# Basic Bioinformatics: Homology, Sequence Alignment, and BLAST

William S. Sanders
Institute for Genomics, Biocomputing, and Biotechnology (IGBB)
High Performance Computing Collaboratory (HPC²)
Mississippi State University

wss2@igbb.misstate.edu

June 3, 2015

## Biology Review:

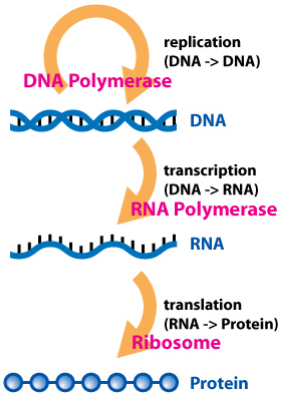
The genome is the genetic material of an organism – it is primarily responsible for heredity and variation within and among species

Genomes can be:

- DNA (deoxyribonucleic acid) → a double helical molecule made up of 4 nucleotides – **A**denine, **G**uanine, **T**hymine, and **C**ytosine
- RNA (ribonucleic acid) → a single stranded molecule also made up of 4 nucleotides – **A**denine, **G**uanine, **U**racil, and **C**ytosine

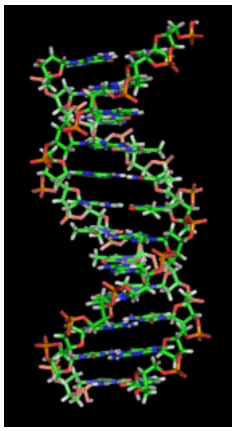
Nucleic acids are translated into:

- Proteins – made up of one or more long chains of amino acids (20 standard amino acids)




The Central Dogma of Molecular Biology

# Biology Review:



- For our purposes, DNA & RNA can be thought of as a character string with an alphabet of ~4 characters:

```
5′-GATTACATGTTTCGGGTACGATGC-3′
3′-CTAATGTACAAAGCCCATGCTACG-5′
```

- Since the 3′ → 5′ strand of DNA is the reverse compliment of the 5′ → 3′ strand, standard convention is to only store the 5′ → 3′ strand:

```
5′-GATTACATGTTTCGGGTACGATGC-3′
```

or

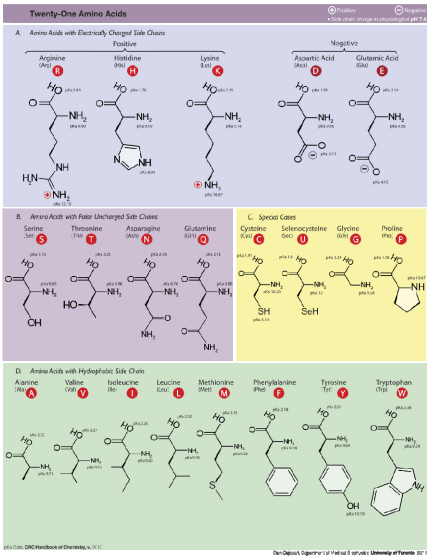
```
GATTACATGTTTCGGGTACGATGC
```

# Biology Review:

- Proteins can be thought of as character strings with an alphabet of ~21 characters:

```
MLLITMATAFMGYVLPWGQMSFWGATV
```

- Protein sequences are represented from the N-terminus (amino-terminus) to their C-terminus (carboxyl-terminus)



# Standard Files Types:

- GenBank ( *.gb| *.genbank)
  - National Center for Biotechnology's (NCBI) Flat File Format (text)
  - Provides a large amount of information about a given sequence record…
  - http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html
- FASTA (*.fasta | *.fa )
  - Pronounced "FAST-A"
  - Simple text file format for storing nucleotide or peptide sequences
  - Each record begins with a single line description starting with ">" and is followed by one or more lines of sequence
- FASTQ (*.fastq | *.fq )
  - Pronounced "FAST-Q"
  - Text based file format for storing nucleotide sequences and their corresponding quality scores
  - Quality scores are generated as the nucleotide is sequenced and correspond to a probability that a given nucleotide has been correctly sequenced by the sequencer

# Biological Sequence Representation:

**FASTA format**

- Can represent nucleotide sequences or peptide sequences using single letter codes

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

**FASTQ format**

- Represents nucleotide sequences and their corresponding quality scores

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

# Biological Sequence Representation:
## FASTQ Format

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

**Line 1** = @ and Sequence identifier *and description (optional)*

**Line 2** = Raw sequence

**Line 3** = + *optionally followed by same sequence id and description*

**Line 4** = Encoded quality values for Line 2 sequence (same number of symbols)

# Biological Sequence Representation:
## FASTQ Format – Quality Scores

- The symbols in Line 4 of a FASTQ file represent the quality of the base at a given position in the sequence
- There are a few different possible encodings of the quality score, dependent on sequencing platform
- Sanger format quality scores range from 0 to 93
- A higher score is better, and the scale is logarithmic:

$$Q_{sanger} = -10 \log_{10} p$$

$$p = 10 * (-Q / 10)$$

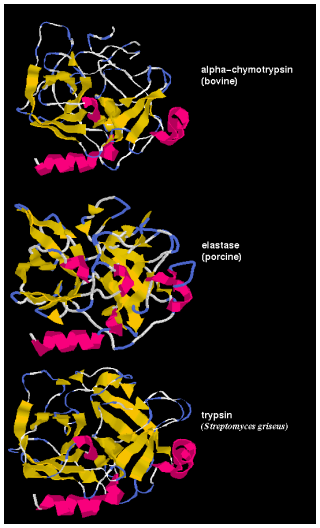
# Biological Sequence Representation:
## FASTQ Quality Scores

- Quality Score Values from Lowest to Highest:

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

| Phred Quality Score | Probability of incorrect base call | Base call accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1,000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |
| 60 | 1 in 1,000,000 | 99.9999% |

# Homology:

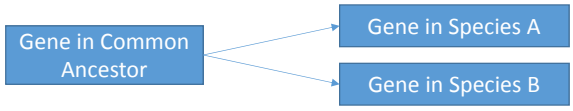



**Homologous Sequences (Homologs)**

A gene related to a second gene by descent from a common ancestral DNA sequence.

# Homology:

- **Orthologous Sequences (Orthologs)** – genes in different species that evolved from a common ancestral gene.




- **Paralogous Sequences (Paralogs)** – genes related by duplication within a genome.


Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original function.

# Homology:

Other "logs" –

- Homoeologs – homologs resulting from the duplication of genes due to a whole genome duplication (WGD) event – more common in plant species
  - Ohnologs – very ancestral homoeologs
- Xenologs – homologs resulting from horizontal gene transfer between two organisms

## Sequence Alignment:

Sequence alignment is the procedure of comparing two (pairwise) or more (multiple) sequences and searching for a series of individual characters or character patterns that are the same in the set of sequences.

- **Global alignment** – find matches along the entire sequence (use for sequences that are quite similar)
- **Local alignment** – finds regions or islands of strong similarity (use for comparing less similar regions [finding conserved regions])

## Sequence Alignment:

```
Sequence 1 = GARVEY
Sequence 2 = AVERY

Global Alignment:

GARVEY-
-A-VERY
```

# Sequence Alignment Example:

| | T | G | T | A | A | G | A | C | G | T | T |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | | | | |
| A | | | | | | | | | | | |
| G | | | | | | | | | | | |
| C | | | | | | | | | | | |
| G | | | | | | | | | | | |
| G | | | | | | | | | | | |
| G | | | | | | | | | | | |
| G | | | | | | | | | | | |

Sequence 1 = TGTAAGACGTT
Sequence 2 = AAGCGGGG

# Why do researchers focus on sequence alignment?

1. To determine possible functional similarity.
2. For 2 sequences:
   a. If they're the same length, are they almost the same sequence? (global alignment)
3. For 2 sequences:
   a. Is the prefix of one string the suffix of another? (contig assembly)
4. Given a sequence, has anyone else found a similar sequence?
5. To identify the evolutionary history of a gene or protein.
6. To identify genes or proteins.

# BLAST (Basic Local Alignment Search Tool):

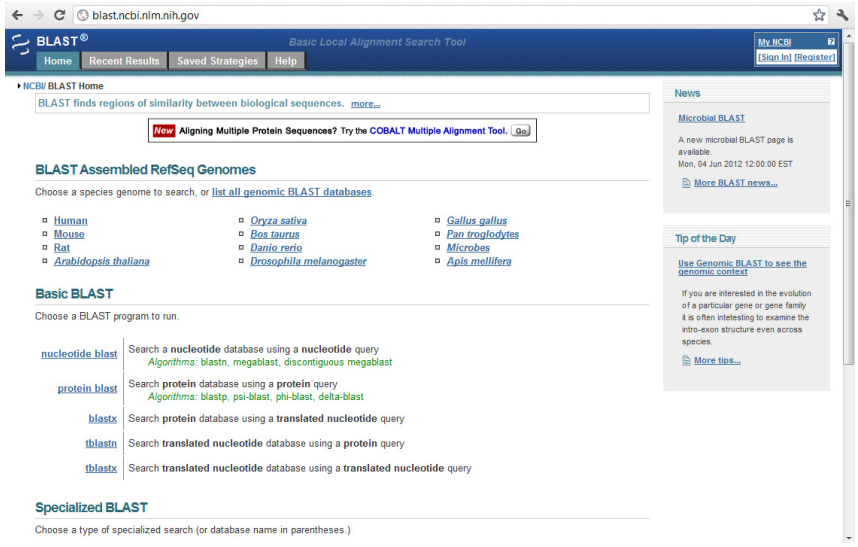
- A tool for determining sequence similarity
- Originated at the National Center for Biotechnology Information (NCBI)
- Sequence similarity is a powerful tool for identifying unknown sequences
- BLAST is fast and reliable
- BLAST is flexible

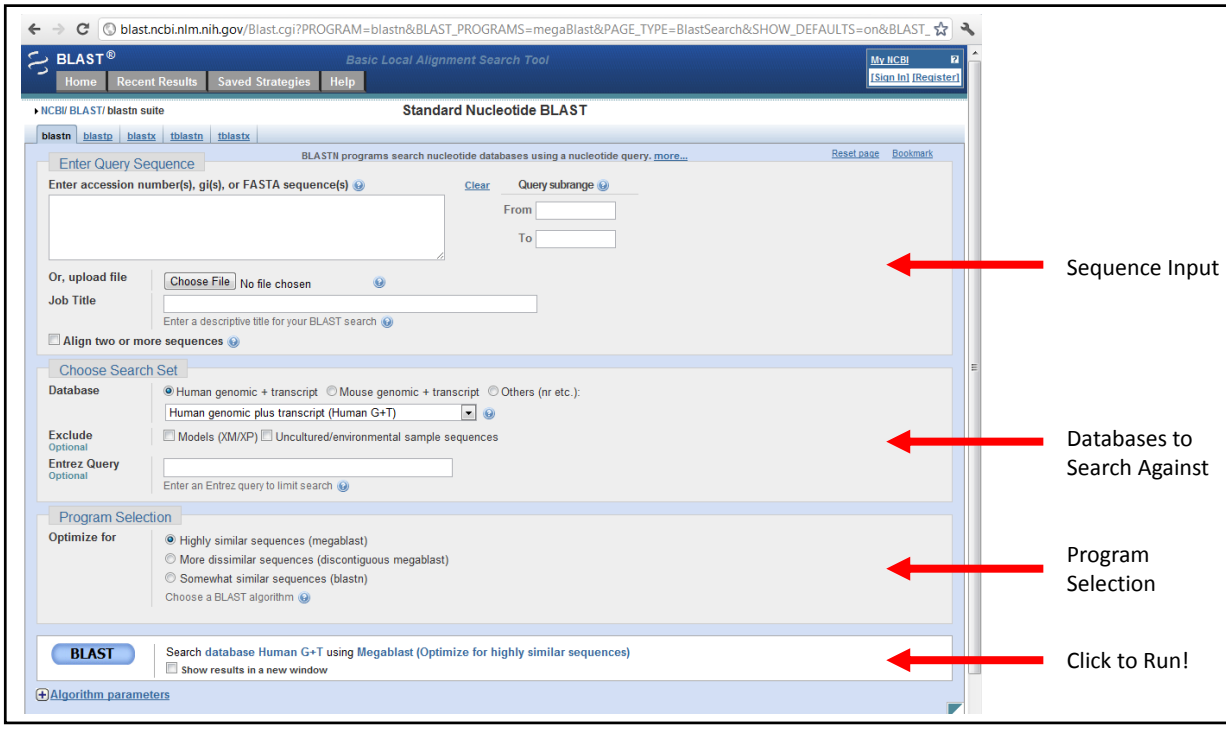
http://blast.ncbi.nlm.nih.gov/

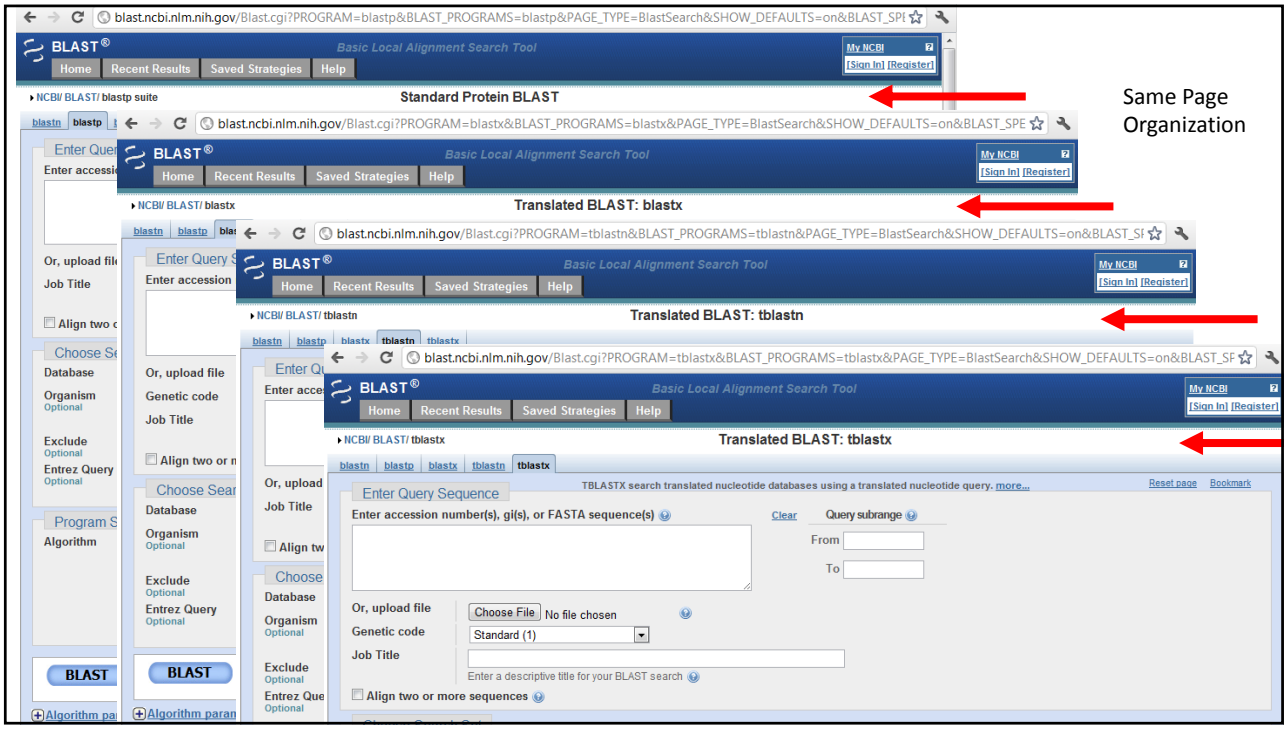
# Standard BLAST Versions:

- **blastn** – searches a nucleotide database using a nucleotide query
  *DNA/RNA sequence searched against DNA/RNA database*
- **blastp** – searches a protein database using a protein query
  *Protein sequence searched against a Protein database*
- **blastx** – search a protein database using a translated nucleotide query
  *DNA/RNA sequence -> Protein sequence searched against a Protein database*
- **tblastn** – search a translated nucleotide database using a protein query
  *Protein sequence searched against a DNA/RNA sequence database -> Protein sequence database*
- **tblastx** – search a translated nucleotide database using a translated nucleotide query
  *DNA/RNA sequence -> Protein sequence searched against a DNA/RNA sequence database -> Protein sequence database*

# NCBI BLAST Main Page:



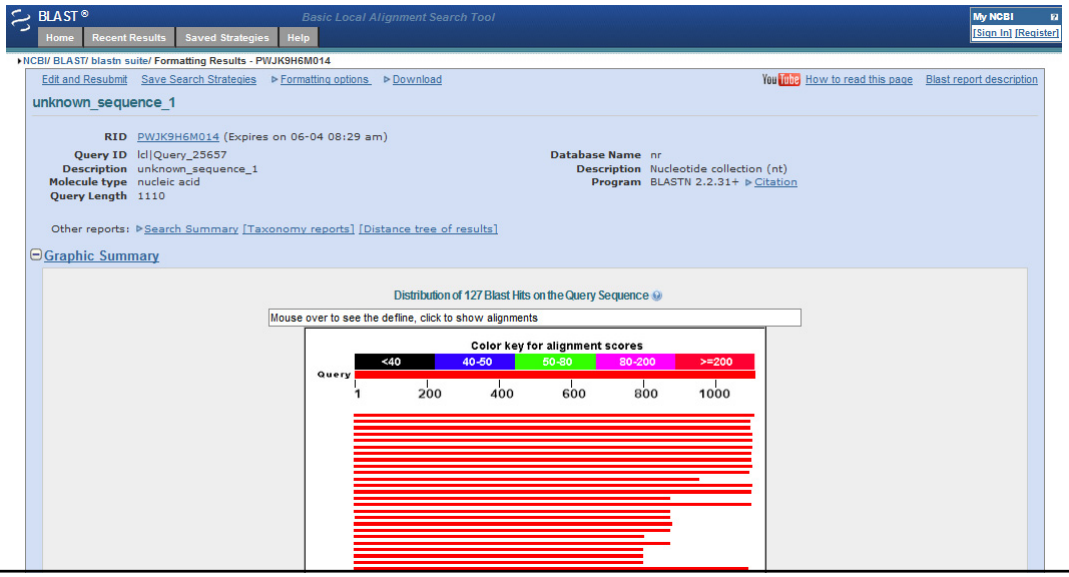



# Today's Example

```
>unknown_sequence_1
TGATGTCAAGACCCTCTATGAGACTGAAGTCTTTTCTACCGACTTCTCCAACATTTCTGCAGCCAAGCAG
GAGATTAACAGTCATGTGGAGATGCAAACCAAAGGGAAAGTTGTGGGTCTAATTCAAGACCTCAAGCCAA
ACACCATCATGGTCTTAGTGAACTATATTCACTTTAAAGCCCAGTGGGCAAATCCTTTTGATCCATCCAA
GACAGAAGACAGTTCCAGCTTCTTAATAGACAAGACCACCACTGTTCAAGTGCCCATGATGCACCAGATG
GAACAATACTATCACCTAGTGGATATGGAATTGAACTGCACAGTTCTGCAAATGGACTACAGCAAGAATG
CTCTGGCACTCTTTGTTCTTCCCAAGGAGGGACAGATGGAGTCAGTGGAAGCTGCCATGTCATCTAAAAC
ACTGAAGAAGTGGAACCGCTTACTACAGAAGGGATGGGTTGACTTGTTTGTTCCAAAGTTTTCCATTTCT
GCCACATATGACCTTGGAGCCACACTTTTGAAGATGGGCATTCAGCATGCCTATTCTGAAAATGCTGATT
TTTCTGGACTCACAGAGGACAATGGTCTGAAACTTTCCAATGCTGCCCATAAGGCTGTGCTGCACATTGG
TGAAAAGGGAACTGAAGCTGCAGCTGTCCCTGAAGTTGAACTTTCGGATCAGCCTGAAAACACTTTCCTA
CACCCTATTATCCAAATTGATAGATCTTTCATGTTGTTGATTTTGGAGAGAAGCACAAGGAGTATTCTCT
TTCTAGGGAAAGTTGTGAACCCAACGGAAGCGTAGTTGGGAAAAAGGCCATTGGCTAATTGCACGTGTGT
ATTGCAATGGGAAATAAATAAATAATATAGCCTGGTGTGATTGATGTGAGCTTGGACTTGCATTCCCTTA
TGATGGGATGAAGATTGAACCCTGGCTGAACTTTGTTGGCTGTGGAAGAGGCCAATCCTATGGCAGAGCA
TTCAGAATGTCAATGAGTAATTCATTATTATCCAAAGCATAGGAAGGCTCTATGTTTGTATATTTCTCTT
TGTCAGAATACCCCTCAACTCATTTGCTCTAATAAATTTGACTGGGTTGAAAAATTAAAA
```

Sequence available for download at:
ftp://ftp.hpc.msstate.edu/outgoing/wss2/

# Results of Our BLAST:




# Analysis of BLAST Results:

- **Max Score** – how well the sequences match
- **Total Score** – includes scores from non-contiguous portions of the subject sequence that match the query
- **Bit Score** – A log-scaled version of a score
  - Ex. If the bit-score is 30, you would have to score on average, about 230 = 1 billion independent segment pairs to find a score matching this score by chance. Each additional bit doubles the size of the search space.
- **Query Coverage** – fraction of the query sequence that matches a subject sequence
- **E value** – how likely an alignment can arise by chance
- **Max ident** – the match to a subject sequence with the highest percentage of identical bases

# Local BLAST Installation:

Executables and documentation available at:

- ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

Documentation: http://www.ncbi.nlm.nih.gov/books/NBK1762/

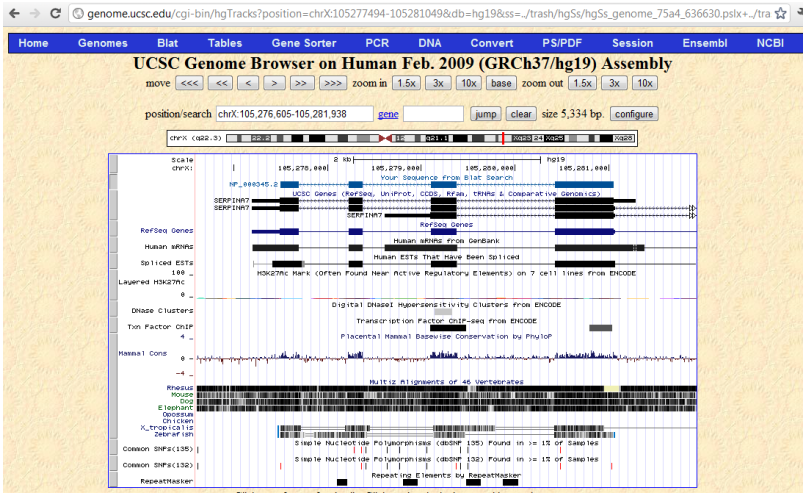
Windows Versions:

- 32-bit - ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.2.30+-win32.exe
- 64-bit - ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.2.30+-win64.exe

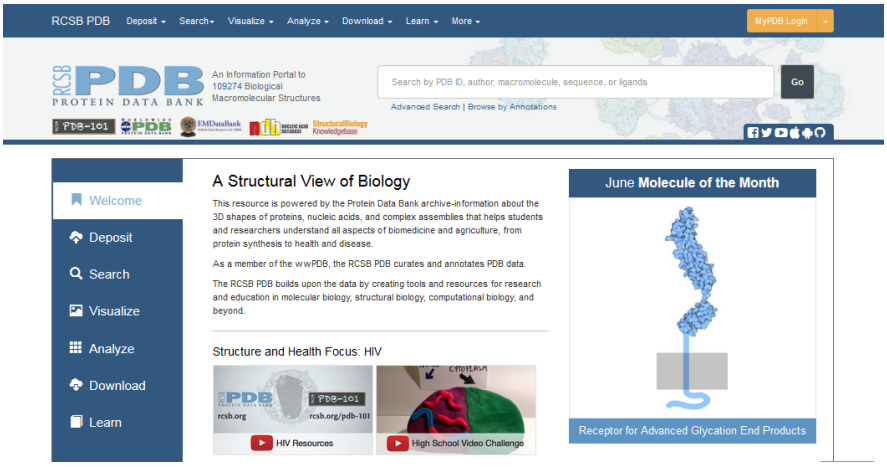
Linux Version:

- ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.2.30+-x64-linux.tar.gz

# Other Online Resources:
# UCSC Genome Browser (genome.ucsc.edu)



# Other Online Resources: Protein Data Bank www.rcsb.org/pdb/home/home.do




# Other Resources:

- EMBL-EBI (European Bioinformatics Institute) - http://www.ebi.ac.uk/
- DDBJ (DNA Data Bank of Japan) - http://www.ddbj.nig.ac.jp/
- NCBI's Sequence Read Archive (SRA) - http://www.ncbi.nlm.nih.gov/sra
- IGBB's Useful Links Page - http://www.igbb.msstate.edu/links.php

Many, many more available online, just search.

# Questions?

?

Contact Info -

- E-mail: wss2@igbb.misstate.edu
- Phone: (662) 325-2839
- Office: A209 Portera HPC2 Building (across Hwy. 182 in the Research Park)