CEITEC

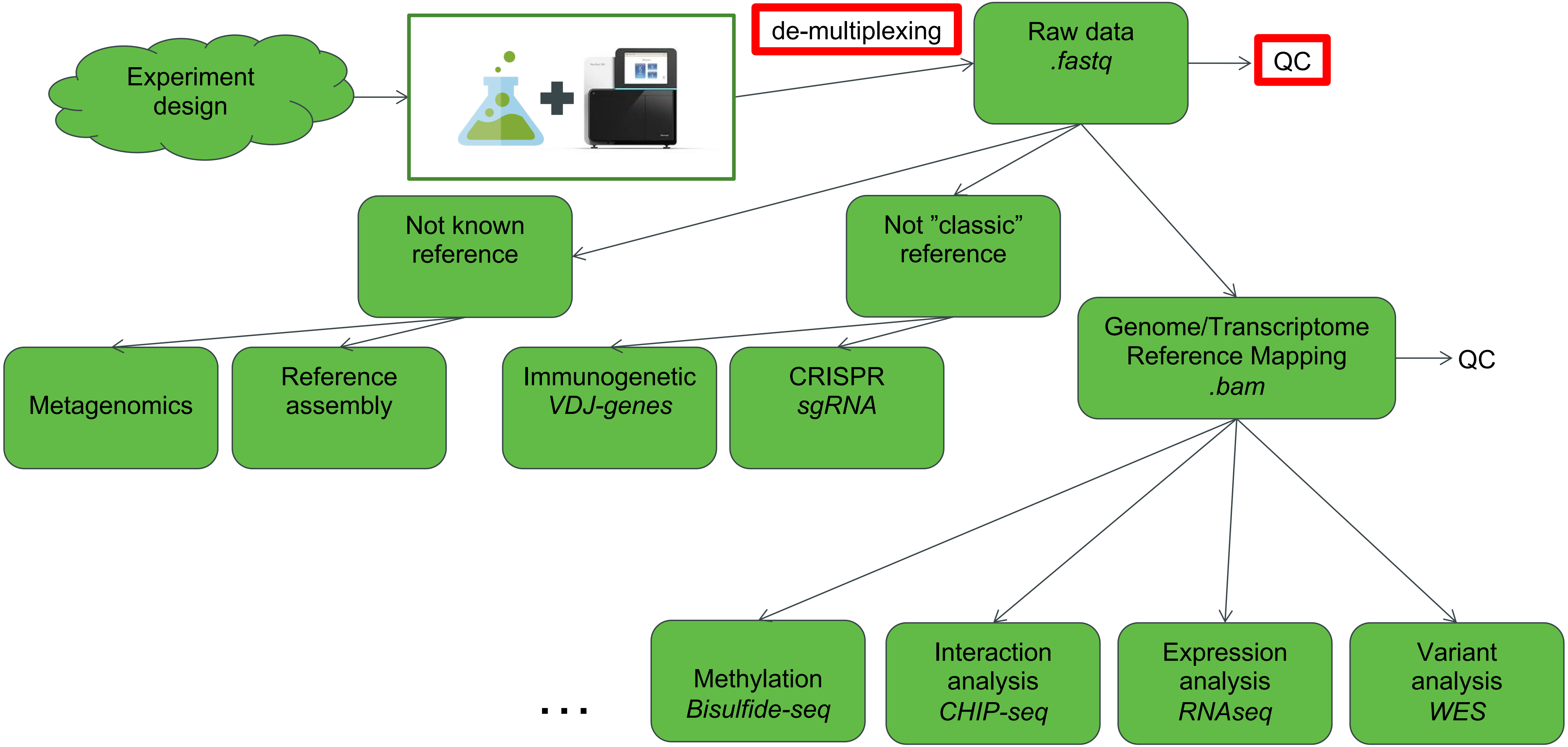
Central European Institute of Technology
BRNO | CZECH REPUBLIC

Modern methods for genome analysis
(PřF:Bi7420)

Lecture 2 : Raw NGS data quality control

Vojta Bystry
vojtech.bystry@ceitec.muni.cz

NGS data analysis



Experiment design

de-multiplexing

Raw data
.fastq

QC

Not known reference

Not ”classic” reference

Genome/Transcriptome Reference Mapping
.bam

QC

Metagenomics

Reference assembly

Immunogenetic
VDJ-genes

CRISPR
sgRNA

. . .

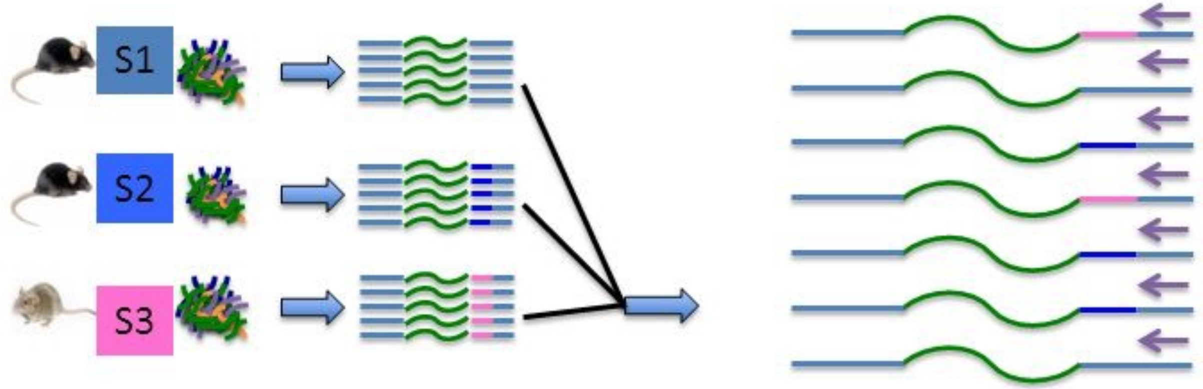
Methylation
Bisulfide-seq

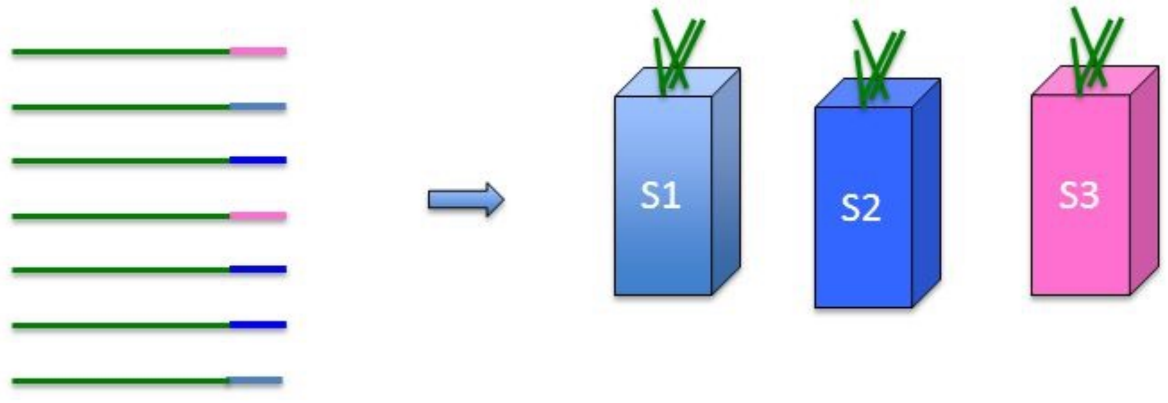
Interaction analysis
CHIP-seq

Expression analysis
RNAseq

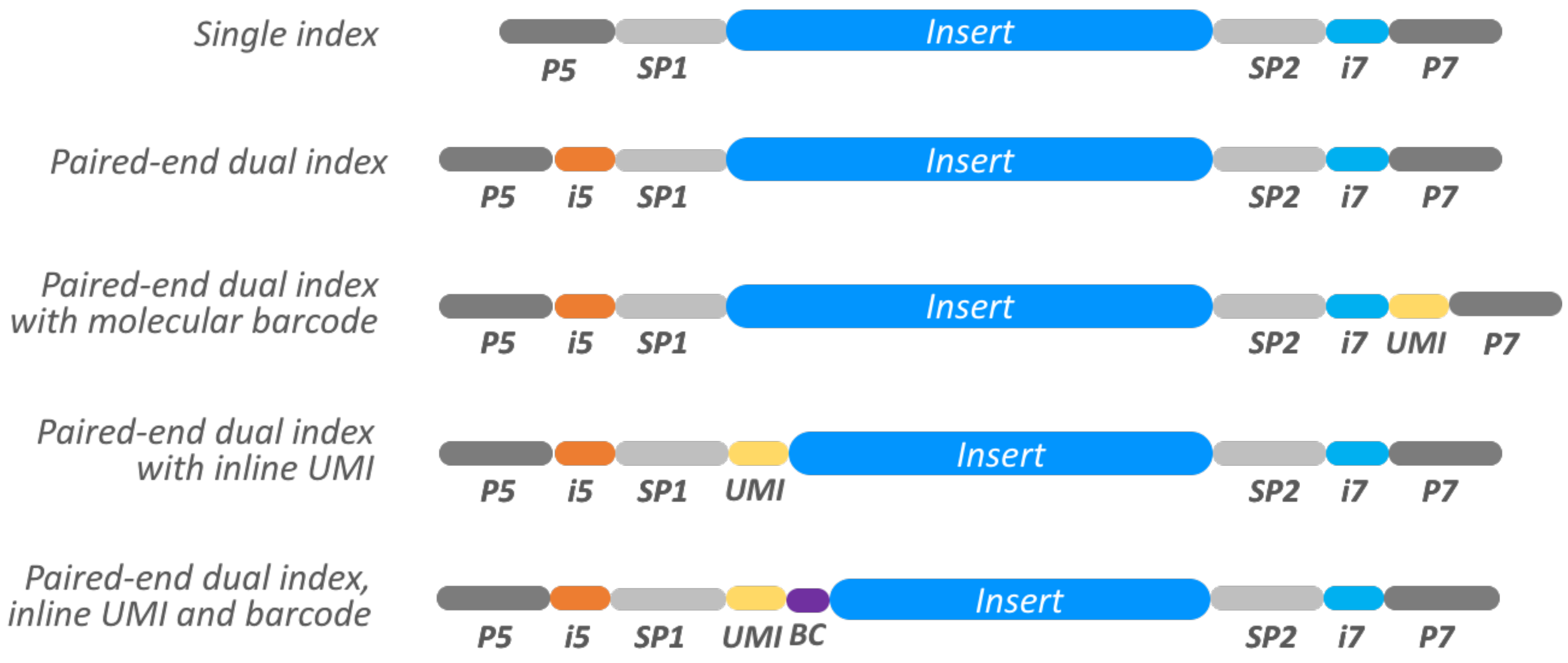
Variant analysis
WES

De-multiplexing





De-multiplexing



Single index

Paired-end dual index

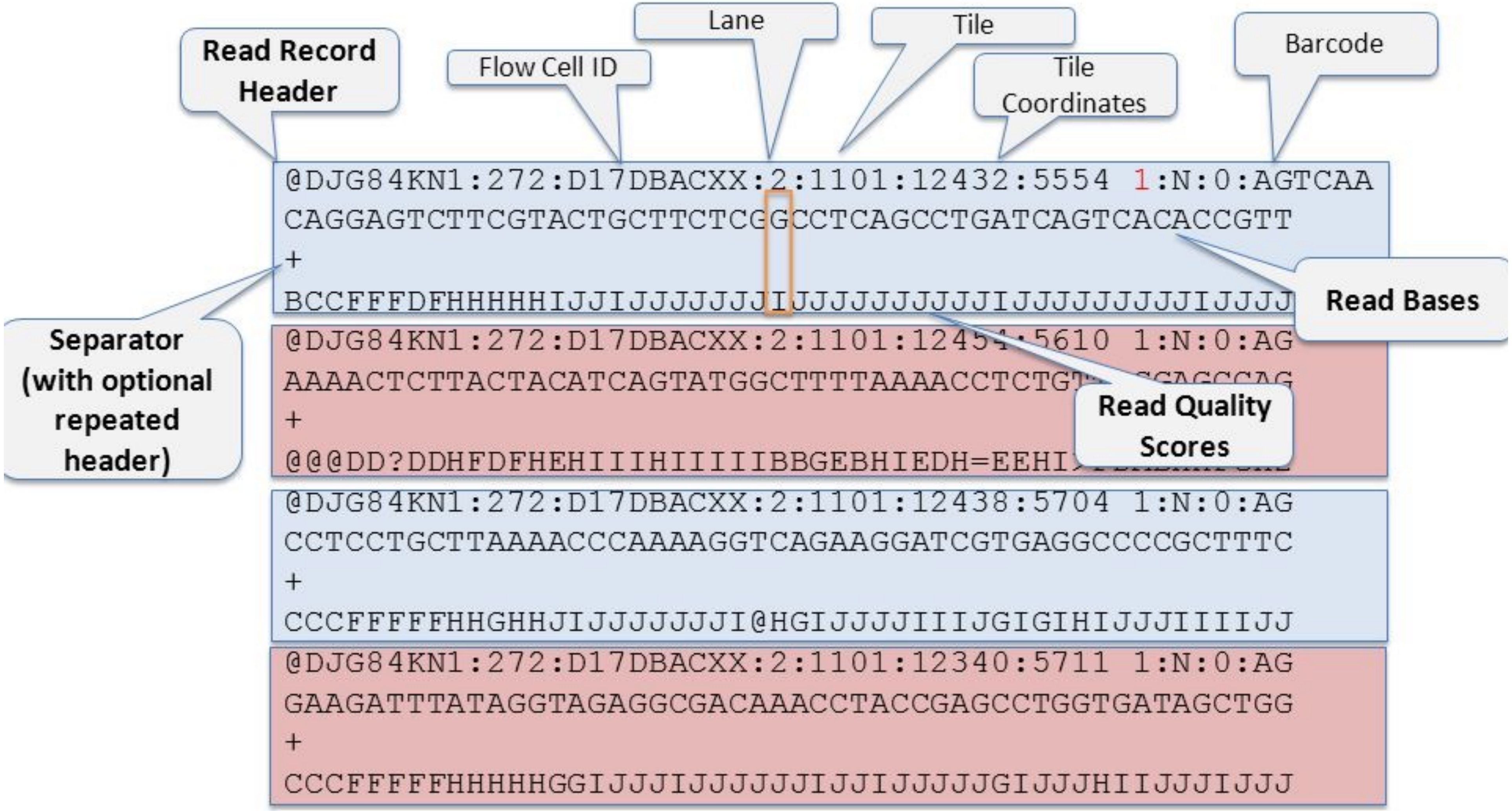
Paired-end dual index with molecular barcode

Paired-end dual index with inline UMI

Paired-end dual index, inline UMI and barcode

- **P5/P7:** *Flow cell binding sequences (platform-specific)*
- **SP1/2:** *Sequencing primer binding sites (common for all libraries)*
- **I5/i7:** *Sample Indexes (specific to a particular library)*
- **UMI:** *Unique molecular index (barcode tag for individual molecules)*
- **BC:** *User-defined barcode (unique per sample, single cell, etc.)*
- **Insert:** *Target DNA or cDNA fragment (library-specific)*

Primary data – fastq file



```
@DJG84KN1:272:D17DBACXX:2:1101:12432:5554 1:N:0:AGTCAA
CAGGAGTCTTCGTACTGCTTCTCGGCCTCAGCCTGATCAGTCACACCGTT
+
BCCFFFDFHHHHHIJJIJJJJJJJIJJJJJJJJJIJJJJJJJJIJJJJ
@DJG84KN1:272:D17DBACXX:2:1101:12454:5610 1:N:0:AG
AAAACTCTTACTACATCAGTATGGCTTTTAAAACCTCTGT
+
@@@DD?DDHFDFHEHIIIHIIIIIBBGEBHIEDH=EEHI>
@DJG84KN1:272:D17DBACXX:2:1101:12438:5704 1:N:0:AG
CCTCCTGCTTAAAACCCAAAAGGTCAGAAGGATCGTGAGGCCCCGCTTTC
+
CCCFFFFFHHGHHJIJJJJJJJI@HGIJJJJIIIJGIGIHIJJJIIIIJJ
@DJG84KN1:272:D17DBACXX:2:1101:12340:5711 1:N:0:AG
GAAGATTTATAGGTAGAGGCGACAAACCTACCGAGCCTGGTGATAGCTGG
+
CCCFFFFFHHHHHGGIJJJIJJJJJJIJJIJJJJJGIJJJHIIJJJIJJJ
```

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 as phred score

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

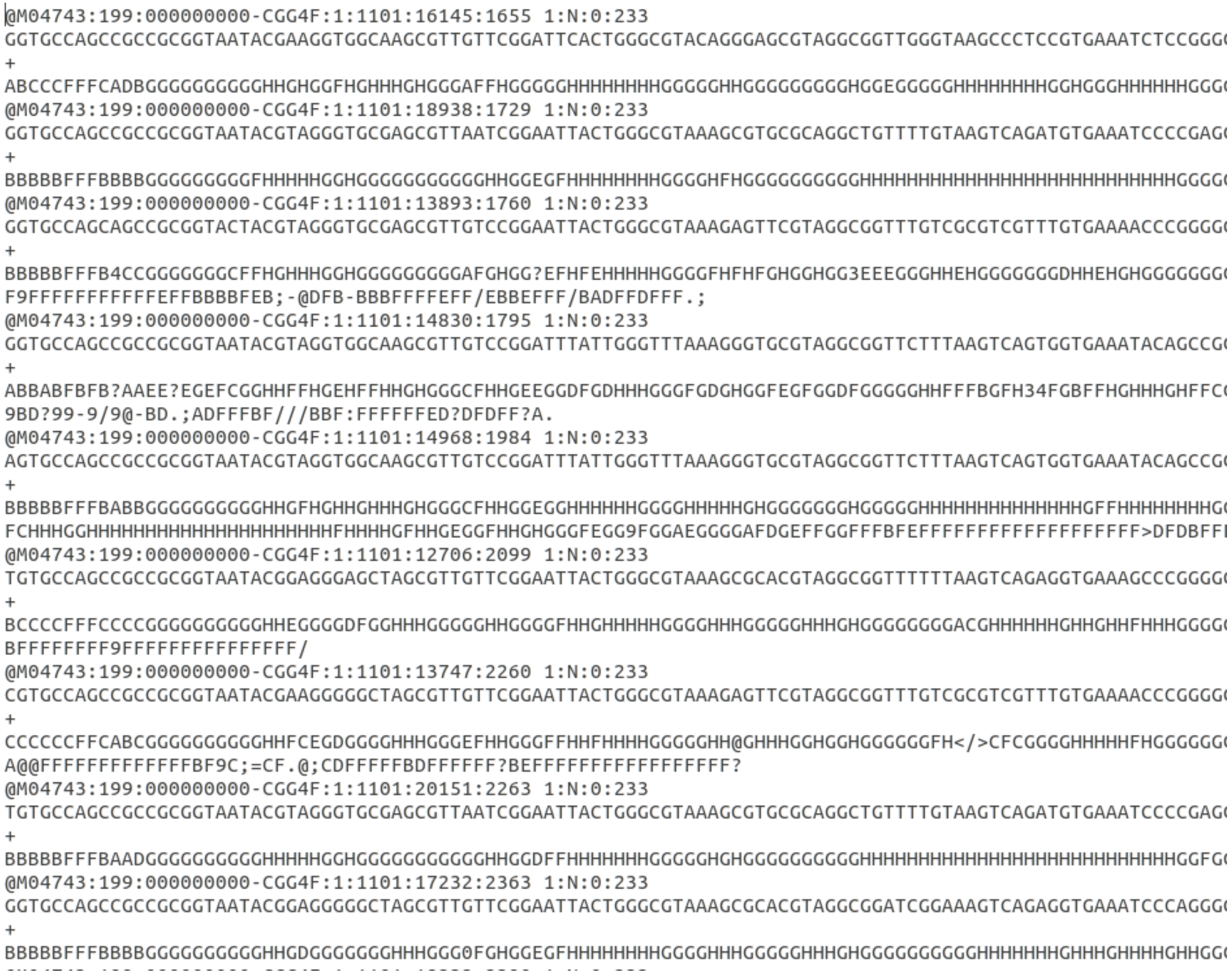
$$Q = -10 \cdot \log_{10} P$$

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

- What the machine things is the quality
- Only account for sequencing errors
- Very good for early problem detection

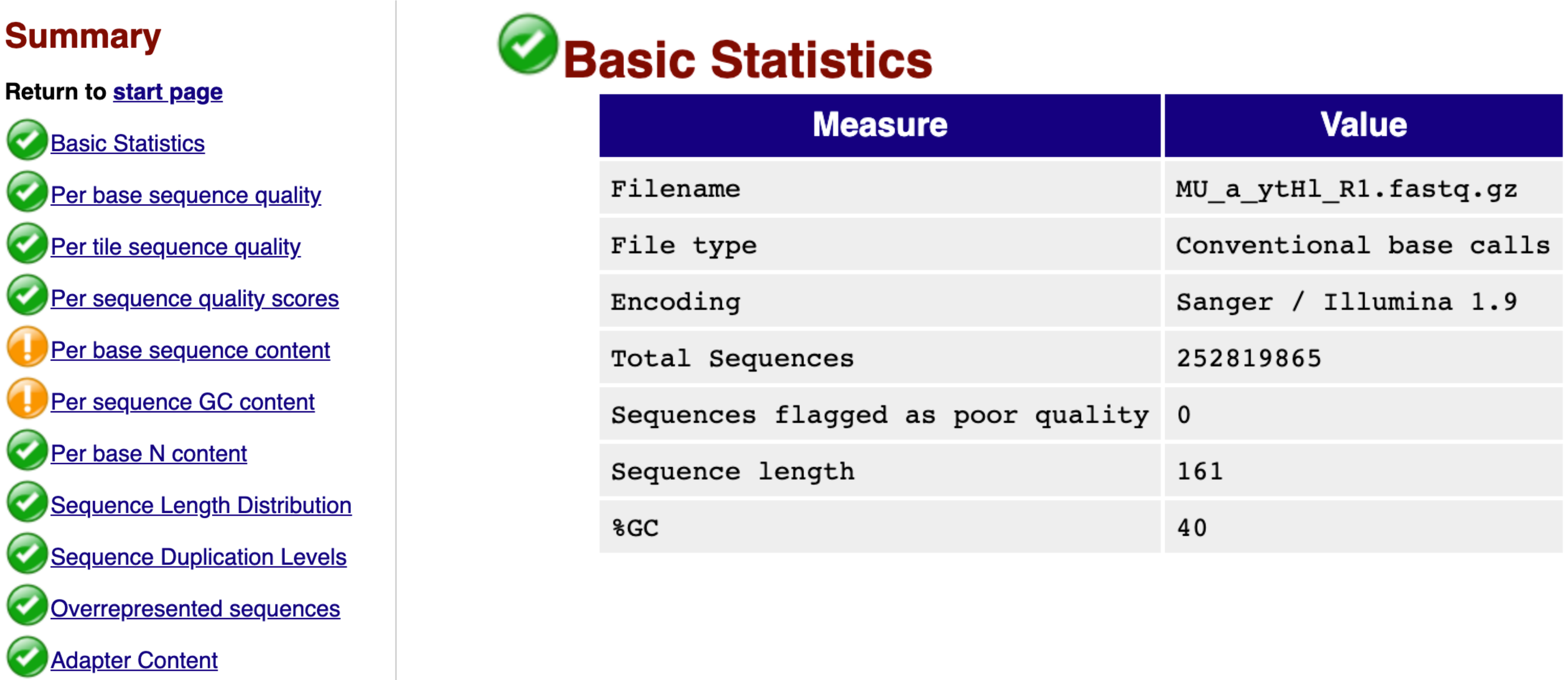
Fastq – quality control

- How can we summarize this?
- What QC can be done?



```
@M04743:199:000000000-CGG4F:1:1101:16145:1655 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGAAGGTGGCAAGCGTTGTTCGGATTCACTGGGCGTACAGGGAGCGTAGGCGGTTGGGTAAGCCCTCCGTGAAATCTCCGGG
+
ABCCCFFFCADBGGGGGGGGGGHHGHGGFHGHHHGHGGGAFFHGGGGGHHHHHHHHGGGGGHHGGGGGGGGGHGGEGGGGGHHHHHHHHGGHGGGHHHHHHGGG
@M04743:199:000000000-CGG4F:1:1101:18938:1729 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGCAGGCTGTTTTGTAAGTCAGATGTGAAATCCCCGAG
+
BBBBBFFFBBBBGGGGGGGGGFHHHHHGGHGGGGGGGGGGGHHGGEGFHHHHHHHHGGGGHFHGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHGGGG
@M04743:199:000000000-CGG4F:1:1101:13893:1760 1:N:0:233
GGTGCCAGCAGCCGCGGTACTACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGTTCGTAGGCGGTTTGTCGCGTCGTTTGTGAAAACCCGGGG
+
BBBBBFFFB4CCGGGGGGGCFFHGHHHGGHGGGGGGGGGAFGHGG?EFHFEHHHHHGGGGFHFHFGHGGHGG3EEEGGGHHEHGGGGGGGDHHEHGHGGGGGGG
F9FFFFFFFFFFFFEFFBBBBFEB;-@DFB-BBBFFFFEFF/EBBEFFF/BADFFDFFF.;
@M04743:199:000000000-CGG4F:1:1101:14830:1795 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGTTCTTTAAGTCAGTGGTGAAATACAGCCG
+
ABBABFBFB?AAEE?EGEFCGGHHFFHGEHFFHHGHGGGCFHHGEEGGDFGDHHHGGGFGDGHGGFEGFGGDFGGGGGHHFFFBGFH34FGBFFHGHHHGHFFC
9BD?99-9/9@-BD.;ADFFFBF///BBF:FFFFFFED?DFDFF?A.
@M04743:199:000000000-CGG4F:1:1101:14968:1984 1:N:0:233
AGTGCCAGCCGCCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGTTCTTTAAGTCAGTGGTGAAATACAGCCG
+
BBBBBFFFBABBGGGGGGGGGGHHGFHGHHGHHHGHGGGCFHHGGEGGHHHHHHGGGGHHHHHGHGGGGGGGHGGGGGHHHHHHHHHHHHHHGFFHHHHHHHHG
FCHHHGGHHHHHHHHHHHHHHHHHHHHHHFHHHHGFHHGEGGFHHGHGGGFEGG9FGGAEGGGGAFDGEFFGGFFFBFEFFFFFFFFFFFFFFFFFFF>DFDBFF
@M04743:199:000000000-CGG4F:1:1101:12706:2099 1:N:0:233
TGTGCCAGCCGCCGCGGTAATACGGAGGGAGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGTTTTTTAAGTCAGAGGTGAAAGCCCGGGG
+
BCCCCFFFCCCCGGGGGGGGGGHHEGGGGDFGGHHHGGGGGHHGGGGFHHGHHHHHGGGGHHHGGGGGHHHGHGGGGGGGGACGHHHHHHGHHGHHFHHHGGGG
BFFFFFFFF9FFFFFFFFFFFFFFFF/
@M04743:199:000000000-CGG4F:1:1101:13747:2260 1:N:0:233
CGTGCCAGCCGCCGCGGTAATACGAAGGGGGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGAGTTCGTAGGCGGTTTGTCGCGTCGTTTGTGAAAACCCGGGG
+
CCCCCCFFCABCGGGGGGGGGGHHFCEGDGGGHHHGGGEFHHGGGFFHHFHHHHGGGGGHH@GHHHGGHGGHGGGGGGFH</>CFCGGGGHHHHHFHGGGGGG
A@@FFFFFFFFFFFFFBF9C;=CF.@;CDFFFFFBDFFFFFF?BEFFFFFFFFFFFFFFFFFF?
@M04743:199:000000000-CGG4F:1:1101:20151:2263 1:N:0:233
TGTGCCAGCCGCCGCGGTAATACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGCAGGCTGTTTTGTAAGTCAGATGTGAAATCCCCGAG
+
BBBBBFFFBAADGGGGGGGGGGHHHHHGGHGGGGGGGGGGGHHGGDFFHHHHHHHGGGGGHGHGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHGGFG
@M04743:199:000000000-CGG4F:1:1101:17232:2363 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGGAGGGGGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGATCGGAAAGTCAGAGGTGAAATCCCAGGG
+
BBBBBFFFBBBBGGGGGGGGGGHHGDGGGGGGGHHHGGG0FGHGGEGFHHHHHHHHGGGGHHHGGGGGHHHGHGGGGGGGGGGHHHHHHHGHHHGHHHHGHHGG
```

FastQC Report



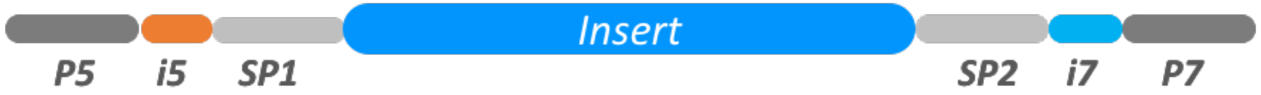
Summary

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

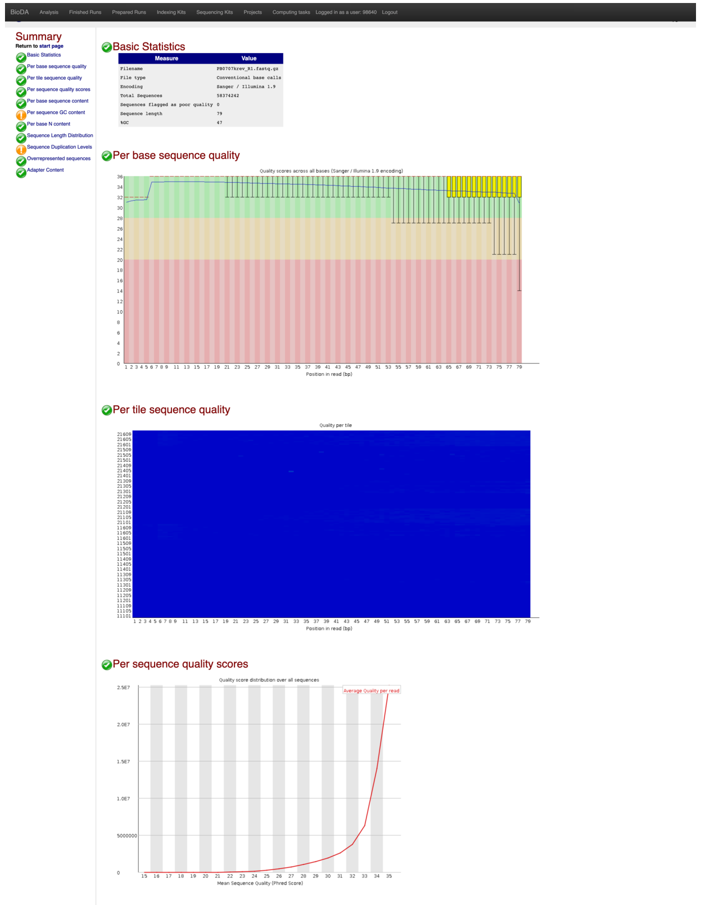
✓ Basic Statistics

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



Fastq – quality control

- Fastqc - tool



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

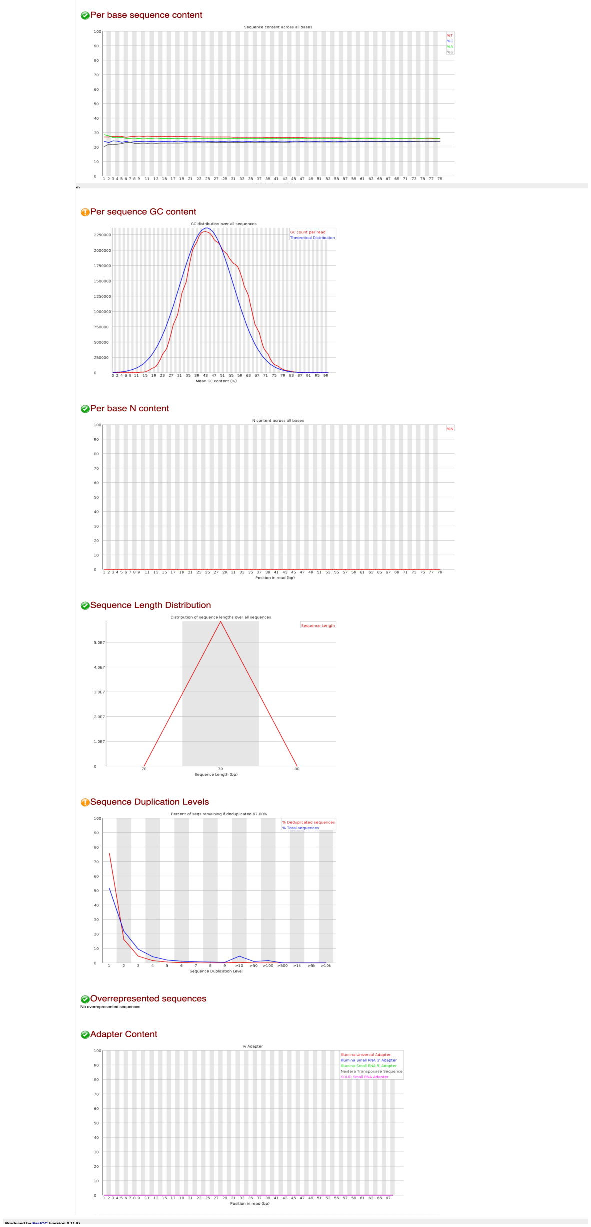
✓ Basic Statistics

Measure	Value
Filename	PB0707krev_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	58374242
Sequences flagged as poor quality	0
Sequence length	79
%GC	47

✓ Per base sequence quality

✓ Per tile sequence quality

✓ Per sequence quality scores


Sequence content across all bases
%T
%C
%A
%G
GC distribution over all sequences
GC count per read
Theoretical Distribution
Mean GC content (%)
N content across all bases
%N
Position in read (bp)
Distribution of sequence lengths over all sequences
Sequence Length
Sequence Length (bp)
Percent of seqs remaining if deduplicated 67.88%
% Deduplicated sequences
% Total sequences
Sequence Duplication Level
% Adapter
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter
Position in read (bp)

✓ Per base sequence content

! Per sequence GC content

✓ Per base N content

✓ Sequence Length Distribution

! Sequence Duplication Levels

✓ Overrepresented sequences

No overrepresented sequences

✓ Adapter Content

Produced by FastQC (version 0.11.8)

CEITEC

@CEITEC_Brno

Thank you for your attention!

Vojta Bystry
vojtech.bystry@ceitec.muni.cz