CEITEC

Central European Institute of Technology

BRNO | CZECH REPUBLIC

Bi7420: Moderní metody pro analýzu genomu

NGS data analysis introduction

Vojta Bystry
vojtech.bystry@ceitec.muni.cz

Plan for **Bi7420**

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

1. Introduction to NGS technology
2. Introduction analysis; NGS Overview
3. DNA resequencing
4. miRNA, lncRNA in cancer - Marek Mráz
5. DNA resequencing, Chip-seq (CLIP-seq)
6. RNA-seq
7. RNA-seq – single cell sequencing

- The plan is open to change

What is NGS?

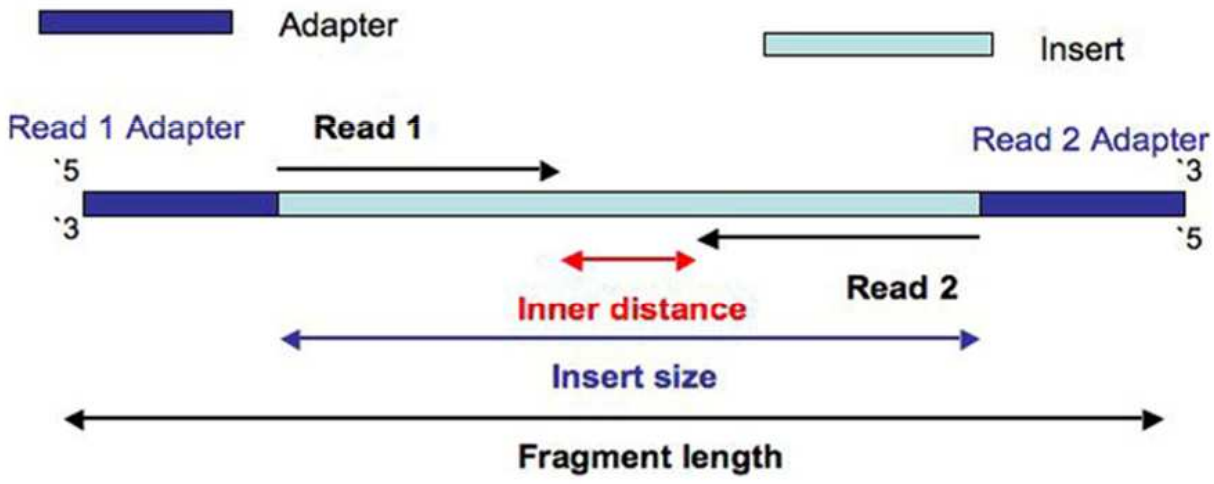
- Next generation sequencing
  - New generation sequencing
  - HTP = High throughput
  - Massively parallel sequencing
- Contrast to Sanger sequencing

What is NGS?

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

What is NGS?

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



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

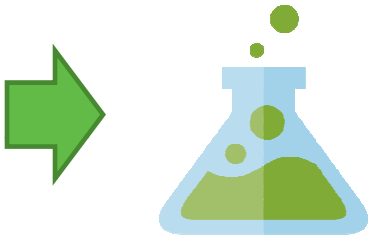
Experimental design

Library preparation

Sequencing

Data analysis

Basic workflow



Experimental design

Library preparation

Sequencing

Data analysis

Why we sequence

What we sequence

How we sequence

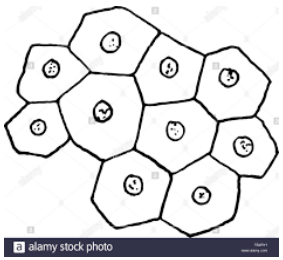
Basic workflow

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

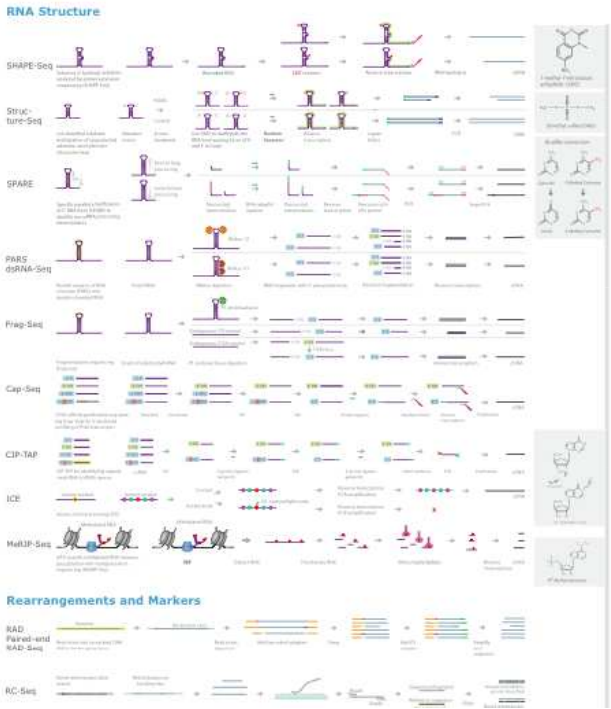
DNA

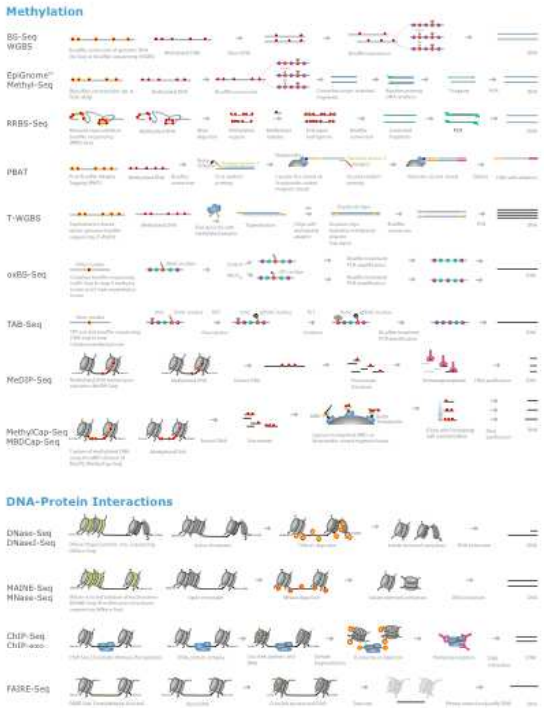
RNA

Select
some parts

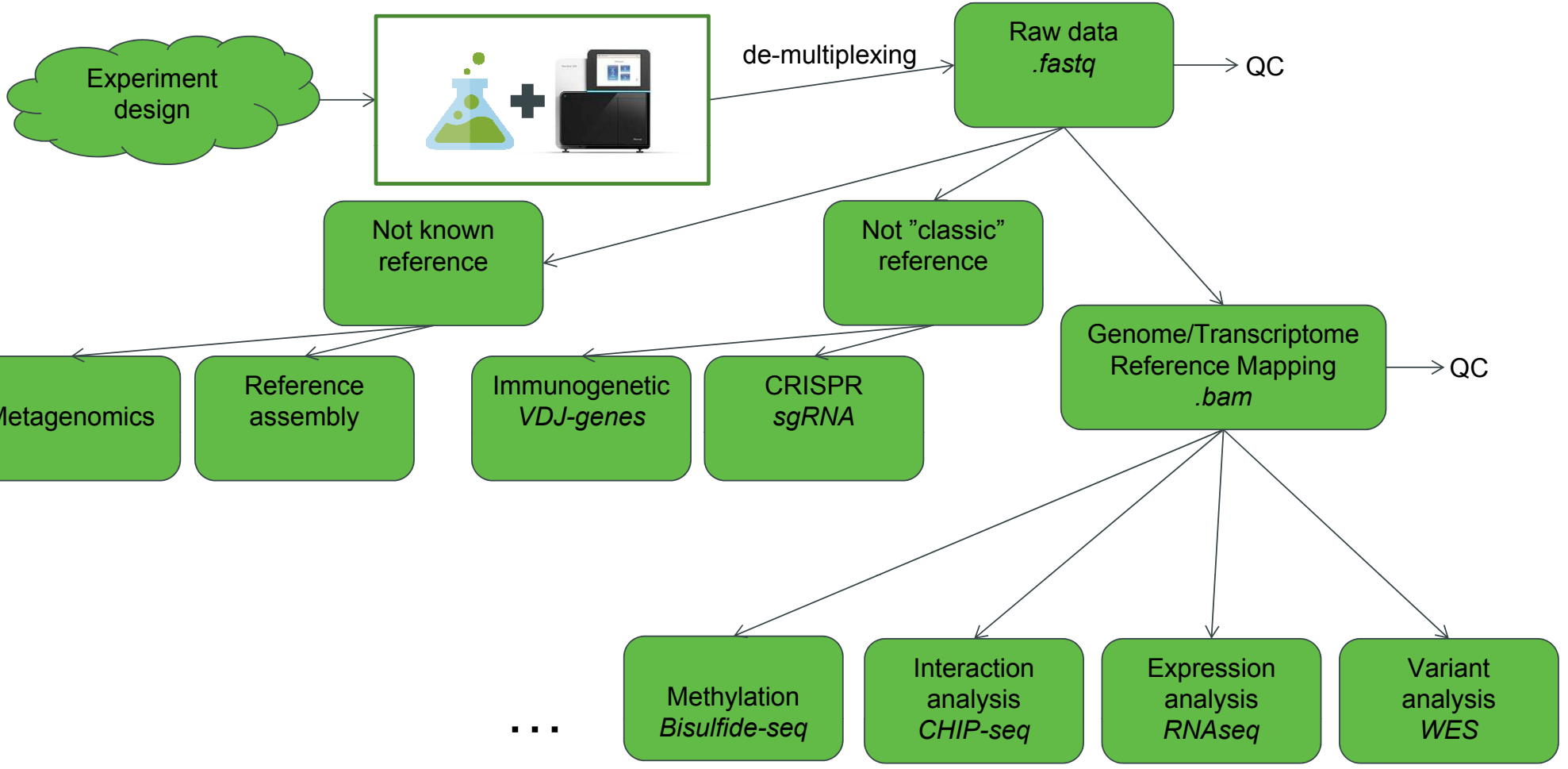
For all you seq...

www.illumina.com/LibraryPrepMethods



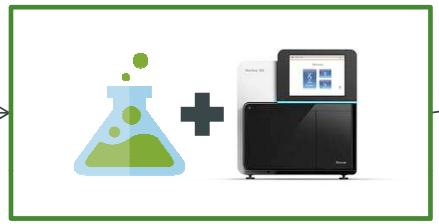


NGS data analysis



CEITEC

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

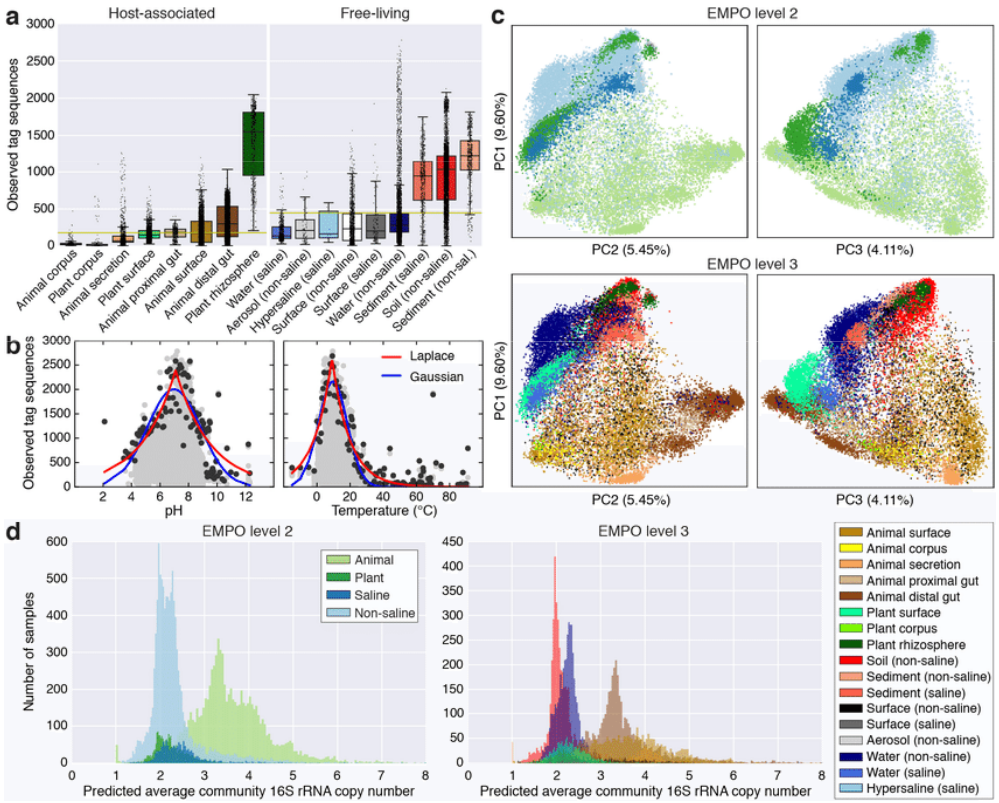
Interaction analysis
CHIP-seq

Expression analysis
RNAseq

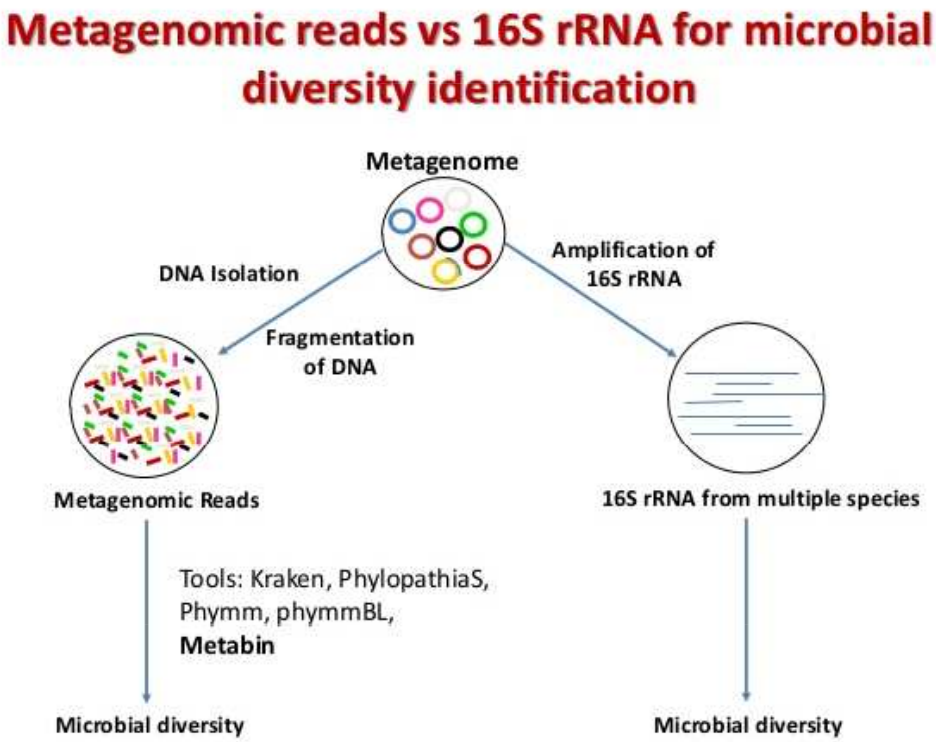
Variant analysis
WES

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

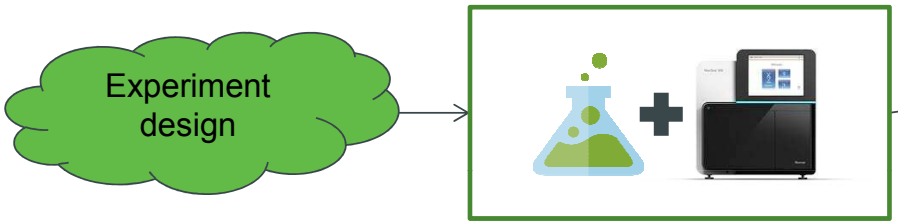


Factors	16S rRNA sequencing	Shotgun Metagenomic Sequencing
Cost	~$50 USD	Starting at ~$150 but price will depend on sequencing depth required
Sample preparation	Similar complexity to shotgun sequencing	Similar complexity to 16S rRNA sequencing
Functional profiling (profile microbial genes)	No (but 'predicted' 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

...

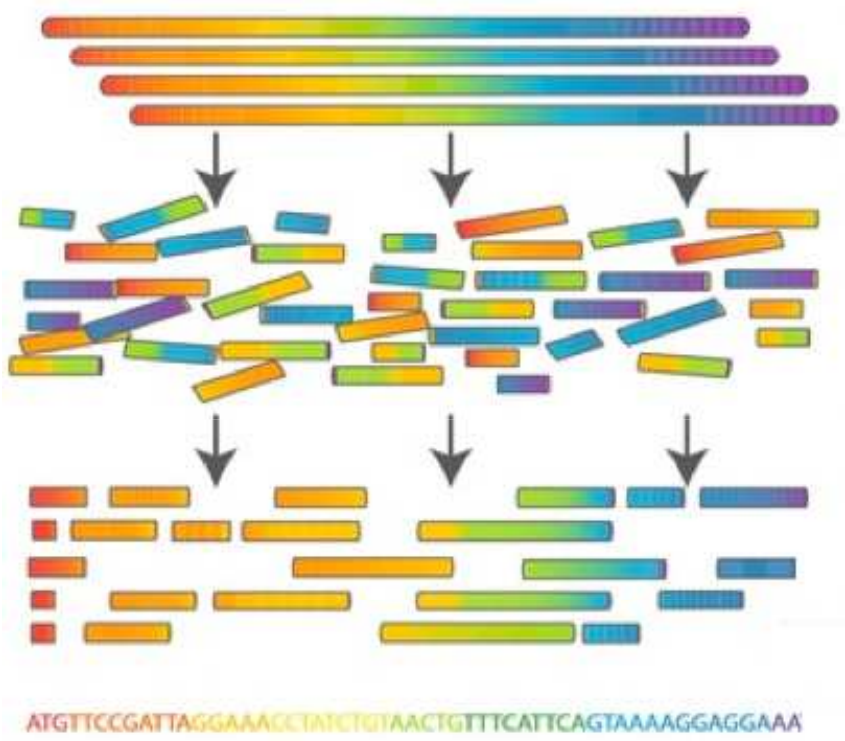
Methylation
Bisulfide-seq

Interaction analysis
CHIP-seq

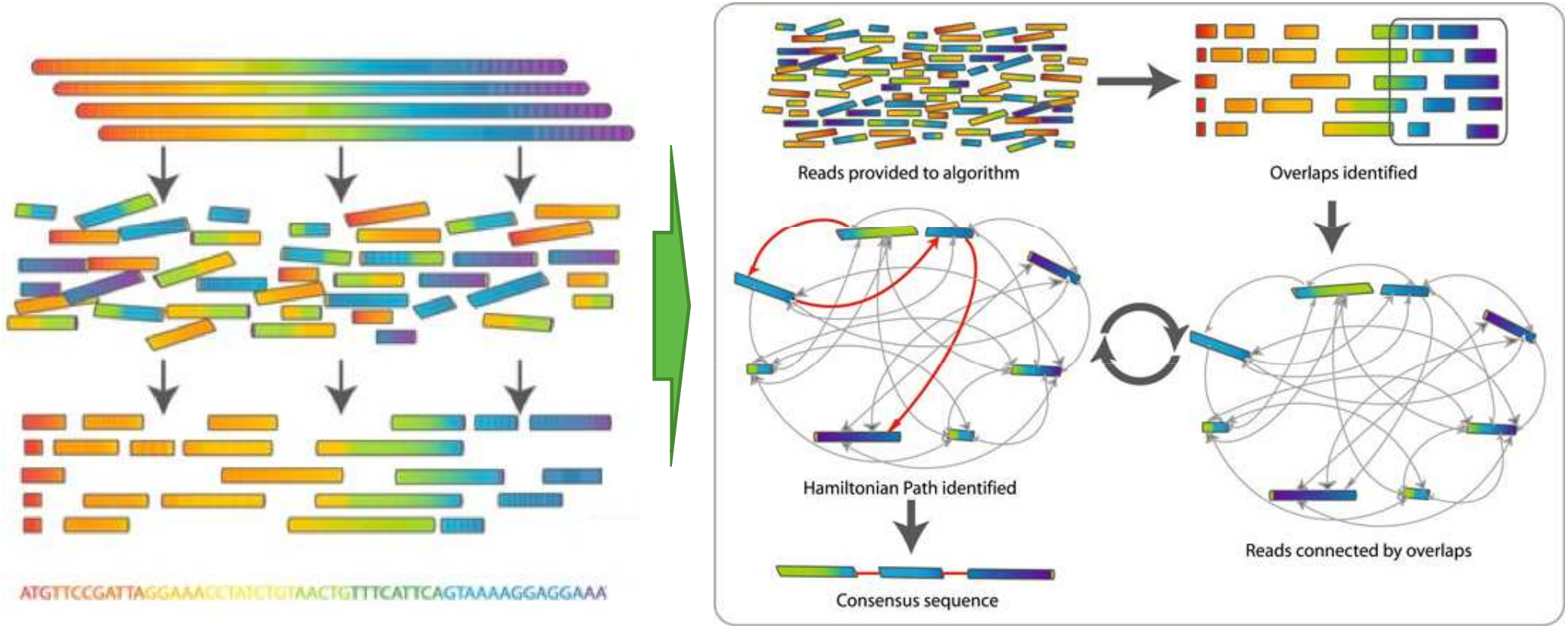
Expression analysis
RNAseq

Variant analysis
WES

Reference Assembly



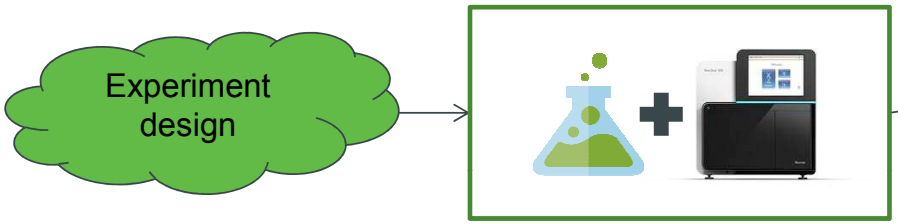
Reference Assembly



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



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

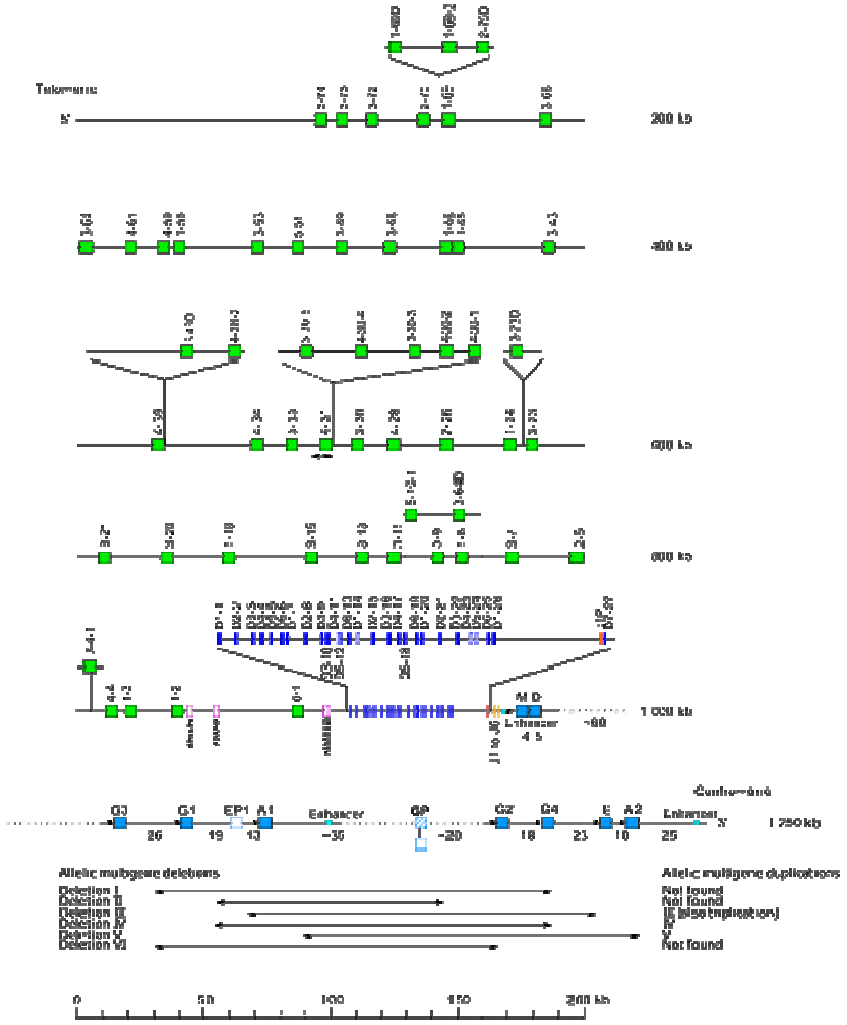
Interaction analysis
CHIP-seq

Expression analysis
RNAseq

Variant analysis
WES

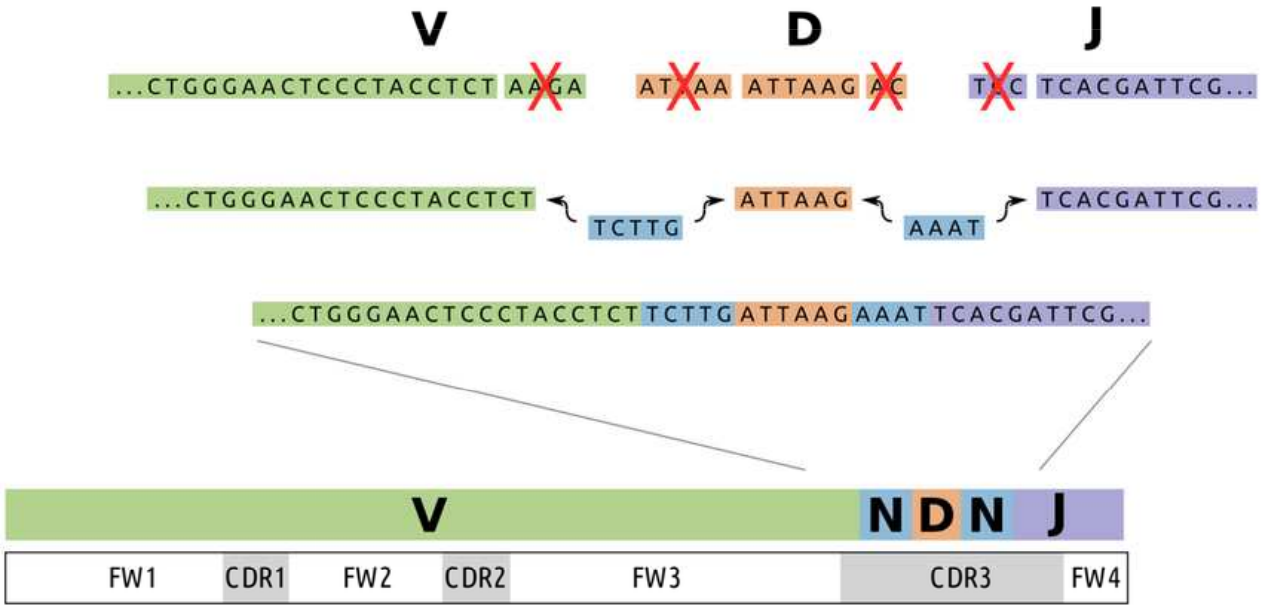
Immunogenetic

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



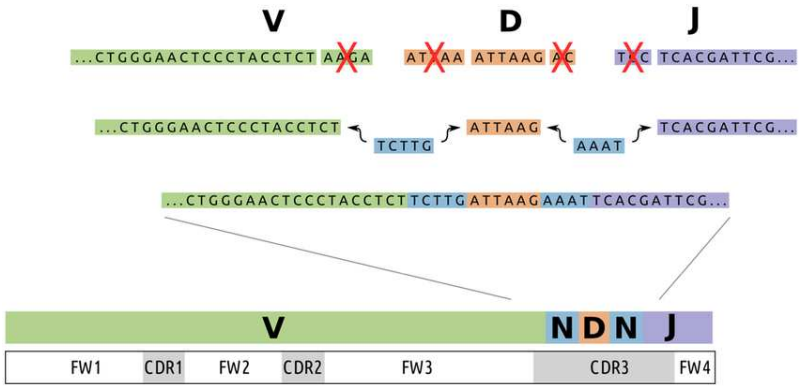
Immunogenetic

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

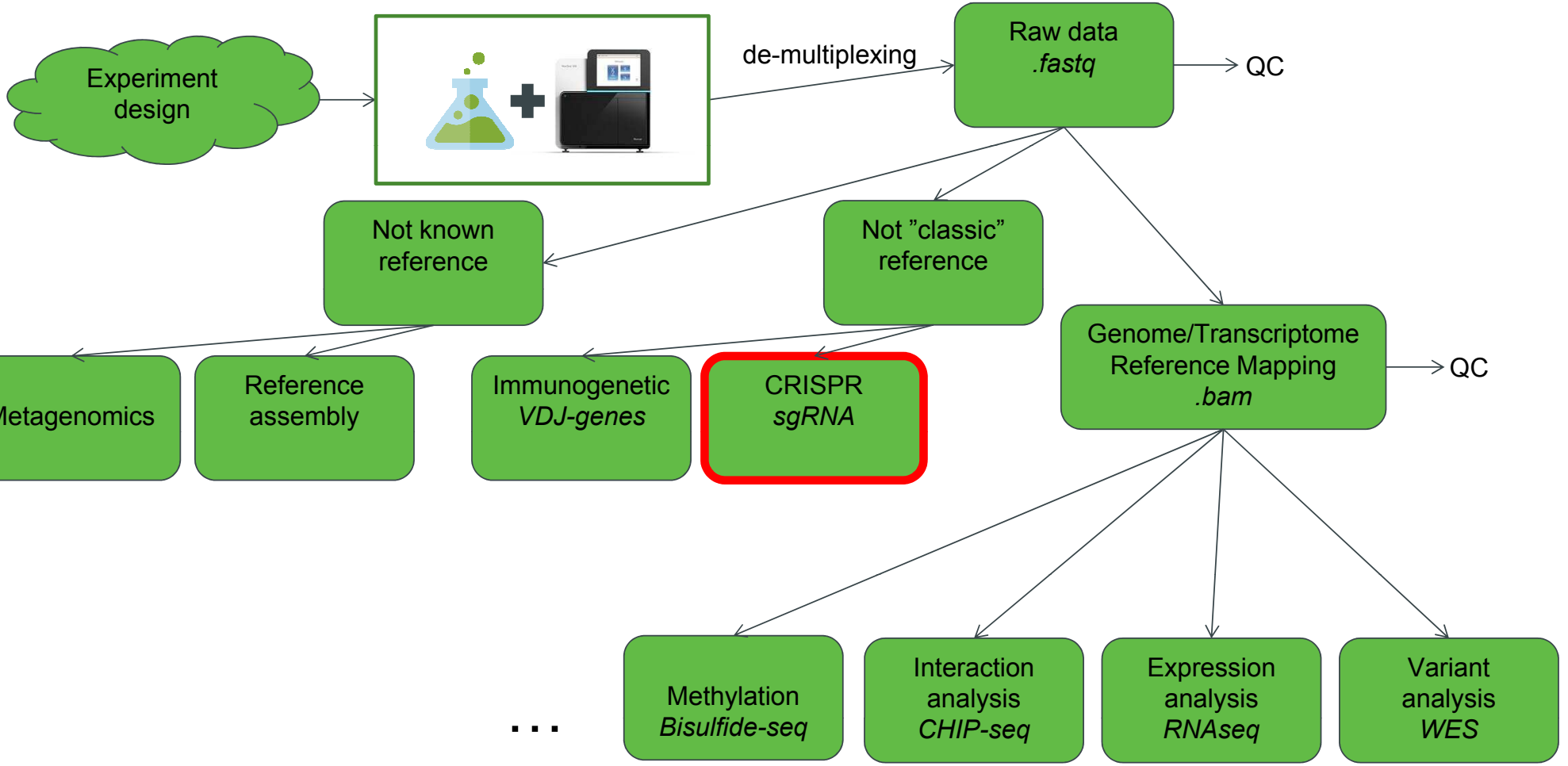


Immunogenetic

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

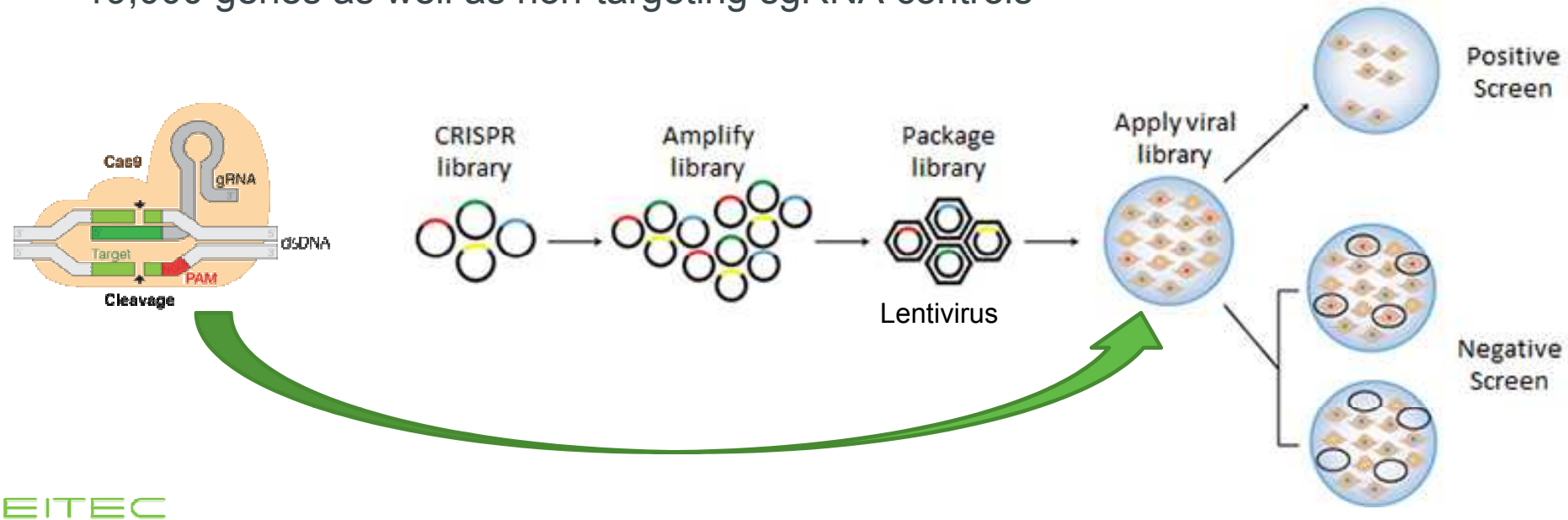


NGS data analysis



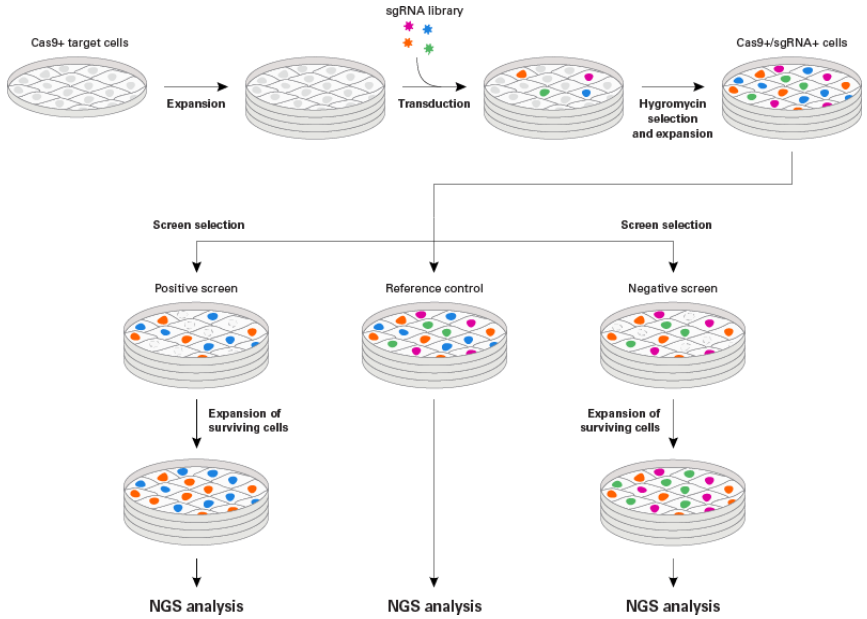
Genome-wide CRISPR-Cas9 knockout screens

- 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



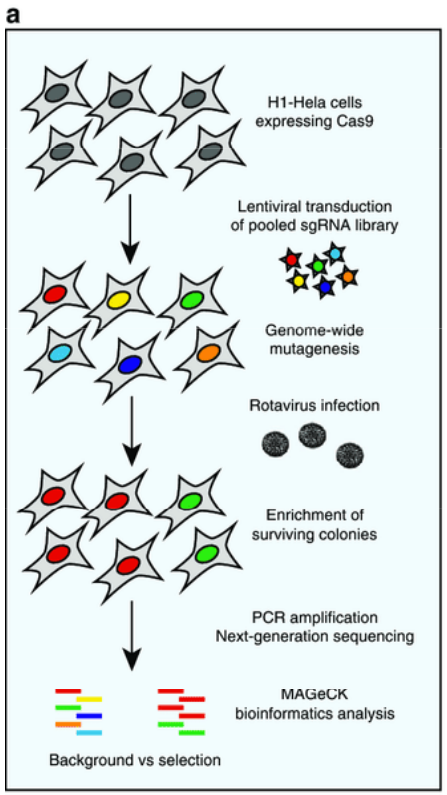
Genome-wide CRISPR-Cas9 knockout screens

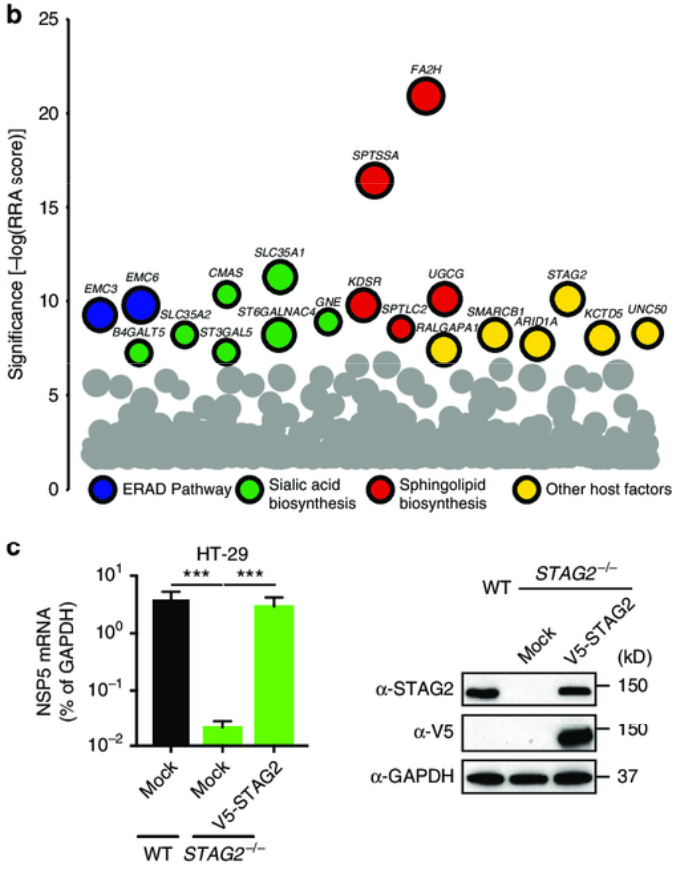
- Screen selection + expansion/enrichment of surviving cells
- NGS sequencing



Genome-wide CRISPR-Cas9 knockout screens

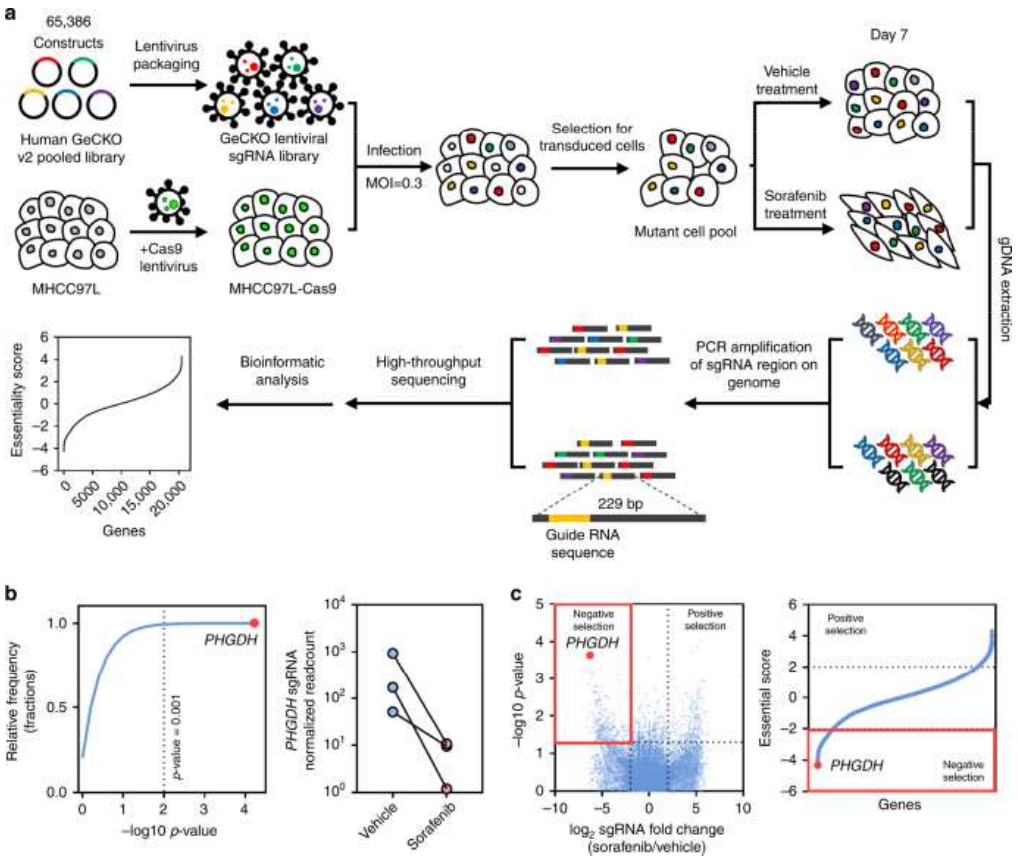
- NGS data analysis
  - Counting cells with different genes KD
  - Counting sgRNA fragments
  - Compare conditions





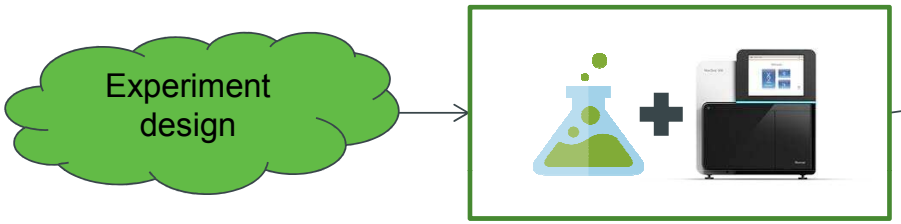
Genome-wide CRISPR-Cas9 knockout screens

- 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



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

...

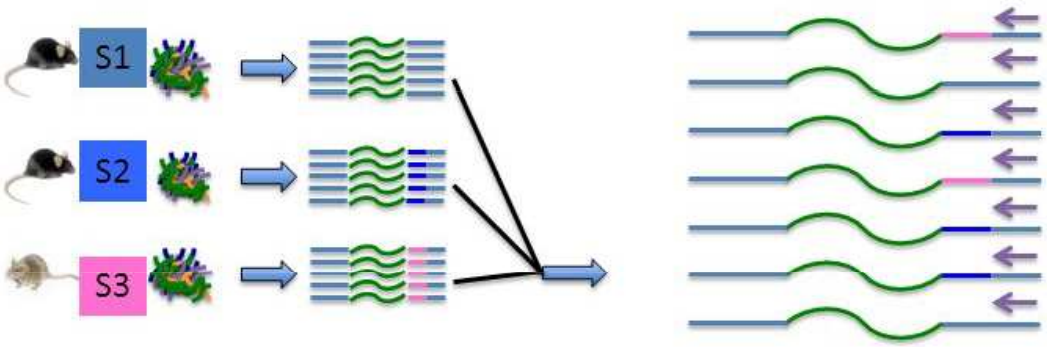
Methylation
Bisulfide-seq

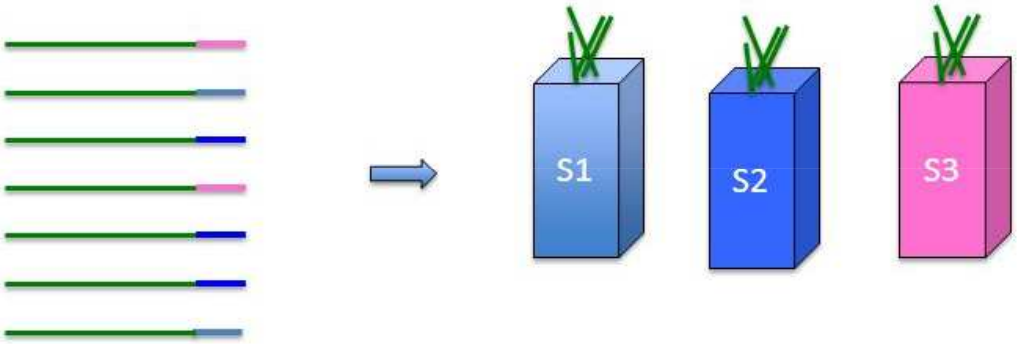
Interaction analysis
CHIP-seq

Expression analysis
RNAseq

Variant analysis
WES

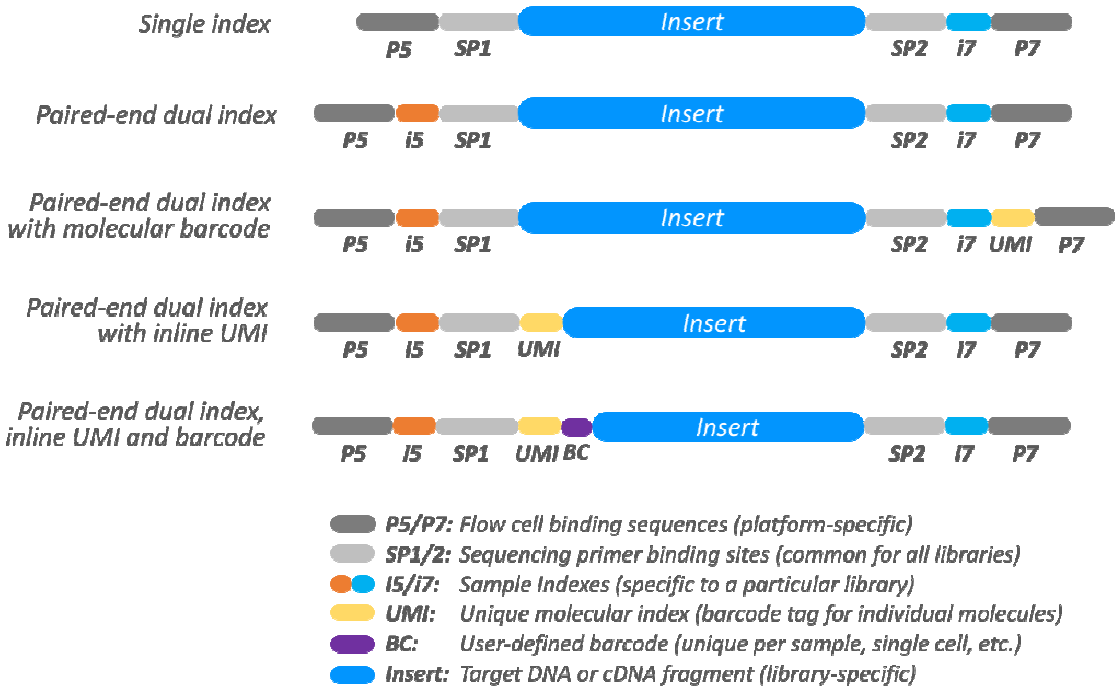
De-multiplexing



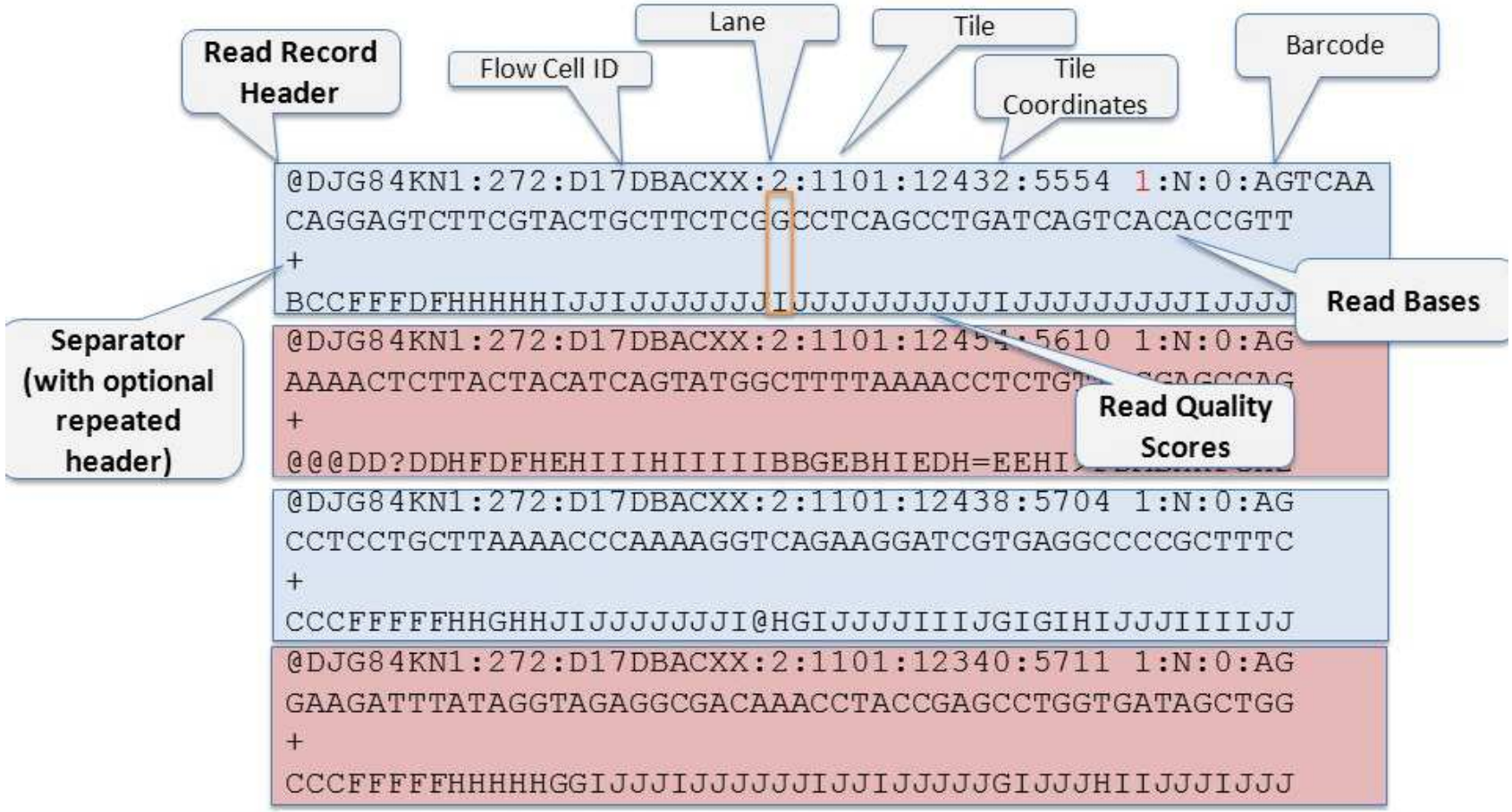


De-multiplexing

- 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

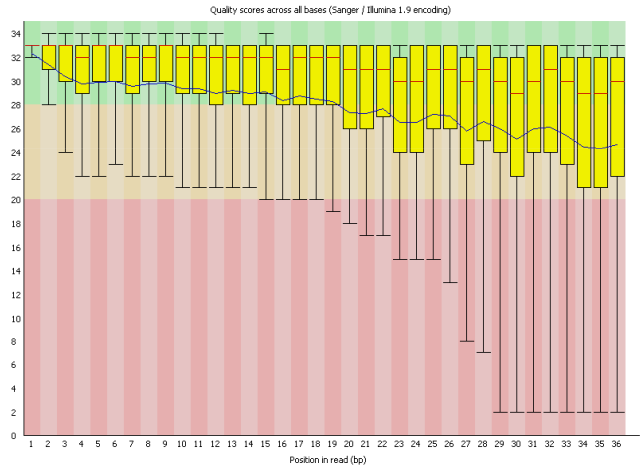
- **Fastq - q stands for quality – coded phred score**

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

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

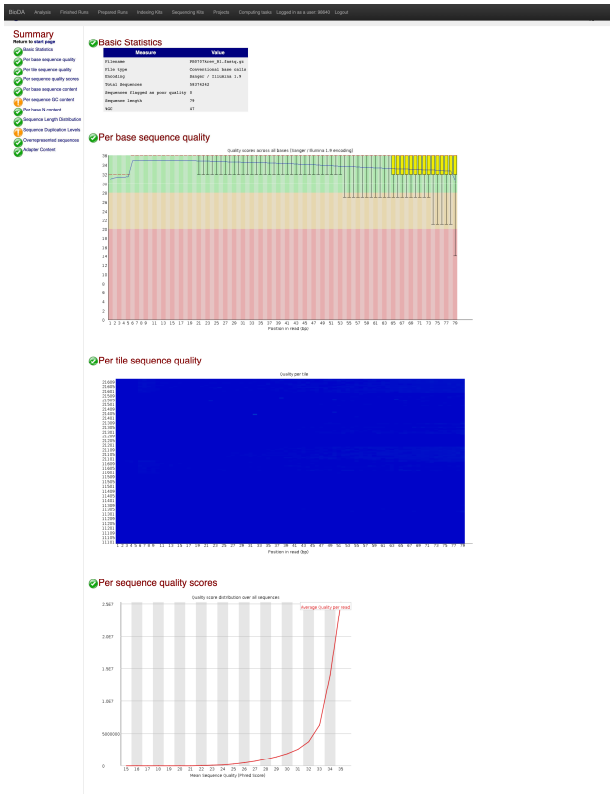
Quality	Error probability
5	31%
10	10%
20	1%
30	0.1%

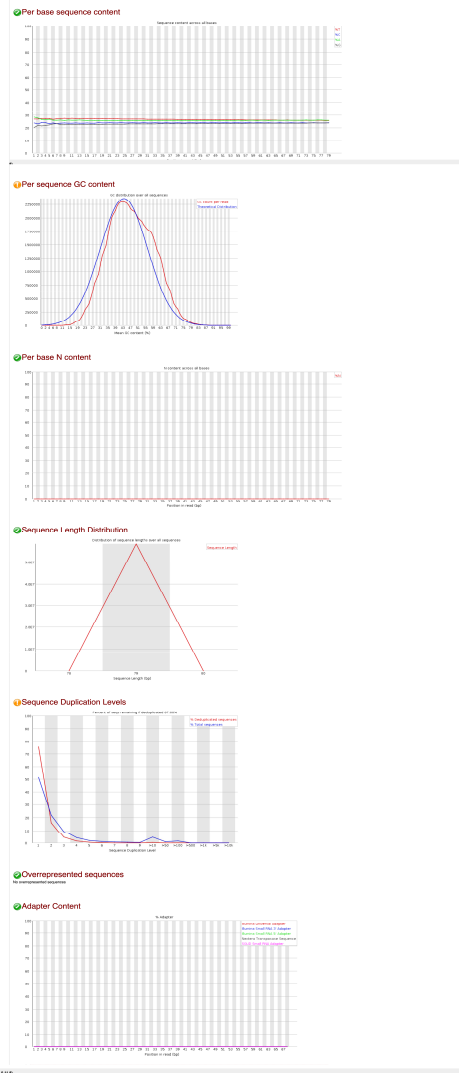
- **Very good for early problem detection**
- **Reasonable for trimming and read filtering**
  - **RNA seq - above phred score 5**



Fastq – quality control

- Fastqc - tool





FastQC Report

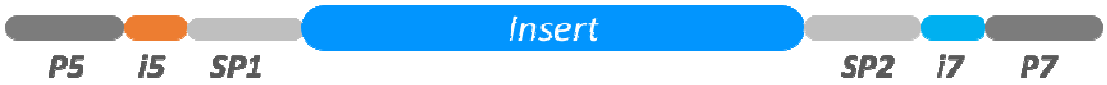
Summary

Return to start page

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	MU_a_ytHl_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	252819865
Sequences flagged as poor quality	0
Sequence length	161
%GC	40



CEITEC

@CEITEC_Brno

Thank you for your attention!

Vojta Bystry
vojtech.bystry@ceitec.muni.cz

CEITEC

www.ceitec.eu