NCBI: Genome

(Máster en Bioinformática, Universidad de Valencia)

David Montaner

17-02-2014

NCBI Genome

Organizes information on genomes:

- genomic sequences
- maps
- chromosomes
- assemblies
- annotations
- . . .

http://www.ncbi.nlm.nih.gov/genome

Search species

Homo sapiens

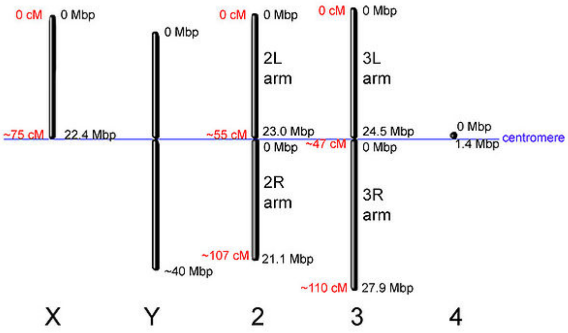
http://www.ncbi.nlm.nih.gov/genome/?term=homo+sapiens
http://www.ncbi.nlm.nih.gov/genome/51

Drosophila melanogaster

http://www.ncbi.nlm.nih.gov/genome/?term=Drosophila+melanogaster
http://www.ncbi.nlm.nih.gov/genome/47

Organization

- chromosomes
- **arms** of chromosomes
- . . . any other molecules



Drosophila melanogaster chromosomes

Data from the National Center for Biotechnology Information (NCBI) and Carvalo (2002)

FPT

All genomes

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

Drosophila melanogaster

ftp://ftp.ncbi.nlm.nih.gov/genomes/Drosophila_melanogaster/

Homo sapiens

ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/

FTP Files

Example:

ftp://ftp.ncbi.nlm.nih.gov/genomes/Drosophila_melanogaster/RELEASE_5_48/CHR_2/

- RefSeq NT_033778: Drosophila melanogaster chromosome **2R**, complete sequence
- RefSeq NT_033779: Drosophila melanogaster chromosome **2L**, complete sequence

Downloads:

```
wget -r ftp://ftp.ncbi.nlm.nih.gov/genomes/Drosophila_melan
```

File Formats

- gbk : GenBank flat file format; meta-data, sequence, and annotations. As in the web

Fasta files (usually for BLAST)

- fna : (FASTA Nucleic Acid file) **chromosomal** sequence
- ffn : (FASTA nucleotide) **coding regions** file
- faa : (FASTA Amino Acid file) **translated** coding regions (proteins)

Annotation files

- gff : GFF3 file containing annotations only (coordinates relative to the **fna** file)

Some other formats

- ptt = Protein Table
- asn = ASN.1 file, print form, replaces .prt
- val = ASN.1 binary format

GFF File Format

- **Generic Feature Format** (Version 3)
- General annotation of fasta files
- Defined by the Sequence Ontology Consortium
- Tabular text file
- Tab separated columns

See specifications at:
http://www.sequenceontology.org/gff3.shtml
Sometimes you will need to build up your GFF file and validate it:
http://modencode.oicr.on.ca/cgi-bin/validate_gff3_online

GFF Fields

1. seqid
2. source
3. type: type of the feature
4. start
5. end
6. score: generally some quality measurement for the alignment (E-values . . .)
7. strand
8. phase
9. Extra attributes: see `http://www.sequenceontology.org/gff3.shtml`

Missing values are indicated with a **dot**

Inspect the files

Count number of sequences in a fasta file:

```
grep ">" NT_033778.fna | wc -l
```

```
grep -n ">" NT_033778.fna
```

BE CAREFUL WITH OVERWRITING:

```
grep >
```

Find a list of the type of features in a GTF file

```
cut -f 3 NT_033778.gff | sort | uniq
```