

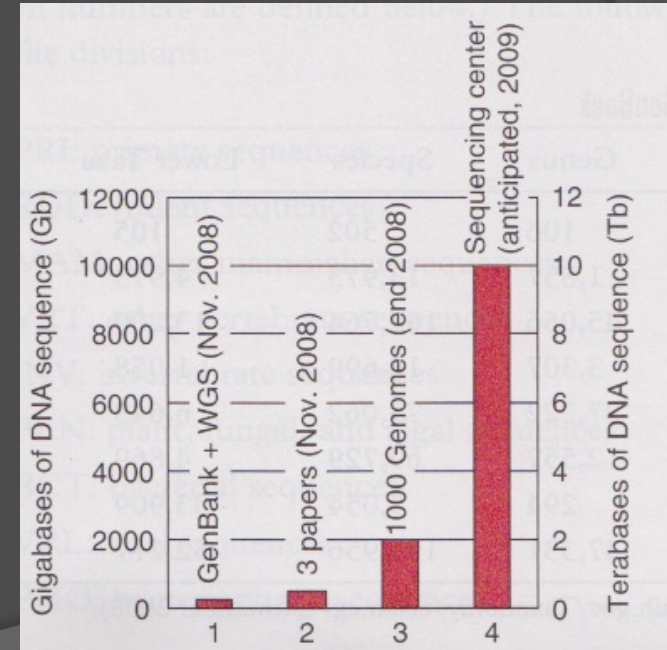
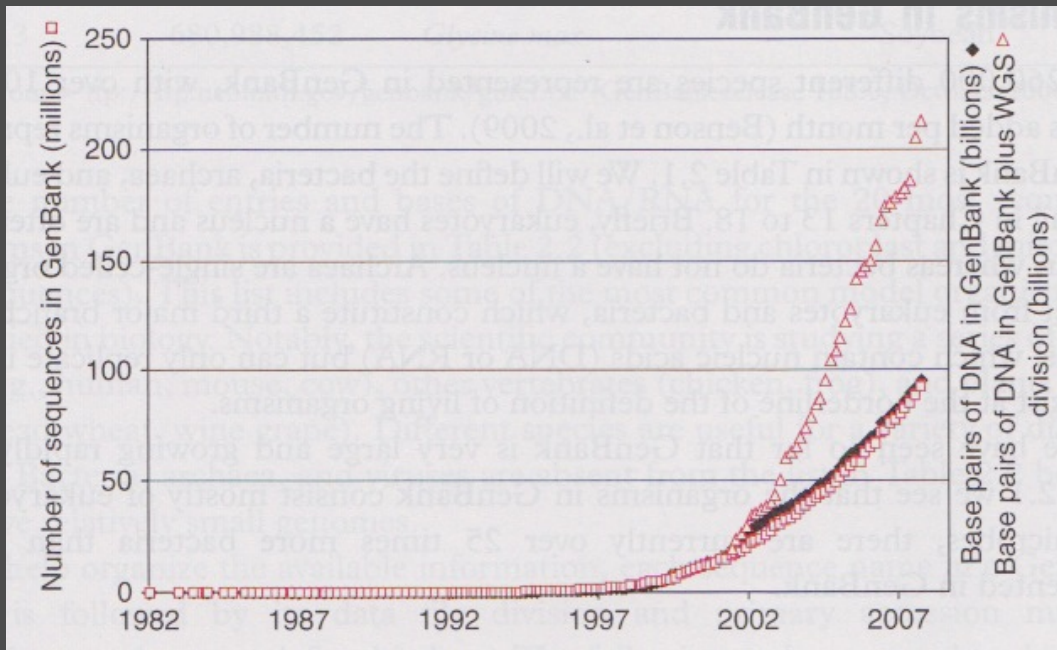
# Biological Databases

Nicholas E. Navin, Ph.D.  
Department of Genetics  
Department of Bioinformatics

TA: Dr. Yong Wang

# History

- The first DNA sequence databases were Genbank (NCBI) and EMBL (Europe) established in 1983
- In 1983 the Genbank database stored just 2000 DNA sequences
- Today it stores > 300 million sequences
- Next-generation sequencing data has resulted in exponential growth of these databases over the last five years



# Types of Databases

## PRIMARY

Stores submitted sequences / archival

- ◉ GenBank
- ◉ dbSNP
- ◉ GEO
- ◉ Sequence Read Archive

## SECONDARY

Curated Databases

- ◉ RefSeq
- ◉ UniProt
- ◉ OMIM
- ◉ Cancer Gene Census



# Biological Databases

## DNA/RNA Sequences

- ◉ Genbank (NCBI)

## Genomic Data

- ◉ UCSC Genome Browser
- ◉ Ensembl (EMBL)
- ◉ Gene Expression Omnibus (GEO)
- ◉ Sequence Read Archive (SRA)

## Cancer Genomics Databases

- ◉ TCGA
- ◉ ICGC
- ◉ COSMIC
- ◉ Oncomine
- ◉ cBio

## Human Variation

- ◉ dbSNP
- ◉ OMIM
- ◉ 1000 genomes Project
- ◉ HapMap

## Gene Ontology and Pathway Databases

- ◉ Panther
- ◉ KEGG
- ◉ Reactome

## Protein Databases

- ◉ Protein Databank (RCSB)
- ◉ PFAM
- ◉ UniProt

## Mutation Prediction Databases

- ◉ SIFT
- ◉ POLYPHEN
- ◉ SeattleSeq

## Genomic Analysis Tools

- ◉ Galaxy
- ◉ Oncomine
- ◉ DAVID
- ◉ Ingenuity (commercial)

Links can be found at:

[http://www.navinlab.com/biodb/biodb/website\\_links.html](http://www.navinlab.com/biodb/biodb/website_links.html)

# Accessing Data from Databases

- ◉ Website Access: Search & Download
- ◉ Biomarts
- ◉ FTP
- ◉ Direct Connection from a Programming Language (MySQL, DAS, API)
- ◉ Direct data access from R or Matlab packages

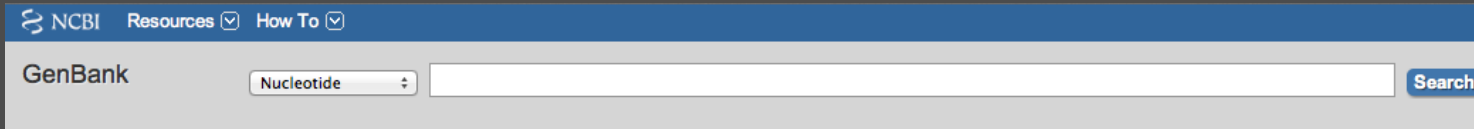
# Cloud-Based Analysis Tools

- ◉ Web or Cloud Based Tools
- ◉ Cloud = Server space that can be used to store data uploaded from your experiments
- ◉ The uploaded data can then be accessed from any device with internet access
- ◉ Programs on the cloud server can then be run on your data
- ◉ Final data or results can then be downloaded



**GALAXY** is an example of a free cloud-based analysis tool for genomic data

# GenBank



NCBI Resources How To

GenBank Nucleotide Search

<http://www.ncbi.nlm.nih.gov/genbank/>

- The original databases for depositing DNA, RNA and Protein sequences
- GenBank Identifiers: NM\_identifier      **NM\_001144919.1**
- Output: GenBank or FASTA file

```
>gi|222144243|ref|NM_001144919.1| Homo sapiens fibroblast growth factor receptor  
2 (FGFR2), transcript variant 9, mRNA  
GGCGGGCGGCTGGAGGAGAGCGCGGTGGAGAGCCGAGCGGGCGGGCGGGTGGCGGAGCGGGCGAGGGAG  
CGCGCGCGGGCCGCCACAAAGCTCGGGCGCCGCGGGGCTGCATGCGGCGTACCTGGCCCCGGCGCGGCGACT  
GCTCTCCGGGCTGGCGGGGGCCGGCCGCGAGCCCCGGGGGCCCCGAGGCCGCGAGCTTGCCCTGCGCGCTCT  
GAGCCTTCGCAACTCGCGAGCAAAGTTTGGTGGAGGCAACGCCAAGCCTGAGTCCTTTCTTCCTCTCGTT  
CCCCAAATCCGAGGGCAGCCCGCGGGCGTCATGCCCGCGCTCCTCCGCAGCCTGGGGTACGCGTGAAGCC  
CGGGAGGCTTGGCGCCGGCGAAGACCCAAGGACCACTCTTCTGCGTTTGGAGTTGCTCCCCGCAACCCCG  
GGCTCGTCGCTTTCTCCATCCCGACCCACGCGGGGCGCGGGGACAACACAGGTGCGCGGAGGAGCGTTGCC  
ATTC AAGTGACTGCAGCAGCAGCGGCAGCGCCTCGGTTCTGAGCCCACCGCAGGCTGAAGGCATTGCGC  
GTAGTCCATGCCCGTAGAGGAAGTGTGCAGATGGGATTAACGTCCACATGGAGATATGGAAGAGGACCGG
```

# GenBank Flat File (GBFF)

```
LOCUS       MUSENH          1803 bp    mRNA           ROD           29-AUG-1997
DEFINITION  Mouse neuroblastoma and rat glioma hybridoma cell line NG108-15
            cell TA20 mRNA, complete cds.
ACCESSION   D25291
VERSION    g1850791
KEYWORDS   neurite extension activity; growth arrest; TA20.
SOURCE     Murinae gen. sp. mouse neuroblastoma-rat glioma hybridoma
            cell line:NG108-15 cDNA to mRNA.
ORGANISM   Murinae gen. sp.
            Eukaryota; Eukaryota; Eukaryotes; Metazoa; Chordata;
            Vertebrata; Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae;
            Murinae.
REFERENCE  1 (sites)
AUTHORS    Tohda,C., Nagai,S., Tohda,M. and Nomura,Y.
TITLE      A novel factor, TA20, involved in neuronal differentiation: cDNA
            cloning and expression
JOURNAL    Neurosci. Res. 23 (1), 21-27 (1995)
MEDLINE    96064354
REFERENCE  3 (bases 1 to 1803)
AUTHORS    Tohda,C.
TITLE      Direct Submission
JOURNAL    Submitted (18-NOV-1993) to the DDBJ/EMBL/GenBank databases. Chihoro
            Tohda, Toyama Medical and Pharmaceutical University, Research
            Institute for Wakan-yaku, Analytical Research Center for
            ethnomedicines; 2630 Sugitani, Toyama, Toyama 930-01, Japan
            (E-mail:CHIHO@med.toyama-mpu.ac.jp, Tel:+81-764-34-2281(ex.2841),
            Fax:+81-764-34-5057)
COMMENT    On Feb 26, 1997 this sequence version replaced gi:793764.
FEATURES   Location/Qualifiers
            source             1..1803
                                /organism="Murinae gen. sp."
                                /note="source origin of sequence, either mouse or rat, has
                                not been identified"
                                /db_xref="taxon:39108"
                                /cell_line="NG108-15"
            misc_signal        156..153
                                /note="AP-2 binding site"
            GC_signal          647..655
                                /note="Sp1 binding site"
            TATA_signal        694..701
            gene              748..1311
                                /gene="TA20"
            CDS                748..1311
                                /gene="TA20"
                                /function="neurite extension activity and growth arrest
                                effect"
                                /codon_start=1
                                /db_xref="PID:d1005516"
                                /db_xref="PID:g193765"
                                /translation="MKRLMVPVSRSLPNSPNVYRFLSWLMLIRVNNLSLISNTLRRR
                                KLRVNPVYTRKRLNIFVLLIPSCRLLMLIYVYRNLLKHWSTVRSRSHSIYRL
                                RPSMRTNILLRCHSYKPKFISHPIYNNVPSMMNLRLGLLSHQSHLDPILRFLMLTIYY
                                RGFNSRSPPLPPNRIRKPNRIKLCR"
            polyA_site         1803
BASE COUNT 507 a      458 c      311 g      527 t
ORIGIN
```

```
1  tcagttttt  tttttttt  tttttttt  tttttttt  tttttttt  ttgattcag
61  tccgtttaca  ttgtgaagt  tcacaggcct  cagtcaacac  aattggactg  ctcaggaat
121  cctccctggg  gaccgcagta  tactggcct  atgaacccaa  gccacctatg  gctaggtagg
181  agaaqtccca  ctgagagct  gactctgga  gaaatgcac  atgctcgtat  cgcattcca
241  eaagggtgac  ctctggccag  agtcagcag  ccgagggttc  tcttcggggc  tctccctca
301  ctgcttgact  ctgctcag  gcgtccatc  tgtgggggga  cgttattgt  attgccttc
361  cattctgac  ggcattgct  ccatttagct  ggaaggggac  agagcctggt  tctctagggc
421  gtttccatgt  gggcctggg  acaatccaaa  agatgggggc  tccaaacac  agatccaga
481  gcccagcgt  tttgttaa  aacactctg  ggggaatga  atgtcacag  ggcgttcag
541  gacaagaac  agctttctg  tcactccat  gagaaccgtc  gcaatcactg  tccgaagag
601  gaggagcca  gaatacagt  gtatggcct  gacgattgcc  cggagagagg  cggagccat
661  ggaagcagaa  agacagaaa  caaccctatt  attaaatt  attaacct  cttcattga
721  cctaccctgc  ccatacaaca  ttctatcag  atgaacatt  ggtcccttc  taggactctg
781  ccaatagct  caatatta  caggtcttt  cttagcaca  caatacaat  cagatacaat
841  aacagccttt  tcaatcagaa  caacattgt  tcgagacgta  aattacgggt  gactatccg
901  atatatacac  gaaaacggag  cctcaatatt  tttttttg  ttattcctc  atgtoggag
961  aggcttatat  tatggatcat  ataatttat  agaaacctga  aacattggag  tacttctact
1021  gtccgagct  atagccacag  catittatag  ctacgtcct  ccagggagac  aaatatcatt
1081  ctgggtgccc  acggttata  caaacctct  atcaaccatc  ccaatatgt  gaacaacct
1141  agtccaagaa  attgagggg  gcttccag  agacaagcc  accctgacc  gattctgc
1201  ttccacttc  atttaacct  ttattatgc  ggcctagca  atgtccacc  tctctctct
1261  ccacgaaca  ggatacaaca  cccaacag  attaaacta  gatcgagata  aaattccatt
1321  tcaccctcac  tacaaccaa  agatartcca  ggtatcccaa  tcaatctct  aattccata
1381  accctagatt  tattttcc  agacataa  gggaccag  caactcaat  accggtat
1441  ccaataaaca  cccaaccca  tattaaccc  gaatgatatt  tctattttg  atcagccatt
1501  ctacgtcca  tcccaataa  actagagggt  gtcttagcct  taatttact  tctctaat
1561  ttagccctaa  taccttctct  tcaactcca  aagcaacgaa  gcttaatt  ccgccaatc
1621  acacaatt  tgcactgat  cctatagcc  aacctacta  tcttaaccgt  aattggggg
1681  caaccagtag  caaccatt  attatcgt  gcaactaga  ctactctca  tctctcaaa
1741  tcacttaat  tcttatcca  atccagaa  ttatcagaa  caaatacta  aaattatc
1801  cat
```

## Header

- Title
- Taxonomy
- Citation

## Features (AA seq)

## DNA Sequence



# UCSC Genome Browser

## UCSC Genome Bioinformatics

[Genomes](#) - [Blat](#) - [Tables](#) - [Gene Sorter](#) - [PCR](#) - [VisiGene](#) - [Session](#) - [FAQ](#) - [Help](#)

<http://genome.ucsc.edu>

### Uses:

- ⦿ Download genomic data
- ⦿ Download annotation tracks
- ⦿ Browse chromosomes, annotations and genomic data
- ⦿ BLAT: search for homologous DNA sequences
- ⦿ Upload and plot custom data
- ⦿ Upload custom annotation tracks

# UCSC Genome Browser

Chrom  
position

## UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

chr10:89,613,175-89,718,512 105,338 bp.

Scale chr10: 50 kb hg18  
Chromosome Band: Chromosome Bands Localized by FISH Mapping Clones 10q23.31  
Recomb Rate: Recombination Rate from deCODE, Marshfield, or Genethon Maps (deCODE default)  
GC Percent: GC Percent in 5-Base Windows  
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics  
RefSeq Genes: PTEN, AK130076  
Simple Nucleotide Polymorphisms (dbSNP build 130)  
Repeating Elements by RepeatMasker  
SINE, LINE, LTR, DNA, Simple, Low Complexity, Satellite, RNA, Other, Unknown

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

track search default tracks default order hide all add custom tracks track hubs configure reverse resize refresh

collapse all Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes. expand all

### Mapping and Sequencing Tracks

refresh

<a href="#">Base Position</a>	<a href="#">Chromosome Band</a>	<a href="#">STS Markers</a>	<a href="#">FISH Clones</a>	<a href="#">Recomb Rate</a>	<a href="#">deCODE Recomb</a>
dense ▾	dense ▾	hide ▾	hide ▾	dense ▾	hide ▾

tracks

Track  
configu  
rations

# UCSC Genome Browser

## Upload Custom Annotation Tracks or Your Own Data

### Add Custom Tracks

clade  genome  assembly

Display your own data as custom annotation tracks in the browser. Data must be formatted in [BED](#), [bigBed](#), [Personal Genome SNP](#), [VCF](#), or [PSL](#) formats. To configure the display, set [track](#) and [browser](#) line attributes. bigWig, BAM and VCF formats must be embedded in a track line in the box below. Publicly available custom

Paste URLs or data:

Or upload:  no file selected

Optional track documentation: Or upload:  no file selected

Click [here](#) for an HTML document template that may be used for Genome Browser track descriptions.

# UCSC Genome Browser

## Download Annotation Tracks

### Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, or to help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#), the [OpenHelix Table Browser tutorial](#) for a narrated presentation of the software features and usage. For more complete information see the [MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [Gene](#) and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence](#)

**clade:**  **genome:**  **assembly:**

**group:**  **track:**

**table:**

**region:**  genome  ENCODE Pilot regions  position

**identifiers (names/accessions):**

**filter:**

**intersection:**

**correlation:**

**output format:**  Send output to  [Galaxy](#)  [GREAT](#)

**output file:**  (leave blank to keep output in browser)

**file type returned:**  plain text  gzip compressed

To reset all user cart settings (including custom tracks), [click here](#).

# Ensembl Genome Browser



[BLAST/BLAT](#) | [BioMart](#) | [Tools](#) | [Downloads](#) | [Help & Documentation](#) | [Blog](#) | [Mirrors](#)

<http://useast.ensembl.org/index.html>

## Identifiers:

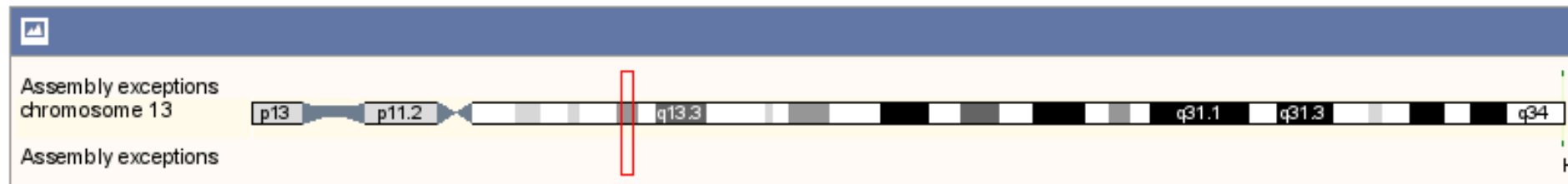
**ENSG00000000133703**    gene  
**ENST00000000530893**    transcript  
**ENSP00000000369497**    protein

## Uses:

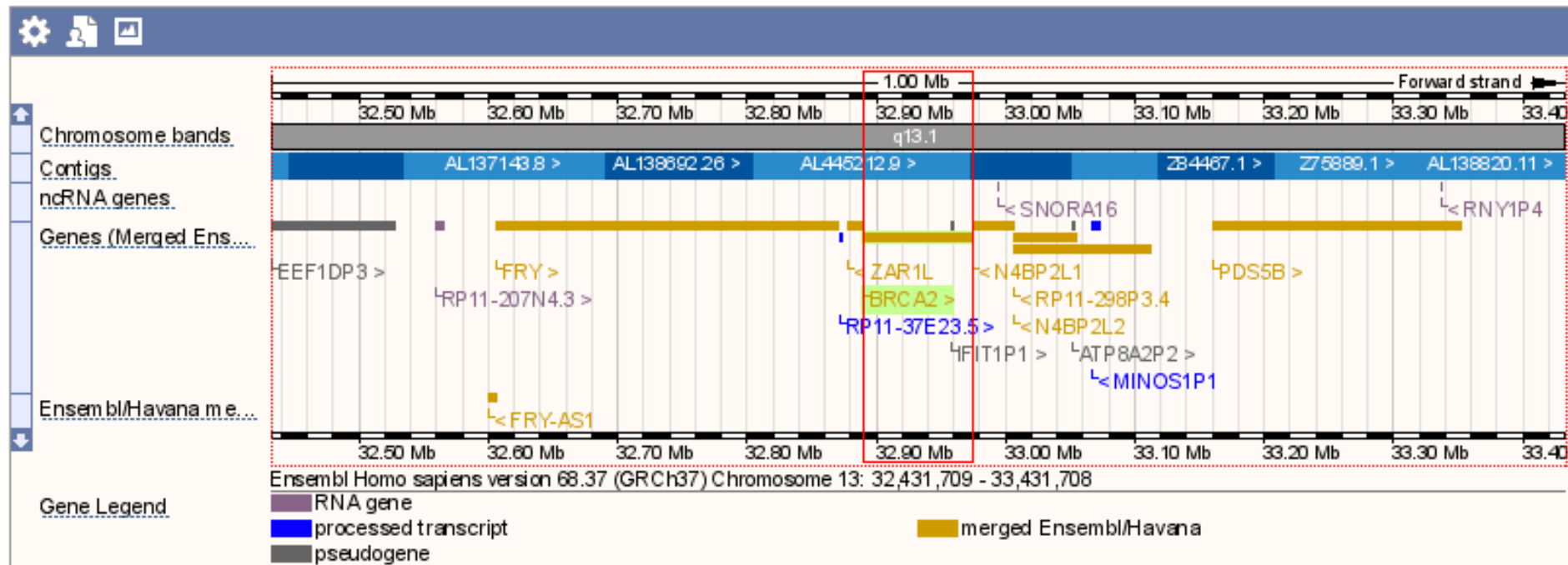
- ◉ Download genomic data
- ◉ Download annotation tracks
- ◉ Browse chromosomes, annotations and genomic data
- ◉ BLAT/BLAST: search for homologous DNA sequences
- ◉ Upload and plot custom data
- ◉ Upload custom annotation tracks

# Ensembl Genome Browser

## Chromosome 13: 32,889,611-32,973,805



## Region in detail



# COSMIC: Catalogue of Somatic Mutations



## Catalogue Of Somatic Mutations In Cancer

<http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/>

(new cosmic site)

What is it?

- ◉ Mutational data deposited from many sequencing studies
- ◉ Includes regular deposits from TCGA, ICGC

Uses:

- ◉ Determine if a specific mutation has been reported before, and in which cancers
- ◉ Does my mutation occur within a protein domain?
- ◉ Find recurrent (driver) mutations
- ◉ Download list of mutations for a gene (or whole genome)

# COSMIC Cancer Genomes



[Home](#) [About](#) [Download](#) [Publications](#) [News](#) [Contact](#) [Help](#)

<http://cancer.sanger.ac.uk/cancergenome/projects/studies/>

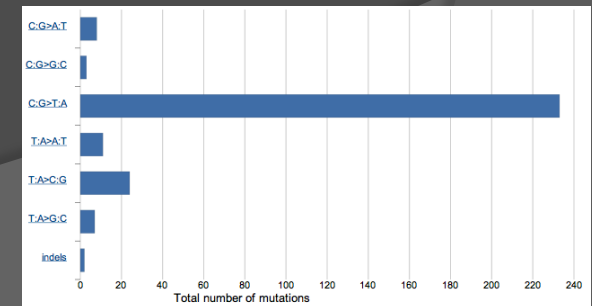
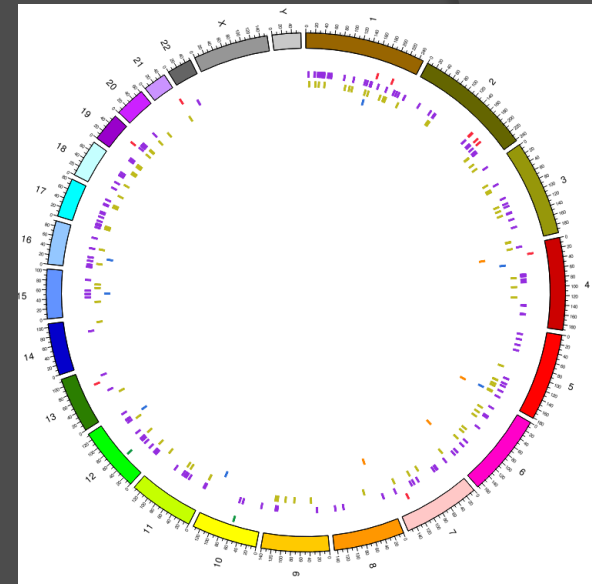
(new cosmic site)

What is it?

- Cancer genome sequencing data

Uses:

- View circos plots of mutations
- Look up transition and transversions mutation spectrums
- Identify mutations present in individual cancer genomes





# ICGC: International Cancer Genome Consortium



International  
Cancer Genome  
Consortium

Data  
Portal

<http://icgc.org>

What is it?

- Cancer genome data for 47 different cancer types
- DNA / RNA Sequencing
- Methylation
- Copy Number Changes

Uses:

- In which cancer types is my gene mutated?
- What classes of mutations occur in my gene?
- FTP download of copy number or expression data
- Controlled Data Access : Download Sequencing Data

# TCGA: The Cancer Genome Atlas

The Cancer Genome Atlas



*Understanding genomics  
to improve cancer care*

<http://cancergenome.nih.gov/>

What is it?

- Cancer genome data for 9 different cancer types
- DNA / RNA Sequencing
- Methylation
- Copy Number Changes

Cancers Types: Brain, Breast, GI, Gynecologic, Head & Neck, Hematologic, Skin, Thoracic, Urologic

Uses:

- In which cancers are my gene of interest mutated?
- What classes of mutations occur in my gene?
- FTP download of copy number or expression data
- Controlled Data Access : Download Sequencing Data

# GALAXY: Cloud-based Genomic Data Analysis Tool



<https://main.g2.bx.psu.edu/>

What is it?

- A collection of UNIX tools and programs for analyzing NGS and genomic data
- A cloud storage server for data analysis

Uses:

- Align sequences to genome
- Manipulate BED and annotation files
- RNA-seq analysis
- Chip-Seq peak calling
- Annotation of variants

...too many applications to mention

# Sequence Read Archive

## SRA

The Sequence Read Archive (SRA) stores raw sequencing data from the next generation of sequencing platforms including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

<http://www.ncbi.nlm.nih.gov/sra>

What is it?

- Data repository for next-generation sequencing files
- Most journals require that sequence data be deposited here before publication

Uses:

- Download FASTQ (raw sequence read files) from papers or studies

# OncoMine



<http://www.oncomine.org/>

What is it?

- Database and Data Analysis Tools for Cancer Samples
- Free for academics, but data cannot be downloaded
- Collection of Genomic Data (Copy Number, Expression and Methylation)

Uses:

- How frequently is my gene amplified or deleted in a cancer type?
- Which genes are most frequently overexpressed in breast cancer?
- Which genes expression correlation or anti-correlations occur in a cancer type?
- Does methylation of a gene result in decreased expression?

# Workshop for Today

<http://www.navinlab.com/biodb>

## Tutorials:

1. UCSC Genome Browser
2. COSMIC
3. Galaxy

Please finish what you can in 1 hour, then we will switch to Dr. Rehan Akbani's lecture and workshop on TCGA data analysis

As your homework assignment, please finish the workshop for today

# Homework Assignment #1

The Homework Assignment #1 will be uploaded to:

<http://www.navinlab.com/bioinfo/bioinfo/homework.html>

On Friday 9/21/2012

The assignment will be due in class on **October 2<sup>nd</sup>**

The format will be 20 questions that cover topics from the first module:

- ◉ Genomics
- ◉ UNIX
- ◉ Perl
- ◉ Biological Databases

Please feel free to contact me ([nnavin@mdanderson.org](mailto:nnavin@mdanderson.org)) or Dr. Yong Wang ([YWang33@mdanderson.org](mailto:YWang33@mdanderson.org)) if you have any questions or need help with the assignment