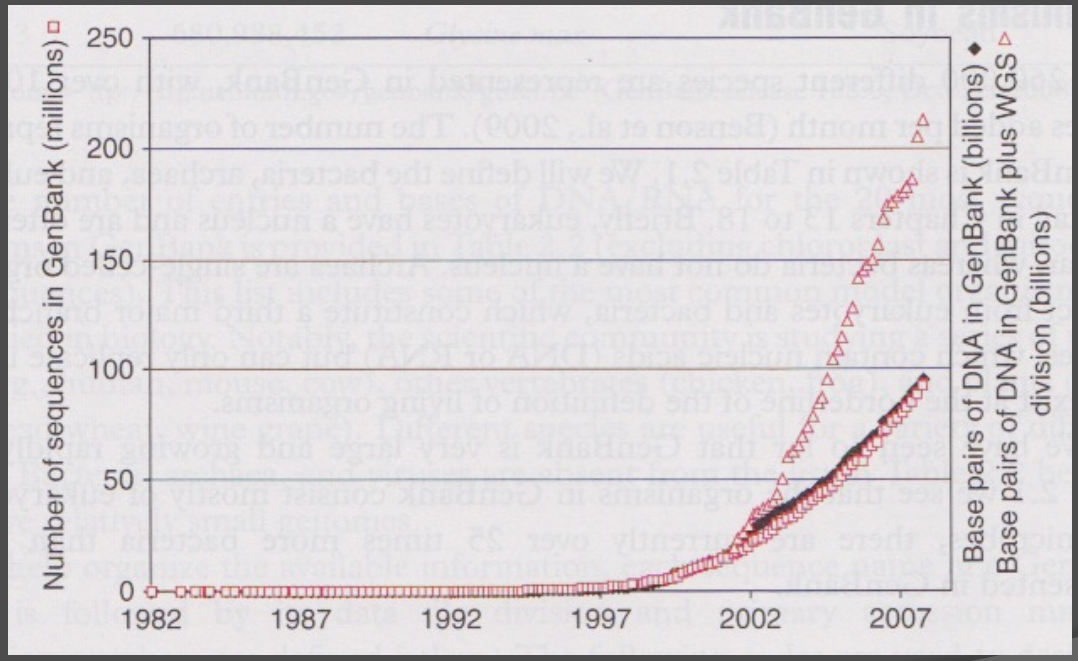
Biological Databases

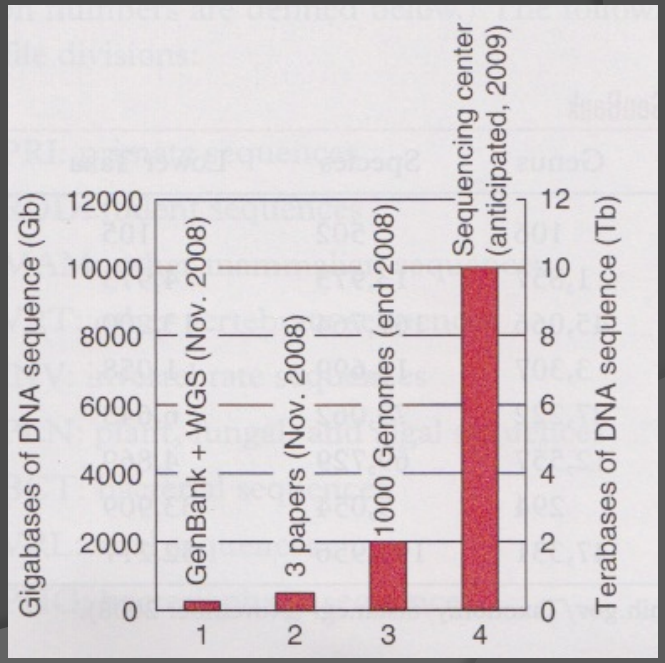
Nicholas E. Navin, Ph.D.
Department of Genetics
Department of Bioinformatics

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

Accessing Data from Databases

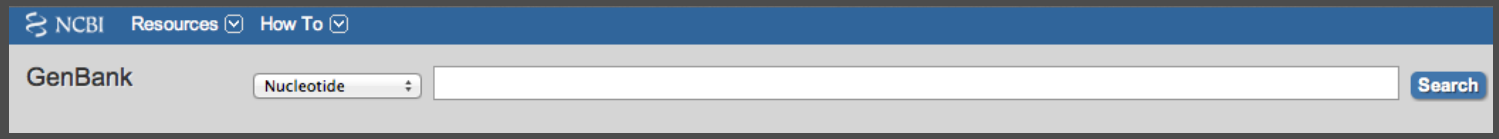
- Website Access: Search & Download
- Biomarts
- FTP
- Direct Connection from a Programming Language (MySQL, DAS, API)
- Direct data access from R or Matlab packages

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



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
GGCGGCGGCTGGAGGAGAGCGCGGTGGAGAGCCGAGCGGGCGGGCGGCGGGTGCGGAGCGGGCGAGGGAG
CGCGCGCGGCCGCCACAAAGCTCGGGCGCCGCGGGGCTGCATGCGGCGTACCTGGCCCGGCGCGGCGACT
GCTCTCCGGGCTGGCGGGGGCCGGCCGCGAGCCCCGGGGGCCCCGAGGCCGCAGCTTGCCTGCGCGCTCT
GAGCCTTCGCAACTCGCGAGCAAAGTTTGGTGGAGGCAACGCCAAGCCTGAGTCCTTTCTTCCTCTCGTT
CCCCAAATCCGAGGGCAGCCCGCGGGCGTCATGCCCGCGCTCCTCCGCAGCCTGGGGTACGCGTGAAGCC
CGGGAGGCTTGGCGCCGGCGAAGACCCAAGGACCACTCTTCTGCGTTTGGAGTTGCTCCCCGCAACCCCG
GGCTCGTCGCTTTCTCCATCCCGACCCACGCGGGGCGCGGGGACAACACAGGTCGCGGAGGAGCGTTGCC
ATTCAAGTGACTGCAGCAGCAGCGGCAGCGCCTCGGTTCCTGAGCCCACCGCAGGCTGAAGGCATTGCGC
GTAGTCCATGCCCGTAGAGGAAGTGTGCAGATGGGATTAACGTCCACATGGAGATATGGAAGAGGACCGG
```

GenBank Flat File (GBFF)

```
LOCUS       MUSNGH      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
NID         g1850791
KEYWORDS    neurite extension activity; growth arrest; TA20.
SOURCE      Murinae gen. sp. mouse neuroblastma-rat glioma hybridoma
            cell_line:NG108-15 cDNA to mRNA.
  ORGANISM  Murinae gen. sp.
            Eukaryotae; mitochondrial 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. Chihiro
            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:CHIHIRO@ms.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"
                     /cell_type="mouse neuroblastma-rat glioma hybridoma"
     misc_signal     156..163
                     /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 extensiion activity and growth arrest
                     effect"
                     /codon_start=1
                     /db_xref="PID:d1005516"
                     /db_xref="PID:g793765"
                     /translation="MMKLWVPSRSLPNSPNHYRSFLSHTLHIRYNNSLFISNTHLSRR
                     KLRVTNPIYTKKSSINIFYLLIPSCRTRLILWIIYIYRNLKWNSTSTVRSHSHSIYRL
                     RPSMRTNIILRCHSYYKPPISHPIYWNNPSRMNLRGLLSRQSHLDPILRFPLHLTIYY
                     RGPSNRSPPLPPNNRIKQPNRIKLRCR"
     polyA_site      1803
BASE COUNT      507 a    458 c    311 g    527 t
ORIGIN
        1 tcagtttttt ttttttttta ttttttttta ttttttttta ttttttttta ttgattcatg
       61 tccgtttaca tttggtaagt tcacaggcct cagtcaacac aattggactg ctcaggaaat
      121 cctccttggt gaccgcagta tacttggcct atgaacccaa gccacctatg gctaggtagg
      181 agaagctcaa ctgtagggct gactttggaa gagaatgcac atggctgtat cgacatttca
      241 catggtggac ctctggccag agtcagcagg ccgagggttc tcttccgggc tgctccctca
      301 ctgcttgact ctgcgtcagt gcgtccatac tgtgggcgga cgttattgct atttgccttc
      361 catttctgtac ggcattgcct ccatttagct ggagagggac agagcctggt tctctagggc
      421 gtttccattg gggcctggtg acaatccaaa agatgaggcc tccaaacacc agaatcagaa
      481 ggcccagcgt atttgtaaaa acacctttctg gtgggaatga atggtacagg ggcgtttcag
      541 gacaaagaac agcttttctg tcactcccat gagaaccgtc gcaatcactg ttccgaagag
      601 gaggagtcca gaatacacgt gtatgggcat gacgattgcc cggagagagg cggagcccat
      661 ggaagcagaa agacgaaaaa cacacccatt atttaaaatt attaacccact cattcattga
      721 cctacctgcc ccatccaaca tttcatcatg atgaaacttt gggtcccttc taggagtctg
      781 cctaatagtc caaatcatta caggtctttt cttagccata cactacacat cagatacaat
      841 aacagccttt tcatcagtaa cacacatttg tcgagacgta aattacgggt gactaatccg
      901 atatatacac gcaaacggag cctcaatatt ttttatttgc ttattccttc atgtcggacg
      961 aggcttatat tatggatcat atacatttat agaaacctga aacattggag tacttctact
     1021 gttcgcagtc atagccacag catttatagg ctacgtcctt ccatgaggac aaatatcatt
     1081 ctgaggtgcc acagttatta caaacctcct atcagccatc ccatatattg gaacaacctt
     1141 agtcgaatga atttgagggg gcttctcagt agacaaagcc accttgaccc gattcttcgc
     1201 tttccacttc atcttaccat ttattatcgc ggccctagca atcgttcacc tcctcttcct
     1261 ccacgaaaca ggatcaaaca acccaacagg attaaactca gatgcagata aaattccatt
     1321 tcacccctac tatacatcaa agatatccta ggtatcctaa tcatattctt aattctcata
     1381 accctagtat tatttttccc agacatacta ggagacccag acaactacat accagctaat
     1441 ccactaaaca ccccacccca tattaaaccc gaatgatatt tcctatttgc atacgccatt
     1501 ctacgctcaa tccccaataa actaggaggt gtcctagcct taatcttatc tatcctaatt
     1561 ttagccctaa taccttttct tcatacctca aagcaacgaa gcctaatatt ccgcccaatc
     1621 acacaaattt tgtactgaat cctagtagcc aacctactta tcttaacctg aattgggggc
     1681 caaccagtag acacccattt attatcattg gccaactagc ctccatctca tacttctcaa
     1741 tcatcttaat tcttatacca atctcaggaa ttatcgaaga caaaatacta aaattatatc
     1801 cat
//
```

Header

- Title
- Taxonomy
- Citation

Features (AA seq)

DNA Sequence

UCSC Genome Browser

http://genome.ucsc.edu

Uses:

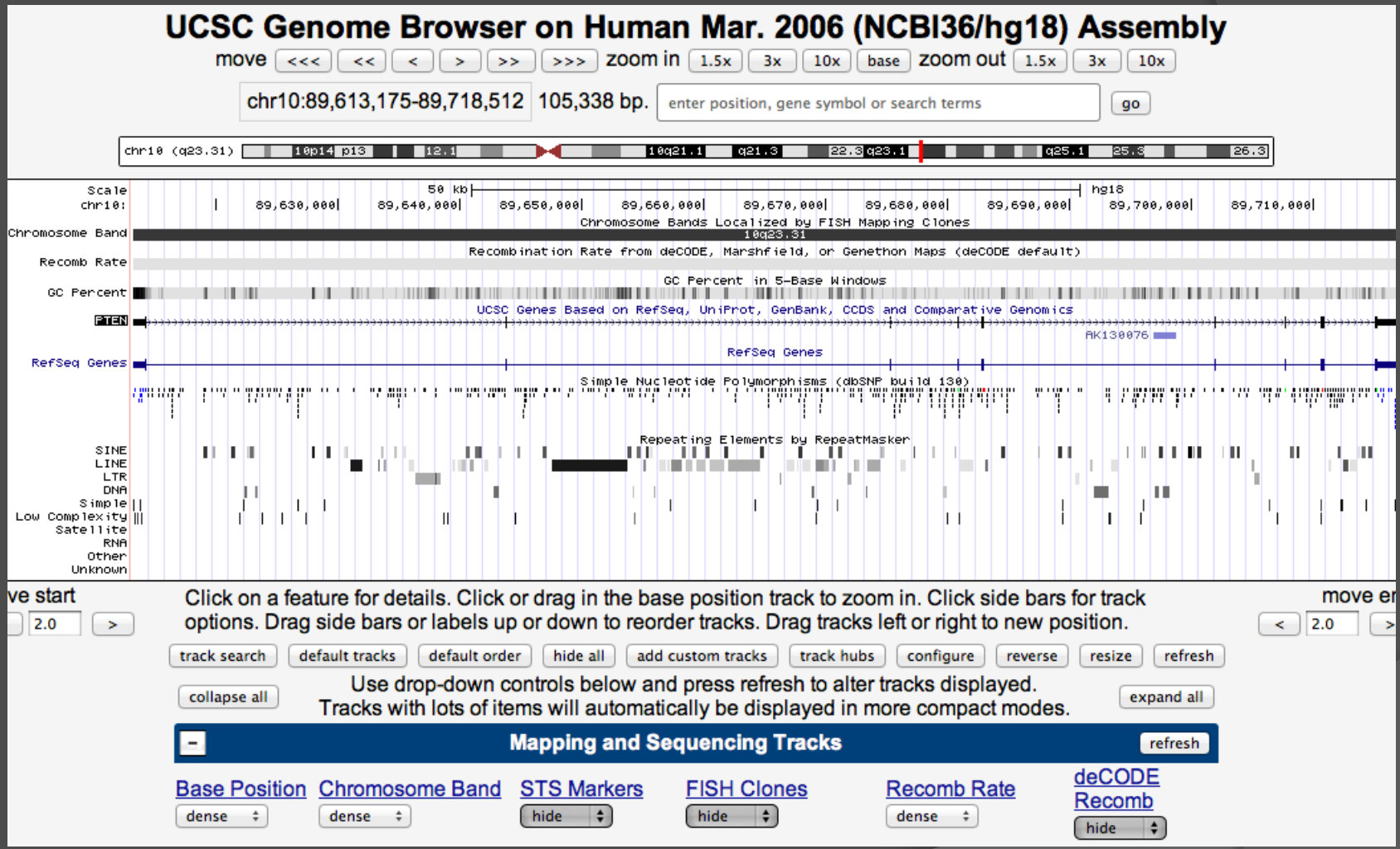
- Download genomic data
- Download annotation tracks
- Browse chromosomes, annotations and genomic data
- BLAT: search for homologous DNA sequences
- Upload and plot custom data
- Upload custom annotation tracks

UCSC Genome Browser

Chrom position

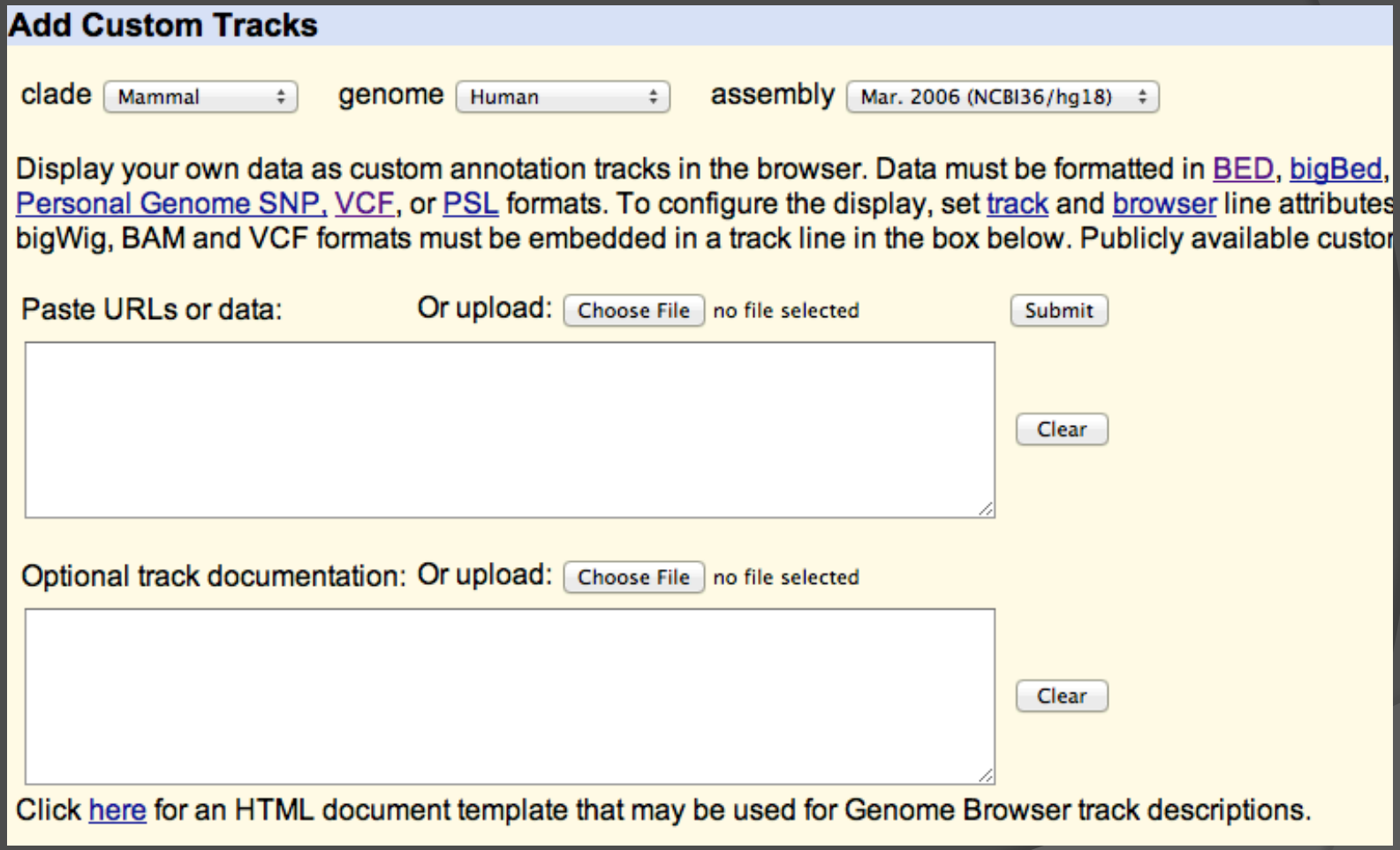
tracks

Track configu rations



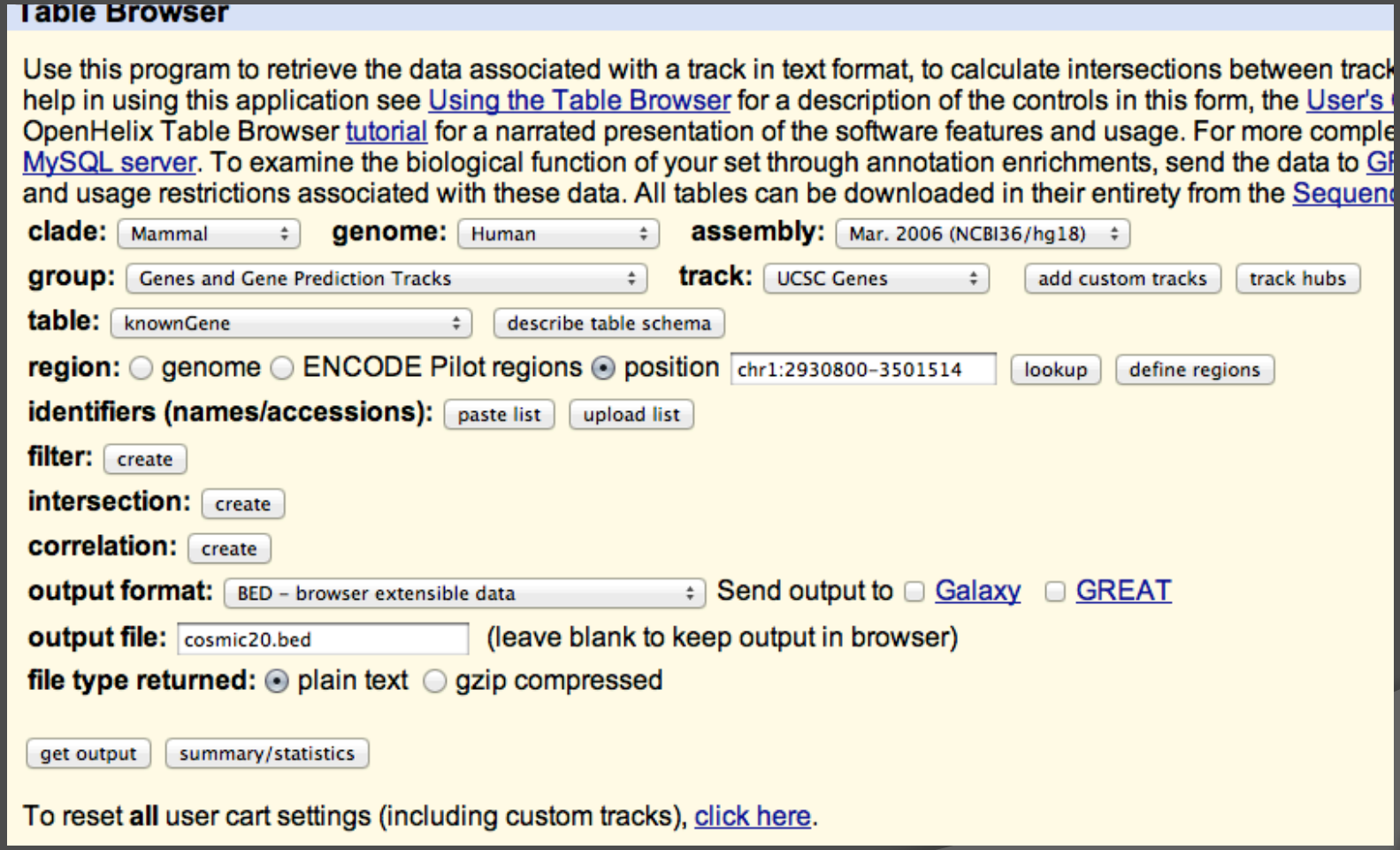
UCSC Genome Browser

Upload Custom Annotation Tracks or Your Own Data

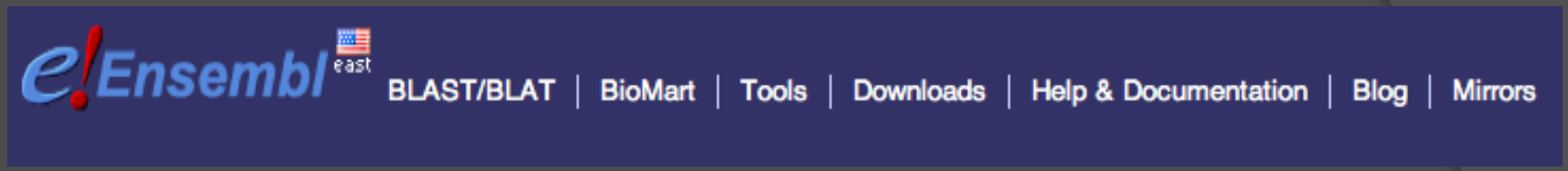


UCSC Genome Browser

Download Annotation Tracks



Ensembl Genome Browser



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

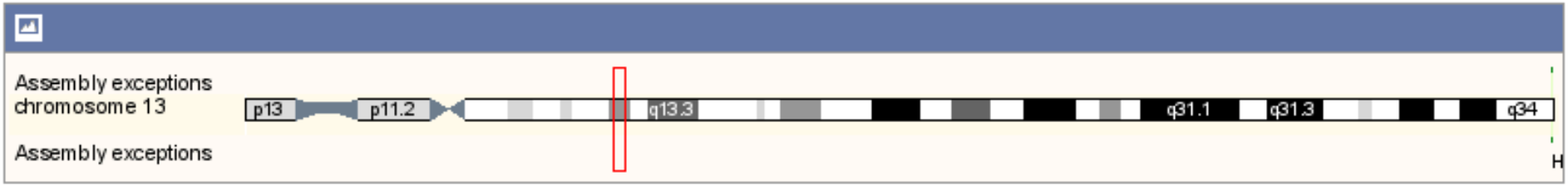
Identifiers:
ENSG00000000133703 gene
ENST00000000530893 transcript
ENSP00000000369497 protein

Uses:

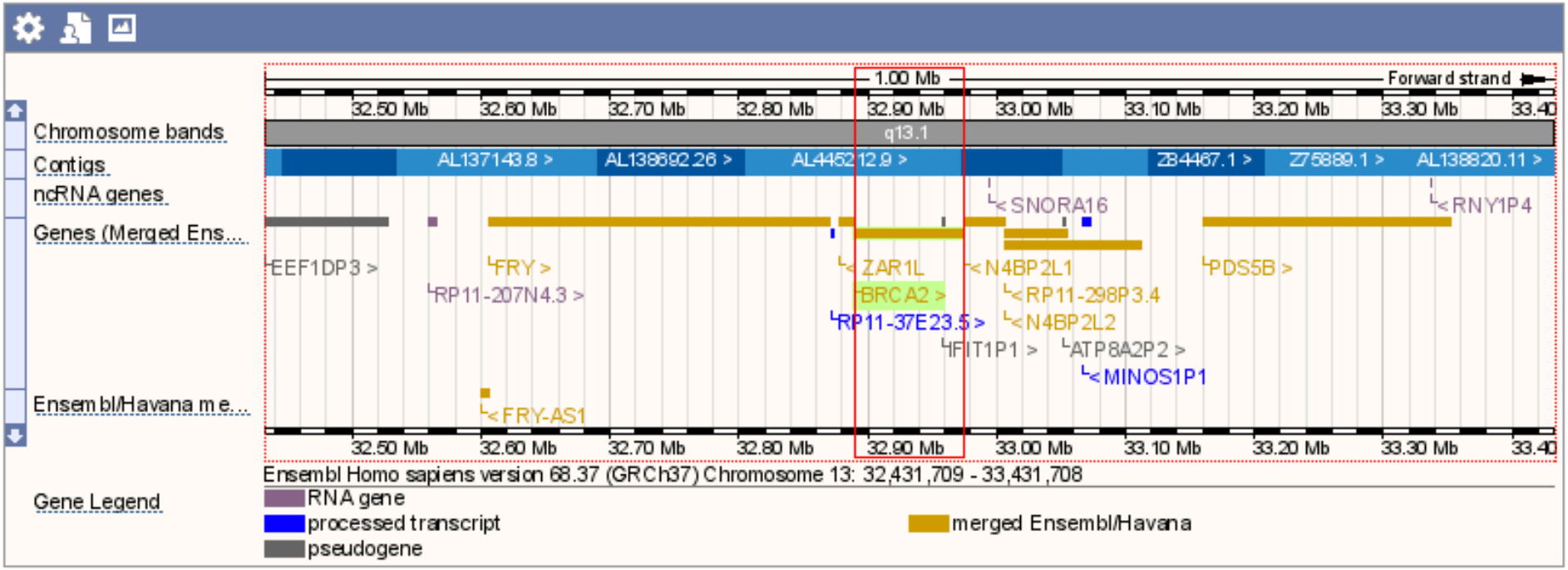
- Download genomic data
- Download annotation tracks
- Browse chromosomes, annotations and genomic data
- BLAT/BLAST: search for homologous DNA sequences
- Upload and plot custom data
- Upload custom annotation tracks

Ensembl Genome Browser

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



Region in detail



COSMIC: Catalogue of Somatic Mutations

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

COSMIC Genomes
Catalogue of somatic mutations in cancer

Home | About | Download | Publications | News | Contact | Help

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

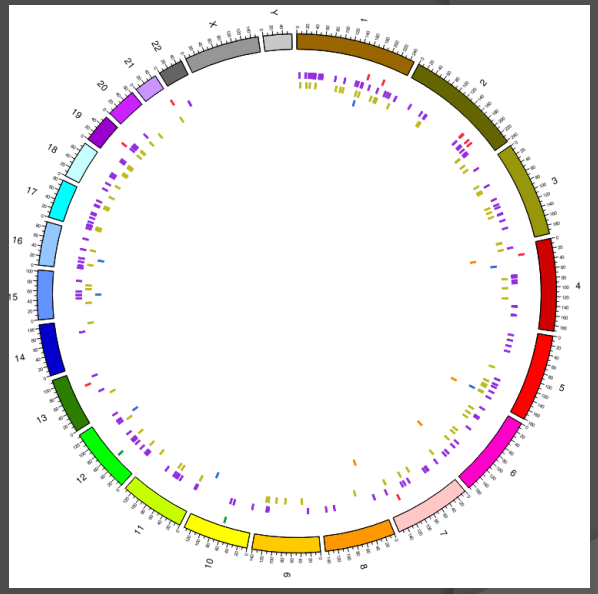
(new cosmic site)

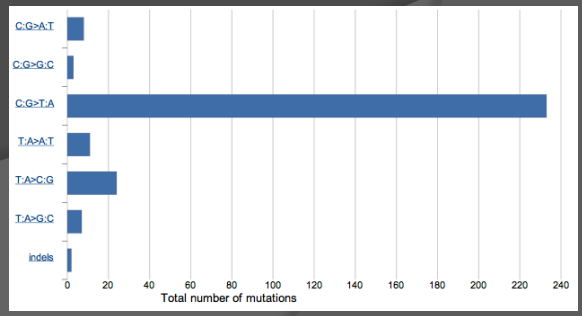
What is it?

- 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

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

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

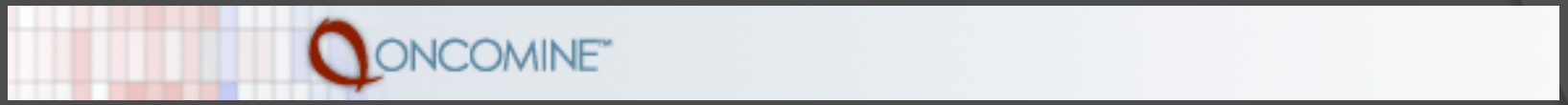
What is it?

- 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 2nd**

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) or Dr. Yong Wang (YWang33@mdanderson.org) if you have any questions or need help with the assignment