CEITEC

Central European Institute of Technology

BRNO | CZECH REPUBLIC

Modern Genomic Technologies (LF:DSMGT01)

Lecture 1 : NGS Overview

Vojta Bystry
vojtech.bystry@ceitec.muni.cz

Introduction to LF: DSMGT01

- NGS data analysis for non-bioinformatics
  - Focus on experiment planning and result interpretation

1. Introduction; NGS Overview
2. DNA resequencing
3. DNA resequencing
4. RNA-seq
5. RNA-seq
6. Chip-seq (CLIP-seq)
7. Everything else ☺ + colloquium

- The plan is open to change - (based on your suggestions and wishes)

What is NGS?

- Next generation sequencing
  - New generation sequencing
  - HTP = High throughput
  - Massively parallel sequencing
- Contrast to Sanger sequencing

What is NGS?

Method	Read length	Accuracy (single read not consensus)	Reads per run	Time per run	Cost per 1 billion bases (in US$)	Advantages	Disadvantages
Single-molecule real-time sequencing (Pacific Biosciences)	30,000 bp (N50); maximum read length >100,000 bases[75][76][77]	87% raw-read accuracy[78]	4,000,000 per Sequel 2 SMRT cell, 100–200 gigabases[75][79][80]	30 minutes to 20 hours[75][81]	$7.2-$43.3	Fast. Detects 4mC, 5mC, 6mA.[82]	Moderate throughput. Equipment can be very expensive.
Ion semiconductor (Ion Torrent sequencing)	up to 600 bp[83]	99.6%[84]	up to 80 million	2 hours	$66.8-$950	Less expensive equipment. Fast.	Homopolymer errors.
Pyrosequencing (454)	700 bp	99.9%	1 million	24 hours	$10,000	Long read size. Fast.	Runs are expensive. Homopolymer errors.
Sequencing by synthesis (Illumina)	MiniSeq, NextSeq: 75–300 bp; MiSeq: 50–600 bp; HiSeq 2500: 50–500 bp; HiSeq 3/4000: 50–300 bp; HiSeq X: 300 bp	99.9% (Phred30)	MiniSeq/MiSeq: 1–25 Million; NextSeq: 130-00 Million; HiSeq 2500: 300 million – 2 billion; HiSeq 3/4000 2.5 billion; HiSeq X: 3 billion	1 to 11 days, depending upon sequencer and specified read length[85]	$5 to $150	Potential for high sequence yield, depending upon sequencer model and desired application.	Equipment can be very expensive. Requires high concentrations of DNA.
Combinatorial probe anchor synthesis (cPAS- BGI/MGI)	BGISEQ-50: 35-50bp; MGISEQ 200: 50-200bp; BGISEQ-500, MGISEQ-2000: 50-300bp[86]	99.9% (Phred30)	BGISEQ-50: 160M; MGISEQ 200: 300M; BGISEQ-500: 1300M per flow cell; MGISEQ-2000: 375M FCS flow cell, 1500M FCL flow cell per flow cell.	1 to 9 days depending on instrument, read length and number of flow cells run at a time.	$5- $120		
Sequencing by ligation (SOLiD sequencing)	50+35 or 50+50 bp	99.9%	1.2 to 1.4 billion	1 to 2 weeks	$60-130	Low cost per base.	Slower than other methods. Has issues sequencing palindromic sequences.[87]
Nanopore Sequencing	Dependent on library preparation, not the device, so user chooses read length (up to 2,272,580 bp reported[88]).	~92–97% single read	dependent on read length selected by user	data streamed in real time. Choose 1 min to 48 hrs	$7-100	Longest individual reads. Accessible user community. Portable (Palm sized).	Lower throughput than other machines, Single read accuracy in 90s.
GenapSys Sequencing	Around 150 bp single-end	99.9% (Phred30)	1 to 16 million	Around 24 hours	$667	Low-cost of instrument ($10,000)	
Chain termination (Sanger sequencing)	400 to 900 bp	99.9%	N/A	20 minutes to 3 hours	$2,400,000	Useful for many applications.	More expensive and impractical for larger sequencing projects. This method also requires the time-consuming step of plasmid cloning or PCR.

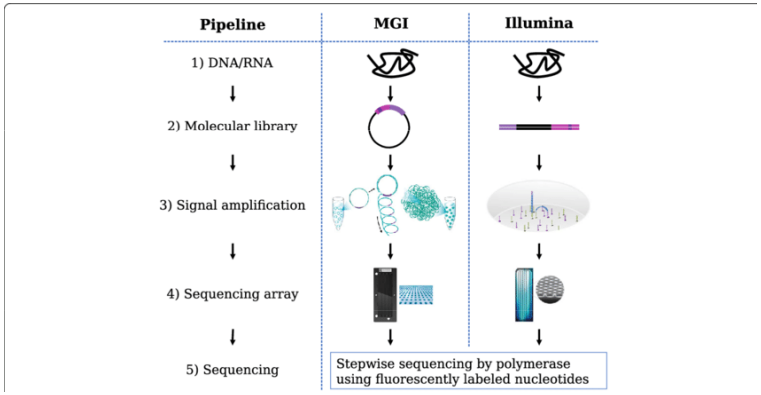
What is NGS?

Method	Read length	Accuracy (single read not consensus)	Reads per run	Time per run	Cost per 1 billion bases (in US$)	Advantages	Disadvantages
Single-molecule real-time sequencing (Pacific Biosciences)	30,000 bp (N50); maximum read length >100,000 bases[75][76][77]	87% raw-read accuracy[78]	4,000,000 per Sequel 2 SMRT cell, 100–200 gigabases[75][79][80]	30 minutes to 20 hours[75][81]	$7.2-$43.3	Fast. Detects 4mC, 5mC, 6mA.[82]	Moderate throughput. Equipment can be very expensive.
Ion semiconductor sequencing)							
Sequencing by synthesis (Illumina)	MiniSeq, NextSeq: 75–300 bp; MiSeq: 50–600 bp; HiSeq 2500: 50–500 bp; HiSeq 3/4000: 50–300 bp; HiSeq X: 300 bp	99.9% (Phred30)	MiniSeq/MiSeq: 1–25 Million; NextSeq: 130-00 Million; HiSeq 2500: 300 million – 2 billion; HiSeq 3/4000 2.5 billion; HiSeq X: 3 billion	1 to 11 days, depending upon sequencer and specified read length[85]	$5 to $150	Potential for high sequence yield, depending upon sequencer model and desired application.	Equipment can be very expensive. Requires high concentrations of DNA.
Combinatorial probe anchor synthesis (cPAS- BGI/MGI)	BGISEQ-50: 35-50bp; BGISEQ-500, MGISEQ-2000: 50-300bp[86]	(Phred30)	BGISEQ-50: 160M; MGISEQ 200: 300M; BGISEQ-500: 1300M per flow cell; MGISEQ-2000: 375M FCS flow cell, 1500M FCL flow cell per flow cell.	1 to 9 days depending on number of flow cells run at a time.			
Sequencing by sequencing)							Slower than other methods. Has issues sequencing palindromic sequences.[87]
Nanopore Sequencing	Dependent on library preparation, not the device, so user chooses read length (up to 2,272,580 bp reported[88]).	~92–97% single read	dependent on read length selected by user	data streamed in real time. Choose 1 min to 48 hrs	$7-100	Longest individual reads. Accessible user community. Portable (Palm sized).	Lower throughput than other machines, Single read accuracy in 90s.
GenapSys Sequencing		99.9% (Phred30)					
Chain termination (Sanger sequencing)	400 to 900 bp	99.9%	N/A	20 minutes to 3 hours	$2,400,000	Useful for many applications.	More expensive and impractical for larger sequencing projects. This method also requires the time-consuming step of plasmid cloning or PCR.

What is NGS?

- Illumina – sequencing by synthesis
- Oxford Nanopore – Nanopore sequencing
- Pacific Bioscience - Single Molecule, Real-Time (SMRT)

- Chinese are coming - BGI DNBSeq platforms

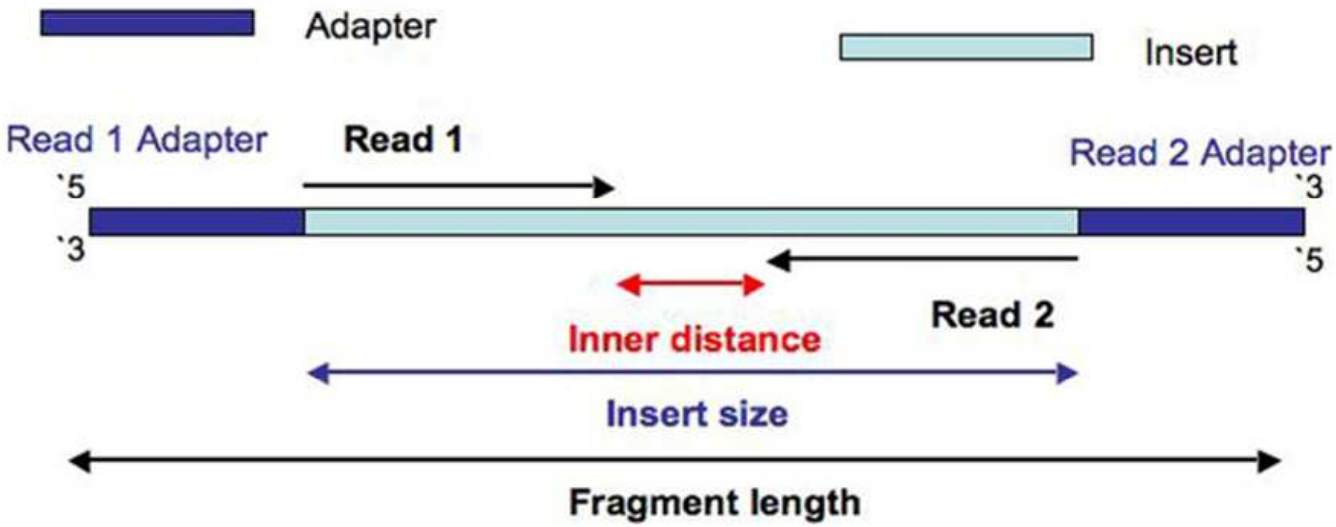


What is NGS?

- Illumina – sequencing by synthesis
- Oxford Nanopore – Nanopore sequencing
- Pacific Bioscience - Single Molecule, Real-Time (SMRT)

Illumina – sequencing by synthesis

- https://www.youtube.com/watch?v=fCd6B5HRaZ8



Raw data

```
>read_no_1
CGGCCTGGAGGCCCTGCAGAACCTGCTGGGCTACAGGTTCGGCGACGAGGG

>read_no_2
GCAGCGTGAGCGCCATCATGGGCAACCCCCAGGTGAAGGCCCACGGCAAGA

>read_no_3
GGGAGACACCCGCACGTGTGGCCCGCATGTATGCTGAGCTCTTCCGCGGAT

>read_no_4
TTTGCCCCGCATCGAGCGGGCTGTGCGGGAAATCCTTCTGGCTGTAGGCGA

>read_no_5
CCTGTGGGGCAAGGTGAACCCCGTGGAGATCGGCGCCGAGAGCCTGGCCAG

>read_no_6
GAGGAGGGCCAGGATCCACCAGAGGAAGGGCCTGCTGTGGTTCATCCCCGC

>read_no_7
CTGCACAGCGACTACAACCTGACCTGGTACAGGAACGGCAGCAACATGCCC

>read_no_8
GTGCTGGGCCTGGCCATCAGCCACTTCCTGCTGGAGCAGTTCCCCGACTAC

>read_no_9
AACCTGGGCGAGTACCTGCTGCTGGGCAAGGGCGAGGAGATGACCGGCGGC

>read_no_10
GTTCCCCGACTACAACGAGGGCGAGCTGAGCAGGCTGAGGAGCGCCATCGT

>read_no_11
CTTCAGCAAGTTCGGCGACCTGAGCAGCGTGAGCGCCATCATGGGCAACCC

>read_no_12
ACCAGAGGAAGGGCCTGCTGTGGTTCATCCCCGCCGCCCTGGAGGACAGCG

>read_no_13
AAGGGCGAGGAGATGACCGGCGGCAGGAGGAAGGCCAGCCTGCTGGCCGAC
```

- 10^5 – 10^10 reads
- 75 – 300Bp
- Could be pair-end

Basic workflow

Experimental design

Library preparation

Sequencing

Data analysis

Basic workflow

Experimental design

Library preparation

Sequencing

Data analysis

Why we sequence

What we sequence

How we sequence

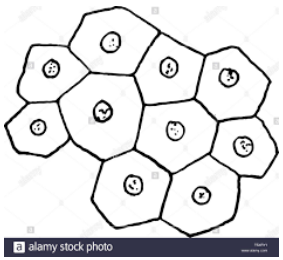
Basic workflow

Experimental design | Library preparation | Sequencing | Data analysis

Why we sequence | **What we sequence** | **How we sequence**

Consultation regarding data analysis is highly advisable.

NGS library preparation



Living material

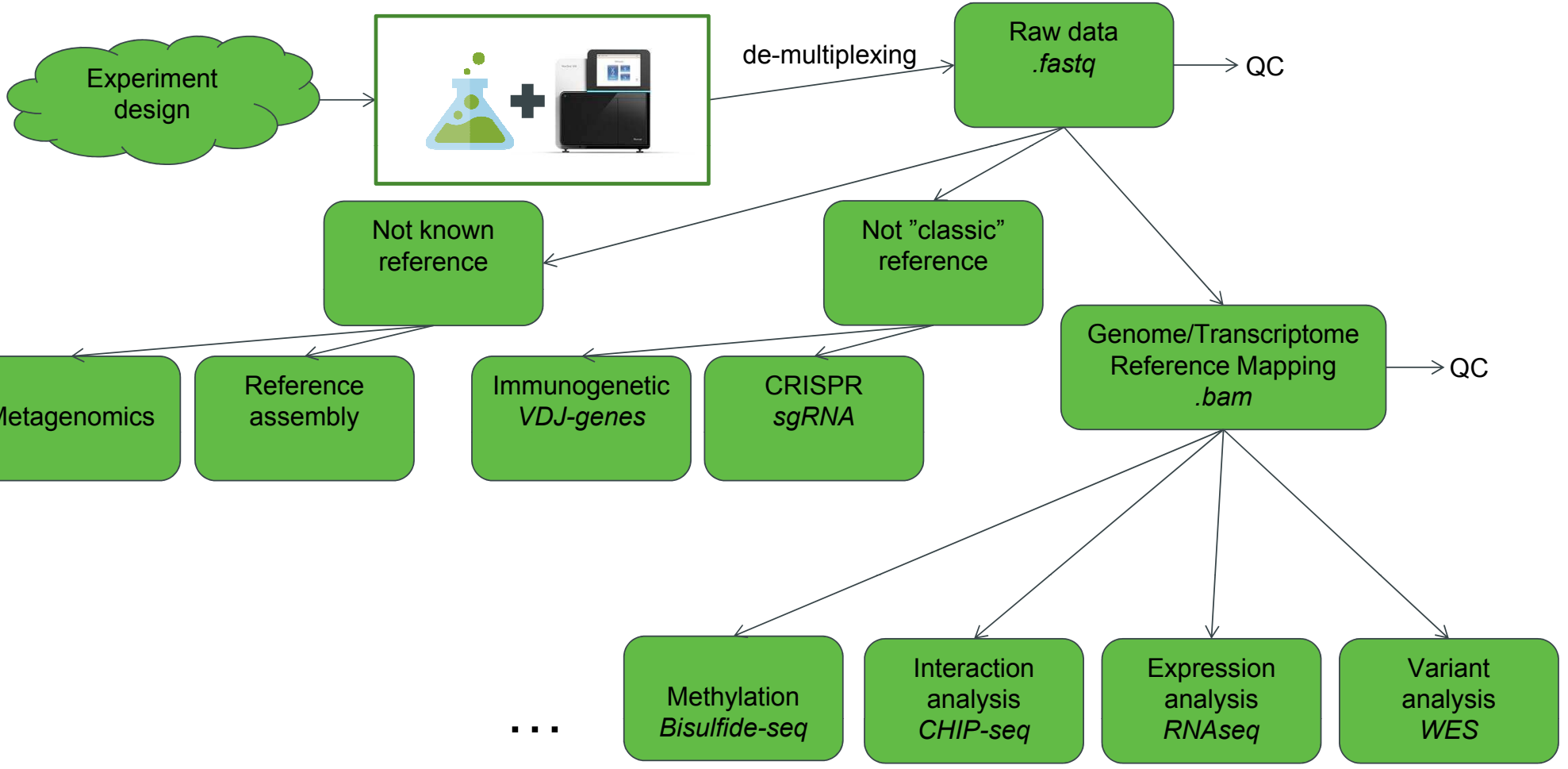
DNA

RNA

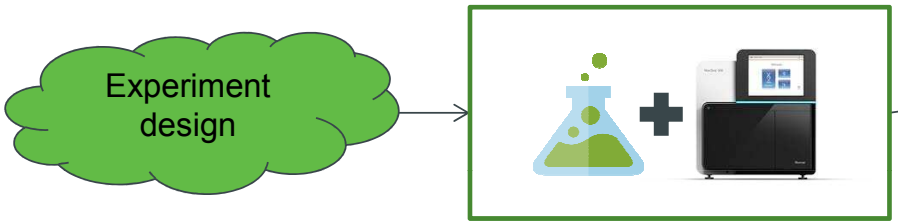
Select
some parts


For all you seq...
www.illumina.com/LibraryPrepMethods
RNA Transcription
ChIRP-Seq
GRO-seq
BRIC-Seq
Bru-Seq
BruChase-Seq
Ribo-Seq
ARTseq™
RIP-Seq
HITS-CLIP
CLIP-Seq
PAR-CLIP
iCLIP
NET-Seq
TRAP-Seq
CLASH
PARE-Seq
GMUCT 1.0
GMUCT 2.0
TIF-Seq
PEAT
SNORE-Seq
RNA Structure
SHAPE-Seq
Struc-ture-Seq
SPARE
PARS
dsRNA-Seq
Frag-Seq
Cap-Seq
CIP-TAP
ICE
MeRIP-Seq
Rearrangements and Markers
RAD
Paired-end
RAD-Seq
RC-Seq
Methylation
BS-Seq
WGBS
EpiGnome™
Methyl-Seq
RRBS-Seq
PBAT
T-WGBS
oxBS-Seq
TAB-Seq
MeDIP-Seq
MethylCap-Seq
MBDCap-Seq
DNA-Protein Interactions
DNase-Seq
DNaseI-Seq
MAINE-Seq
MNase-Seq
ChIP-Seq
ChIP-exo
FAIRE-Seq

NGS data analysis



NGS data analysis



Experiment design

de-multiplexing

Raw data
.fastq

QC

Not known reference

Not ”classic” reference

Metagenomics

Reference assembly

Immunogenetic
VDJ-genes

CRISPR
sgRNA

Genome/Transcriptome Reference Mapping
.bam

QC

...

Methylation
Bisulfide-seq

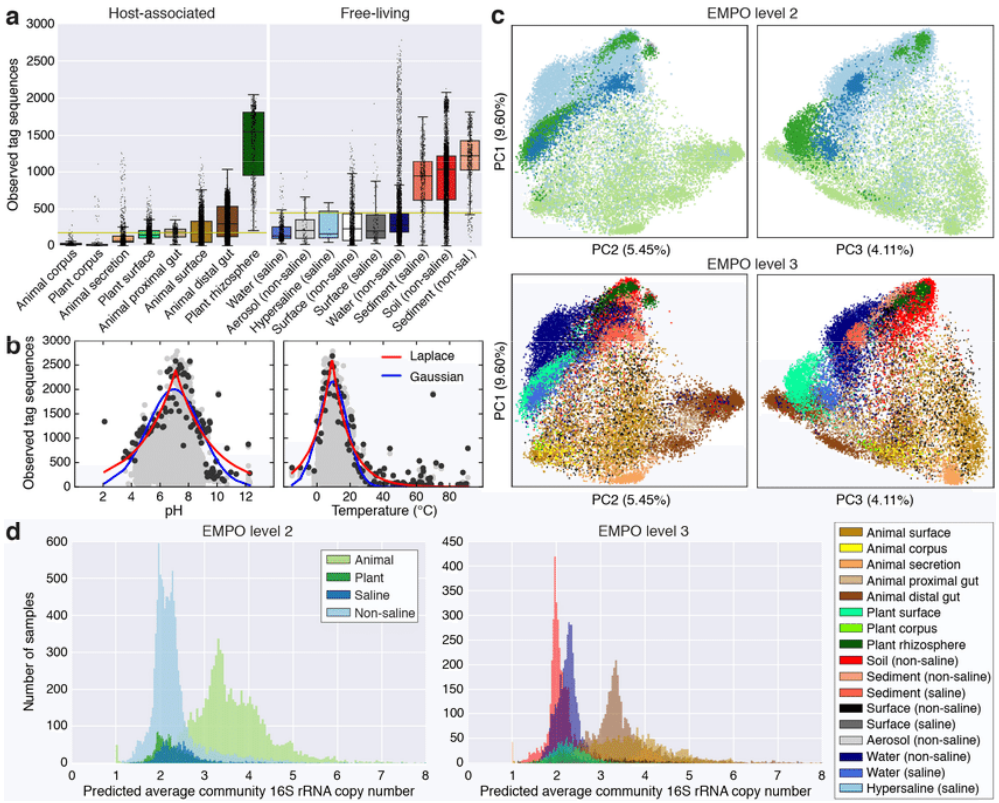
Interaction analysis
CHIP-seq

Expression analysis
RNAseq

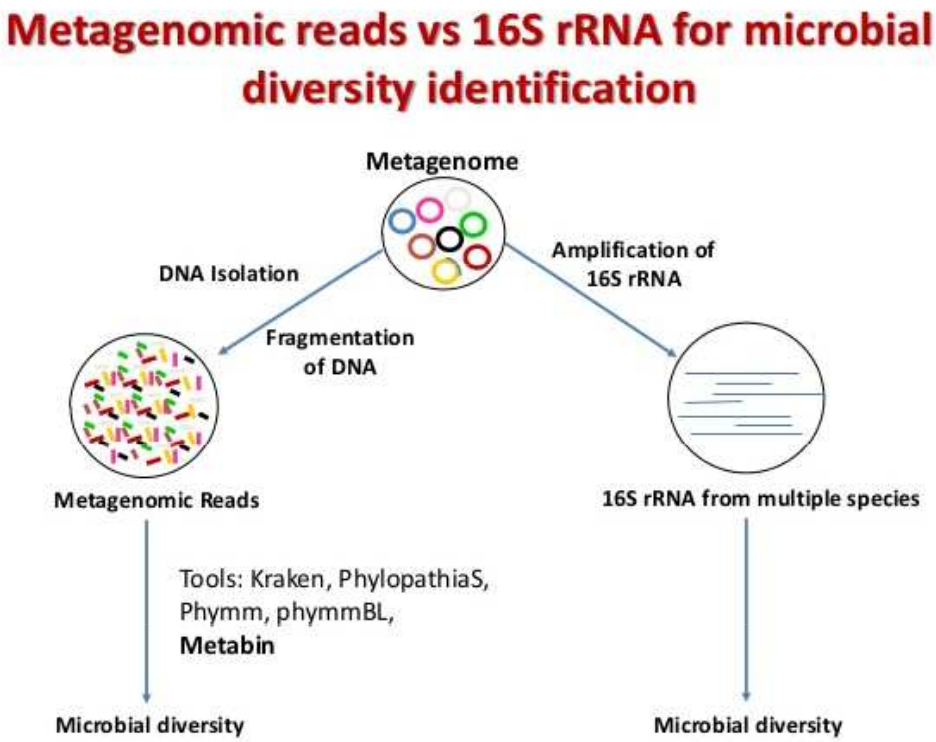
Variant analysis
WES

Metagenomics

- Environmental statistics about populations
  - alpha, beta, gamma diversity
  - identify known bacterial species
  - eventually functional profiling
    - E.g. antimicrobial resistance genes
- Sequencing techniques
  - 16S rRNA sequencing
  - Shotgun metagenomic sequencing



Metagenomics – 16S rRNA vs. Shotgun

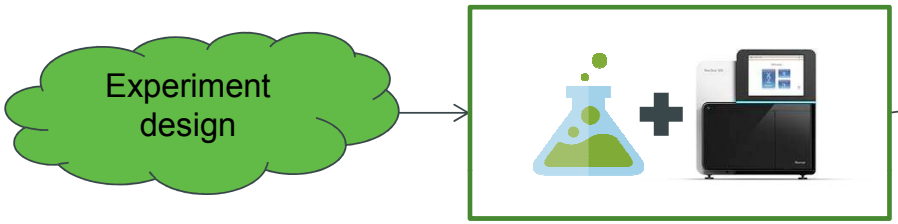


Factors	16S rRNA sequencing	Shotgun Metagenomic Sequencing
Cost	~$50 USD	Starting at ~$150 but price will depend on sequencing depth required
Sample preparation	Similar complexity to shotgun sequencing	Similar complexity to 16S rRNA sequencing
Functional profiling (profile microbial genes)	No (but 'predicted' functional profiling is possible)	Yes (but it only reveals information on functional potential)
Taxonomic resolution: Genus, species, strain?	Bacterial genus (sometimes species); dependent on region(s) targeted	Bacterial species (sometimes strains and single nucleotide variants, if sequencing is deep enough)
Taxonomic coverage	Bacteria and archaea	All taxa, including viruses
Bioinformatics requirements	Beginner to intermediate expertise	Intermediate to advanced expertise
Databases	Established, well-curated	Relatively new, still growing
Sensitivity to host DNA contamination	Low (but PCR success depends on the absence of inhibitors and the presence of a detectable microbiome)	High , varies with sample type (but this can be mitigated by calibrating the sequencing depth)
Bias	Medium to high (retrieved taxonomic composition is dependent on selected primers and targeted variable region)	Lower (while metagenomics is "untargeted", experimental and analytical biases can be introduced at various stages)

Metagenomics – 16S rRNA vs. Shotgun

- Study Examples
  - **Assessment of the bacterial microbiome of Amazonian soil**
    - 16S rRNA sequencing may provide more taxonomic resolution
  - **Changes in microbiome composition and antimicrobial gene carriage following fecal transplant**
    - shotgun sequencing to assess both compositional and functional differences
  - **Daily fluctuations in gut microbiome following 2 week dietary fiber intervention**
    - shotgun sequencing to assess both compositional and functional differences

NGS data analysis



Experiment design

de-multiplexing

Raw data
.fastq

QC

Not known reference

Not "classic" reference

Metagenomics

Reference assembly

Immunogenetic
VDJ-genes

CRISPR
sgRNA

Genome/Transcriptome Reference Mapping
.bam

QC

...

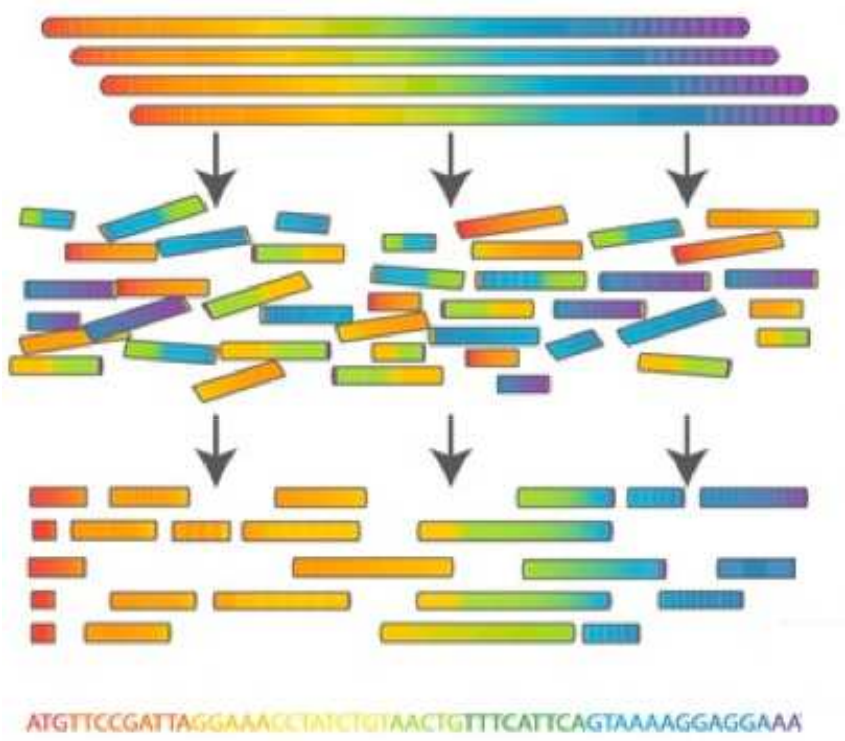
Methylation
Bisulfide-seq

Interaction analysis
CHIP-seq

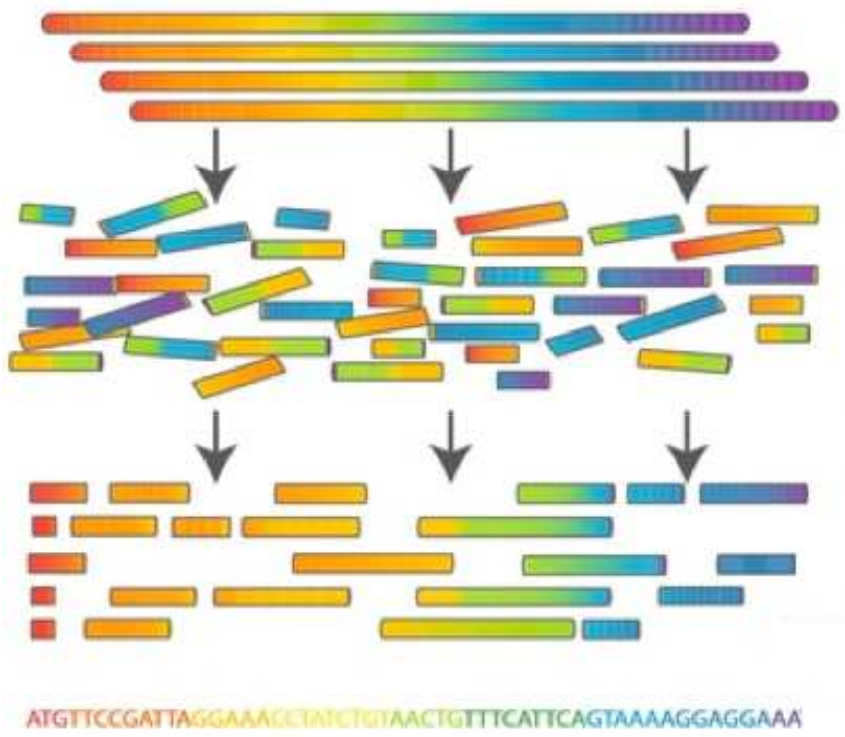
Expression analysis
RNAseq

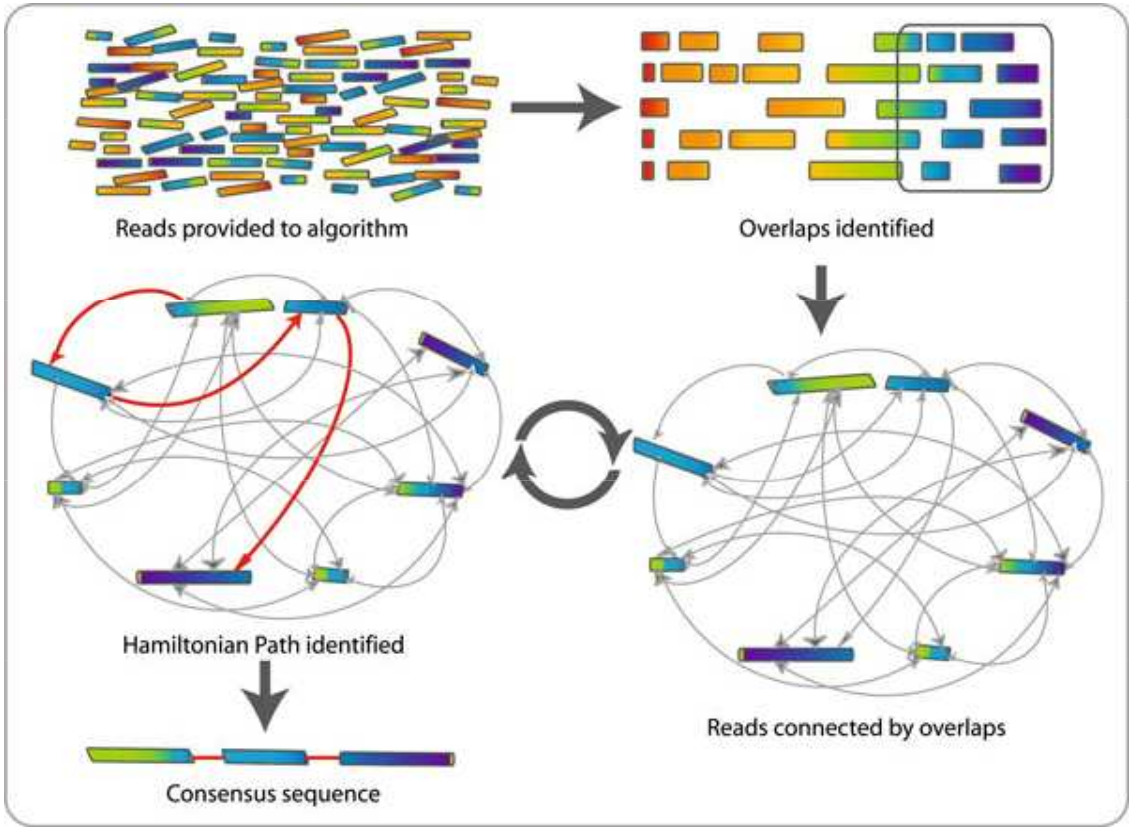
Variant analysis
WES

Reference Assembly



Reference Assembly

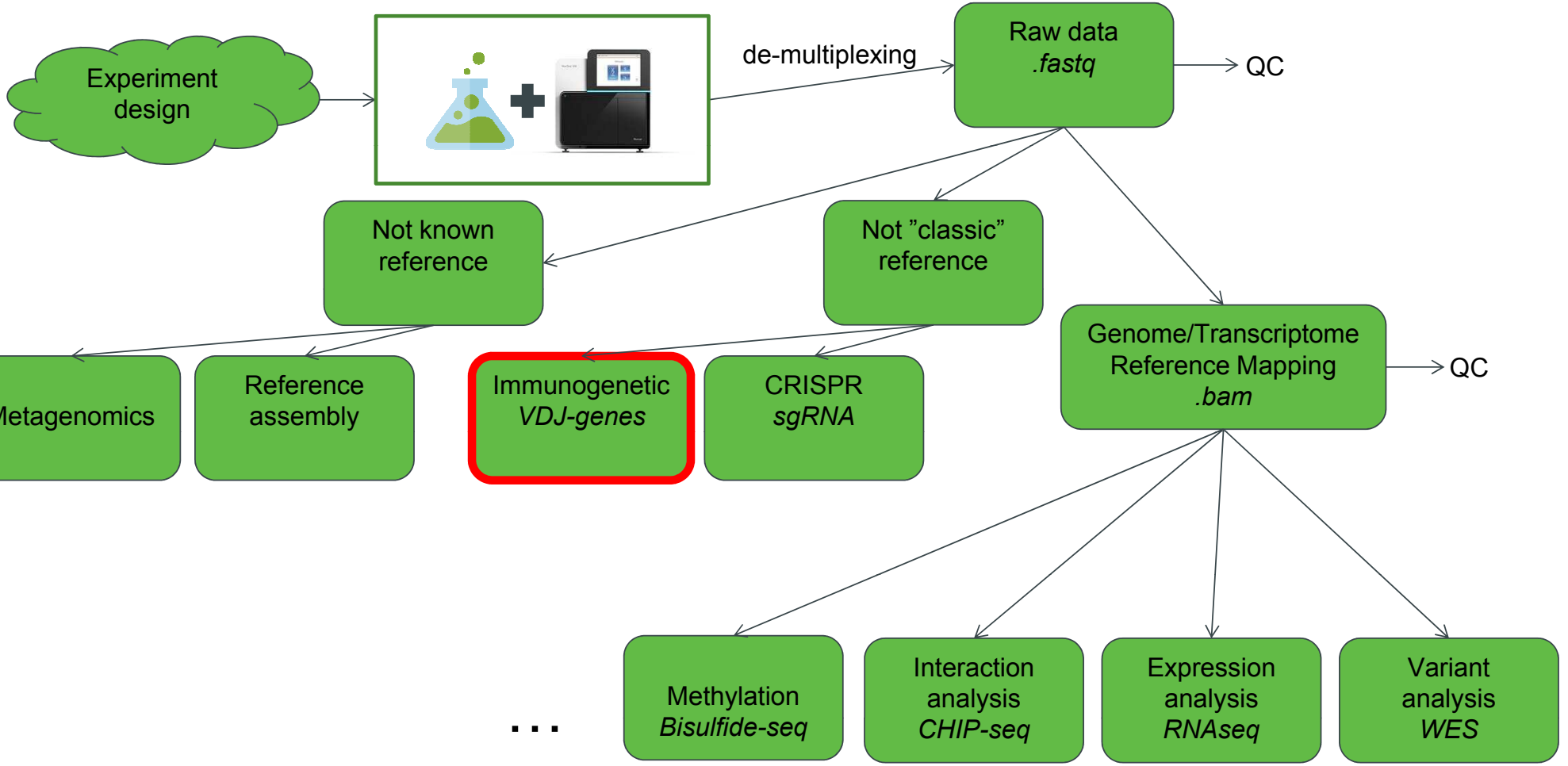




Reference Assembly

- Genome – DNA – *very hard and costly*
- Transcriptome – RNA
- Multiple sequencing types highly beneficial
  - Pair-end
  - Long reads
  - Mate-pairs
- Similar reference helpful – assembly by homology

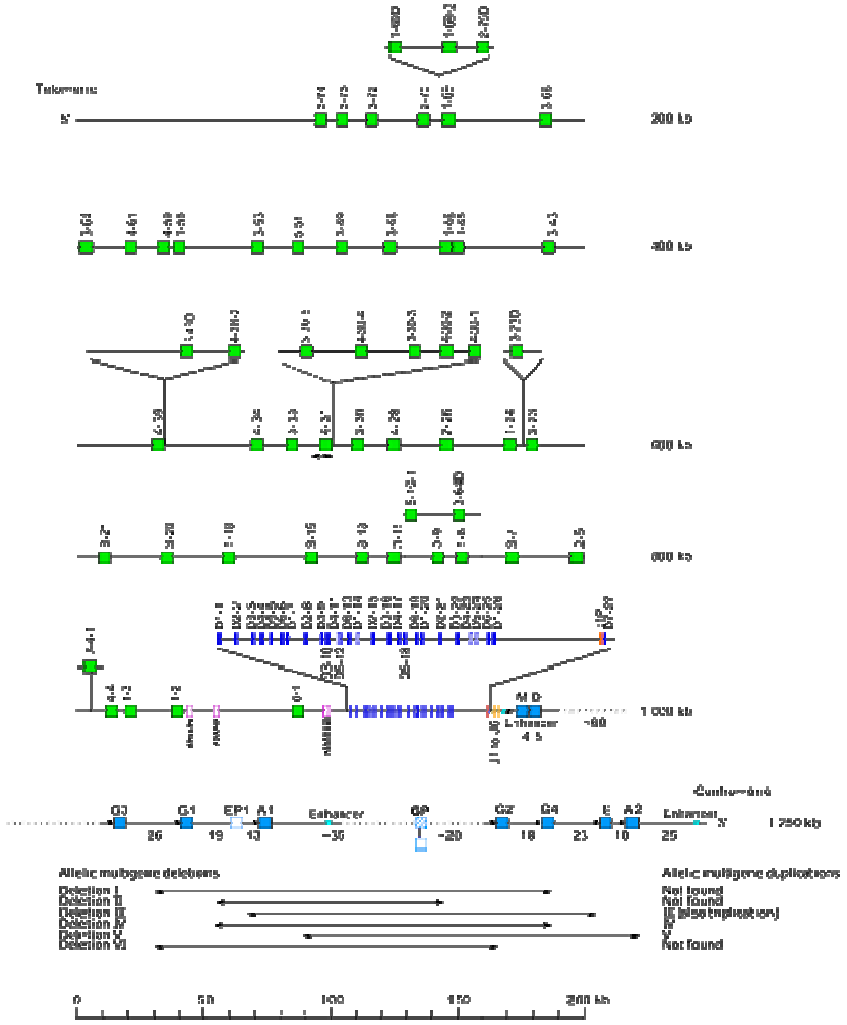
NGS data analysis



CEITEC

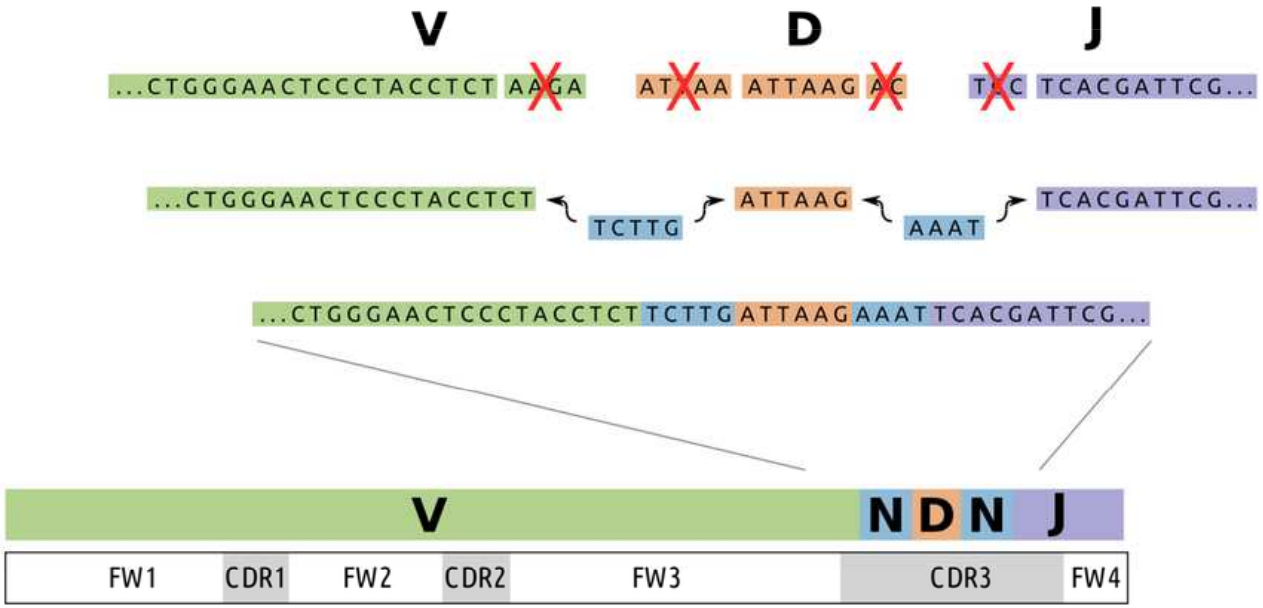
Immunogenetic

- T-cell receptor , Immunoglobulin – (B-cell)
- Gene rearrangement during cell maturation
  - VDJ recombination



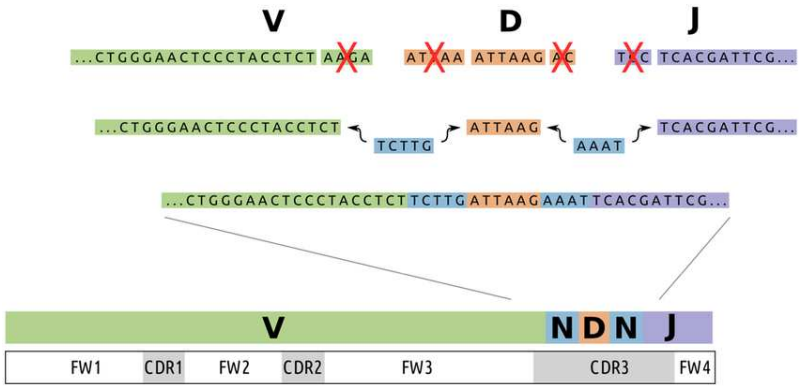
Immunogenetic

- T-cell receptor , Immunoglobulin – (B-cell)
- Gene rearrangement during cell maturation
  - VDJ recombination

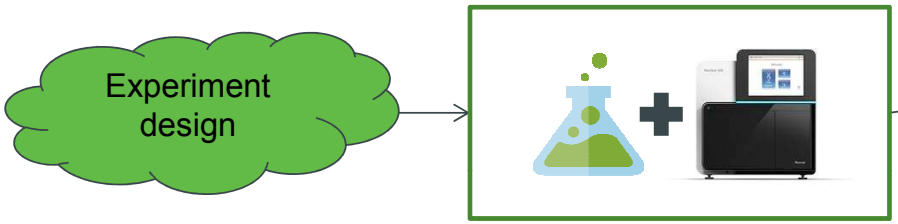


Immunogenetic

- Different cell populations
  - Clonal studies
  - Repertoire usage
- Main usage – blood malignancies (leukemias)



NGS data analysis



Experiment design

de-multiplexing

Raw data
.fastq

QC

Not known reference

Not "classic" reference

Metagenomics

Reference assembly

Immunogenetic
VDJ-genes

CRISPR
sgRNA

Genome/Transcriptome Reference Mapping
.bam

QC

...

Methylation
Bisulfide-seq

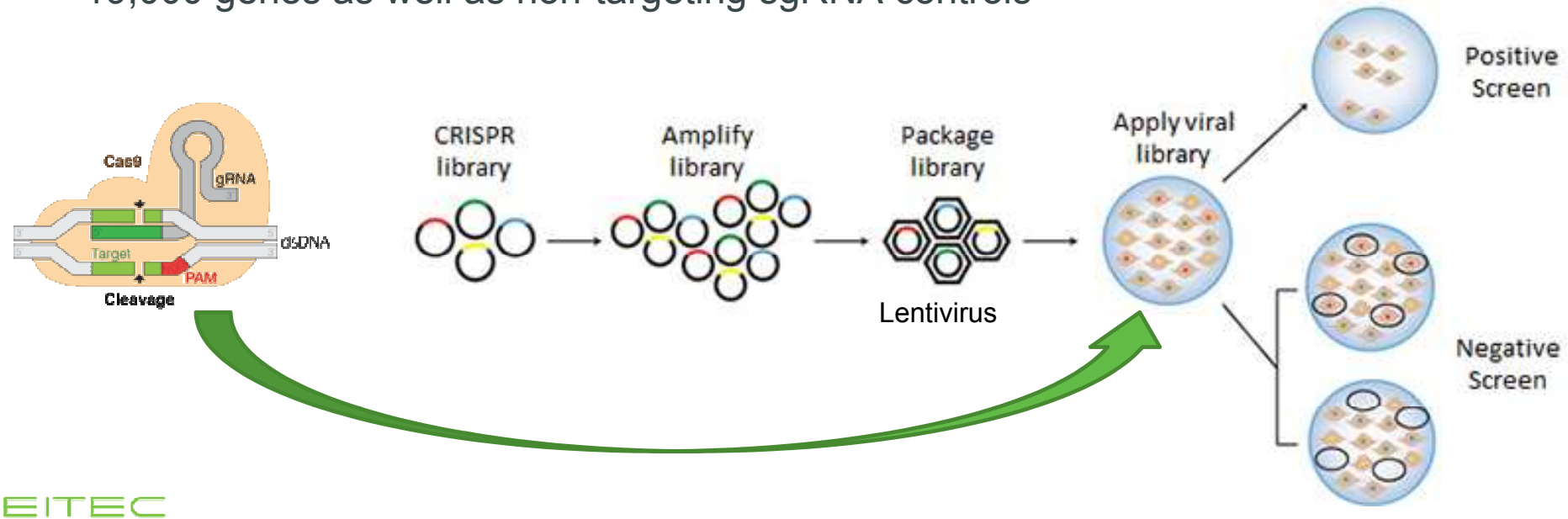
Interaction analysis
CHIP-seq

Expression analysis
RNAseq

Variant analysis
WES

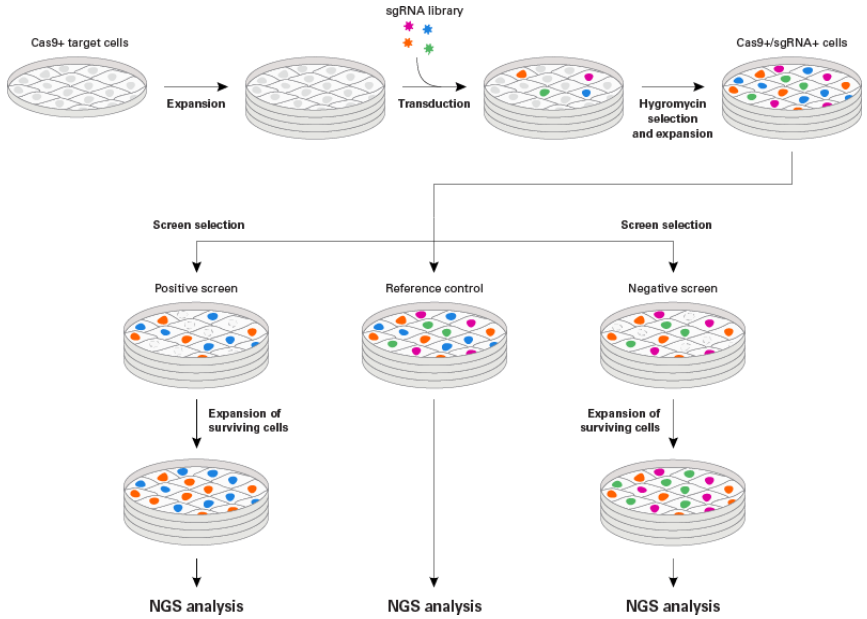
Genome-wide CRISPR-Cas9 knockout screens

- Cas9 (CRISPR associated protein 9) is a protein which plays a vital role in the immunological defense of certain bacteria against DNA viruses
- sgRNA libraries
  - Each sgRNA knockout specific gene
  - 76,000 guide RNAs (sgRNAs) with four highly active guides per gene, targeting about 19,000 genes as well as non-targeting sgRNA controls



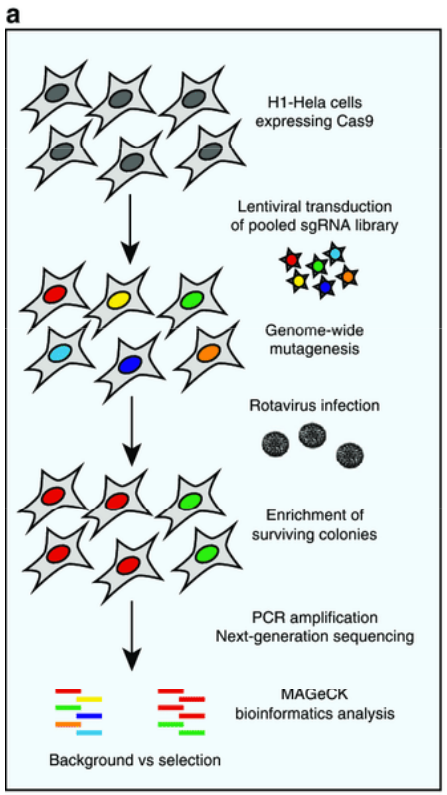
Genome-wide CRISPR-Cas9 knockout screens

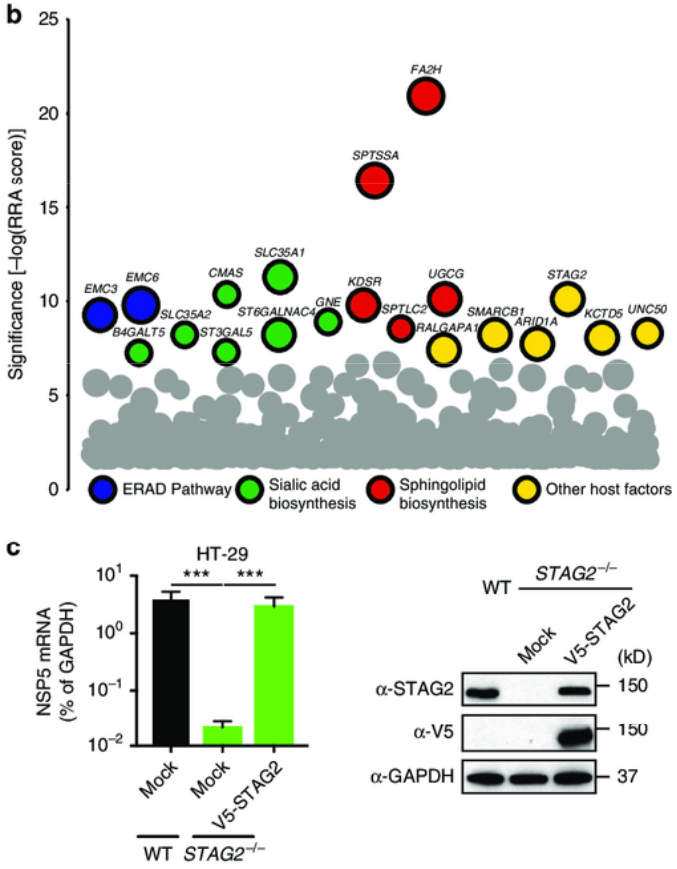
- Screen selection + expansion/enrichment of surviving cells
- NGS sequencing



Genome-wide CRISPR-Cas9 knockout screens

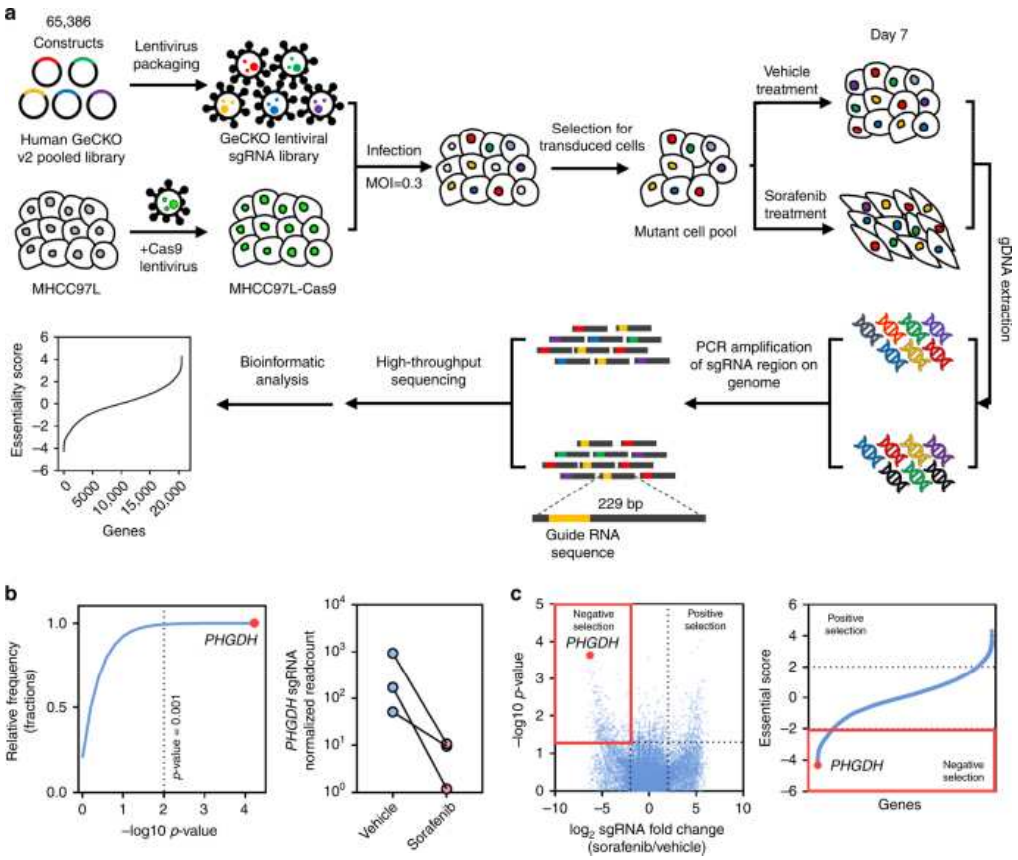
- NGS data analysis
  - Counting cells with different genes KD
  - Counting sgRNA fragments
  - Compare conditions





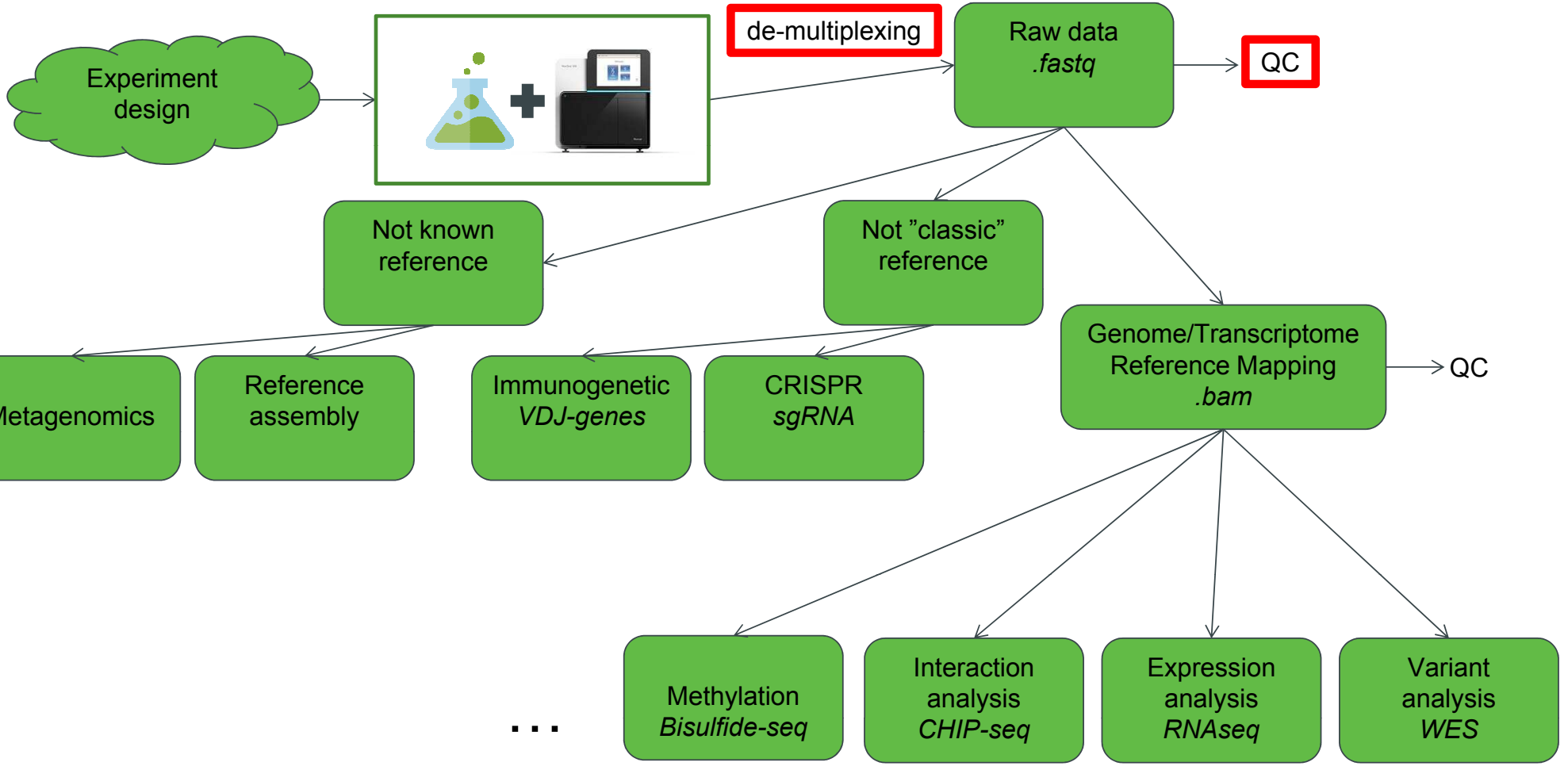
Genome-wide CRISPR-Cas9 knockout screens

- Example study

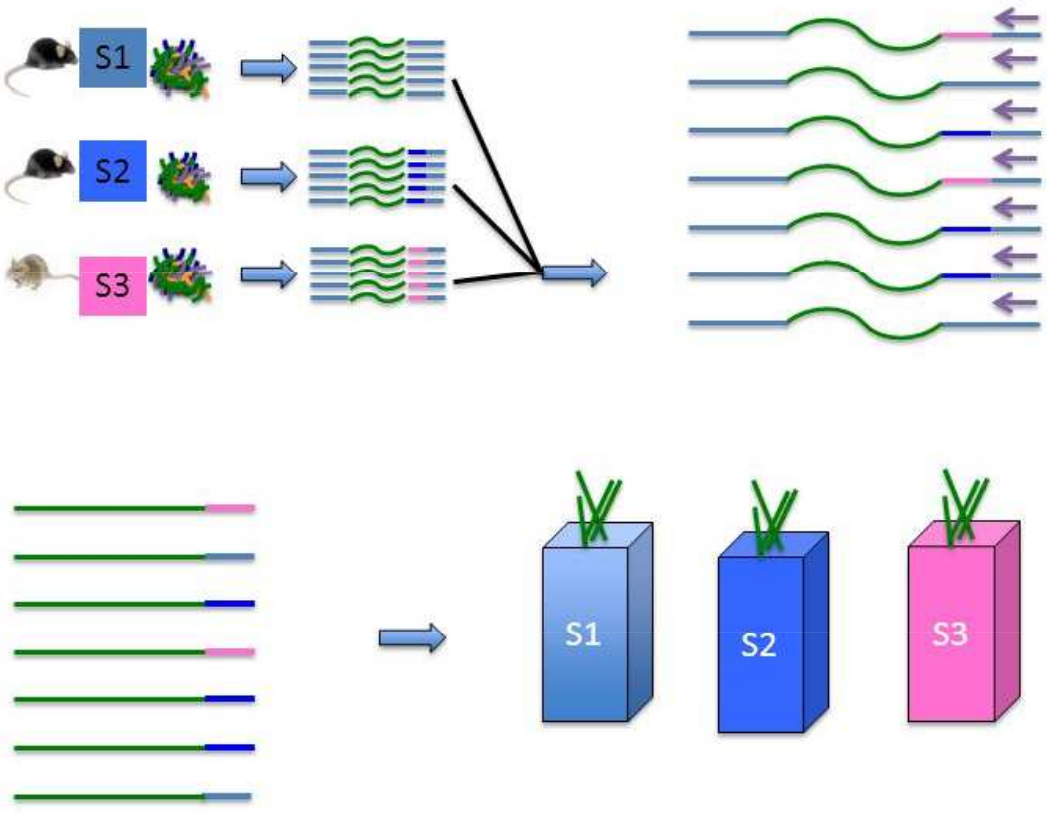


Wei, L., Lee, D., Law, CT. *et al.* Genome-wide CRISPR/Cas9 library screening identified PHGDH as a critical driver for Sorafenib resistance in HCC. *Nat Commun* **10,** 4681 (2019). https://doi.org/10.1038/s41467-019-12606-7

NGS data analysis

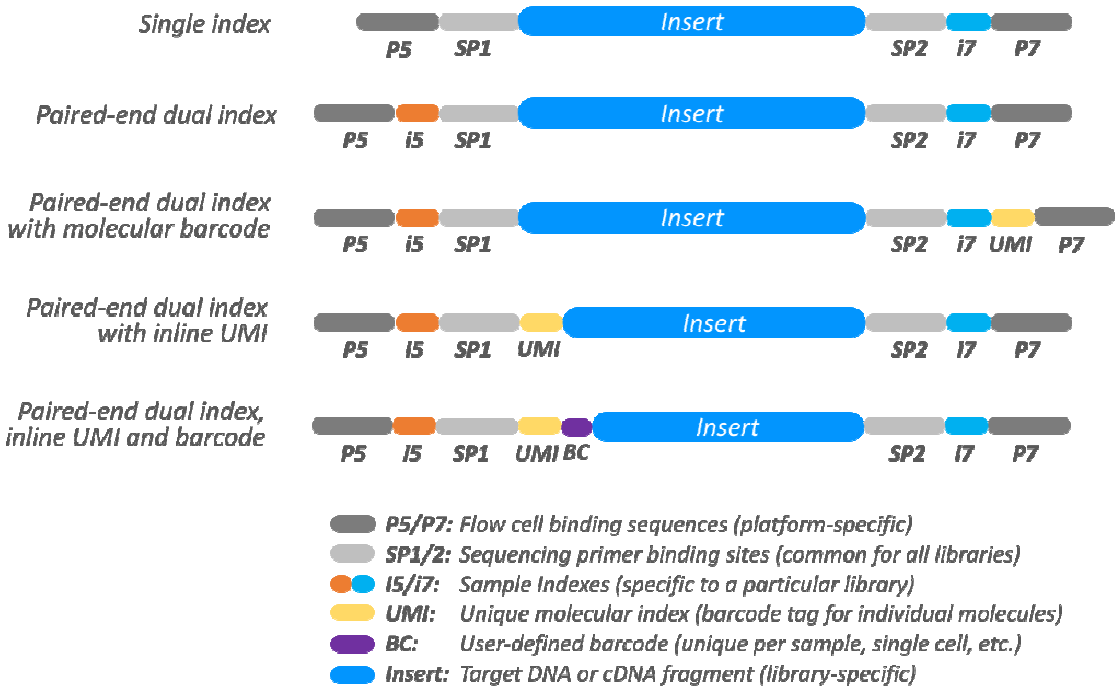


De-multiplexing

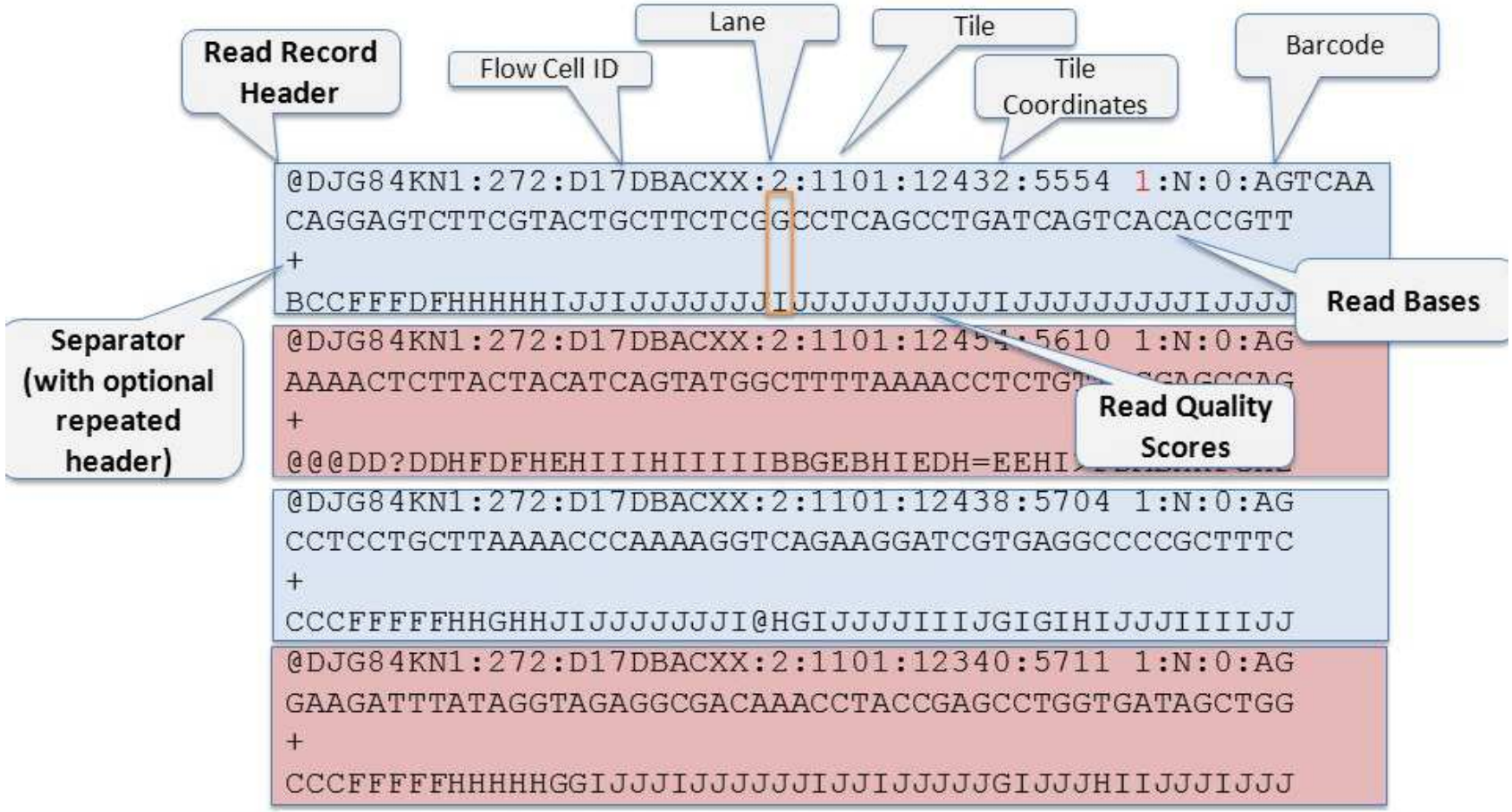


De-multiplexing

- Bcl2fastq tool
  - *Needs sample sheet with indexes*
  - *Number of barcode mismatches*
    - *Check undetermined*



Primary data – fastq file



NOTE: for paired-end runs, there is a second file with one-to-one corresponding headers and reads.

Fastq format - quality

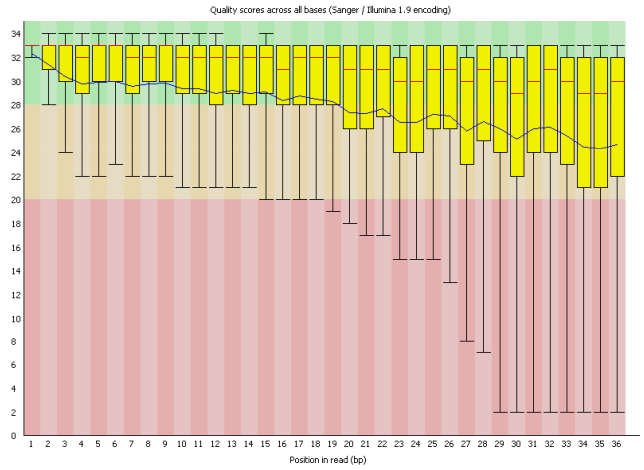
- **Fastq - q stands for quality – coded phred score**

```
CFFFFEFFGCEEGECFGGGGAFF87@E:++6C<++3:,8,33,,:,,,:,,:,,,
```

$$Q = -10 \cdot \log_{10} P$$

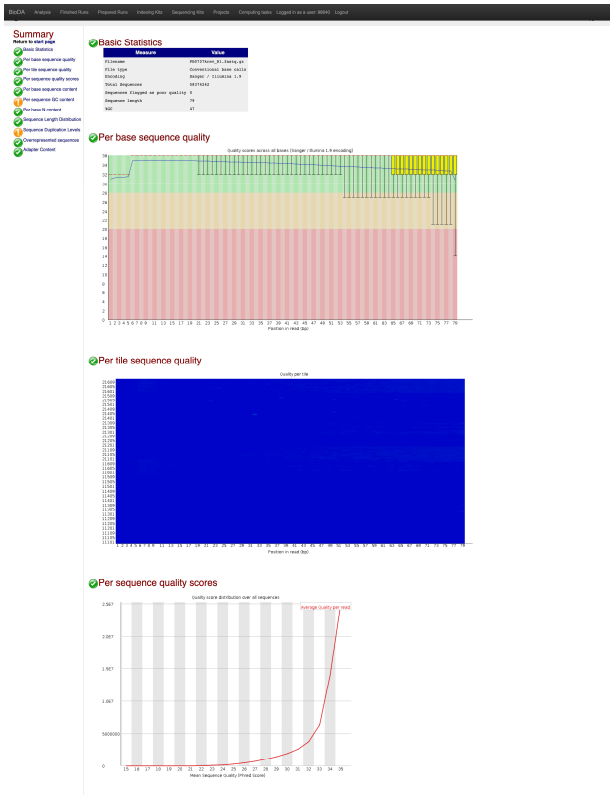
Quality	Error probability
5	31%
10	10%
20	1%
30	0.1%

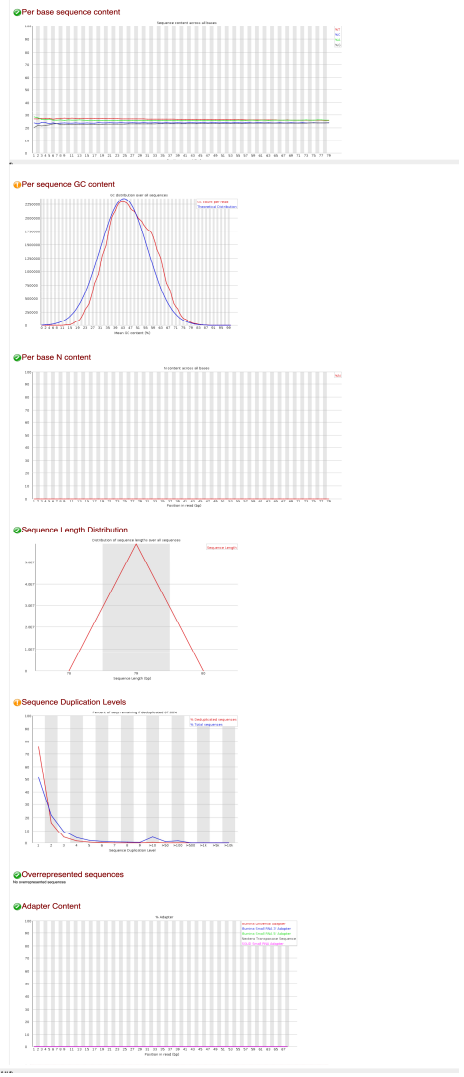
- **Very good for early problem detection**
- **Reasonable for trimming and read filtering**
  - **RNA seq - above phred score 5**



Fastq – quality control

- Fastqc - tool





CEITEC

@CEITEC_Brno

Thank you for your attention!

Vojta Bystry
vojtech.bystry@ceitec.muni.cz

CEITEC

www.ceitec.eu