In silico processing of Metagenomic data

Rohita Sinha

Dept of Food Science & Technology, UNL

Metagenomics

A giant leap

Culture based analysis

Massively parallel DNA sequencing

Sequencing metagenomic samples

DNA extracted from environmental samples

Amplification of 16S rRNA genes

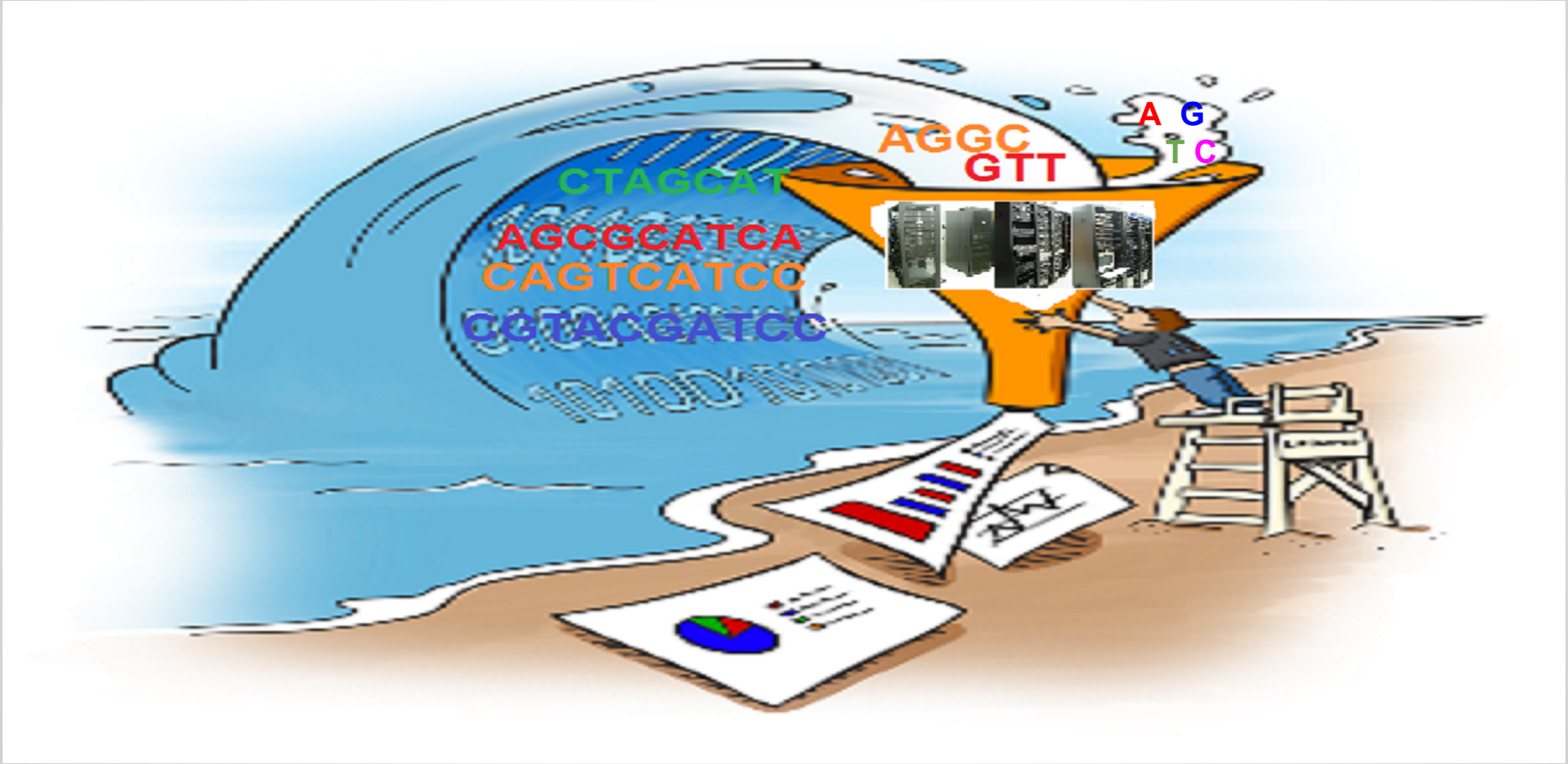
1- Data size in Mbs/sample
2- Well established data processing protocols.
3- Quick profiling of bacterial diversity.
4- No information about the protein content of bacterial population.

Complete sequencing of genetic materials

1- Data size in GBs/sample
2- Data processing protocols are still evolving.
3- Relatively slower than 16S rRNA processing (quick profilers like MetaPhlAn also exist).
4- Detailed analysis protein content of the bacterial population is also possible.

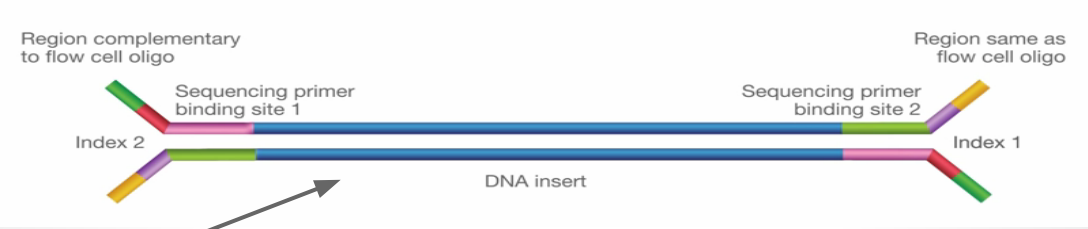
High resolution of information

Computational resources & skills are essential to analyze metagenomic data



Data pre-processing

Nextera library prep + Illumina Sequencing



Structure of our final DNA insert, ready to be sequenced

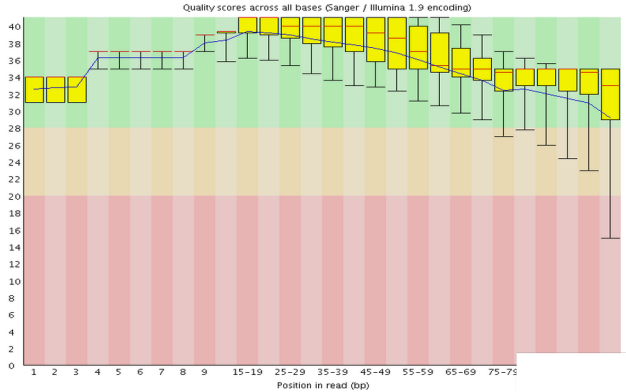
5' Primer-1----Read1AGATGTGTATAAGAGACAG----------Primer- Read2^{C} ----Primer2^{C} **3'**
3' Primer-1^{C} ---- Read1-Primer^{C} ----------GACAGAGAATATGTGTAGA-Read2 ------Primer2 **5'**

1- Shorter insert length may lead to sequencing of "sequencing primers", hence removal of those fragments are necessary.

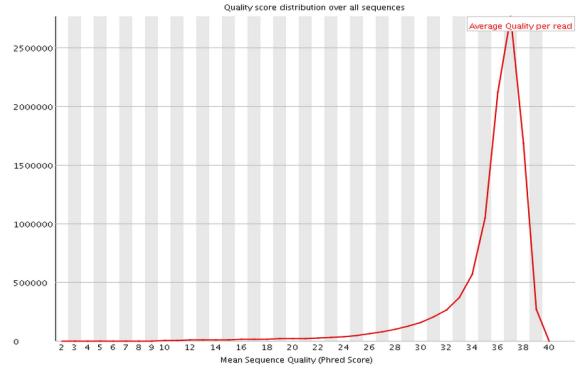
2- Illumina sequencer tends to produce low quality base call towards 3' end of the reads. These low quality bases should be removed before the analysis.

Data pre-processing (Fastqc report)

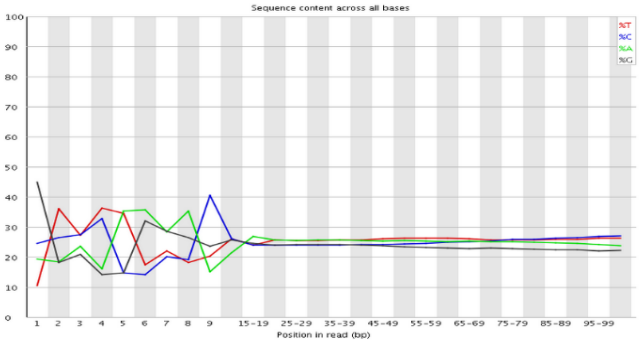
Typical quality score distribution of our samples



Typical read based quality score distribution of our samples



Typical base distribution of our samples



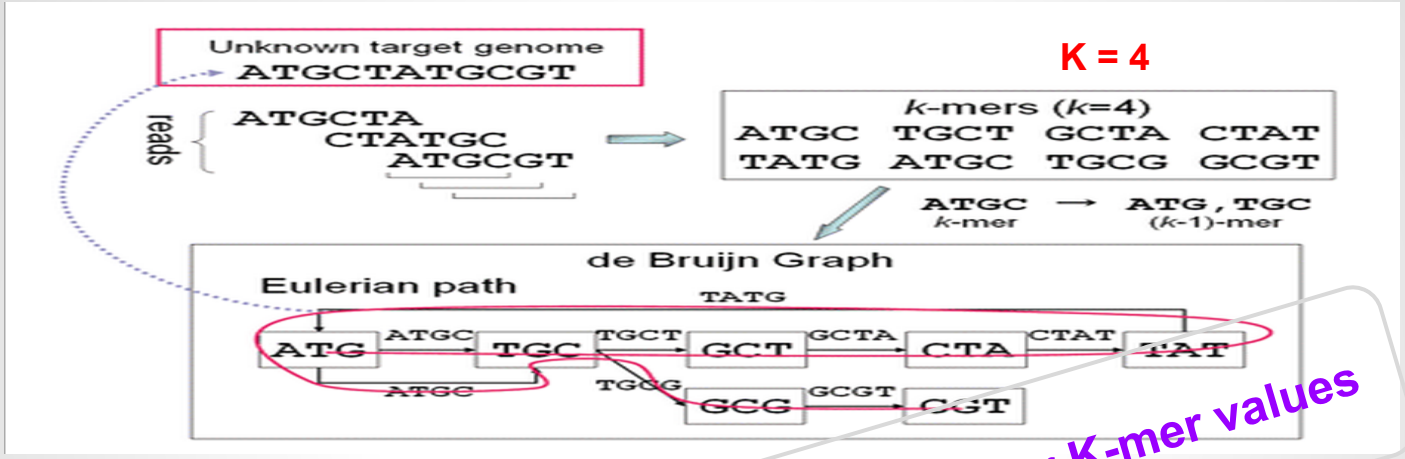
Data analysis (metagenome assembly)

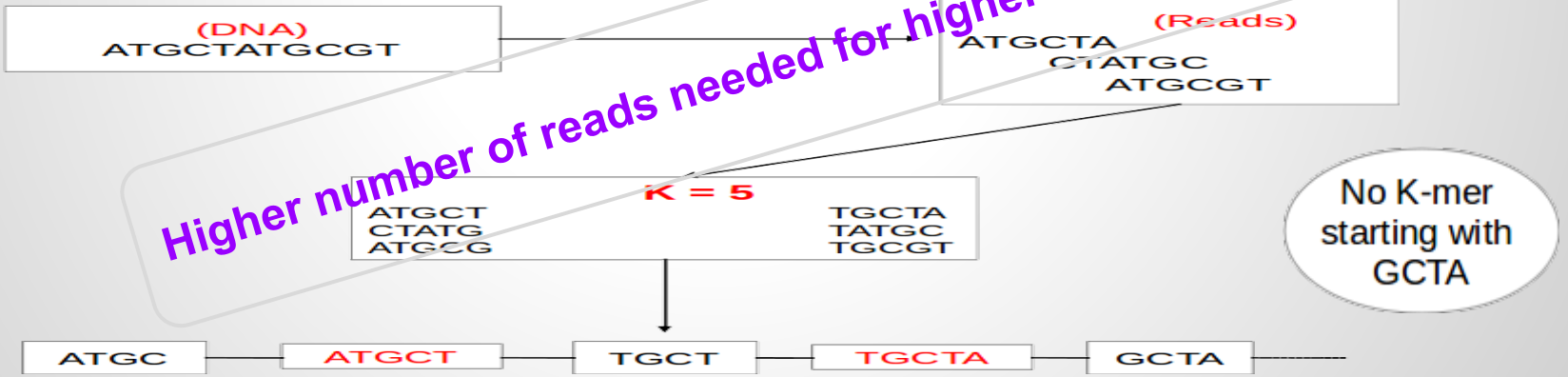
- Assembly terms:
  - Contigs: Long contiguous chain of nucleotides (ATGC) generated through an assembly operation are called contigs.
  - Scaffolds: Two contigs are concatenated by multiple ‘Ns’, when paired-end reads strongly suggest the proximity of contigs.

- Majority of metagenome assemblers use ‘de Bruijn’ graph to generate contigs.

- Output of a ‘de Bruijn’ graph based assemblers, overly depends on its single input parameter (k-mer size).

Metagenome assembly (de Bruijn graph)





Taxonomic annotation of contigs

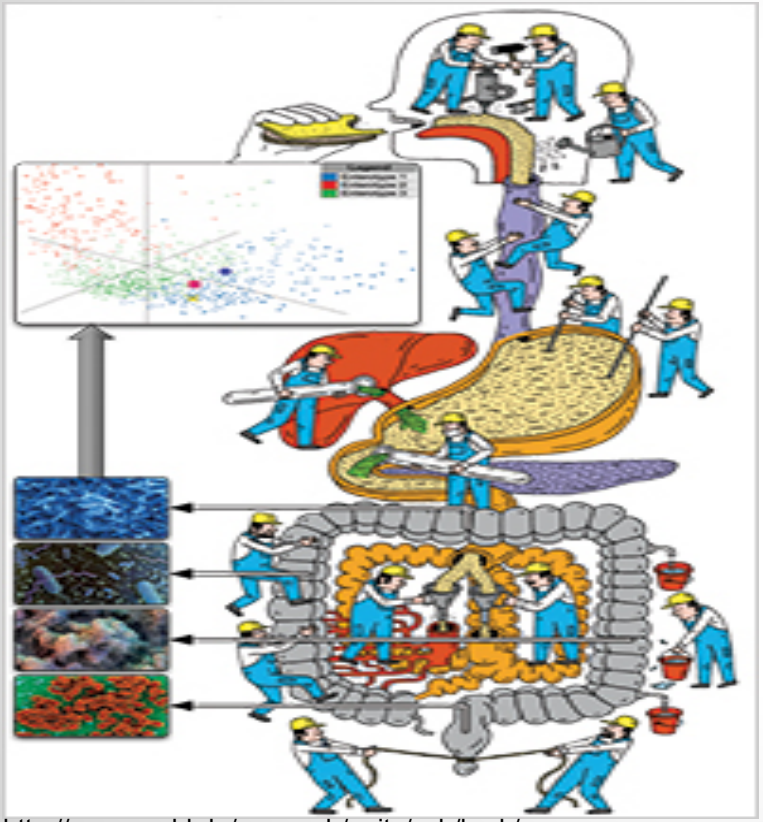
1- Simple “BLASTn” based analysis of a good quality contigs may be classified easily if:

- All the blast hits goes to single reference sequence (same GI)
- All the blast hits goes to multiple reference sequence (multiple GIs) having same genus

2- K-mer based methods like kraken, MyTaxa can be used to classify those contigs which could not be classified by “BLASTn”.

3- Pre-processed reads can be aligned to taxonomically annotated cotnigs to calculate the microbial abundance profiles.

Biological importance of the functional profiling of a microbiome



http://www.embl.de/research/units/scb/bork/

Microbial system is in perfect symbiosis with our intrinsic systems

It is governed by the biochemical potential of the microbiome (protein content)

Therefore, correct functional profiling of the microbiome is essential to understand the delicate biochemical interaction between microbes & their environment

Functional annotation of metagenomic data