Biological Data Formats

Malay (mbasu@kumc.edu)

Contents

1 Sources of data

The major sources of data are sequences databases:

1. NCBI: http://www.ncbi.nlm.nih.gov
2. EBI: http://www.ebi.ac.uk/
3. ENSEMBL: http://www.ensembl.org/index.html
4. UCSC Genome Browser: https://genome.ucsc.edu/

2 Human Reference Genome

1. Genome Reference Consortium. http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/
2. Gencode (ENSEMBL). http://www.gencodegenes.org/
3. Illumina Igenomes: https://support.illumina.com/sequencing/sequencing_software/igenome.html
4. GATK specific datasets (GATK resource bundle): https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle
5. 10X genomics reference genomes for cell ranger.

3 FASTA format

The most common file format for sequence files.

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

4 GenBank format

Sample GenBank record: http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

5 Annotation

Annotation is a way to provide extra information over the raw sequence. Some sequence file formats by design have the annotaion built into the formation. An example is GenBank. Most commonly, annotaion comes in separate files. The files generally of two types:

1. GTF

```
381 Twinscan    CDS 380 401 .   +   0   gene_id "001"; transcript_id "001.1";
381 Twinscan    CDS 501 650 .   +   2   gene_id "001"; transcript_id "001.1";
381 Twinscan    CDS 700 707 .   +   2   gene_id "001"; transcript_id "001.1";
```

2. GFF

   The file format specification can be found here: http://useast.ensembl.org/info/website/upload/gff.html?redirect=no

6 Problem 1

Bert Voglestein in a Science paper published in 2013 (PMID: 23539594) reported a list of Tumor Suppressor genes and Oncogenes. The list is available in the `data` folder as `vogelstein_tsg.txt`. Use the UNIPROT REST API get the protein IDs for the genes.

A stardard from the the API is like this

https://rest.uniprot.org/uniprotkb/search?query=reviewed:true+AND+organism_id:9606+AND+gene:BRCA1&format=tsv&fields=accession,reviewed

The URL needs to be quoted to use with `wget` to avoid `&` and also the first line returned is a header which can be removed using `grep -v Entry`.

7 Problem 2

Download the Swissprot FASTA file from the UNIPROT website (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz). Write a script to extract the sequences corresponding to the IDs created in Problem 1 from this file.