



CEITEC

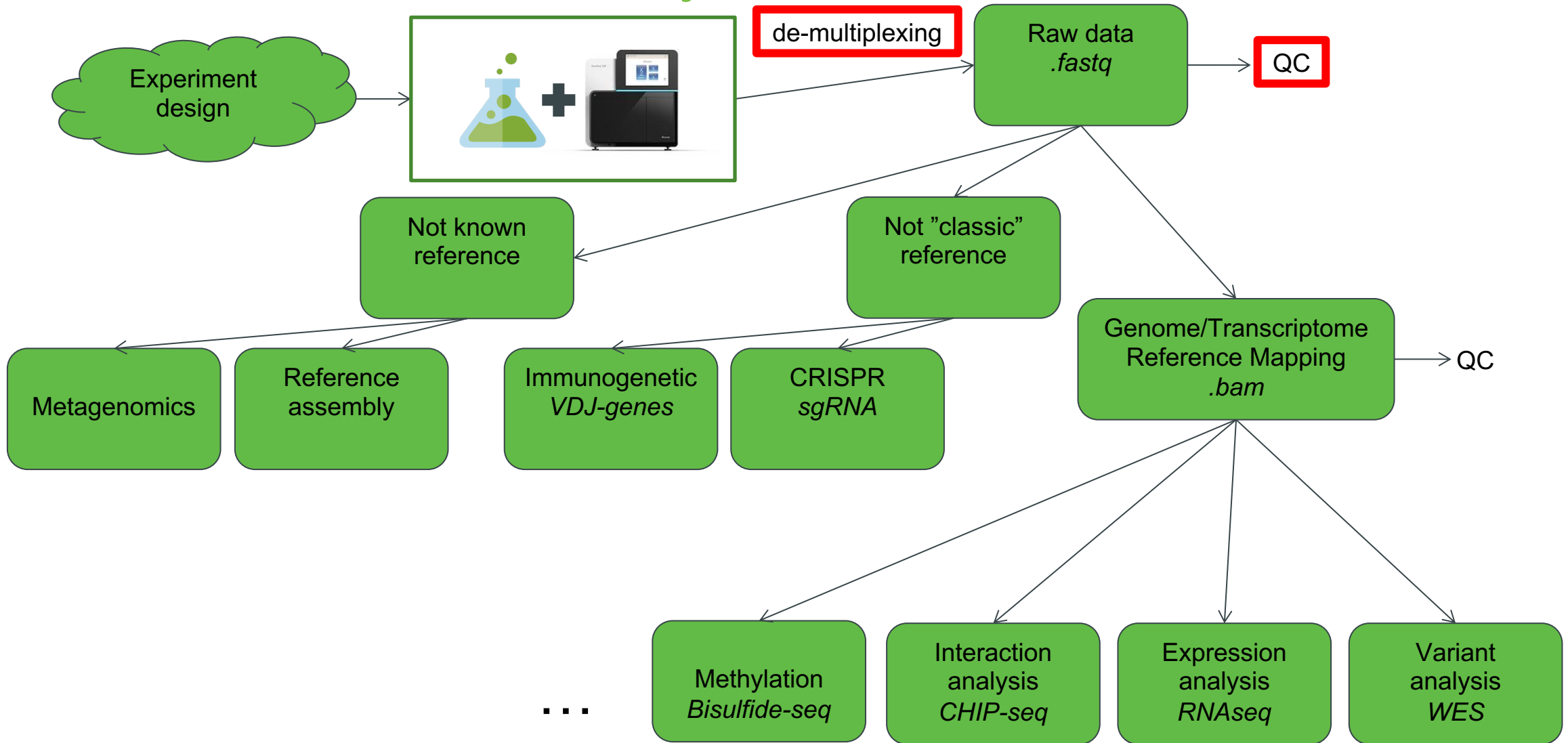
Central European Institute of Technology
BRNO | CZECH REPUBLIC

**Modern methods for genome analysis
(PřF:Bi7420)**

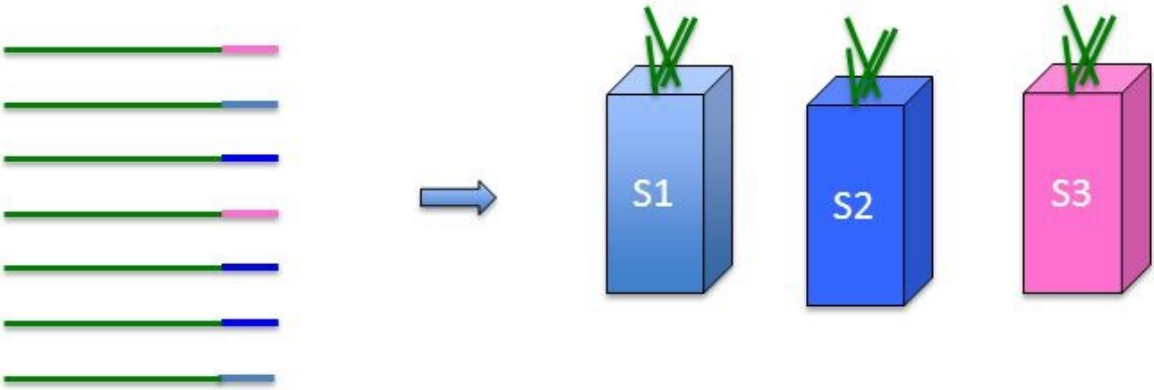
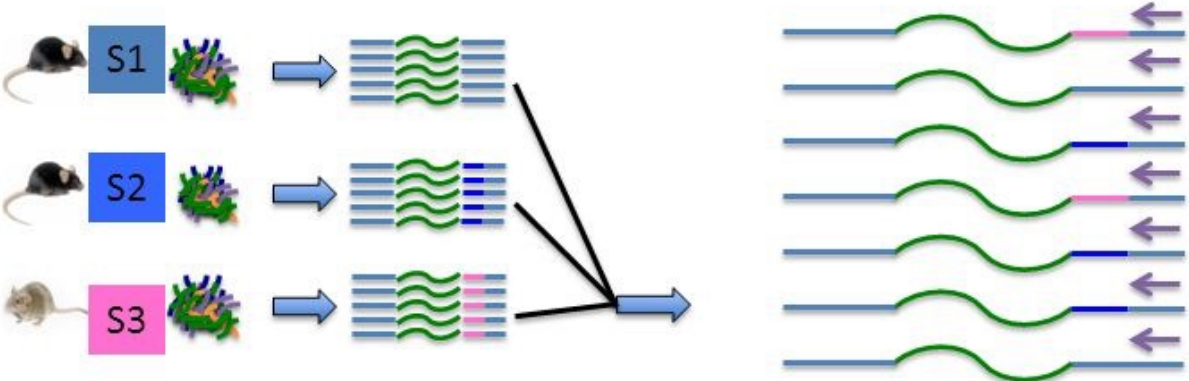
Lecture 2 : Raw NGS data quality control

Vojta Bystry
vojtech.bystry@ceitec.muni.cz

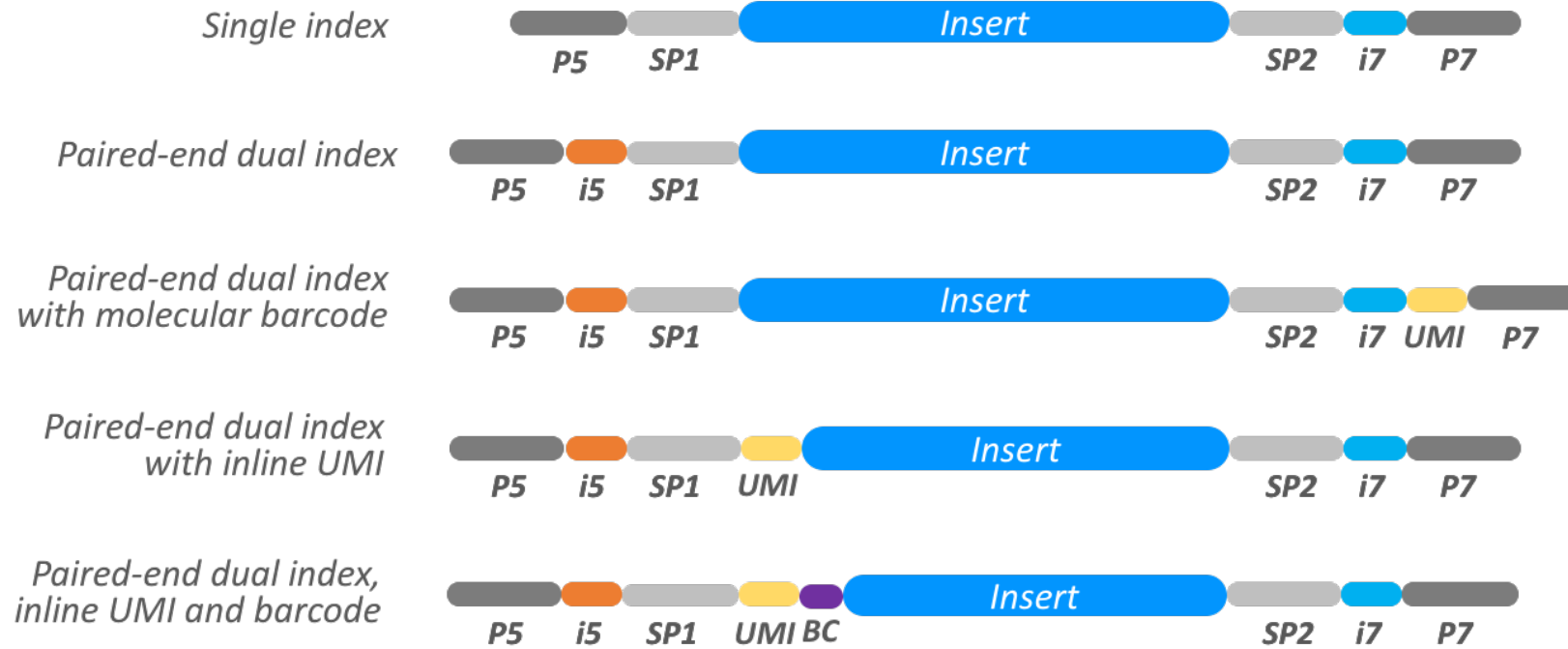
NGS data analysis



De-multiplexing



De-multiplexing



- 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)

Fastq format - quality

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

CFFFFEFGCEEGECFGGGGAFF87@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 thinks 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
GGTGCCAGCCGCCGCGGTAATACGAAGGTGGCAAGCGTTGTTCCGATTACTGGGCGTACAGGGAGCGTAGGCGGTTGGGTAAGCCCTCCGTGAAATCTCCGGG
+
ABCCFFFCADBGGGGGGGGGGHGHGGFHHGHGGGAFFHGGGGHHHHHHHHGGGGHGGGGGGGGHGGEGGGGGHHHHHHHHGGHGGHHHHHHGGG
|M04743:199:000000000-CGG4F:1:1101:18938:1729 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGTAGGGTGGCAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCAGGCGTGTGTTGTAAGTCAGATGTGAAATCCCCGAG
+
BBBBBFFB4CCGGGGGGGCFHHHHHGGHGGGGGGGGHGGEGFHHHHHHHHGGGGHFGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGGGG
|M04743:199:000000000-CGG4F:1:1101:14968:1984 1:N:0:233
GGTGCCAGCCGCCGCGTACTACGTAGGGTGGCAGCGTTGTCCGGAATTACTGGGCGTAAAGAGTTCGTAGGCGGTTTGTCCGCGTCTTTGTGAAAACCCGGGG
+
BBBBBFFB4CCGGGGGGGCFHHHHHGGHGGGGGGGAFGHGG?EFHFEHHHHHGGGGFHFHFGHGGHGG3EEEGGGHHEGGGGGGDHEHGGGGGGGG
F9FFFFFFFFFFFFBFBFBFB;~@DFB-BBBFFFEFF/EBBEFFF/BADFFDFFF.;
|M04743:199:000000000-CGG4F:1:1101:14830:1795 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGATTATTGGGTTTAAAGGGTGCAGTAGGCGGTTCTTTAAGTCAGTGGTAAATACAGCCG
+
ABBABFBFB?AAEE?EGEFCGGHHFFHGEHFFHHGHGGGCFHHGEEGGDFGDHGGGGFDDGGHGGFEGFGGDFGGGGHHFFBGFH34FGBFFHGHGHGFFC
9BD?99-9/9@-BD.;ADFFBF//BBF:FFFFFFED?DFDFF?A.
|M04743:199:000000000-CGG4F:1:1101:14968:1984 1:N:0:233
AGTGCCAGCCGCCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGATTATTGGGTTTAAAGGGTGCAGTAGGCGGTTCTTTAAGTCAGTGGTAAATACAGCCG
+
BBBBBFFBABBGGGGGGGGGGHGHGGFHHGHGGGGCFHHGGEGHHHHHHGGGGHHHHHHGGGGGGGGGGGGHHHHHHHHHHHHHGFHHHHHHHHGG
FCHHHGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGHGGEGFHHGHGGGFEGG9FGGAEGGGGAFDGEFFGGFFBFEFFFFFFFFFFFFFFFFFFFF>DFDBFFI
|M04743:199:000000000-CGG4F:1:1101:12706:2099 1:N:0:233
TGTGCCAGCCGCCGCGGTAATACGGAGGGAGCTAGCGTTGTTCCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGTTTTTTAAGTCAGAGGTGAAAGCCCGGGG
+
BCCCCFFCCCGGGGGGGGGHHEGGGDFGGHHGGGGGGHGGGGFHHGHHHHHGGGGHHHHGGGGHHHHGGGGGGGGHGGGGGGGACGHHHHHHGHGHGHHHHGGGG
BFFFFFFFF9FFFFFFFFFFFFFFFF/
|M04743:199:000000000-CGG4F:1:1101:13747:2260 1:N:0:233
CGTGCCAGCCGCCGCGGTAATACGAAGGGGCTAGCGTTGTTCCGGAATTACTGGGCGTAAAGAGTTCGTAGGCGGTTTGTCCGCGTCTTTGTGAAAACCCGGGG
+
CCCCCFFCABCGGGGGGGGGHGFCEGGGGHHGGGEFHHGGGFHHFHHHHHHGGGGHH@GHHHGGHGGHGGGGGGFH</>CFCGGGGHHHHHFGGGGGG
A@@FFFFFFFFFFFFBF9C;=CF.@;CDFFFFFFBDFFFFFFF?BEFFFFFFFFFFFFFFFFFFFF?
|M04743:199:000000000-CGG4F:1:1101:20151:2263 1:N:0:233
TGTGCCAGCCGCCGCGGTAATACGTAGGGTGGCAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGATCGGAAAGTCAGAGGTGAAATCCCCGAG
+
BBBBBFFBAAADGGGGGGGGGGHHHHHGGHGGGGGGGGGGHGGDFHHHHHHHHGGGGGGHGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHGGFG
|M04743:199:000000000-CGG4F:1:1101:17232:2363 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGGAGGGGCTAGCGTTGTTCCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGATCGGAAAGTCAGAGGTGAAATCCAGGGG
+
BBBBBFFB4CCGGGGGGGCFHHHHHGGHGGGGGGGGGGHGGDFHHHHHHHHGGGGGGHGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGGG
```

FastQC Report

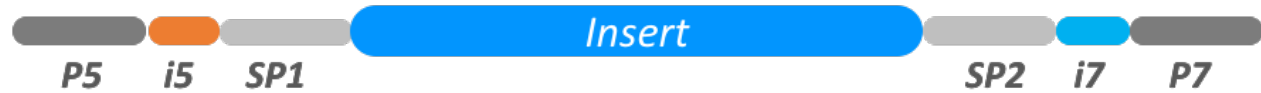
Summary

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](#)

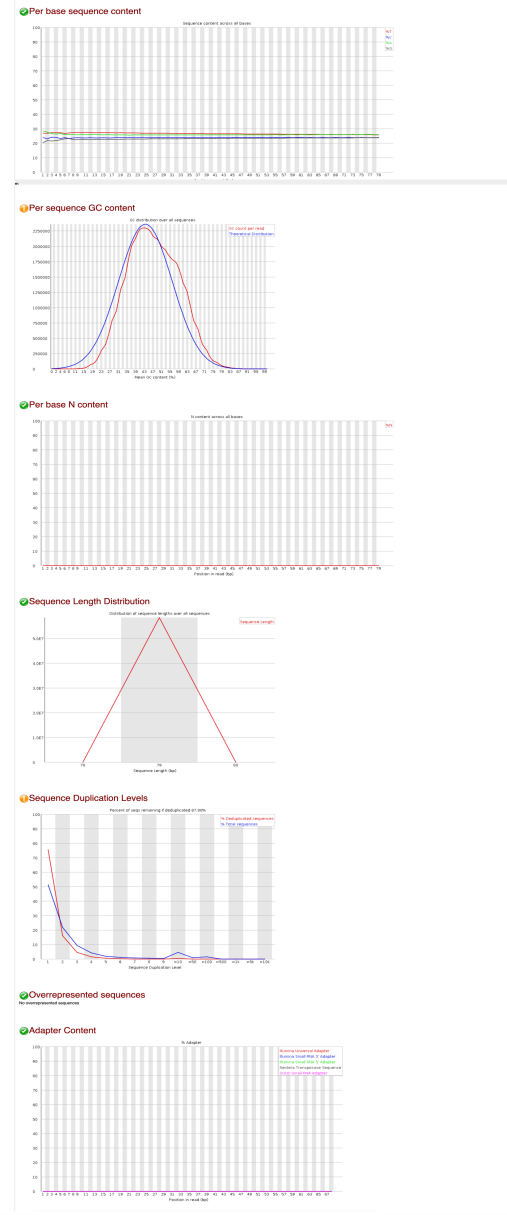
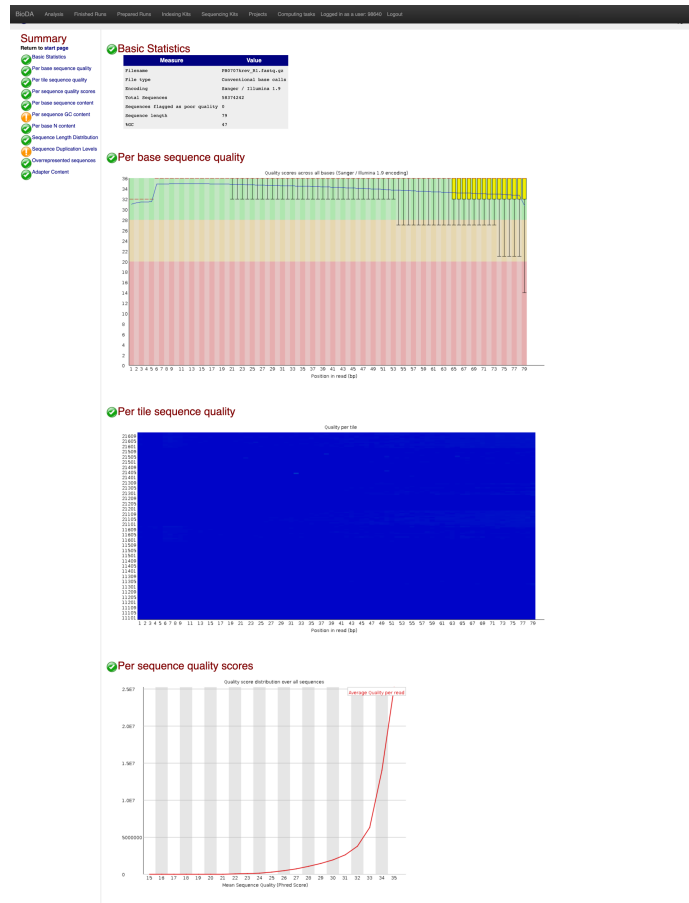
✔ Basic Statistics

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





CEITEC



@CEITEC_Brno

Thank you for your attention!

