METAGENOME &
QIIME (QUANTITIATIVE INSIGHT INTO MICROBIAL ECOLOGY)

Presented by:
Kumar Arvind
Research Scholar
Dept. of Genomic Science
Central University of Kerala

Metagenomics

"The application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species"

- Kevin Chen and Lior Pachter

- Also referred as

  Environmental genomics,

  Ecogenomics,

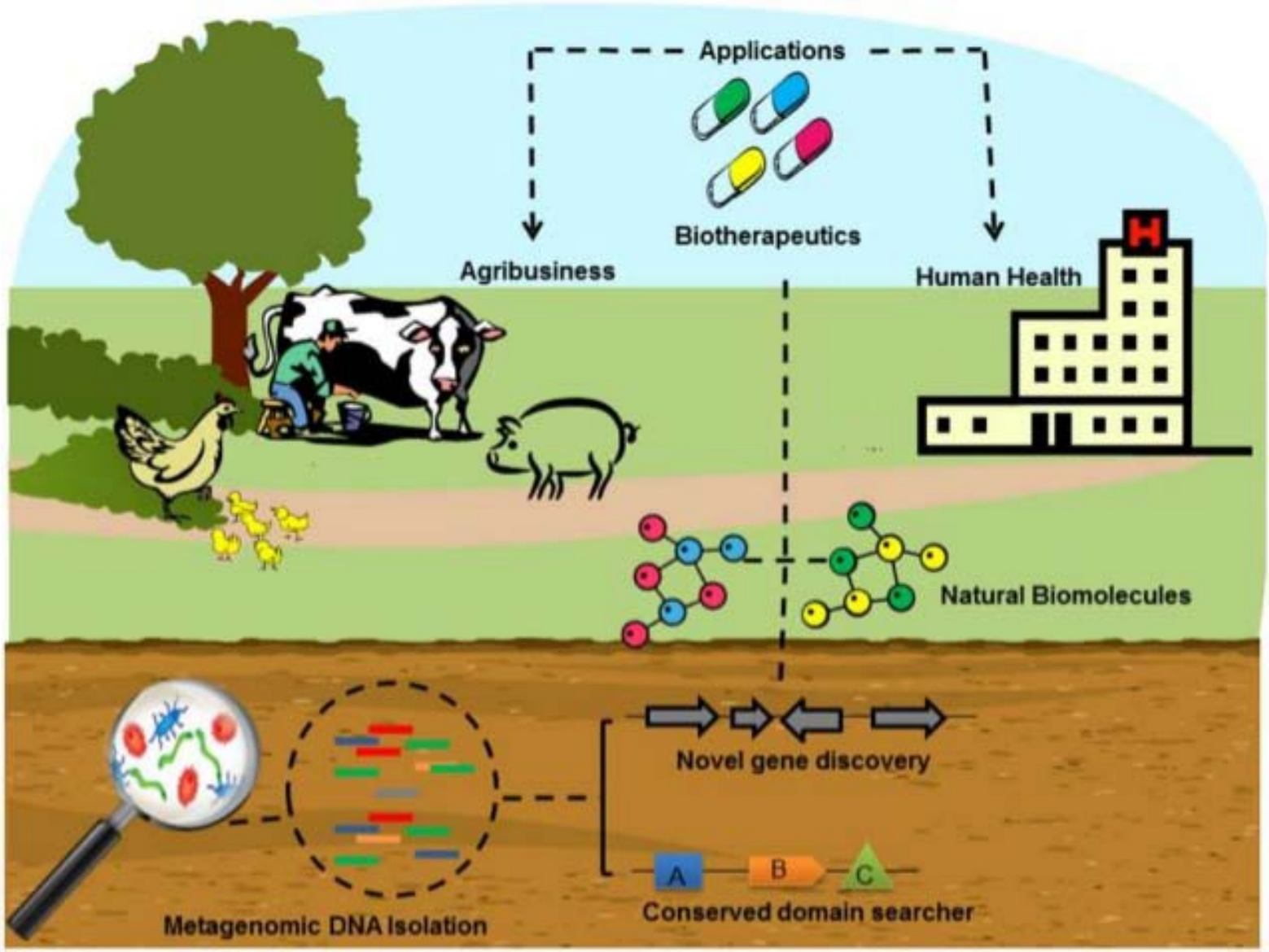
  community genomics.

- The term "**metagenomics**" was first used by

  Jo Handelsmann, Jon Clardy, Robert M. Goodman.

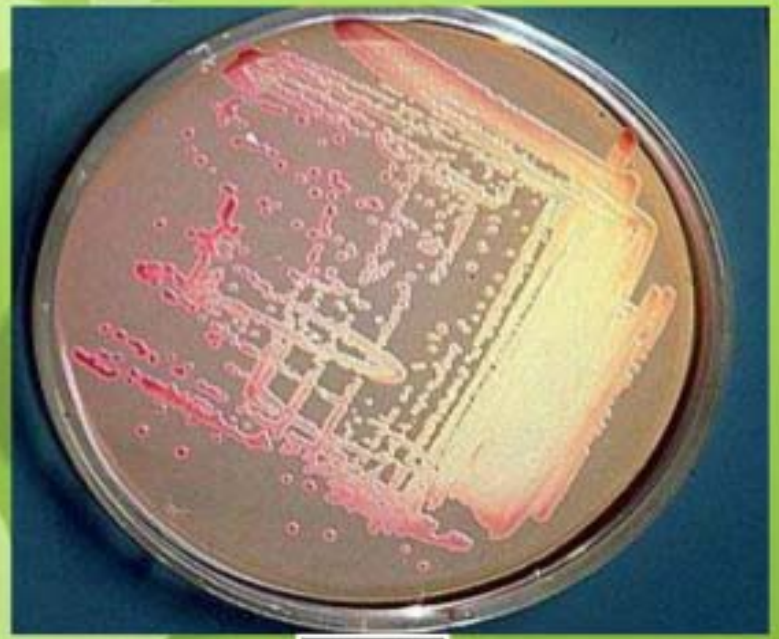
Why Metagenomics ?

- Microbes, as communities, are key-players in maintaining environmental stability.
- Investigate microbes in their natural environment.
- High-throughput gene-level studies of communities.
- Resource for development of novel genes, enzymes and Chemical compounds.


Applications
Biotherapeutics
Agribusiness
Human Health
H
Natural Biomolecules
Novel gene discovery
A
B
C
Conserved domain searcher
Metagenomic DNA Isolation

Genomics and Metagenomics

Isolate



Traditional microbial genomics

Community

Metagenomics

Which variable region to choose one ?

- Experimental design
- Sampling
- Sample processing
- Sequencing
- Sequence processing

Which NGS Platform ?

Platform Features

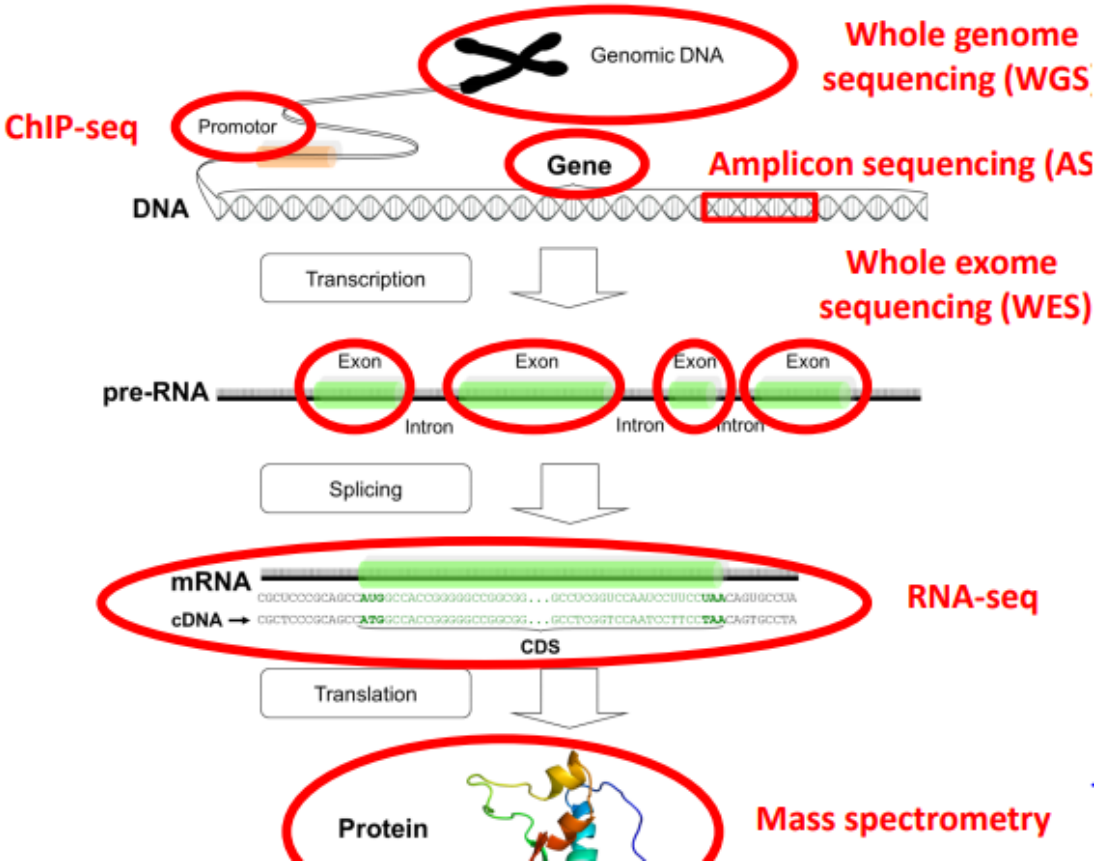






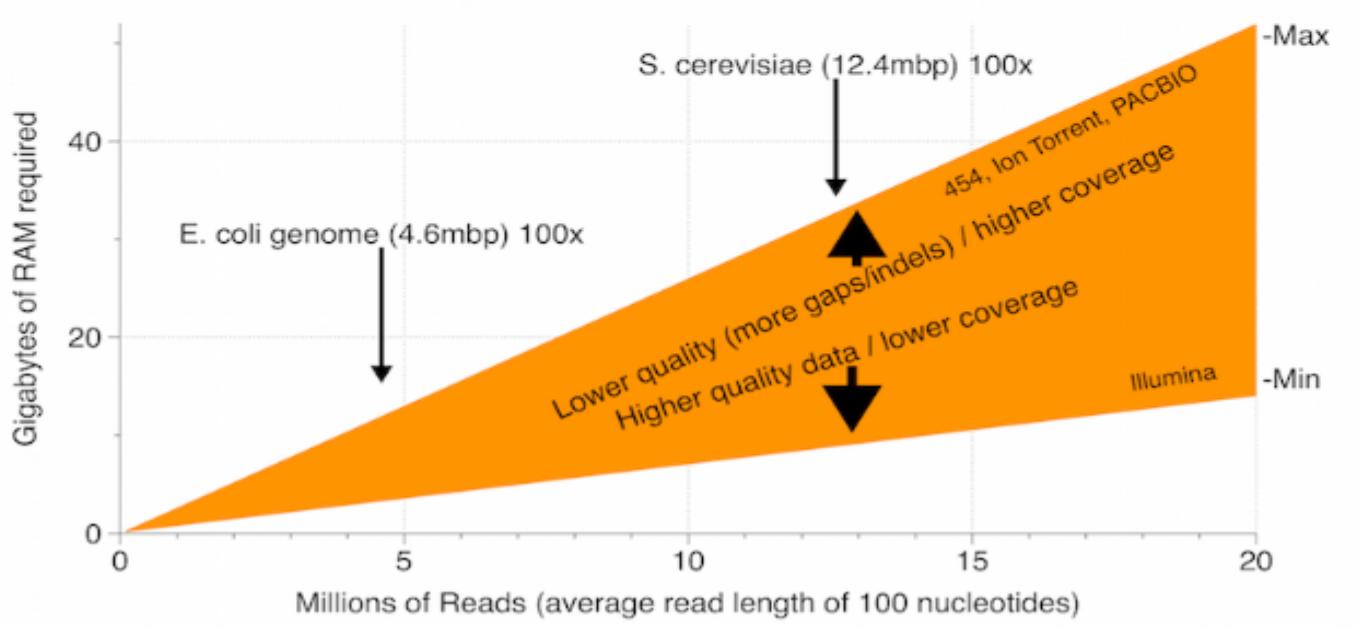
Feature	HiSeq2500 - Highoutput	HiSeq2500 – Rapid mode	MiSeq	PacBio RSII
Number of reads	150-180M/lane	100-150M/lane	12-15M (v2) 20-25M (v3)	50-80K/SMRT cell
Read length	2 x 100 bp	2 x 150 bp	2 x 300 bp (v3)	~ 10-20 kb
Yield per lane (PF data)	up to 35 Gb	up to 45Gb	up to 15 Gb	up to 0.4 Gb
Instrument Time	~12-14 days	~2 days	~2 days	~2 hours
Pricing per Gb	$59 (PE100)	$53 (PE150)	$108 (PE300)	$697

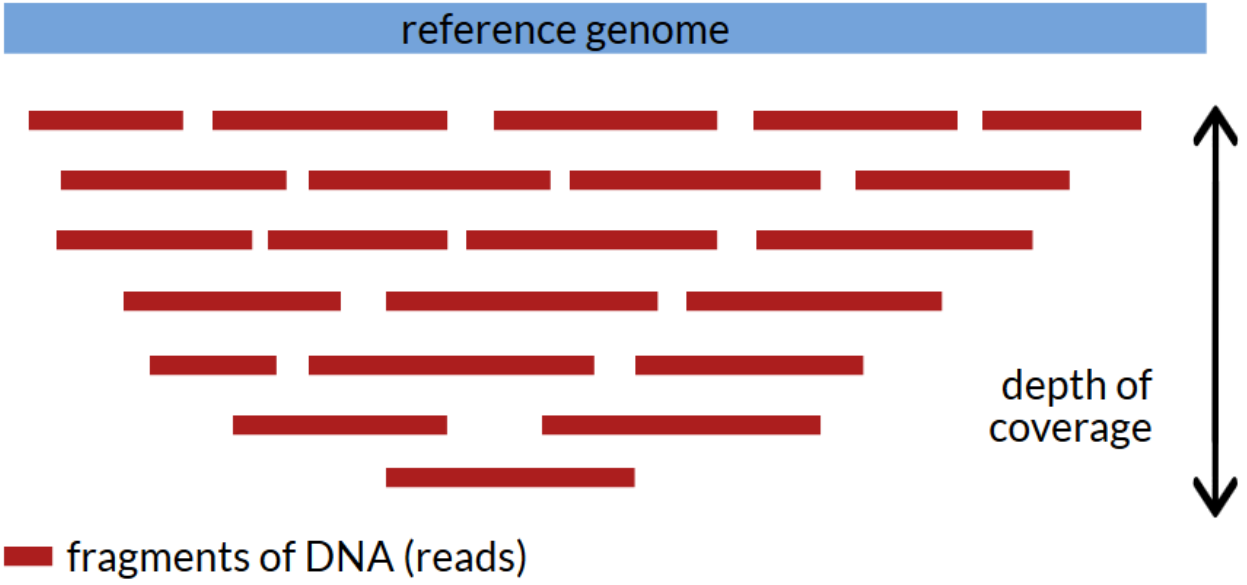
What do we want to sequence?



General rules of thumb for RAM requirements:

- Assembling data with higher coverage depth will require more RAM – aim for coverage between 50 and 100x
- Assembling lower quality data, with more miscalls, indels and gaps, will require more RAM
- Doubling the size of your dataset (total nucleotides) will roughly double RAM requirements
- For illumina data roughly 1 GB of RAM will be required to assemble a data set of 1 million reads (with an average read length of 100 nucleotides)

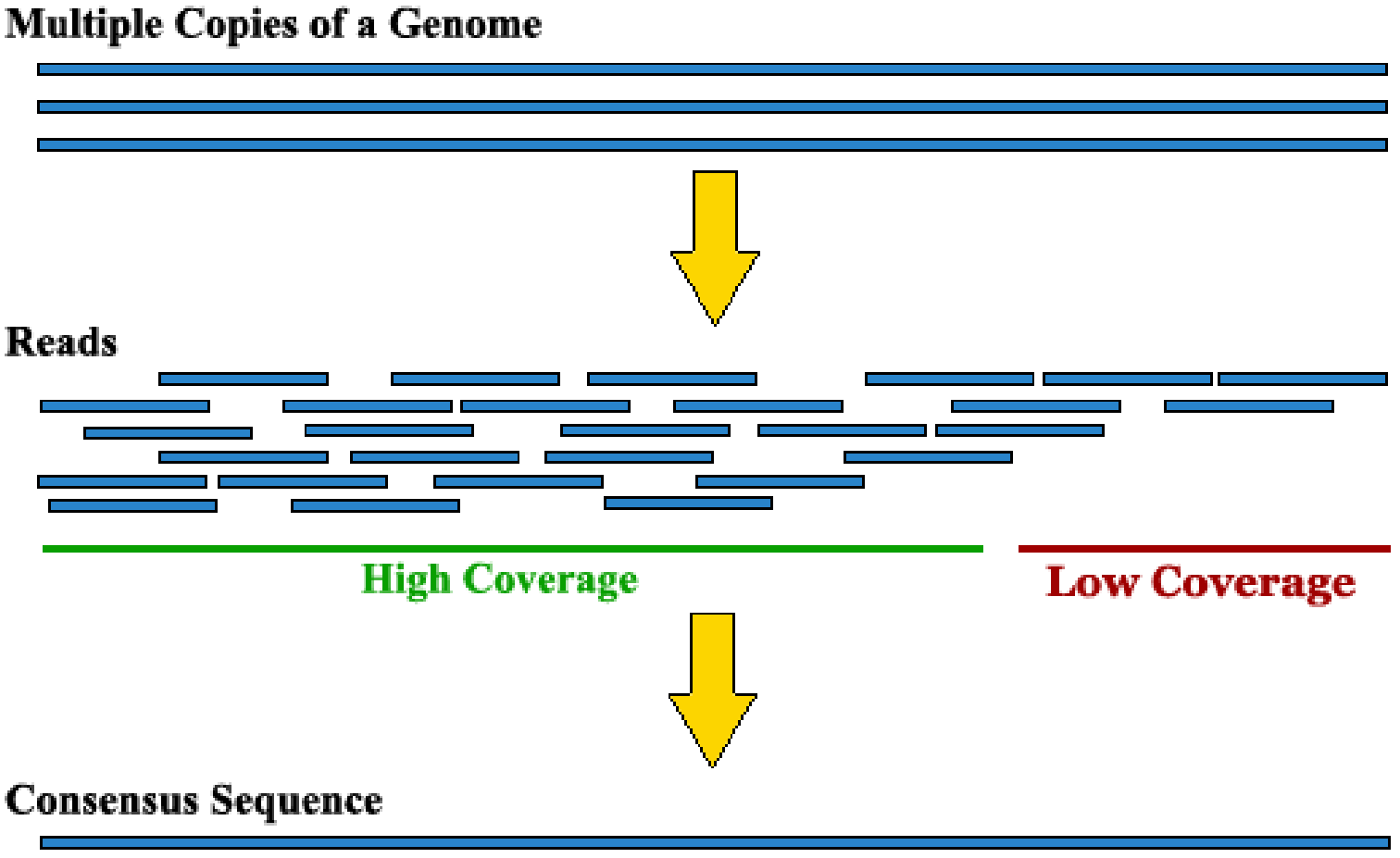


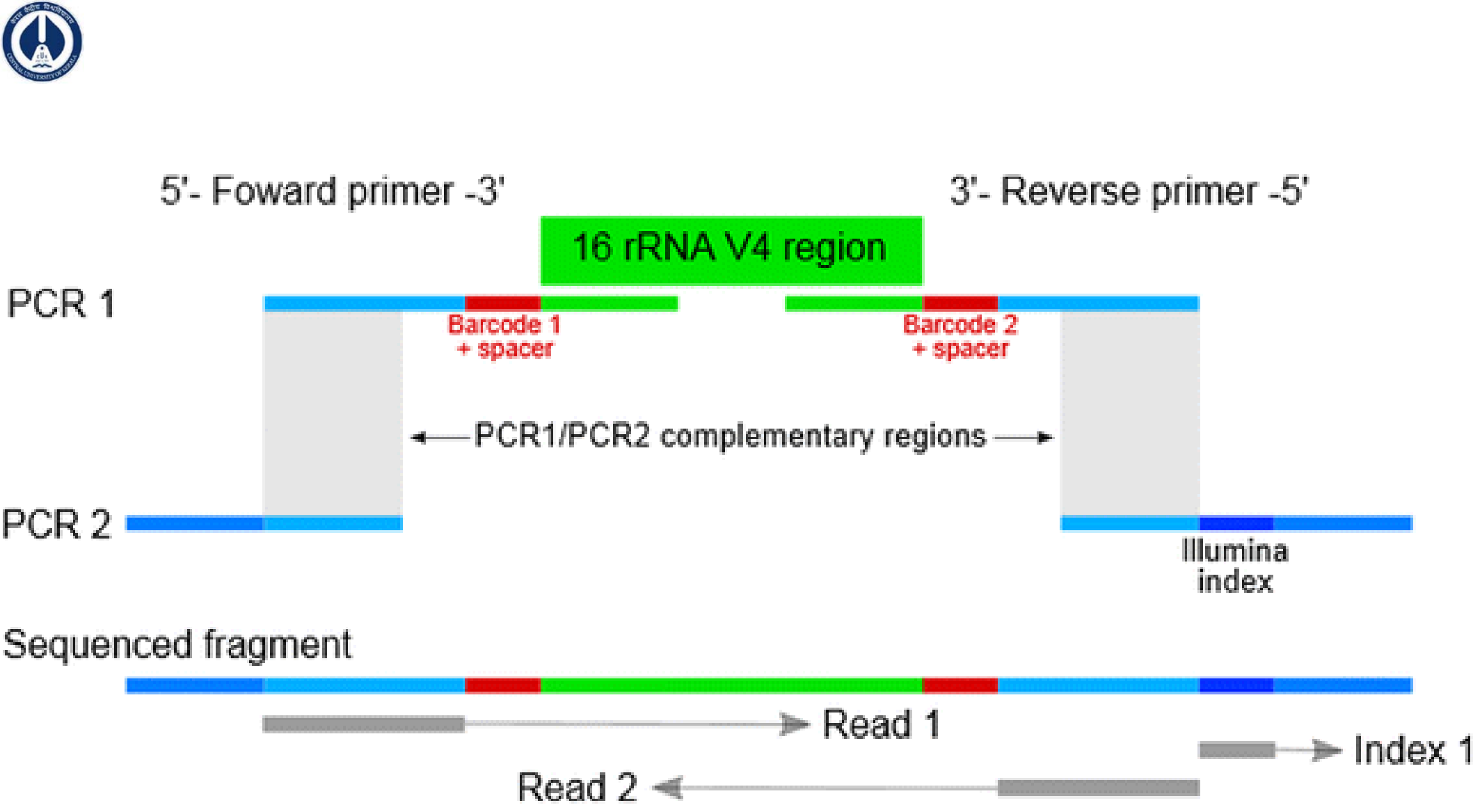


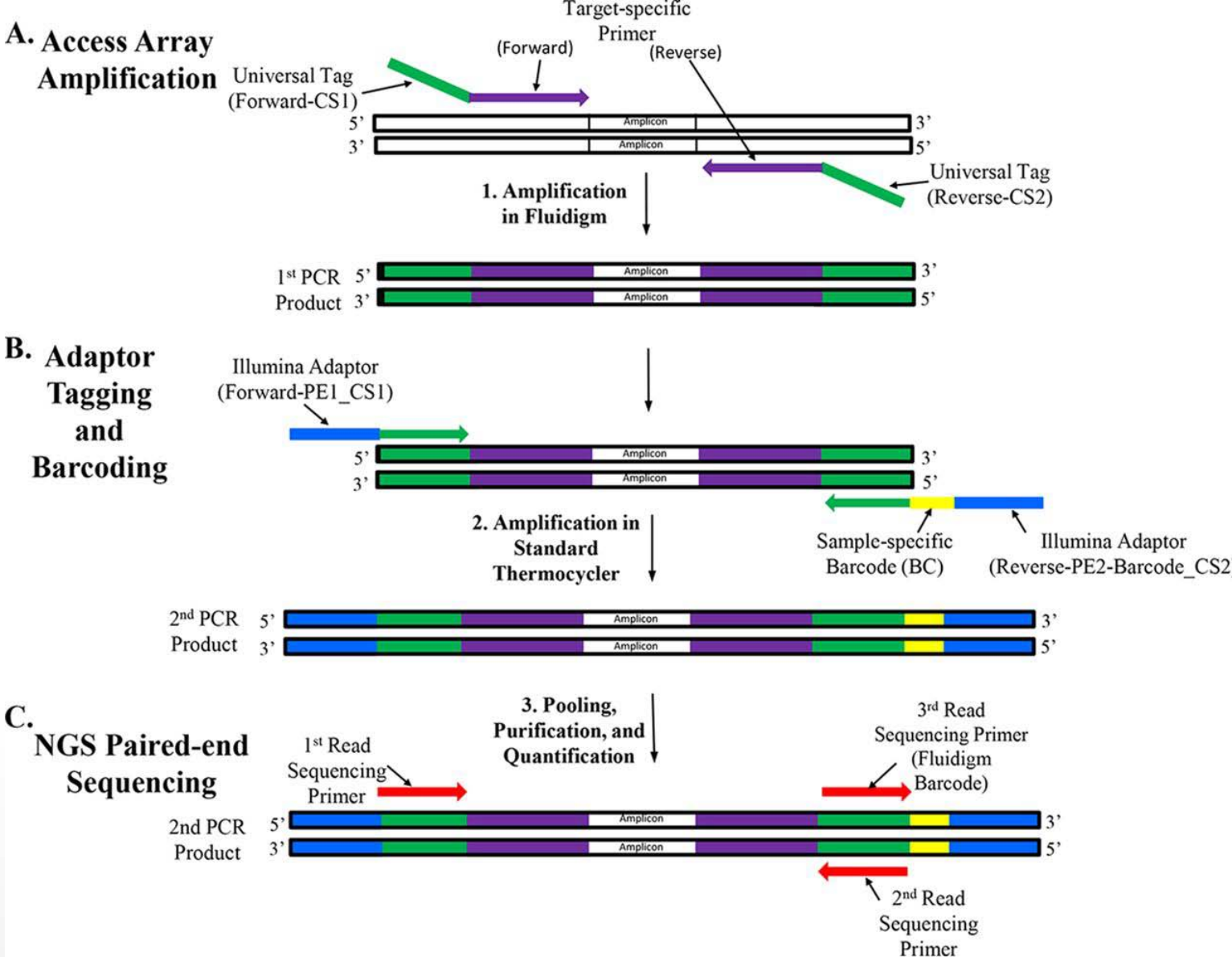
% coverage is used, e.g., in sequence mapping.

x coverage (or -fold coverage is used to describe the sequencing depth.

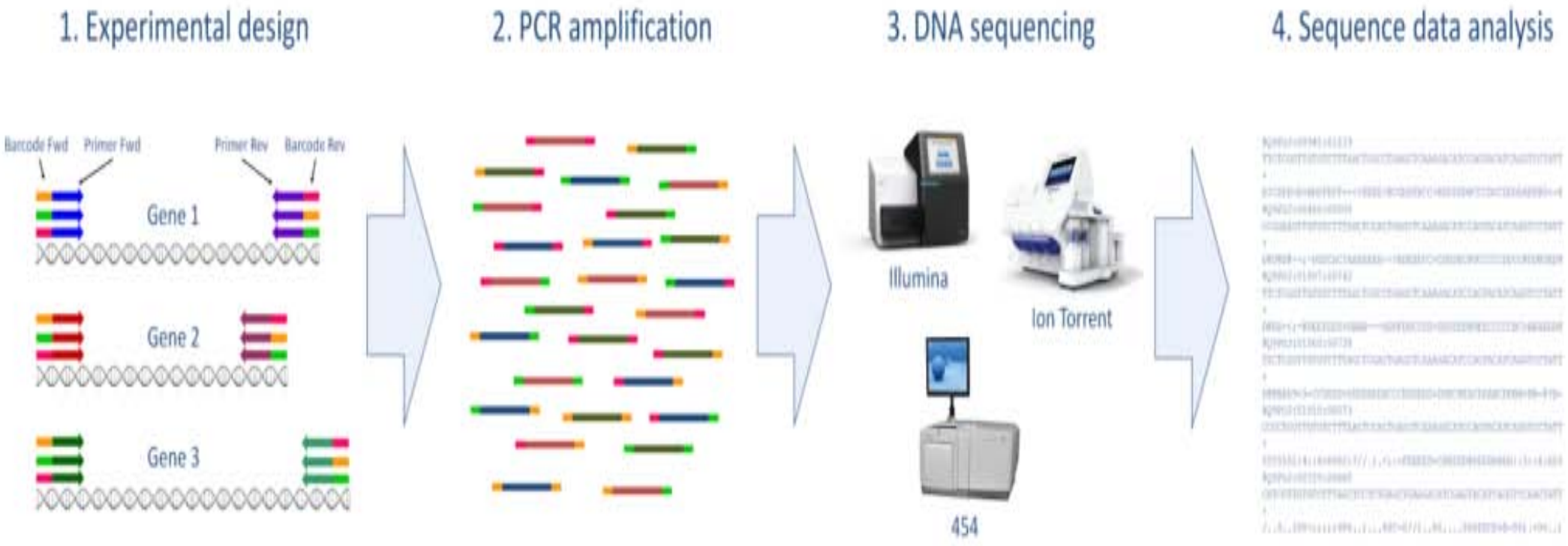
For example, if your genome has a size of 10 Mbp and you have 100 Mbp of sequencing data that is assembled to said 10 Mbp genome, you have 10x coverage.

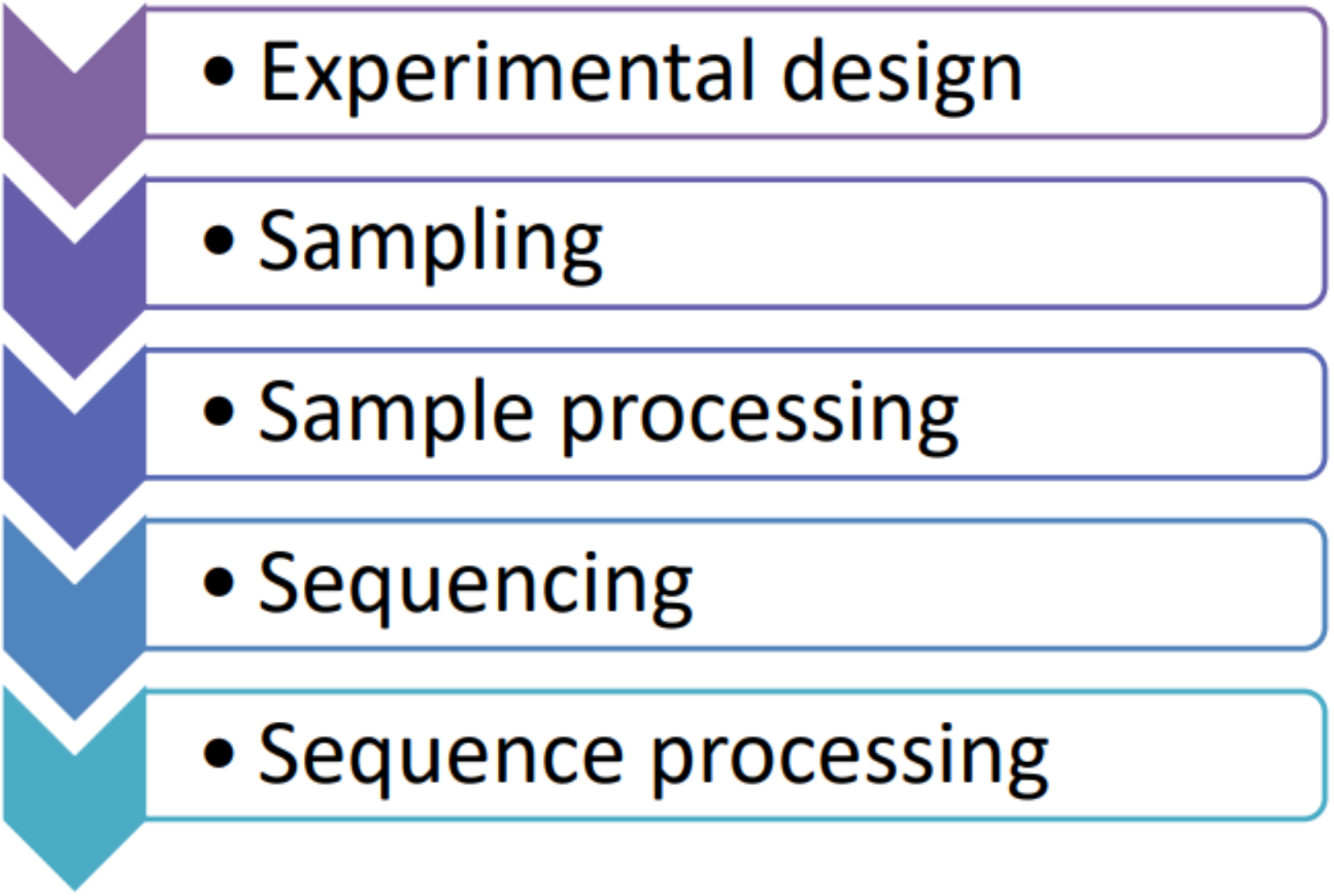

Multiple Copies of a Genome
Reads
High Coverage
Low Coverage
Consensus Sequence


5'- Foward primer -3'
3'- Reverse primer -5'
16 rRNA V4 region
PCR 1
Barcode 1
+ spacer
Barcode 2
+ spacer
PCR1/PCR2 complementary regions
PCR 2
Illumina
index
Sequenced fragment
Read 1
Index 1
Read 2


A. Access Array Amplification
Target-specific Primer
(Forward)
(Reverse)
Universal Tag (Forward-CS1)
5'
3'
Amplicon
Amplicon
3'
5'
Universal Tag (Reverse-CS2)
1. Amplification in Fluidigm
1st PCR Product
5'
3'
Amplicon
Amplicon
3'
5'
B. Adaptor Tagging and Barcoding
Illumina Adaptor (Forward-PE1_CS1)
5'
3'
Amplicon
Amplicon
3'
5'
2. Amplification in Standard Thermocycler
Sample-specific Barcode (BC)
Illumina Adaptor (Reverse-PE2-Barcode_CS2)
2nd PCR Product
5'
3'
Amplicon
Amplicon
3'
5'
C. NGS Paired-end Sequencing
3. Pooling, Purification, and Quantification
1st Read Sequencing Primer
3rd Read Sequencing Primer (Fluidigm Barcode)
2nd PCR Product
5'
3'
Amplicon
Amplicon
3'
5'
2nd Read Sequencing Primer

Amplicon Sequencing 4 steps workflow




• Experimental design
• Sampling
• Sample processing
• Sequencing
• Sequence processing

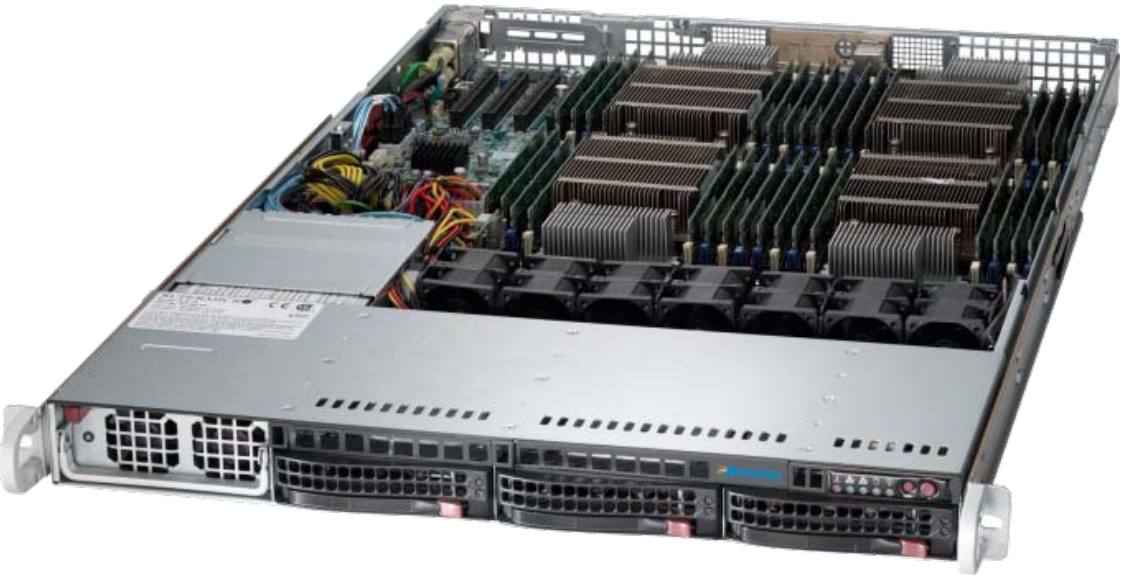
QIIME

Integrated Pipeline of Third-Party Tools

Required Configuration

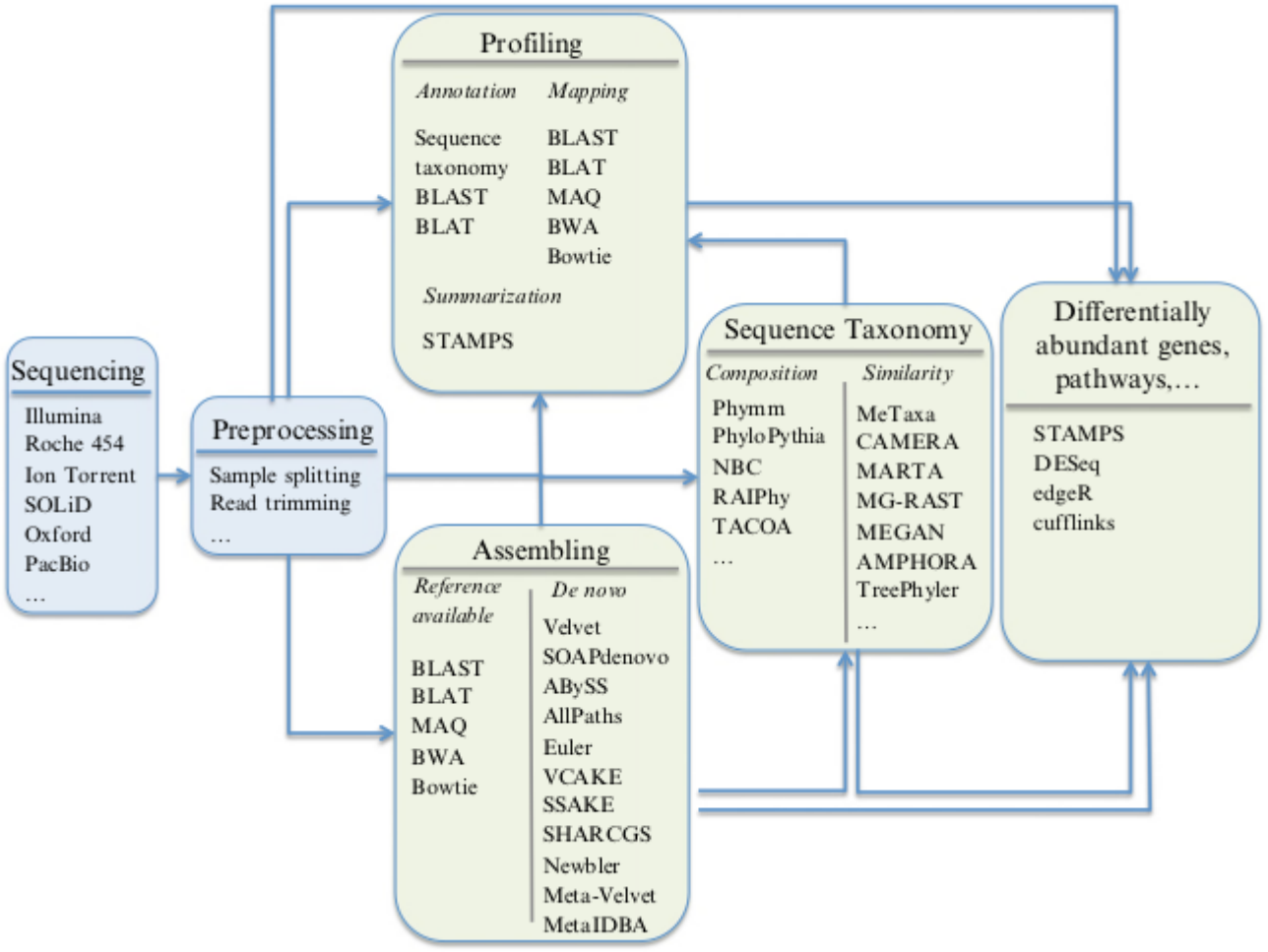
Configuration (To start with) **:**

4~8GB DDR3 Corsair + Intel i5 or i7 +
500GB~1TB HD is okay.



Required Configuration:

Intel Xeon E5-2640V2/ 2 GHz (8-Core),
128 GB DDR3 SDRAM,
1TB Standard Serial ATA 600 HDD 7200 rpm.


Sequencing
Illumina
Roche 454
Ion Torrent
SOLiD
Oxford
PacBio
…
Preprocessing
Sample splitting
Read trimming
…
Profiling
Annotation
Sequence taxonomy
BLAST
BLAT
Mapping
BLAST
BLAT
MAQ
BWA
Bowtie
Summarization
STAMPS
Assembling
Reference available
BLAST
BLAT
MAQ
BWA
Bowtie
De novo
Velvet
SOAPdenovo
ABySS
AllPaths
Euler
VCAKE
SSAKE
SHARCGS
Newbler
Meta-Velvet
MetaIDBA
Sequence Taxonomy
Composition
Phymm
PhyloPythia
NBC
RAIPhy
TACOA
…
Similarity
MeTaxa
CAMERA
MARTA
MG-RAST
MEGAN
AMPHORA
TreePhyler
…
Differentially abundant genes, pathways,…
STAMPS
DESeq
edgeR
cufflinks

Upstream Analysis Steps

Raw data processing

1) Join forward and reverse reads

merge forward and reverse reads (multiple samples)
multiple_join_paired_ends.py -i IlluminaPairedReads -o JoinedReads

Result: one folder per sample (each containing a file: fastqjoin.join.fastq)

Join Unpaired Ends

- The first step when using QIIME is to join the unpaired ends of the fastq files.

- Two reads are supplied for each sample: a forward read and a reverse read.

- The multiple_join_paired_ends.py script will join these two reads together for multiple samples.

- The output from this script will be a folder with three files for every one sample. The file named ‘fastqjoin.join.fastq’ is the file containing the successfully paired files.

Quality filter

filter out low base quality and rename samples

split_libraries_fastq.py -i *sequence-files* --sample_ids *new-sample-names* -o SEQ/ -q 19
--barcode_type 'not-barcoded'

 # Example (sample-list is separated by comma without space behind comma)

split_libraries_fastq.py -i JoinedReads/SampleA_L001_R1_001/fastqjoin.join.fastq**,**
JoinedReads/SampleB_L001_R1_001/fastqjoin.join.fastq
--sample_ids SampleA,SampleB -o SEQ/ -q 19
--barcode_type 'not-barcoded'

-o SEQ/ - output: save results to folder "SEQ"
-q 19 - accept base quality Phred >= Q20
--barcode_type 'not-barcoded' - barcode not present in sequence (already removed)

check total number of sequences in file seqs.fna

```
cat SEQ/seqs.fna | grep '>' | wc -l
```

QIIME's default Key tools : -

- uclust (Edgar, 2010). Used for OTU picking.
- usearch (Edgar, 2010). Used for OTU picking and chimera checking.
- RDP classifier (Wang et al., 2007). Used for taxonomy assignment.
- GreenGenes database (DeSantis et al., 2006) used as a reference database
- for taxonomy assignment and reference-based OTU picking.
- PyNAST (Caporaso, Bittinger, et al., 2010). Used for multiple sequence
- alignment.
- UniFrac (Lozupone & Knight, 2005). Used as a phylogenetic metric for beta-diversity analysis.

OTU picking

```
echo 'pick_otus:enable_rev_strand_match True' > otu_settings.txt
```

```
# run OTU clustering
pick_open_reference_otus.py -i $PWD/SEQ/seqs.fna -o $PWD/OTU/ -p $PWD/otu_settings.txt
```

Get OTU & read count per sample

```
biom summarize-table -i OTU/otu_table_mc2_w_tax_no_pynast_failures.biom --qualitative
                     -o stats_OTUs_per_sample.txt
```

```
# Get number of reads per sample
biom summarize-table -i OTU/otu_table_mc2_w_tax_no_pynast_failures.biom
                     -o stats_reads_per_sample.txt
```

Convert OTU table into text file

```
# Convert OTU .biom file to .tsv text file
biom convert -i otu_table_mc2_w_tax_no_pynast_failures.biom
             -o otu_table_mc2_w_tax_no_pynast_failures.tsv
             --to-tsv --header-key taxonomy
```

Convert OTU tables into relative abundance taxa tables

1) OTU's annotated by the same taxa are merged and converted to relative abundances

2) Relative taxa abundances are visualized as bar plots

```
summarize_taxa_through_plots.py -i OTU/otu_table_mc2_w_tax_no_pynast_failures.biom -o Taxa/ -m metadata.csv
```