**Genomic Expression Basics**

Dr. S. Ravichandran, BIFX550

saka.ravi@gmail.com

GEO Datasets example

Let us explore the GEO dataset, SRP033351

And the corresponding run SRR1039508

https://www.ncbi.nlm.nih.gov/sra/SRR1039508/

http://www.ebi.ac.uk/ena/data/view/SRR1039508

http://www.ebi.ac.uk/ena/data/search?query=SRR1039508

NCBI allows only to download the experiment in a compressed format not in fastq format. If you want to download the fastq format directly then go to ENA.

You will see two files. Remember the experiment is a paired-end sequencing.

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR103/008/SRR1039508/SRR1039508_1.fastq.gz

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR103/008/SRR1039508/SRR1039508_2.fastq.gz

Here is the beginning of the file SRR1039508_1.fastq, which contains the first read of the pair for the experiment:

```
@SRR1039508.1 HWI-ST177:290:C0TECACXX:1:1101:1225:2130/1

CATTGCTGATACCAANNNNNNNNGCATTCCTCAAGGTCTTCCTCCTTCCCTTACGGAATTACA

+

HJJJJJJJJJJJJJJ#########00?GHIJJJJJJJIJJJJJJJJJJJJJJJJJHHHFFFFFD

@SRR1039508.2 HWI-ST177:290:C0TECACXX:1:1101:1311:2131/1

CCCTGGACTGCTTCTTGAAAAGTGCCATCCAAACTCTATCTTTGGGGAGAGTATGATAGAGAT

+

HJJJJJJJJJJJJJJJJJIIJIGHIJJJJJJJJJJJJJJJJJJJJJJGHHIDHIJJHHHHHHF

@SRR1039508.3 HWI-ST177:290:C0TECACXX:1:1101:1414:2163/1

TCGATCCATCGATTGGAAGGCACTGATCTGGACTGTCAGGTTGGTGGTCTTATTTGCAAGTCC

+

HJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJGJJIGHIJJBGIJCGIAHIJHHHHHHHFFFFF
```
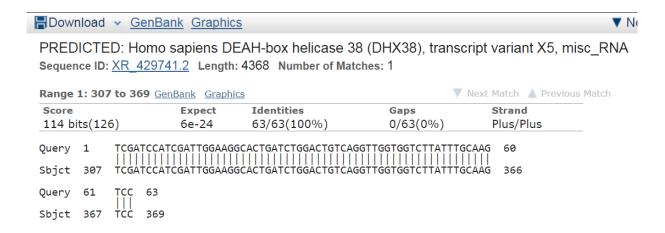
Here we see three reads, with names, sequence, +, then quality score.

If you take the third read and do a BLASTN, you will be able to see where this came from

PREDICTED: Homo sapiens DEAH-box helicase 38 (DHX38), transcript variant X5, misc_RNA

Sequence ID: XR_429741.2 Length: 4368Number of Matches: 1
BLASTN with DB: Human G+T (2 databases)

Download ∨  GenBank  Graphics                                                    ▼ N

PREDICTED: Homo sapiens DEAH-box helicase 38 (DHX38), transcript variant X5, misc_RNA
Sequence ID: XR_429741.2  Length: 4368  Number of Matches: 1

Range 1: 307 to 369 GenBank  Graphics          ▼ Next Match  ▲ Previous Match

| Score | Expect | Identities | Gaps | Strand |
|---|---|---|---|---|
| 114 bits(126) | 6e-24 | 63/63(100%) | 0/63(0%) | Plus/Plus |

```
Query   1    TCGATCCATCGATTGGAAGGCACTGATCTGGACTGTCAGGTTGGTGGTCTTATTTGCAAG  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct   307  TCGATCCATCGATTGGAAGGCACTGATCTGGACTGTCAGGTTGGTGGTCTTATTTGCAAG  366

Query   61   TCC  63
             |||
Sbjct   367  TCC  369
```

The first three reads in the second file are as follows (look at the id from the first and compare it to the second; Also look in the first line for /1 for the first read and /2 for the second read indicating that these two reads are from paired end sequencing runs):

@SRR1039508.1 HWI-ST177:290:C0TECACXX:1:1101:1225:2130/2

CAGATGAGGCGTGTTGGCCAGAGAGCCATTGTCAACAGCAGAGATGNNNNNNNNNNNNNNAATCC

+

HJJJJJJJJJJHIIIJJJJJJJJJJJJJJJJJJJJJJJJHIJIJHII#################

@SRR1039508.2 HWI-ST177:290:C0TECACXX:1:1101:1311:2131/2

TACTCCGGAGAACAGATGGGATTCCCTAGGAGACCCTTGAGGGAAAAGGGAGCCCCAATCTCT

+

FJJJJJJJFHEHJJJHIIJJGGIIJJGIIJGJHJJJJJHGIJJIGIHHHHFFFDDDDDDDDDE

@SRR1039508.3 HWI-ST177:290:C0TECACXX:1:1101:1414:2163/2

TCGCTCTCTCCGTTTCAGGGAAGCCAGCAAGTCCAGTCCGAGTAATGAAGGGCGGGGAGCAGG

+

HJJJJJJJJJJJJJJJJJJJJIJJJJJJJJJJJIJJJJJJFHIJFHJJJJJIHHHFDDDDDDDDD

Note the first one is from the left portion and the corresponding read from the second file could be the one from the lower right.