# BISC-577: Homework # 1

Due on Tuesday, March 31, 2015

**Saket Choudhary**
**2170058637**

# Contents

## Question # 1

SRA is a publically available archive of biological sequencing datasets. By making raw data and the associated metadata available, SRA aims to make genomics reproducible. New discoveries are possible using the existing datasets.

## Question # 2

FASTQ is ASCII-based format for storing sequences along with their quality scores. FASTQ originated from FASTA format. FASTA files do not store quality scores. Eash FASTQ record consists of typically four lines and a .fastq/.fq file is typically a collection of different records. The first line consists of a unique identifier. The second line is sequence[A/T/C/G/N] where N stands for base not sequenced. The third line is a separator '+' and the fourth line is simply ascii encoded quality scores. These scores are an indicative of how sure the sequencer is that a particular base is infact that and not anything else or noise.

## Question # 3

Accession: SRR1287226 Run: SRR1287226 Metadata associated includes the the experimental design, platform used and library preparation strategy. This sample seems to be missing information about the experimental design. It is not evident if the samples came from "before" or "after" correction as mentioned in the abstract.

## Question # 4

Aspera took close to 1260 seconds while wget/ftp took close to 2740 seconds. wget/ftp relies on a 'handshaking' method(TCP) to ensure reliable data transfer. Metadata exchanges take place at different points to ensure the packet transmiited is received by the client. Aspera relies on UDP, avoiding the 'handshake' overhead though still maintaining data integrity at the application layer.

## Question # 5

SRA filesize: 2919274711 byes
FastQ size: 12965100984

## Question # 6

[Question # 6] continued on next page...

## Question # 7

Sequencing adapters are known sequenced used for making the DNA fragments ligate to primers and bind to the flow channel more efficiently. The sequencer might often treats the ligated sequence as a shotgun sequence.
To remove sequencing adapters, *cutadapt* https://github.com/marcelm/cutadapt
cutadapt was used to performe trimming both $5'$ and $3'$ trimming: $5'$ adapter sequence: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
$s'$ adapter sequence: CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT
The plot generated using 'fastqc' didn't show presence of any sequencing adapters, though there were a few overrepresented sequences and hence trimming resulted in all the reads passing the filter. *cutadapt* produces additional statistics about the presence of nucleotides in percentage preceeding the adapter and uses a partial matching scheme to search for possible primers. The trimming was performed with at most 10% error in paired-end mode. Though the whole of adapter was not present(0 reads contained the whole adapter sequence), some 3 length sequences were over-represented and were trimmed resulting in few output reads being less than 100bp long. A total of 660417 + 18529 reads had overrepreented 3-mer coming from the adapter and were trimmed. *cutadapt* takes into consideration over-representation by taking the base case to be the probability of observing that $k - mer$ in a random sequence.
The nucleotide bar plot will change drastically if all the reads have some part of the adapter sequence. The 'C' and 'G' plot in particular should ideally decrease in height post trimming, since they seem to be abundant in the adapters.

continued on next page...