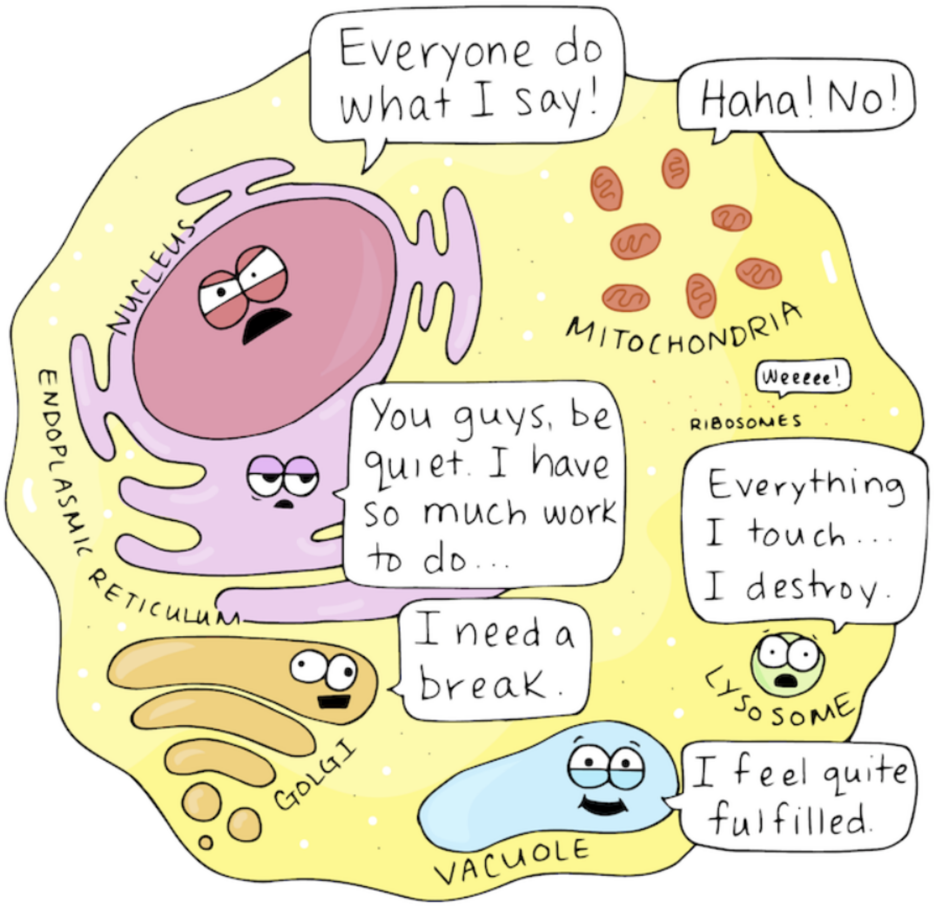
CS123A Bioinformatics Module 1 – Week 2 – Presentation 2

Leonard Wesley

Computer Science Dept

San Jose State Univ



If organelles could talk.

Agenda

- Accession Numbers
- Entrez DB
- Ensembl

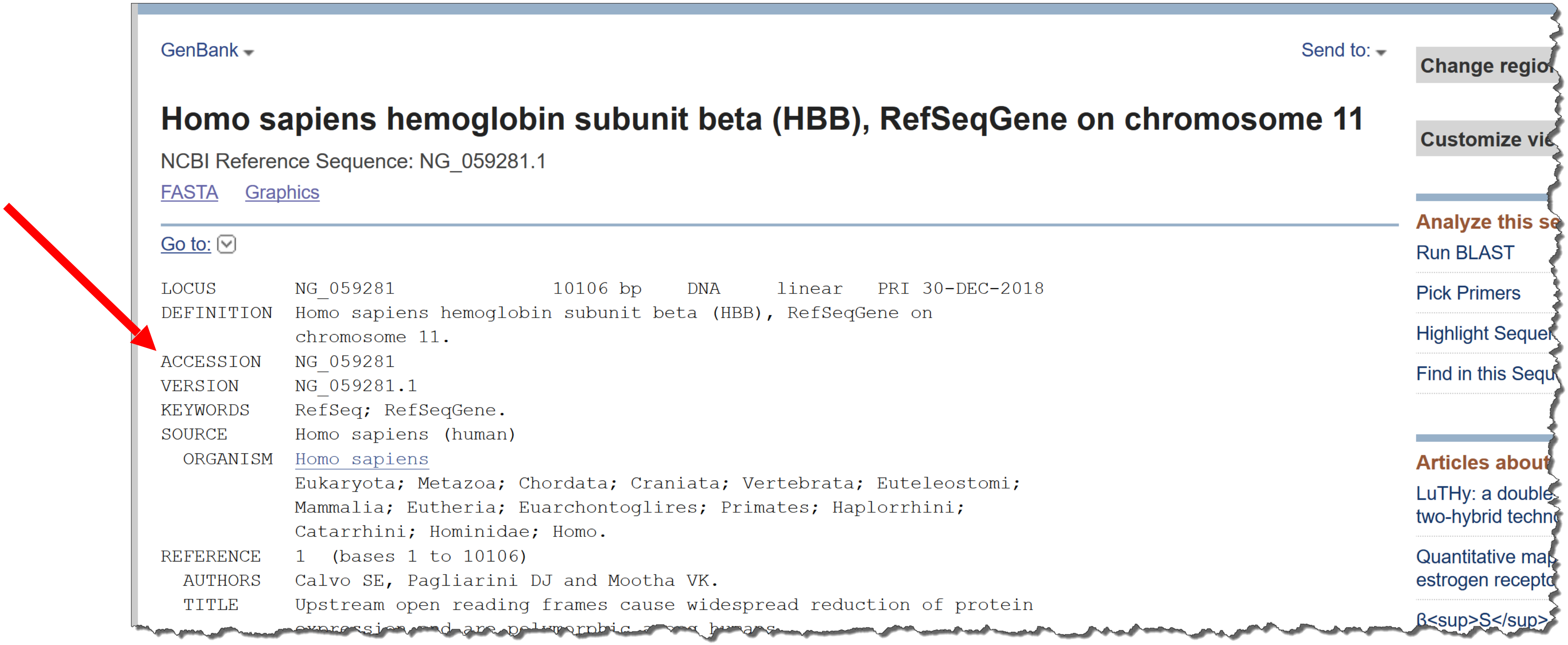
Accession (ACC) Numbers

- 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.

- 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



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  Homo sapiens
            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 among humans.
```

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.

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)

- **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 Welcome 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.