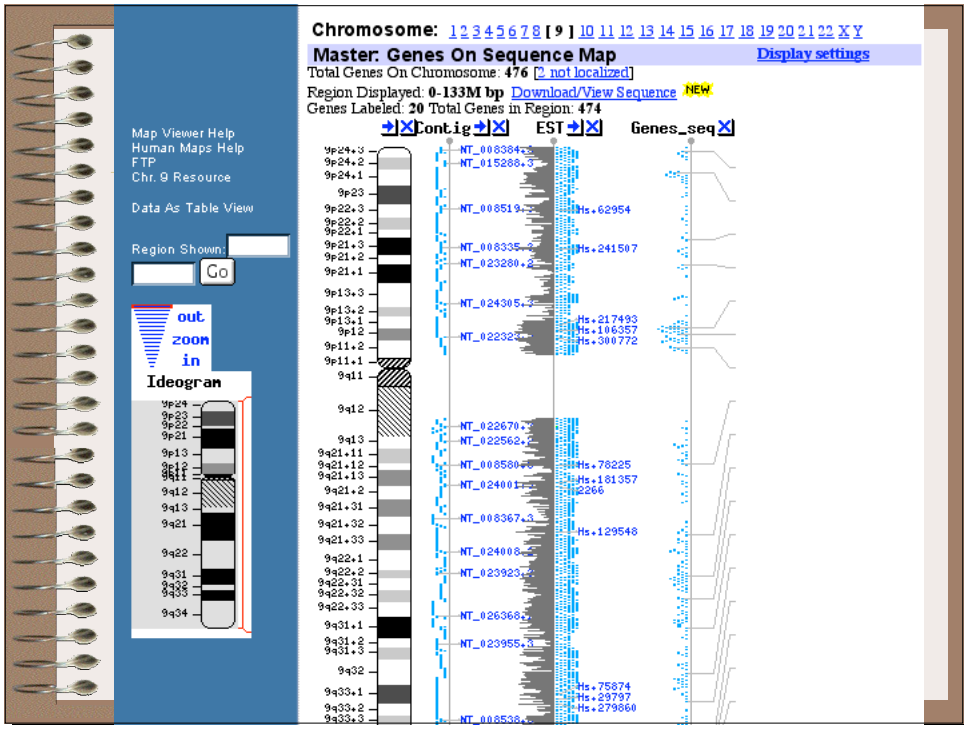
NCBI Display options

A. Genes

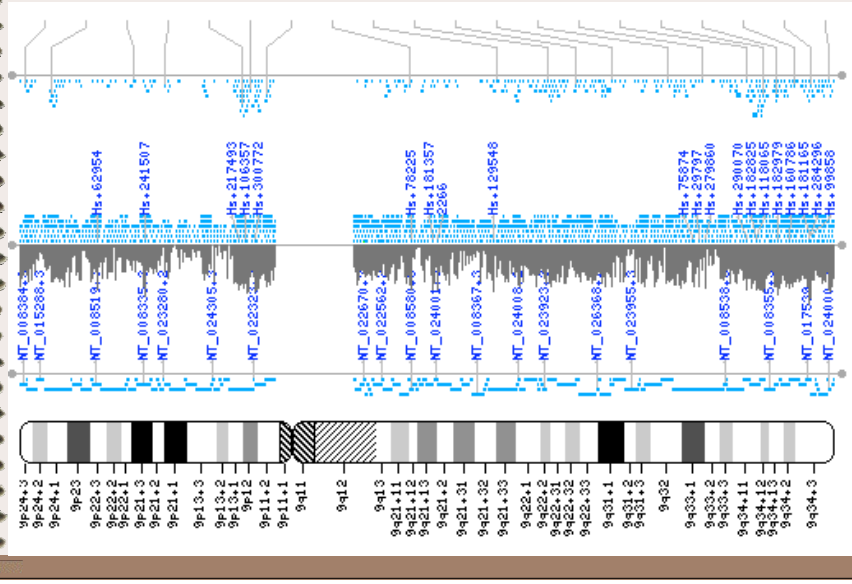
B. Variations, juxtaposed with genes

C. Several STS maps, juxtaposed with genes





NCBI Chromosome 9



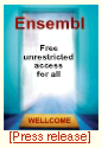
project Ensembl

The Sanger Centre

EBI

Human Genome Server

About Ensembl v1.0



Ensembl is a joint project between EMBL - EBI and the Sanger Centre to develop a software system which produces and maintains automatic annotation on eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust.

[Press release]

With Ensembl you can ...

- Download all data, free, without constraints
- Search the DNA from the human genome
- Browse chromosomes
- Find genes, SNPs and mouse genome matches
- Look for proteins and protein families

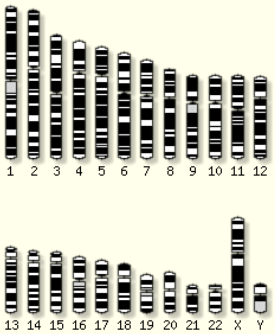
How Do I ...?

- Find genomic sequences similar to my protein sequence?
- Look up a positional marker and examine candidate disease genes in the region?
- Find the expression profile of a gene?
- More ...

Ensembl provides ...

- Identification of 90% of known human genes in the genome sequence
- Prediction of 10,000 additional genes, all with supporting evidence
- Connections to other resources worldwide, leveraging many public genomic databases and tools
- This website, www.ensembl.org, facilitates public access to this data by offering a web-based genome browser.

Browse a Chromosome



Ensembl Links

- News
- Download
- BLAST
- SSAHA
- Docs
- Dev

- **New:** Ensembl Mouse server

Help

Help — Click on any help icon to pop up a context-sensitive help window.

---

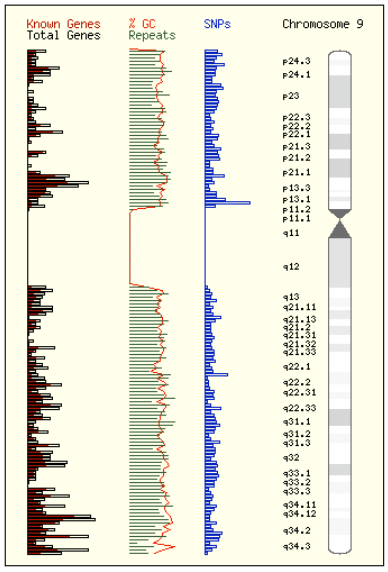
project Ensembl *MapView*

The Sanger Centre

EBI

Home ▲ News ▲ BLAST ▲ Disease Browser ▲ Docs ▲ Download ▲

Find Marker | Lookup [e.g. RH9632, D1S2806] Help



Chromosome 9

Known Ensembl Genes: 565 **SNPs:** 45992
Novel Ensembl Genes: 478 **Length:** 141263275 bp

Change Chromosome

Chromosome: 9 Lookup

Jump to Contigview

Click anywhere on the chromosome ideogram or one of the feature distribution plots to jump to a contig-level view of features at that point.

Alternatively, you can jump to contigview between any two landmark markers on this chromosome:

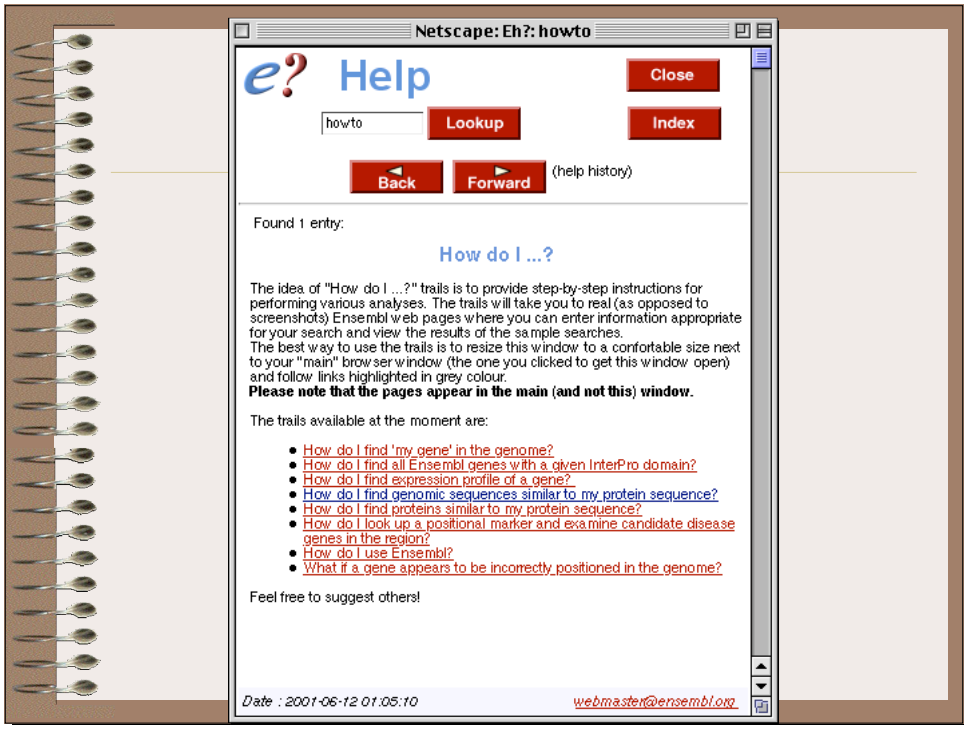
Between: D9S2065 (at 2.19 Mb)

and: D9S2065 (at 2.19 Mb)

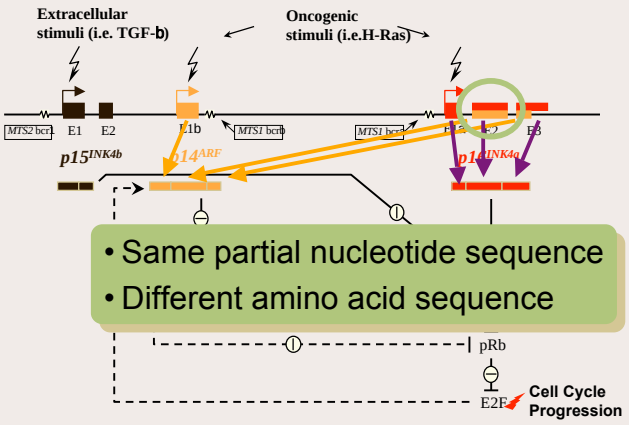
Lookup

OMIM Diseases

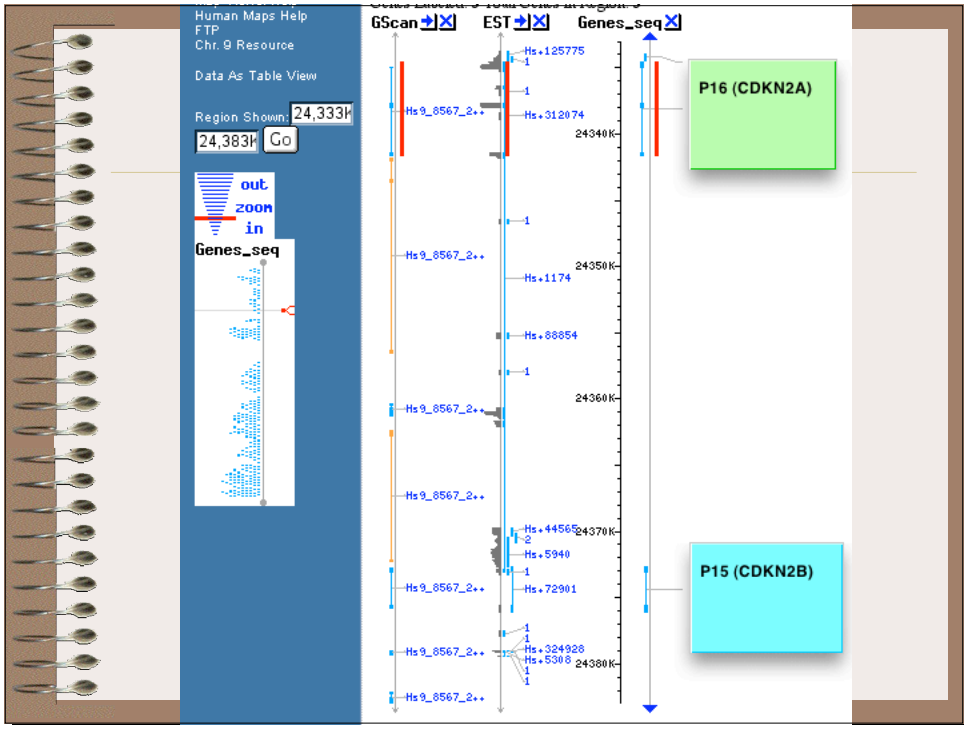
Browse OMIM Diseases on this chromosome.

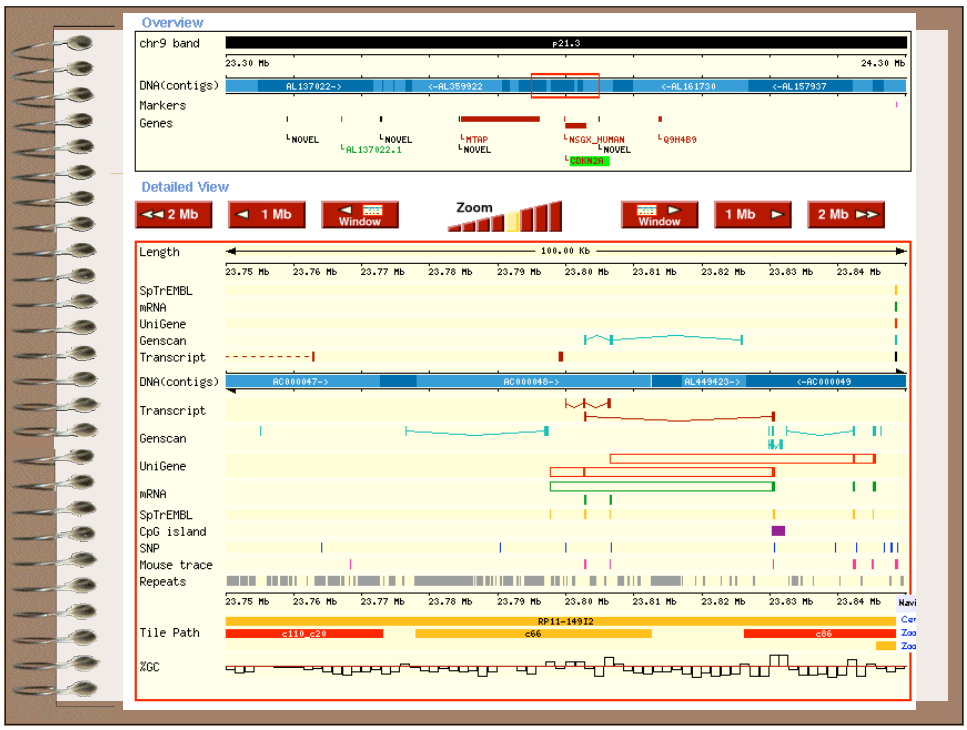


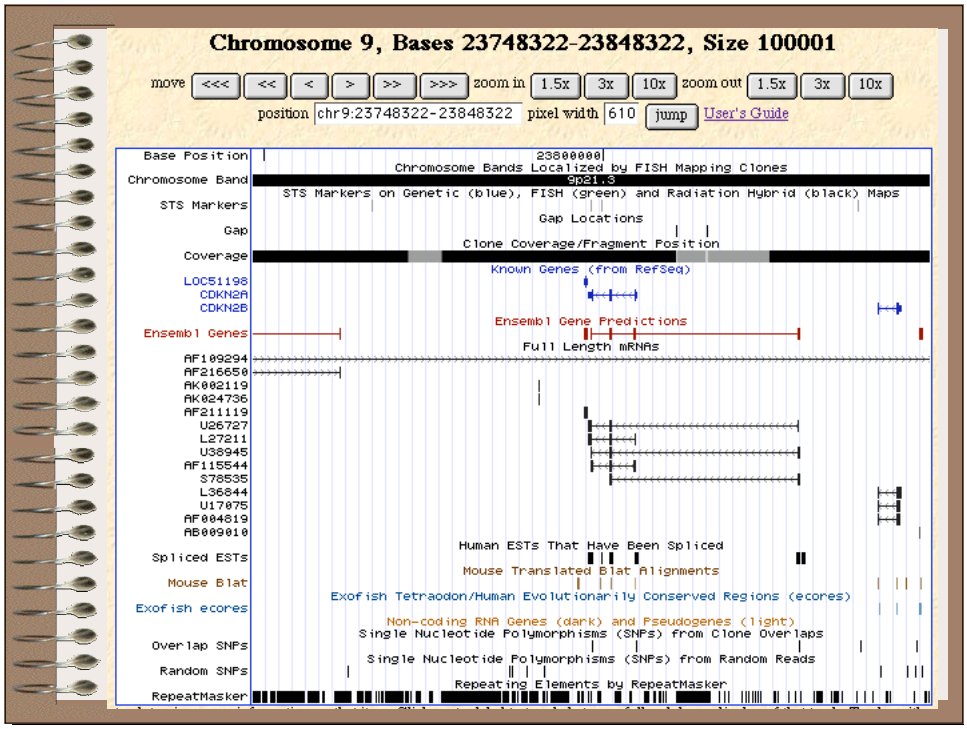
9p21 gene cluster is a nexus of the Rb and p53 pathways



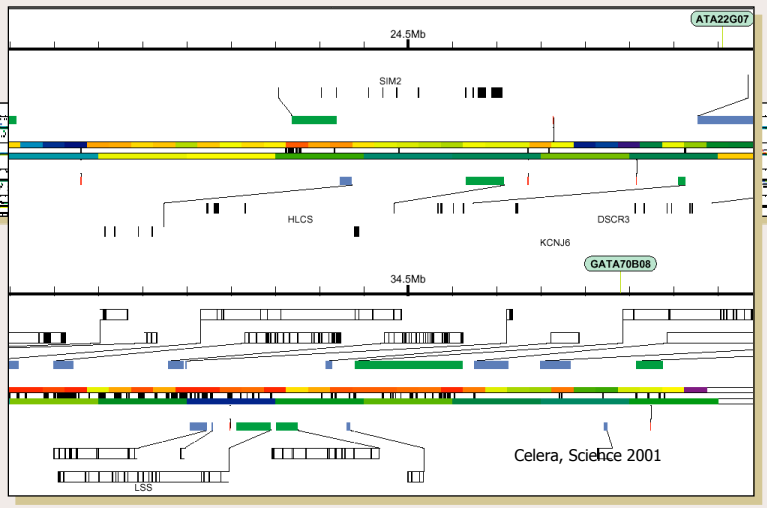
- Same partial nucleotide sequence
- Different amino acid sequence


Human Maps Help
FTP
Chr. 9 Resource
Data As Table View
Region Shown: 24,333K
24,383K Go
out
zoom
in
Genes_seq
GScan
EST
Genes_seq
P16 (CDKN2A)
P15 (CDKN2B)
24340K
24350K
24360K
24370K
24380K


Overview
chr9 band
p21.3
23.30 Mb
24.30 Mb
DNA(contigs)
Markers
Genes
Detailed View
2 Mb
1 Mb
Window
Zoom
Window
1 Mb
2 Mb
Length
100.00 Kb
SpTrEMBL
mRNA
UniGene
Genscan
Transcript
DNA(contigs)
Transcript
Genscan
UniGene
mRNA
SpTrEMBL
CpG island
SNP
Mouse trace
Repeats
Tile Path
%GC

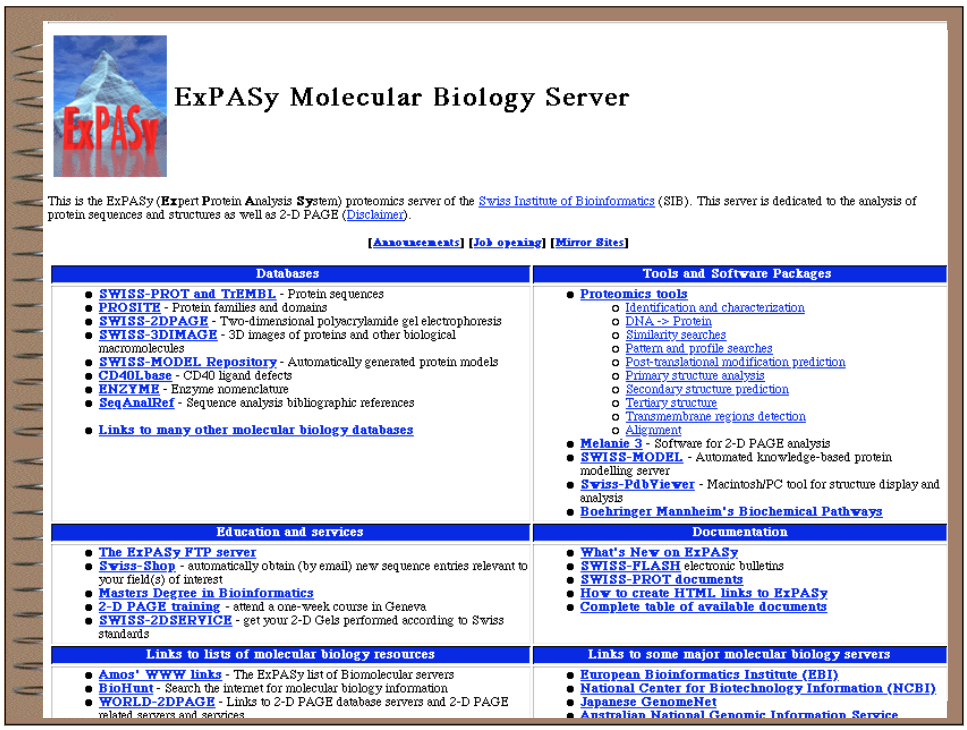


Celera Chromosome 21



Celera, Science 2001

Beyond the Genome



ExPASy Molecular Biology Server

This is the ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB). This server is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer).

[**Announcements**] [**Job opening**] [**Mirror Sites**]

Databases

- **SWISS-PROT and TrEMBL** - Protein sequences
- **PROSITE** - Protein families and domains
- **SWISS-2DPAGE** - Two-dimensional polyacrylamide gel electrophoresis
- **SWISS-3DIMAGE** - 3D images of proteins and other biological macromolecules
- **SWISS-MODEL Repository** - Automatically generated protein models
- **CD40Lbase** - CD40 ligand defects
- **ENZYME** - Enzyme nomenclature
- **SeqAnalRef** - Sequence analysis bibliographic references
- **Links to many other molecular biology databases**

Tools and Software Packages

- **Proteomics tools**
  - Identification and characterization
  - DNA -> Protein
  - Similarity searches
  - Pattern and profile searches
  - Post-translational modification prediction
  - Primary structure analysis
  - Secondary structure prediction
  - Tertiary structure
  - Transmembrane regions detection
  - Alignment
- **Melanie 3** - Software for 2-D PAGE analysis
- **SWISS-MODEL** - Automated knowledge-based protein modelling server
- **Swiss-PdbViewer** - Macintosh/PC tool for structure display and analysis
- **Boehringer Mannheim's Biochemical Pathways**

Education and services

- **The ExPASy FTP server**
- **Swiss-Shop** - automatically obtain (by email) new sequence entries relevant to your field(s) of interest
- **Masters Degree in Bioinformatics**
- **2-D PAGE training** - attend a one-week course in Geneva
- **SWISS-2DSERVICE** - get your 2-D Gels performed according to Swiss standards

Documentation

- **What's New on ExPASy**
- **SWISS-FLASH** electronic bulletins
- **SWISS-PROT documents**
- **How to create HTML links to ExPASy**
- **Complete table of available documents**

Links to lists of molecular biology resources

- **Amos' WWW links** - The ExPASy list of Biomolecular servers
- **BioHunt** - Search the internet for molecular biology information
- **WORLD-2DPAGE** - Links to 2-D PAGE database servers and 2-D PAGE related servers and services

Links to some major molecular biology servers

- **European Bioinformatics Institute (EBI)**
- **National Center for Biotechnology Information (NCBI)**
- **Japanese GenomeNet**
- **Australian National Genomic Information Service**

Physical Properties

Prediction of Physical Properties

- Compute pI/MW http://www.expasy.ch/tools/pitool.html
- MOWSE http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse
- PeptideMass http://www.expasy.ch/tools/peptide-mass.html
- TGREASE ftp://ftp.virginia.edu/pub/fasta/
- SAPS http://www.isrec.isb-sib.ch/software/SAPSform.html

Prediction of Protein Identity Based on Composition

- AACompIdent http://www.expasy.ch/tools/aacomp/
- AACompSim http://www.expasy.ch/tools/aacsim/
- PROPSEARCH http://www.embl-heidelberg.de/prs.html

Motifs and Patterns

- BLOCKS http://blocks.fhcrc.org
- Pfam http://www.sanger.ac.uk/Software/Pfam/
- PRINTS http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/PRINTS.html
- ProfileScan http://www.isrec.isb-sib.ch/software/PFSCANform.html

Protein Structure

Prediciton of Secondary Structure and Folding Classes

- nnpredict http://www.cmpharm.ucsf.edu/_nomi/nnpredict.html
- PredictProtein http://www.embl-heidelberg.de/predictprotein/
- SOPMA http://pbil.ibcp.fr/
- Jpred http://jura.ebi.ac.uk:8888/
- PSIPRED http://insulin.brunel.ac.uk/psipred
- PREDATOR http://www.embl-heidelberg.de/predator/predatorinfo.html

Prediction of Specialized Structures or Features

- COILS http://www.ch.embnet.org/software/COILSform.html
- MacStripe http://www.york.ac.uk/depts/biol/units/coils/mstr2.html
- PHDtopology http://www.embl-heidelberg.de/predictprotein
- SignalP http://www.cbs.dtu.dk/services/SignalP/
- TMpred http://www.isrec.isb-sib.ch/ftp-erver/tmpred/www/TMPREDform.html

Structure Prediction

- DALI http://www2.ebi.ac.uk/dali/
- Bryant-Lawrence ftp://ncbi.nlm.nih.gov/pub/pkb/
- FSSP http://www2.ebi.ac.uk/dali/fssp/
- UCLA-DOE http://fold.doe-mbi.ucla.edu/Home
- SWISS-MODEL http://www.expasy.ch/swissmod/SWISS-MODEL.html
- TOPITS http://www.embl-heidelberg.de/predictprotein/