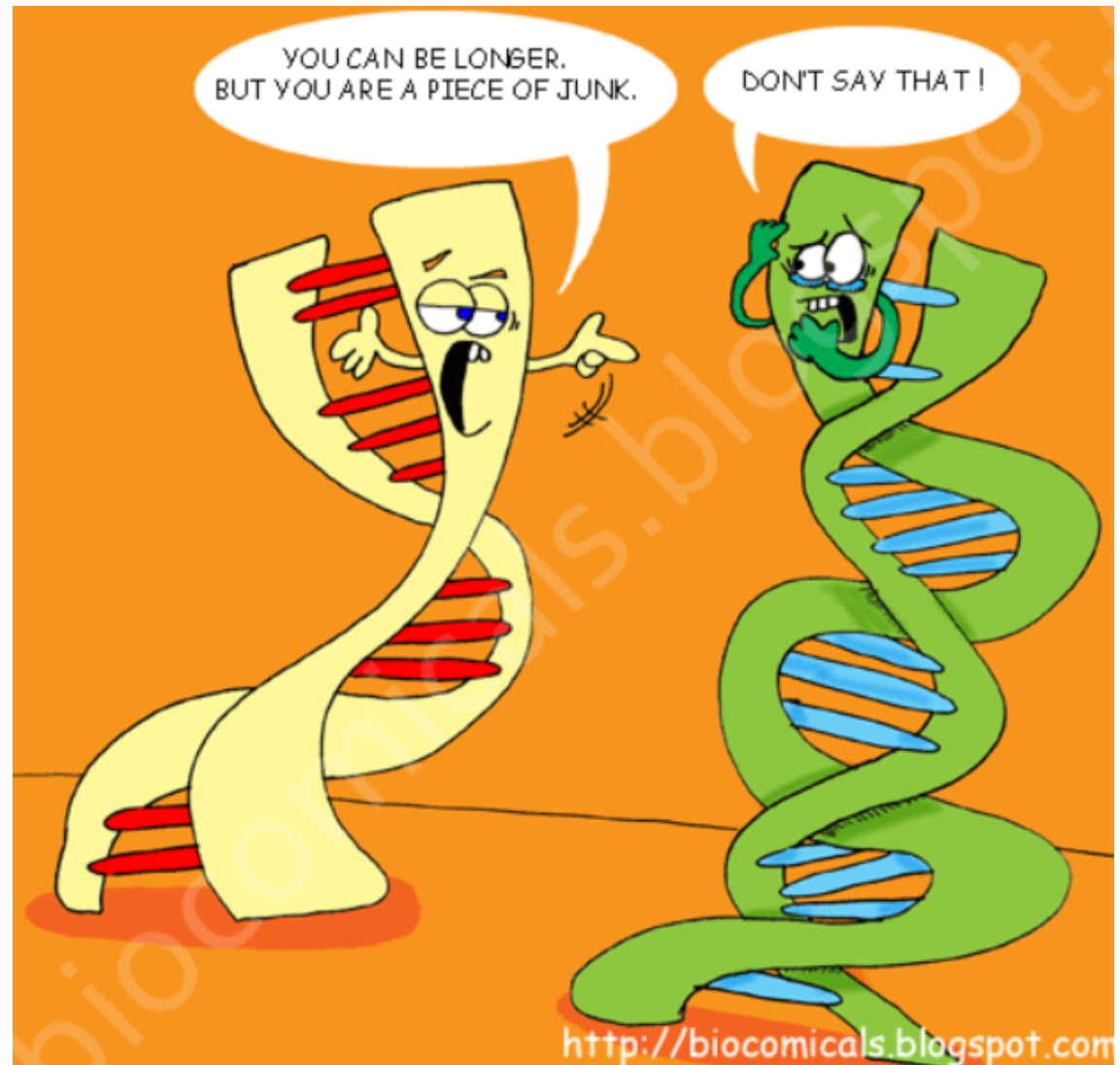


# CS123A Bioinformatics Module 1 – Week 3 – Presentation 1

Leonard Wesley  
Computer Science Dept  
San Jose State Univ



# Agenda

- Organizing Info About Bioinformatics DBs
- Finish GenBank DB
- Accession Numbers
- Entrez DB
- Ensembl

# Organizing Information RE Bioinformatics DBs

Desired Info	NCBI (US)	EMBL (Europe)	DDBJ (Japan)
Gene & Chromosome Loc on chromosome, ...	NCBI/Gene <a href="https://www.ncbi.nlm.nih.gov/gene/">https://www.ncbi.nlm.nih.gov/gene/</a>	EMBL/EBI <a href="https://www.embl.org/">https://www.embl.org/</a> or <a href="https://www.ebi.ac.uk/">https://www.ebi.ac.uk/</a>	...
Gene & Nucleotide Seq & ACC, FASTA, ...	<a href="https://www.ncbi.nlm.nih.gov/nucleotide/">https://www.ncbi.nlm.nih.gov/nucleotide/</a>	"	...
Protein	...	...	...
:	:	:	:
detailed information on specific genes.	Entrez <a href="https://www.ncbi.nlm.nih.gov/search/">https://www.ncbi.nlm.nih.gov/search/</a>	N/A	N/A
Info on genes, mRNA, proteins on specific species		ENSEMBL <a href="https://www.ensembl.org/index.html?redirect=no">https://www.ensembl.org/index.html?redirect=no</a>	

## Lets Get Info About HBB

The screenshot shows the NCBI Nucleotide database search results for the query 'HBB'. The browser address bar shows the URL: <https://www.ncbi.nlm.nih.gov/nucleotide/?term=HBB>. The search results page displays a summary of 9989 nucleotide sequences. A highlighted box contains the text: "See [HBB hemoglobin subunit beta](#) in the Gene database" and "hbb reference sequences [Genomic \(2\)](#) [Transcript \(1\)](#) [Protein \(1\)](#)". Below this, the first result is listed: "1. [Synthetic construct Homo sapiens clone CCSBHm\\_00010626 HBB \(HBB\) mRNA, encodes complete protein](#)". The result details include "573 bp linear other-genetic", "Accession: KR710229.1", and "GI: 823670799". The right sidebar shows "Results by taxon" with a list of organisms: "Chlorocebus sabaeus (983)", "Peromyscus maniculatus (459)", "Salmo salar (321)", "Homo sapiens (262)", "Mus musculus (261)", and "All other taxa (7702)".

- NCBI: <http://www.ncbi.nlm.nih.gov/>
- GenBank: <https://www.ncbi.nlm.nih.gov/nucleotide/> or <https://www.ncbi.nlm.nih.gov/gene/>
- Enter HBB into the search box ...

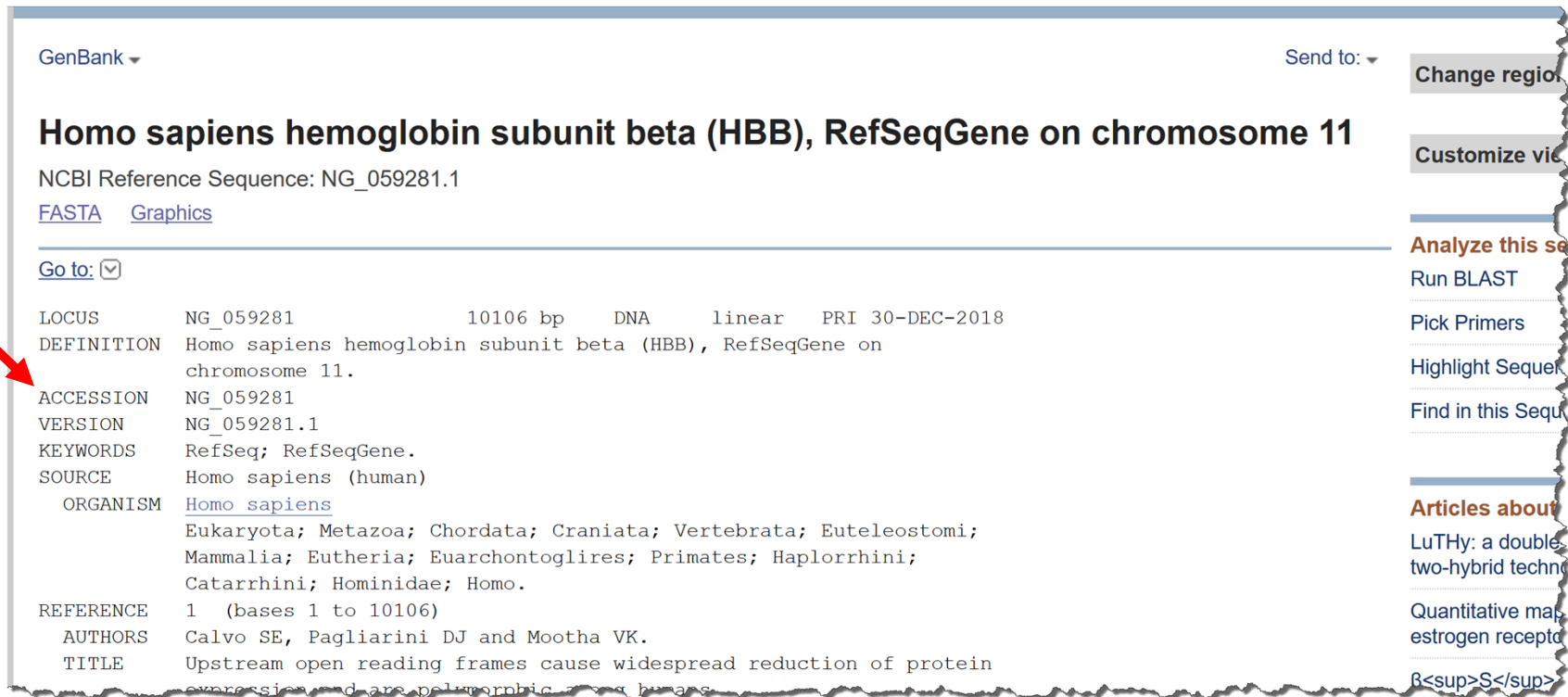
# Find The following Info For BRAF In GenBank

- Gene ID:
- Chromosome #:
- Last Updated:
- Location:
- Length:
- Number Exons:
- Any Synonyms, If so, what are they:
- From FASTA file, what is the start position of the first coding region, i.e., how many nucleotides from the first nucleotide does AUG start:

# Accession (ACC) Numbers

- The International Nucleotide Sequence Database Collaboration DDBJ/EMBL/GenBank all receive sequence submissions, assign accessions, and exchange data so that all three groups represent the total collection.
- ACCs are unique identifier given to a [DNA](#) or [protein](#) sequence record to allow for tracking of different versions of that sequence record and the associated sequence over time in a single data repository.
- Nucleotides
  - 1 letter + 5 numerals
  - 2 letter + 6 numerals
  - 2 letter + 8 numerals
- Proteins
  - 3 letter + 5 numerals
  - 3 letter + 7 numerals
- WGS
  - 4 letters + 2 numerals for WGS assembly version + 6 or more numerals
  - 6 letters + 2 numerals for WGS assembly version + 7 or more numerals
- Def of ACC Prefix letters: <https://www.ncbi.nlm.nih.gov/Sequin/acc.html>

# HBB Accession Number



GenBank Send to: ▼ Change region Customize view

## Homo sapiens hemoglobin subunit beta (HBB), RefSeqGene on chromosome 11

NCBI Reference Sequence: NG\_059281.1

[FASTA](#) [Graphics](#)

[Go to:](#) ☒

LOCUS	NG_059281	10106 bp	DNA	linear	PRI 30-DEC-2018
DEFINITION	Homo sapiens hemoglobin subunit beta (HBB), RefSeqGene on chromosome 11.				
ACCESSION	NG_059281				
VERSION	NG_059281.1				
KEYWORDS	RefSeq; RefSeqGene.				
SOURCE	Homo sapiens (human)				
ORGANISM	<a href="#">Homo sapiens</a> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.				
REFERENCE	1 (bases 1 to 10106)				
AUTHORS	Calvo SE, Pagliarini DJ and Mootha VK.				
TITLE	Upstream open reading frames cause widespread reduction of protein expression and are polymorphic across humans				

**Analyze this sequence**  
[Run BLAST](#)  
[Pick Primers](#)  
[Highlight Sequence](#)  
[Find in this Sequence](#)

**Articles about**  
[LuTHy: a double two-hybrid technique](#)  
[Quantitative map of estrogen receptor](#)  
[B<sup>S</sup>](#)

# Find The Refseq ACC for BRAF & HIV

- Find ACC, FASTA seq in GenBank ( <https://www.ncbi.nlm.nih.gov/> )
- Find ACC, FASTA seq in EMBL-EBI UK ( <https://www.ebi.ac.uk/> )



## Entrez DB \*

- Entrez Gene provides detailed information on specific genes. It is a searchable database, by ACC number or keyword, which pulls information from RefSeq genomes. The genes can be viewed in several formats and there are many links to other Entrez databases and external links.
- Go to <https://www.ncbi.nlm.nih.gov/search/>
- Enter NG\_059281 the ACC # for HBB
- Can get to same & related info via NCBI Gene/Nucleotide.

\* <https://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html>

## Entrez DB (cont.)

- More details on using these functions are in the Entrez help document and FAQ pages.
- **Examples** (from [http://www.ncbi.nlm.nih.gov/books/NBK21085/#ch19.How\\_to\\_Query\\_Gene](http://www.ncbi.nlm.nih.gov/books/NBK21085/#ch19.How_to_Query_Gene))
- **Example 1:** Find all Gene records from fungi that have expression data in UniGene or GEO.
  - Go to NCBI Entrez search link on previous slide.
  - Enter `fungi[organism] AND ( "gene unigene"[filter] OR "gene geo"[filter])` into the search box. Click SEARCH.

# Ensembl

- Ensembl is a joint scientific project between the European Bioinformatics Institute and the Wellcome Trust Sanger Institute, which was launched in 1999 in response to the imminent completion of the Human Genome Project. This database provides a centralized resource for genes, mRNAs, and proteins of our own species and other vertebrates.
- In Ensembl, you can take advantage of the descriptor fields. To do this, you can first select a species from the dropdown menu, or search all species, by keyword. The keyword can be the name of a gene, the abbreviation for a gene, or a chromosomal location.
- Examples of each are: 1) gene, insulin; 2) abbreviation, BRCA2, or 3) chromosomal location, X:100,000 .. 200,000.

# Ensembl Exercise

- Go to ENSEMBL at <https://www.ensembl.org/index.html?redirect=no>
- Select Human as the organism
- For HBB and BRAF find:
  - Chromosome #
  - Location of gene
  - ENSEMBL ACC/ENSG # and version
  - # Exons

# Peek Ahead: Bioinformatics DB Assignment

From the list below, choose an  
ACC number to investigate:

Mammalian	NM_001122.2, NM_000041, X06359.1, M35551.1, AF080219, NP_001035568.2, AJ005203.1, L13593, AAH05707, NP_001002824
Bacterial/Viral	AAC07338, YP_002814319.1, YP_002868778.1, FJ445749.1, NP_659661, YP_093952, YP_001121138.1, NP_862693.1, YP_002650826.1, FJ445749.1
Plant	X56734, NM_112711.3, NM_128034.2, NM_148101.1, NM_001073934.1, XM_002339373.1, XM_002337972.1, XM_002325952.1, XM_001703606.1, XM_001703306.1
Other Eukaryotes	BM104655, AY039235.1, NM_212859.2, NM_131264.1, NM_001009901.1, NM_001137660.1, AB050623.1, FJ460570.1, NM_001029448.2, NM_171059.3

1. Is your sequence a protein sequence or nucleotide sequence?
2. What species or organism does the sequence come from?
3. What are the associated references, either journal articles or submission references?
4. When was the sequence submitted? Has the sequence been revised?
5. How long is the sequence?
6. Can you find all the related sequences/info? For example, if you have a DNA/mRNA sequence can you find related organisms, or earlier versions (if not version 1)
7. Are there any unique features associated with your sequence?