

# Genome and DNA Sequence Databases

BME 110/BIOL 181

CompBio Tools

Todd Lowe

March 31, 2009

# Admin

- Reading:
  - Chapters 1 & 2
- Notes available in PDF format on-line (see class calendar page):

<http://www.soe.ucsc.edu/classes/bme110/Spring09/BME110-calendar.html>

# Class Objectives

- Use bioinformatics to investigate biological problems
- Identify bioinformatics resources online and/or available for local use
- Understand bioinformatics methods in order to use tools appropriately for a problem of interest
- Interpret significance of results
- Generate quality figures to illustrate biological hypotheses or conclusions

# Today's topics

- Basic rules for doing bioinformatics on the web
- Types of Databases
- Searching PubMed for scientific papers
  
- Sequencing
- Basic tools for genomics
- Genome Databases

# Bioinformatics on the Web

- Golden Rules:
  - Use published databases and methods
    - Supported and maintained
    - Trusted by community
  - Document what you've done
    - Sequence identification numbers
    - Server, database, program versions
    - Program parameters
  - Assess reliability of results
    - Understand and use reported confidence measures
    - Compare results of multiple servers
    - Do results support/conflict with other available data?

# Most Basic Database: Sequence Repositories

- Three major sequence repositories
  - NCBI
    - National Center for Biotechnology Information
    - [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)
  - EBI
    - European Bioinformatics Institute
    - [www.ebi.ac.uk](http://www.ebi.ac.uk)
  - DDBJ
    - DNA Data Bank of Japan
    - [www.ddbj.nig.ac.jp](http://www.ddbj.nig.ac.jp)
- Same sequence information in all three
- Different tools for searching and retrieval

# Major NCBI Databases

- Genbank / Entrez – nucleotide / protein search
- PubMed / Medline – journal publication search
- Genomes – full genomes, info & sequence
- Taxbrowser – full species' taxonomy
- GEO – Gene Expression Omnibus
- Many other smaller db's...

# NAR Database Index

- Great collation of biological databases, Nucleic Acids Research Database Issue

<http://www3.oup.co.uk/nar/database/c/>

More than 1000 Databases (!!)

(+110 in last year)

Sorted alphabetically & by category

Books date quickly, use on-line collections like this (or Google) to get most current information

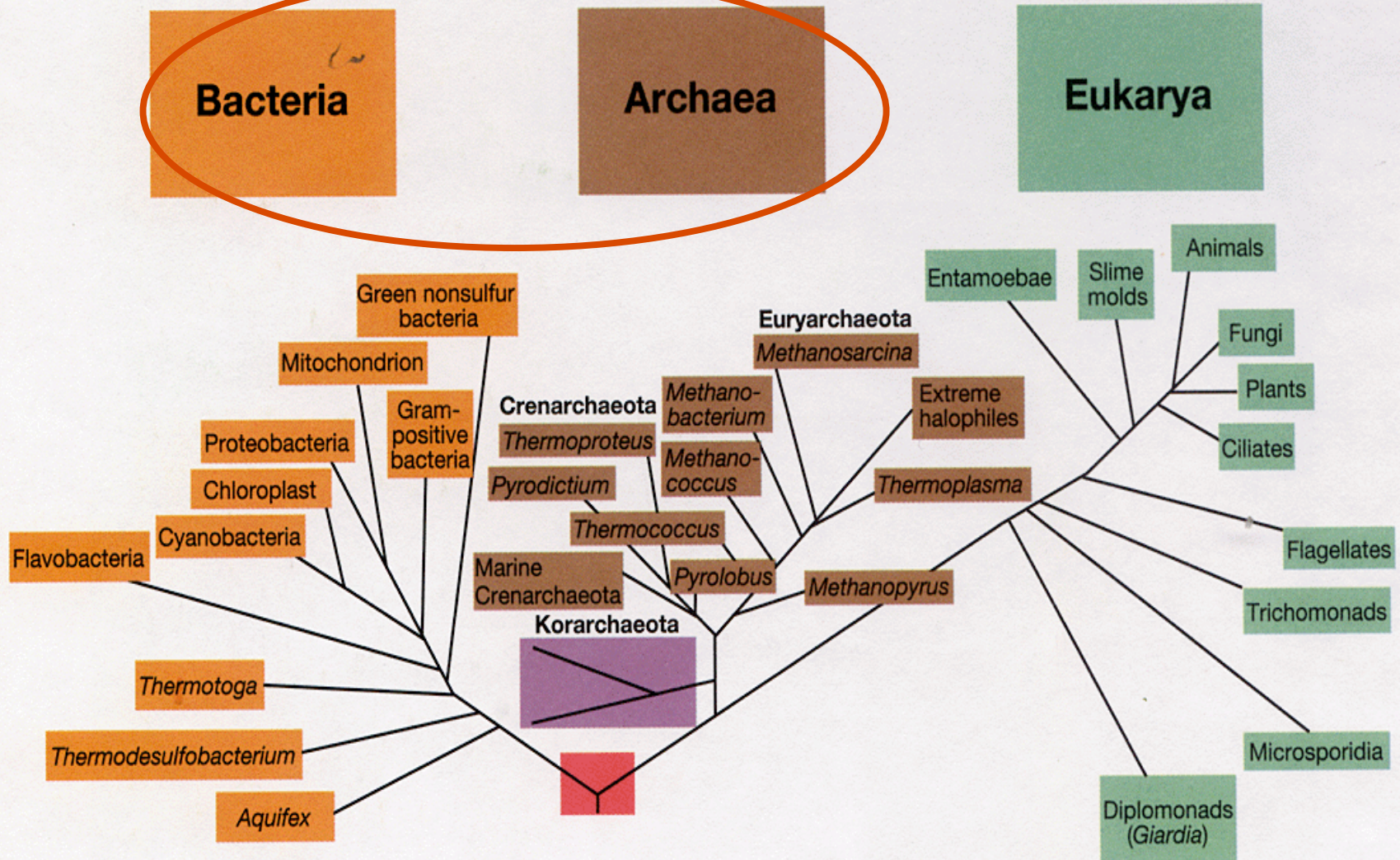
# Literature: PubMed @ NCBI

- What-- access to Medline
  - Primarily biomedical, molecular biology & biochemistry journals
- Searching – Entrez search engine
  - Logical operators
  - Field Delimiters
  - What's related
  - PMIDs
  - Returns article title, authors, & abstract
- PubMed Central – free access to many *full-text* scientific papers
- UCSC Electronic Journal access
- Special procedure for getting access to journal articles off-campus

# Genes & Genomes

# Three Domains of Life

## Prokaryotes



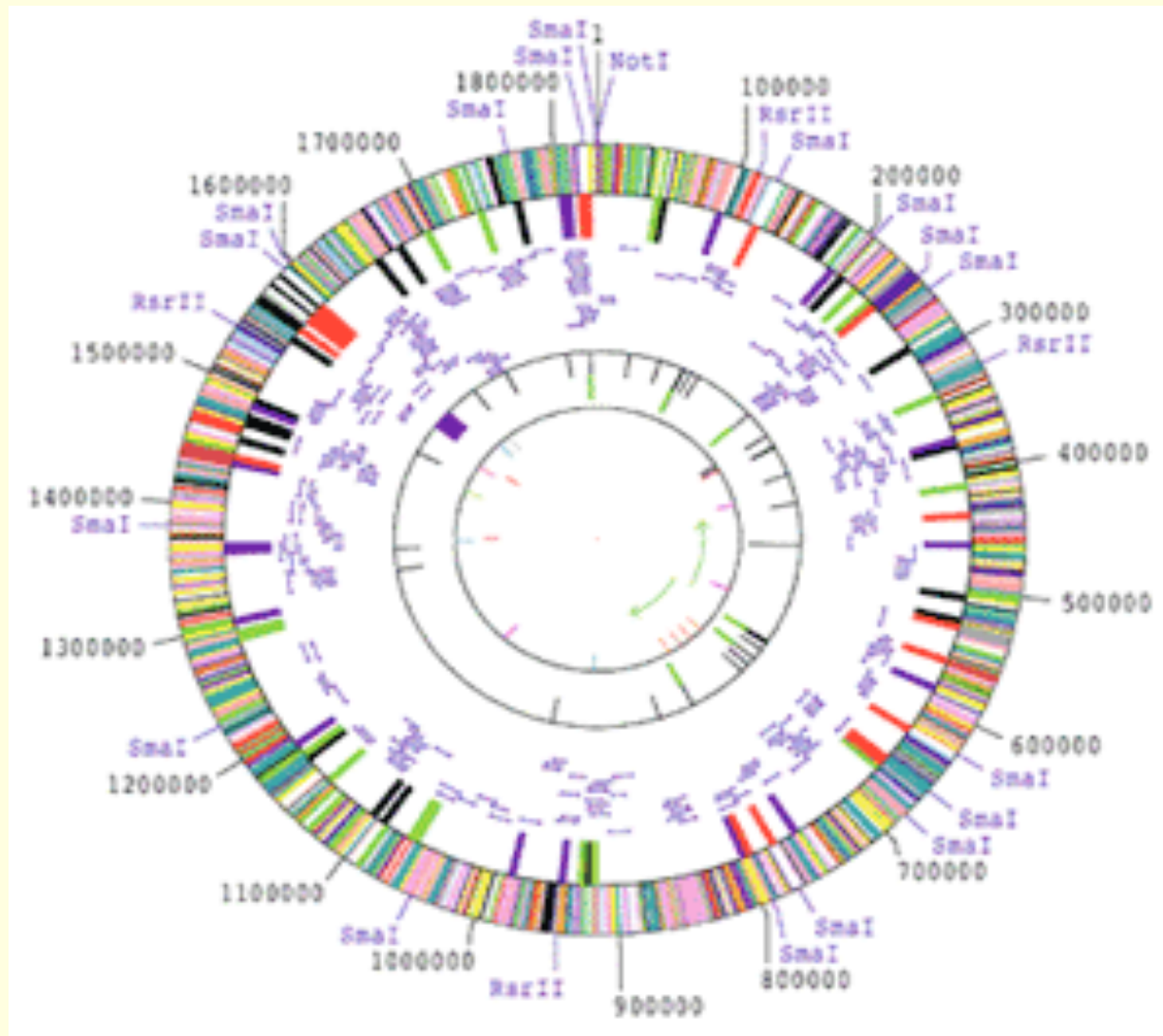
# What is a Genome?

- A complete set of instructions for life encoded in DNA
- Organized in chromosomes
  - prokaryotes generally have one main circular chromosome
  - eukaryotes have multiple linear chromosomes
- Instructions are generally in the form of **genes**, and are the “unit” of heredity

# Why Sequence a Genome?

- We wish to understand how the entire cell / organism works
  - thousands of complex gene interactions
  - complete “parts list” is first step to understanding how parts work together as a whole
- Economy of scale - faster, cheaper to sequence all genes at once, than one at a time by many different researchers

First fully sequenced Organism: *H. Influenza*  
The Institute for Genome Research (TIGR) - 1995



# Human Genome Project

- In 1980's, initial discussion to sequence human genome (first key meeting here at UCSC!)
  - Began: 1990
  - Planned finish: 2005
- Original Estimates:
  - ~100,000 genes
  - 3 billion nucleotides, projected cost \$300 Million
  - less than 5% of genome codes for genes
  - Early opposition to sequencing 95% “junk”, taking money away from basic research
- Final Results:
  - Draft completed June 2000, “Finished” April 2003
  - Final cost ~ \$3 billion
  - ~25,000-30,000 genes now
  - Up to 50% of genome is transcribed with possible biological function

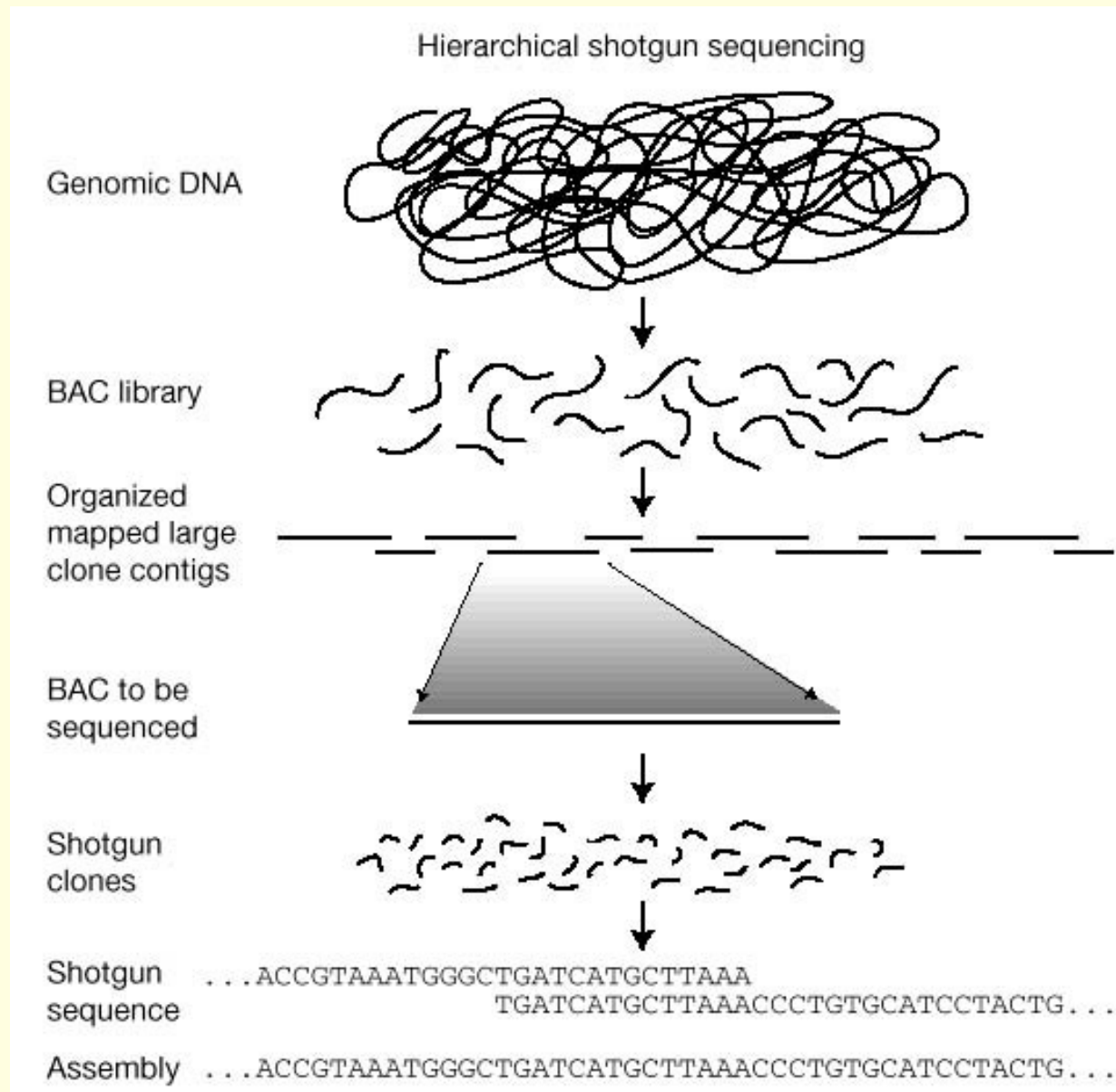
# Preparing for the Human Sequencing...

- Mapping the human genome
- Practice: sequencing model organisms

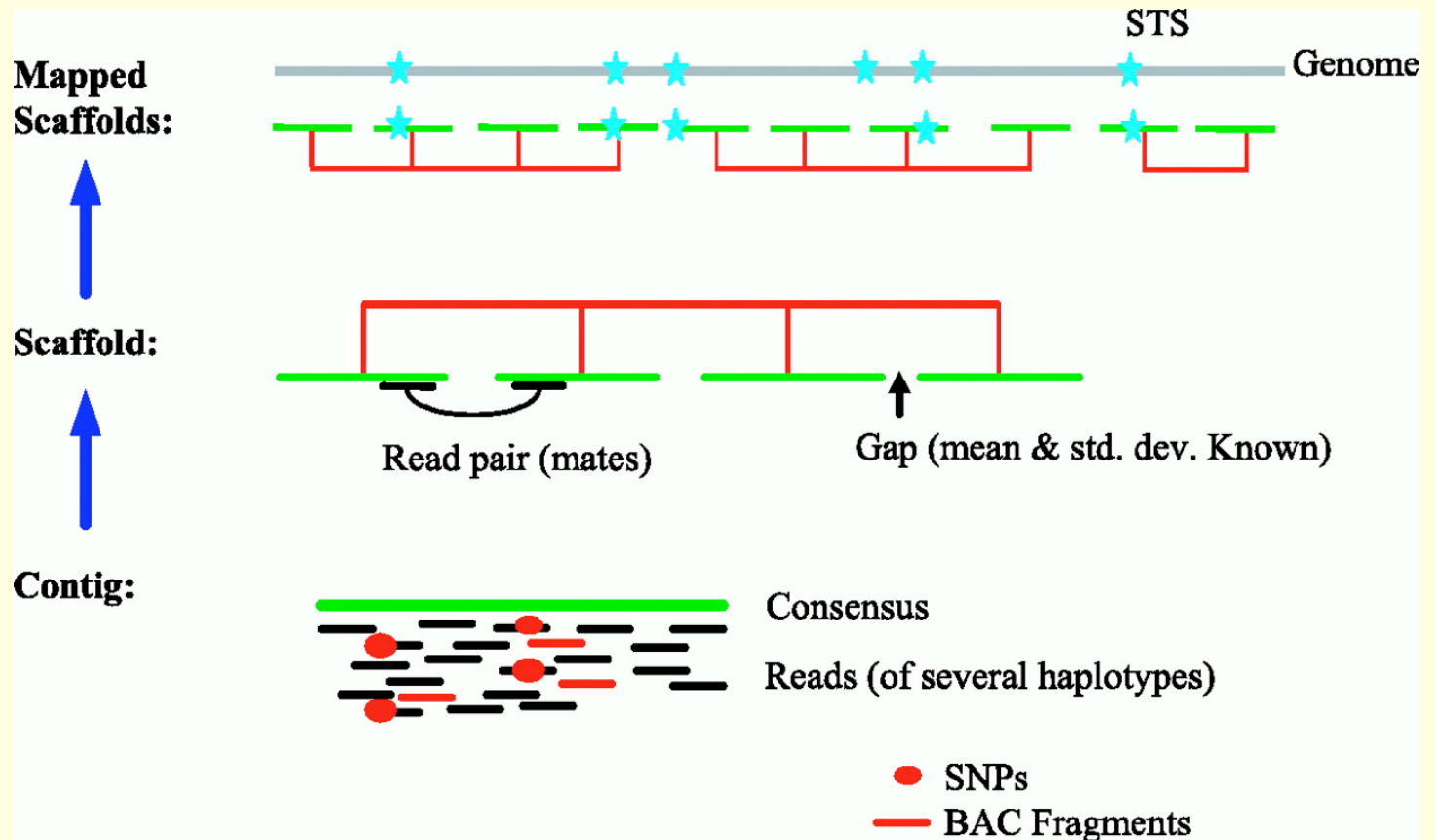
	Size (Mb)	Finished
– Baker's yeast ( <i>S. cerevisiae</i> )	12.5	1996
– <i>E. coli</i>	4.5	1997
– Roundworm ( <i>C. elegans</i> )	100	1998
– Fruit fly ( <i>D. melanogaster</i> )		160
2000		
– Plant weed ( <i>A. thaliana</i> )	120	2000

# Sequencing Genomes: Strategy #1

- Original “Top-Down” Strategy
- Deliberate, small chance for major errors



## Sequencing Genomes: Strategy #2



WGS (Whole Genome Shotgun) – “Bottom Up” Strategy

- No ordering of a clone library - straight to sequencing to build “scaffolds”
- Mapping data (STS’s) used to place scaffolds

# Race to the “Finish” – June 25, 2000

- Change in public’s strategy to assemble meant ordered clone libraries would not be finished
- Alternate method needed for putting the pieces together
- UC Santa Cruz to the rescue!
  - David Haussler & Jim Kent, in just a few months, created a clever program and network of 100 Dell desktop computers to assemble the genome by Celera’s target date

# Sequencing Technologies

## Established (“Long” reads)

Capillary sequencing: ABI 3700 and MegaBASE machines allow first rapid automated sequencing in tiny capillary tubes

- 600-800 bases at a time x 384 wells = ~ 268,000 bases decoded / run
- Bulk of human genome decoded on these

## New Technologies (Massively parallel, short reads)

454 / Roche Sequencers 100-350 bases, 100 Million bases / run

Solexa / Illumina Genetic Analyzer: ~35 bases, >1.3 Billion bases / run

ABI SOLiD Sequencing: 35 base pieces, >3 Billion bases / run

Chemical decoding is no longer major cost – re-assembling all the pieces with computers *is* new major cost

# Current Genome Challenges

- \$1,000 Genome Prize – new goal set in 2003 by J. Craig Venter Science Foundation to make personalized medicine available to all (\$500,000 prize)
- Archon X Prize – build a machine that can sequence 100 people in 10 days or less, for \$10,000/genome or less (\$10 Million prize): <http://genomics.xprize.org/>

# Computational Biology Tools

- Human Genome: 3 billion A/C/T/G's  
staring at you -  
Now what??
- Have we seen this sequence before?  
Similarity searching
- Discovering completely new genes  
The "Gene Finders"
- Finding a role for each new gene:  
Functional genomics

# One sequence “read” from the human genome:

>ctg14072

```
CATGGAAACCCCANAAAAACATGAAATGCATACCGAACTACAAAAAAGG
AAAATAAATATAAACACATTCCTAAAACCTTAAAAATGAAGGAGATTCAGA
CAGTCCCTCCTGGTAAAATGTGAAATTGCACCCAGCTGCAGCAGCTACT
GTAAATATCCAAGGAATCAGTTTTTAAGTGTTTGGGGATCCCAGGGATCCC
TGCAAAGCACTCAGGATTTTAAACATTAAGCTCACAAATTACAGCAGCTGG
CCGGGCACAGTGGCTCACGCCCGTAATCCCAGCACTTTTGGAGGCCGAGG
CAGGTGGATCACCTGAGGTCTCCACTAAAAATACAAAAAACTAGCCAGGG
TGTGTGGCGGACATCTGTAATCCCAGCCACTTCGGGGGCTGAGGCAGGAG
AATCACTTGAACCCGGGAGGTGGAGGTTGCATTGAGCTGACGTTATGCCA
TTGCACTCCGGCCTGNGCAACAGAGAGAACTTCATCTCTAACTACTAAT
TACAGCAACCAACAGGCCTCTAGGTTAGTTACCACCCTAACCTTTTCGTT
CGAGATTTTCAAACCACCTTGAACGTGGGTATTTTTTTGTGGGTCCTTTAT
CTTCATTCATTAATCACATTATCAGACATTCCCTGAGTGGCCTGGTTCTG
TATACATGCTGAAGCTTCCAAATCAACCGTCCGTTTGGCTTCCCACAAC
```

- Where are the genes?

# Annotating a Genome

## Goals:

1. Note the positions of any known or predicted genes
2. Give as much information about function of genes as possible (and certainty of information)

Purpose is to help biologists make connections between their work and the sequence you are annotating

# What's in a Genome?

1. Genes
2. Introns (eukaryotes only)
3. Gene regulatory sequences (promoters, enhancers, silencers, suppressors)
4. Structural elements (telomeres, centromeres)
5. Repetitive DNA
6. Pseudogenes
7. Everything else (important??)

# Two Types of Genes

## 1) Protein-coding

- Purpose: make proteins
- Common Pattern:  
[Start codon] [codon 1] [codon 2] [...] [Stop codon]

## 2) Non-protein Coding RNAs (ncRNAs):

- Purpose: make a functional RNA
- No common pattern shared by all ncRNAs

# Protein Gene Finders

- Based on statistical analysis of DNA for
  - Start codon [ATG] (sometimes TTG/GTG)
  - Stop codon [TGA], [TAG], [TAA]
  - Codon “frequencies”
  - Splice junctions - sequence motifs (patterns) at borders between exons and introns

# Differences Between Eukaryotic and Prokaryotic Genes/Genomes

- Prokaryotes (Bacteria + Archaea)
  - Usually no introns, so no need to detect introns; ORFs ~ genes
  - Multiple genes are often organized into *operons* –functionally related genes that are co-transcribed in one long mRNA
  - Usually a single large circular chromosome (0.5-5 Mbp); often with some small circular DNA elements called *plasmids*
  - 70-95% of genome codes for genes
- Eukaryotes
  - Genes broken up into “exons” and spliced out “introns”; complicates accurate gene prediction
  - Generally, minority of DNA in genome codes for genes
  - Multicellular eukaryotes have much more complex regulation

# Graphical Genome Browsing

- NCBI – Source for sequences and “default” annotation for all genomes, as well as browsers  
[www.ncbi.nlm.nih.gov/](http://www.ncbi.nlm.nih.gov/)
- UCSC Genome Browser – human & many other genomes  
[genome.ucsc.edu](http://genome.ucsc.edu) AND [archaea.ucsc.edu](http://archaea.ucsc.edu)
- Ensembl – major European human genome browser, similar goals as UCSC browsers  
[www.ebi.ac.uk/ensembl/](http://www.ebi.ac.uk/ensembl/)
- JGI’s IMG – Newest, fanciest microbial genome browser and genome comparison tools  
<http://img.jgi.doe.gov/cgi-bin/pub/main.cgi>

# What Genomes Are Available?

List of completed genomes increases almost every week

GOLD Website: Listing of finished and “in progress” genomes

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

961 published, completed genomes

4600+ genome sequencing projects in progress

# Getting Sequences: NCBI

- Go to <http://www.ncbi.nlm.nih.gov/>
- Choose “Nucleotide” or “Protein”
- Type in query, same rules as for making PubMed queries (logical operators, limits, etc)
- Or, to get a genome, choose “Genome Project” and type species name

# Genbank File Format

- Completely annotated sequence

```
LOCUS NC_000854 1669695 bp DNA circular BCT 07-APR-2003
DEFINITION Aeropyrum pernix, complete genome.
ACCESSION NC_000854
SOURCE Aeropyrum pernix
ORGANISM Aeropyrum pernix Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales;
        Desulfurococcaceae; Aeropyrum.
FEATURES Location
            /Qualifiers source 1..1669695
            /organism="Aeropyrum pernix"
            /db_xref="taxon:56636"
gene complement(213..938)
            /gene="APE0001"
CDS complement(213..938)
            /gene="APE0001" /codon_start=1
            /product="hypothetical protein"
BASE COUNT 360022 a 473378 c 466849 g 369446 t
ORIGIN
1   aaataataat aaaaattaag tgactcatgc attatcctac gaggtaaaaa tatggtataa
61   attgtcccag actaccatca atttagggac aatagtgttt aagggatggc cttcggagct
121  ggcagctcgc gggttcaaac tcgcgtaggg cccgagttct agttatagtt gcggtggattt
. . .
```

# FASTA File Format

- Mostly sequence, little description
- General format often used for web server analysis tools

```
>Seq_name      Description                               (all on first line)
AGTACGGACCAGACAGGCCGATAGGACG
AGGCCGATAGGACGAGGCCGATAGGACG
CGTTA
>Next_seq_name  Description
ACCGATTACCGA
```

# Download A Whole Genome!

- NCBI Genomes

`ftp://ftp.ncbi.nih.gov/genomes/`

- What do all these files mean?

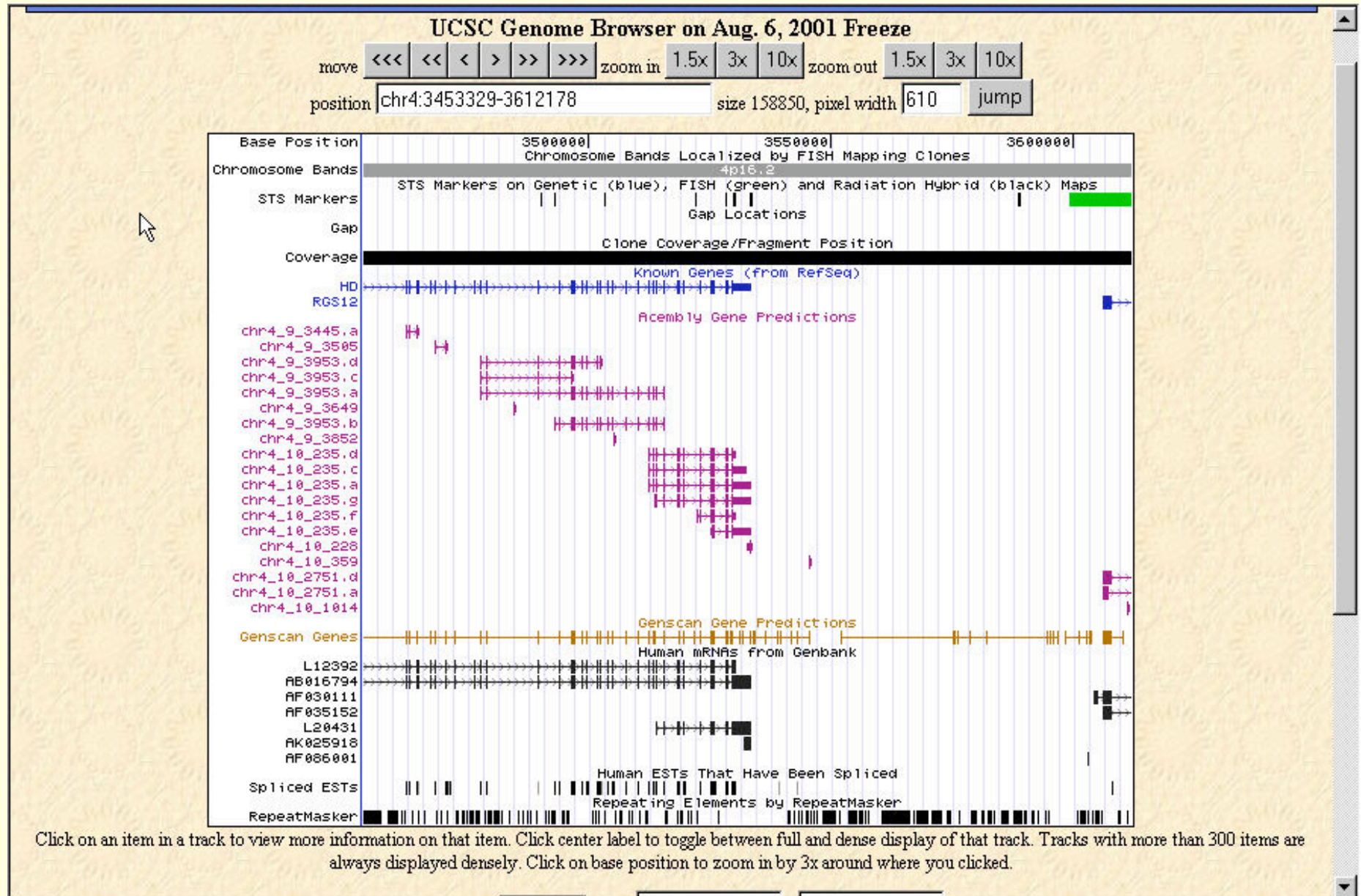
“.faa” FASTA format predicted proteins – all predicted proteins

“.fna” FASTA predicted genes – all predicted genes (DNA)

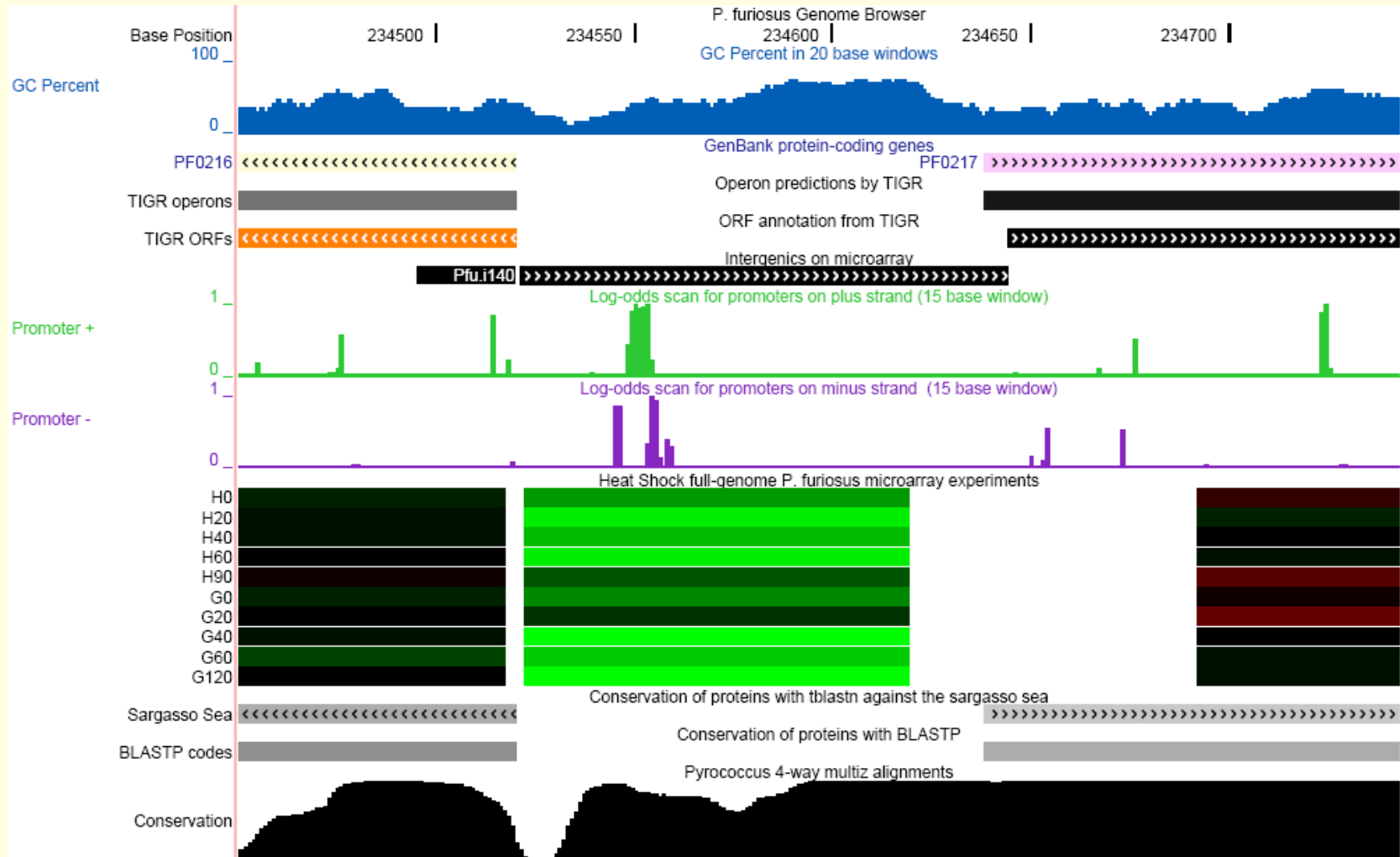
“.fna” FASTA format, whole genome (DNA)

“.gbk” Genbank format, whole genome with complete annotation

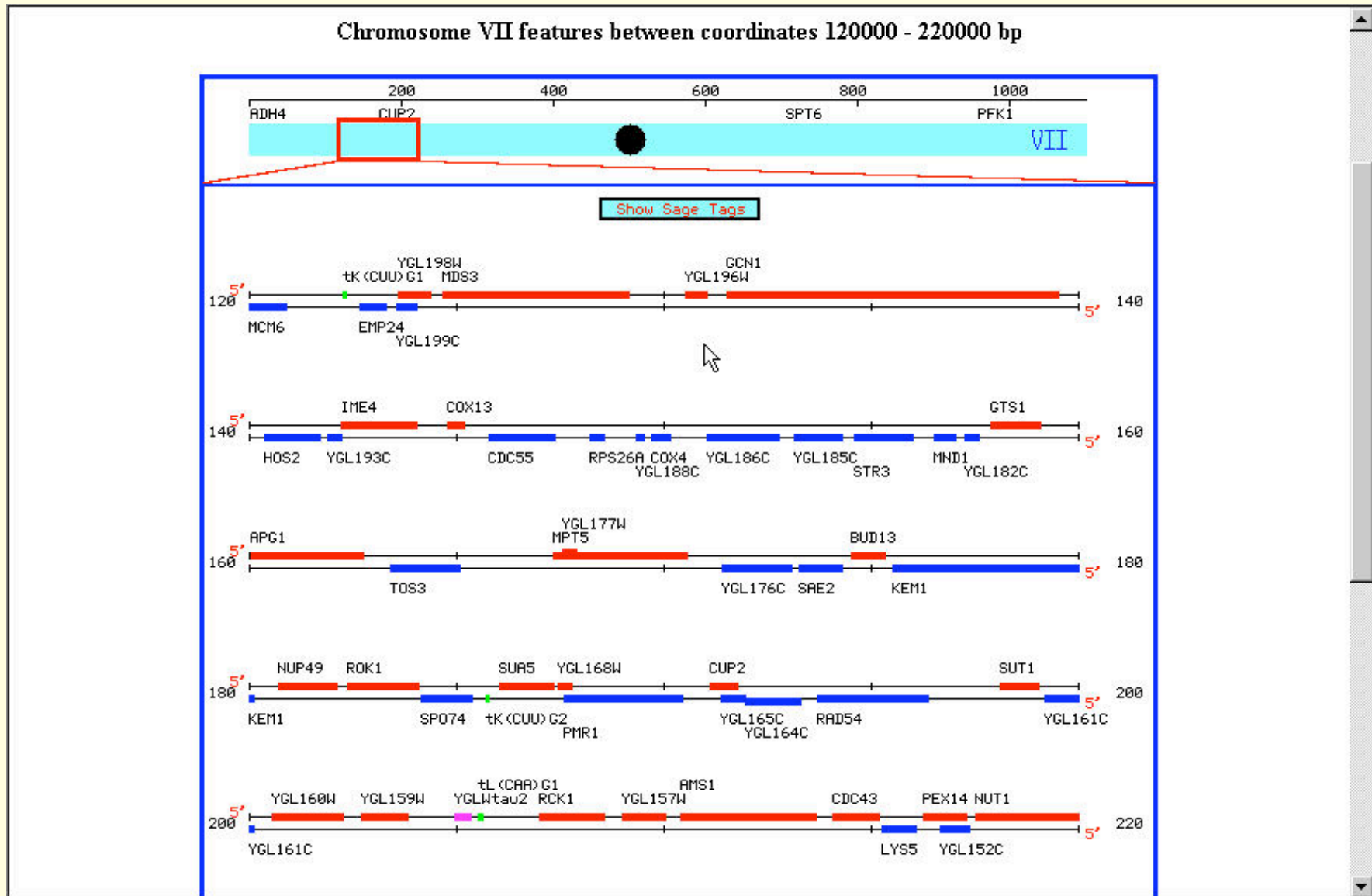
# Graphical Displays of Genome Information



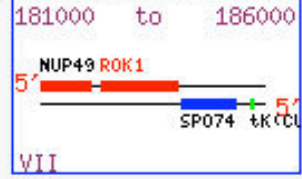
# UCSC Archaeal Genome Browser



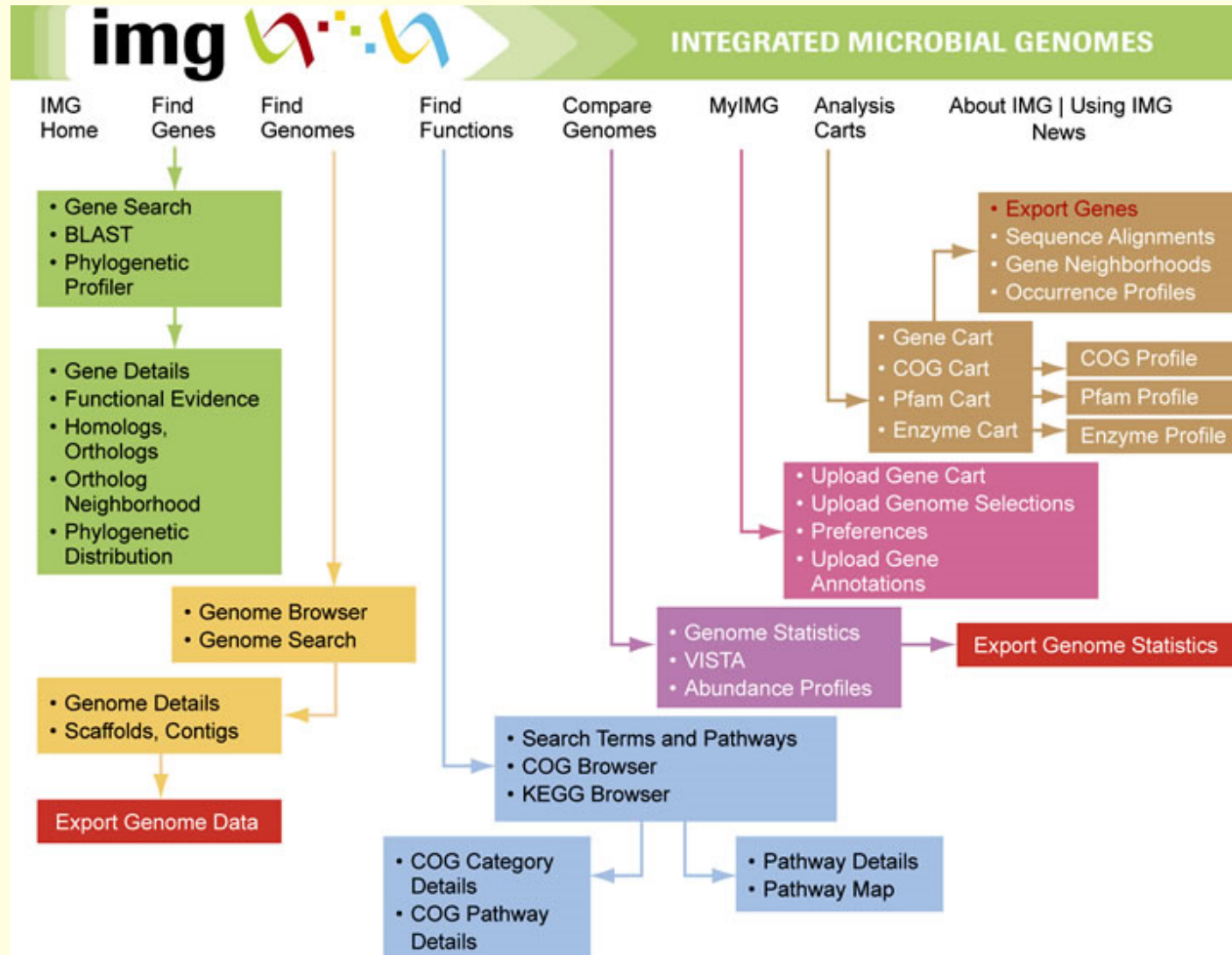
# Yeast Genome Database (SGD)



# Yeast Gene Entry (ROK1)

ROK1 BASIC INFORMATION		ROK1 RESOURCES	
Standard Name	ROK1	<p>Click on map for expanded view</p>  <p>181000 to 186000 5' NUP49 ROK1 5' SP074 tKCCU VII</p>	
Systematic Name	YGL171W	<b>Literature</b> Gene_Info Lit. Guide <input type="button" value="View"/>	
Feature Type	ORF	<b>Retrieve Sequences</b> DNA (w/ introns) <input type="button" value="Retrieve"/>	
<b>GO Annotations</b>	<a href="#">ROK1 GO evidence and references</a>	<b>Sequence Analysis Tools</b> BLASTP <input type="button" value="Analyze"/>	
Molecular Function	<ul style="list-style-type: none"><li><a href="#">ATP dependent RNA helicase</a></li></ul>	<b>Maps and Displays</b> Chr. Features Map <input type="button" value="View"/>	
Biological Process	<ul style="list-style-type: none"><li><a href="#">35S primary transcript processing</a></li><li><a href="#">mRNA splicing</a></li></ul>	<b>Comparison Resources</b> Worm Homologs <input type="button" value="View"/>	
Cellular Component	<ul style="list-style-type: none"><li><a href="#">nucleolus</a></li></ul>	<b>Functional Analysis</b>	
Description	contains domains found in the DEAD protein family of ATP-dependent RNA helicases; high-copy suppressor of kern1 null mutant		
Phenotype	<ul style="list-style-type: none"><li>Old format: Null mutant is inviable</li><li>Systematic deletion: <a href="#">inviable</a></li></ul> <p><a href="#">More Phenotype Details for ROK1</a></p>		
Position	<a href="#">ChrVII: coordinates 182394 to 184088</a> <a href="#">Old format Sequence details</a>		

# Joint Genome Institute's Microbial Genome Browser: IMG



# Demonstrations

- NCBI databases
- UCSC genome browser