MUNI | RECETOX

Research infrastructure

Basics of Sequencing Technologies

PřF:E4014 Projekt z Matematické biologie a biomedicíny - biomedicínská bioinformatika
FI:IV110 Project in Sequence Analysis
FI:IV114 Projekt z bioinformatiky a systémové biologie

Vojtěch Bartoň
vojtech.barton@recetox.muni.cz

RECETOX, Masaryk University

September 29, 2024

Table of Contents

Sequencing

DNA Sequencing

DNA sequencing is the process of determining the nucleic acid sequence – the order of nucleotides in DNA.

Examples

Question: What's it good for?

Sequencing Technology

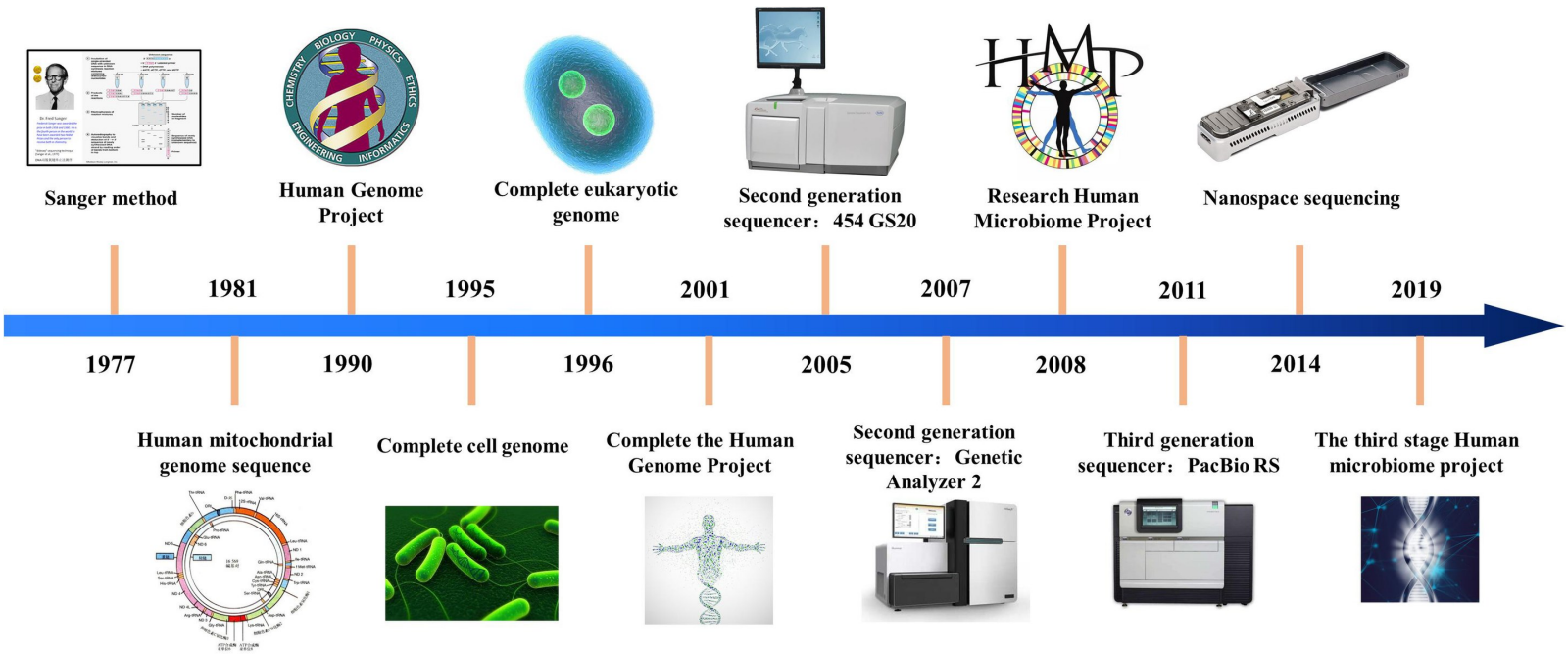


Figure: History of sequencing technology[1]

Illumina sequencing

- NextGeneration sequencing technology
- Sequencing by synthesis
- Utilizing PCR
- Widely used

Principle

https://youtu.be/fCd6B5HRaZ8?si=0Np6Q6pX4236HnvN

Oxford Nanopore

- Third generation of sequencing technology
- Sequencing by ion stream disruption (electricity)
- Long reads, real-time
- Squiggle

Principle

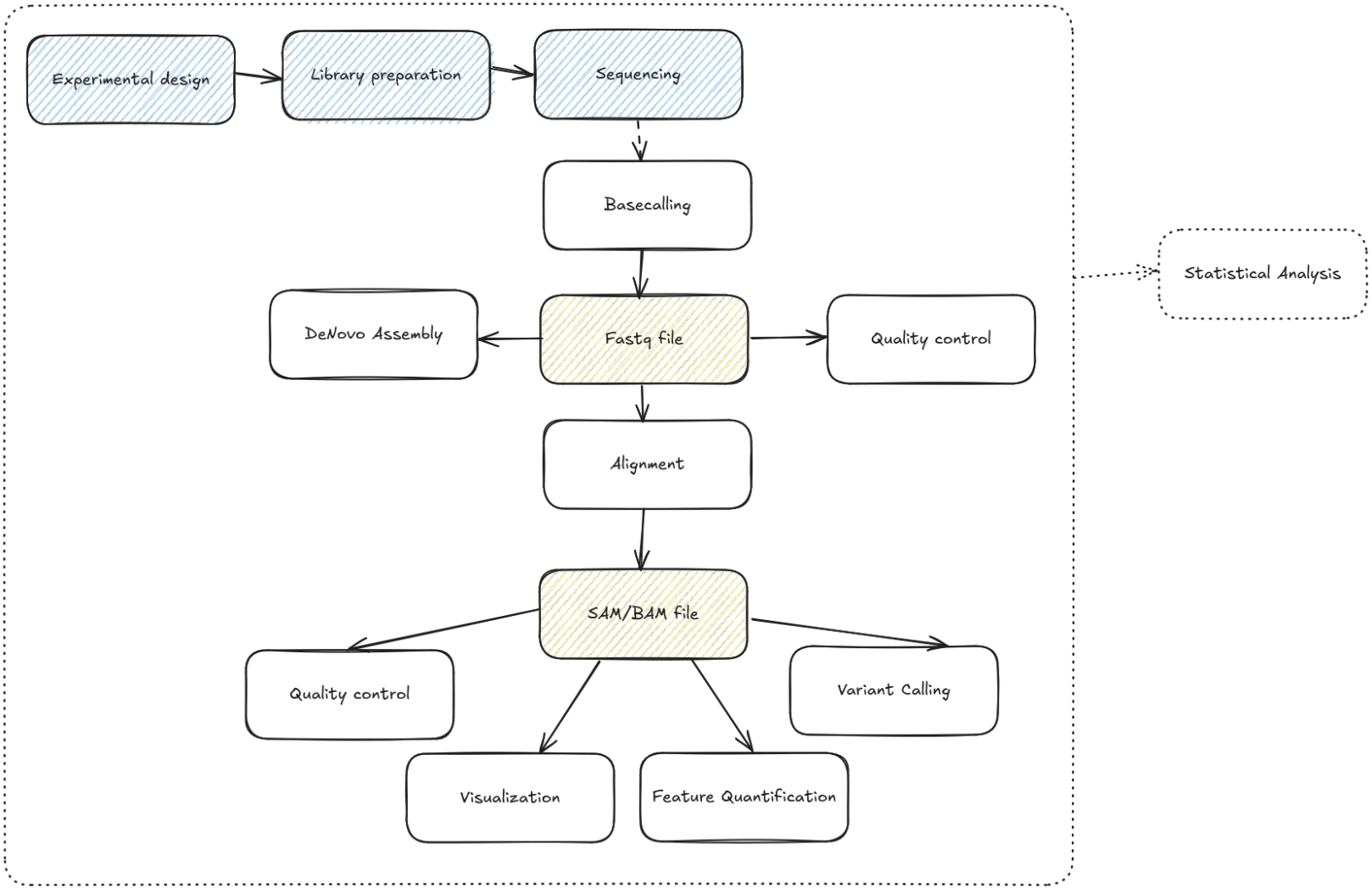
https://youtu.be/RcP85JHLmnI?si=k732mK9liWV3gw5d

Comparison

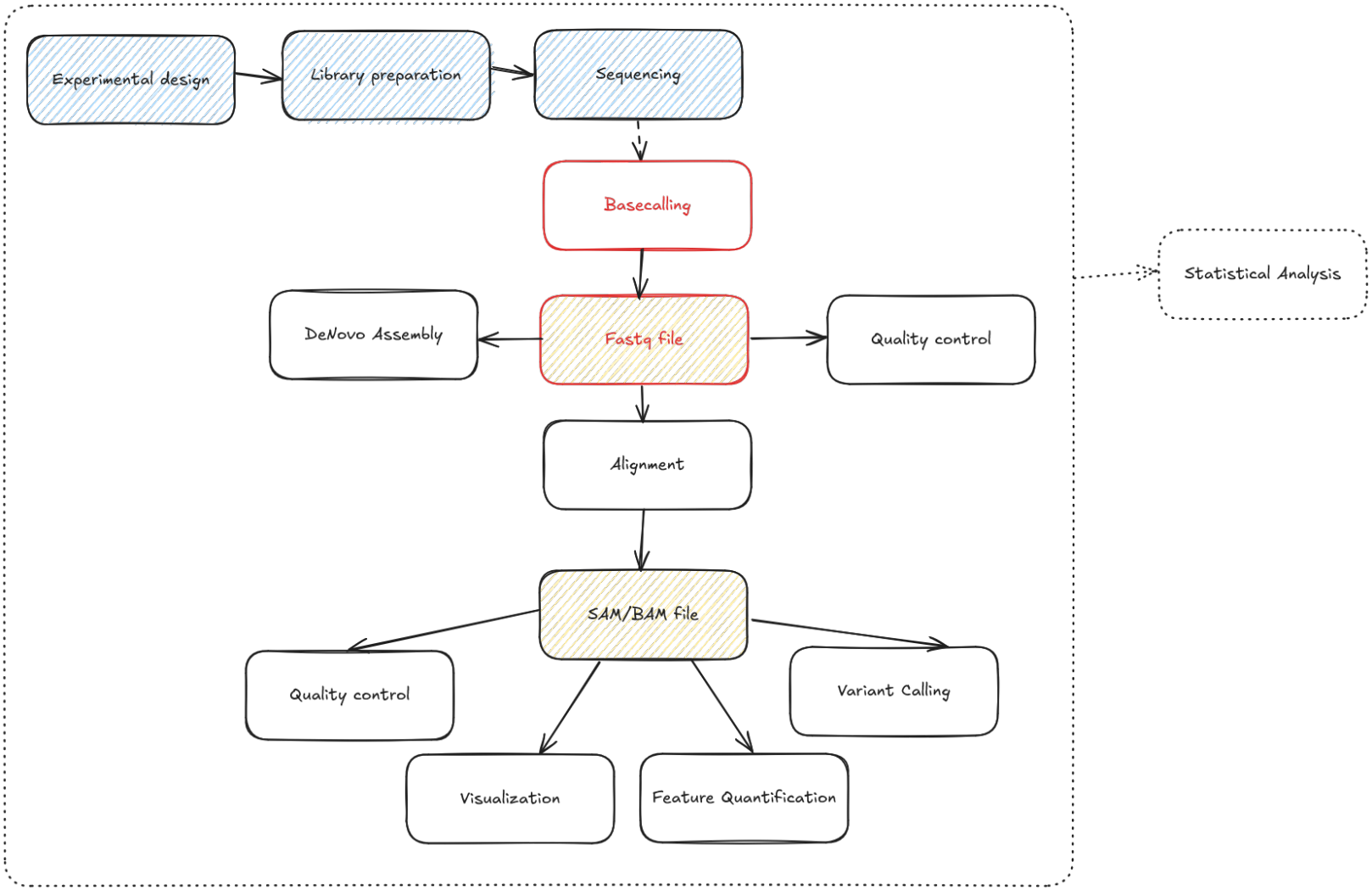
	Illumina	Oxford Nanopore
Read length	< 600 bp	< 2 Mbp
Accuracy	99 %	87-98 %
Price per Gbp	$ 40-60 (NextSeq) $ 10-35 (NovaSeq)	$ 50-200 (minION) $ 20-40 (PromethION)
Real-time		✓
Epigenomics	(Special chemistry)	✓

Table: Comparison of technologies

General workflow



Basecalling



Basecalling

Definition

Basecalling is the process of converting raw sequencing signals into a nucleotide sequence (A, T, C, G).

Examples

Question: How is basecalling done for Illumina and Oxford Nanopore?

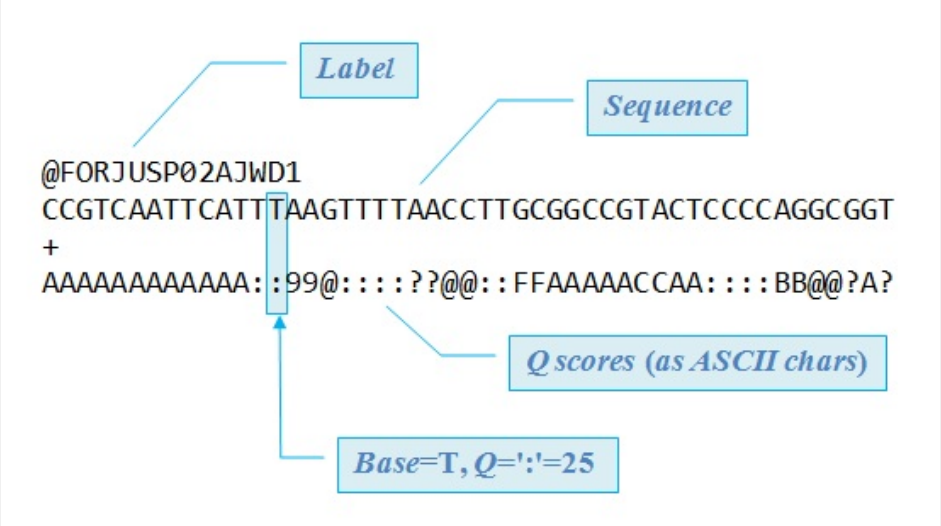
Fastq format

The FASTQ format is a text-based file format used to store both the raw sequence data and the corresponding quality scores from sequencing. Each entry consists of four lines:

1. Sequence identifier starting with @.
2. Raw nucleotide sequence (A, T, C, G).
3. + symbol, sometimes followed by the same identifier.
4. PHRED quality scores encoded as ASCII characters corresponding to each nucleotide in the sequence.

Fastq format

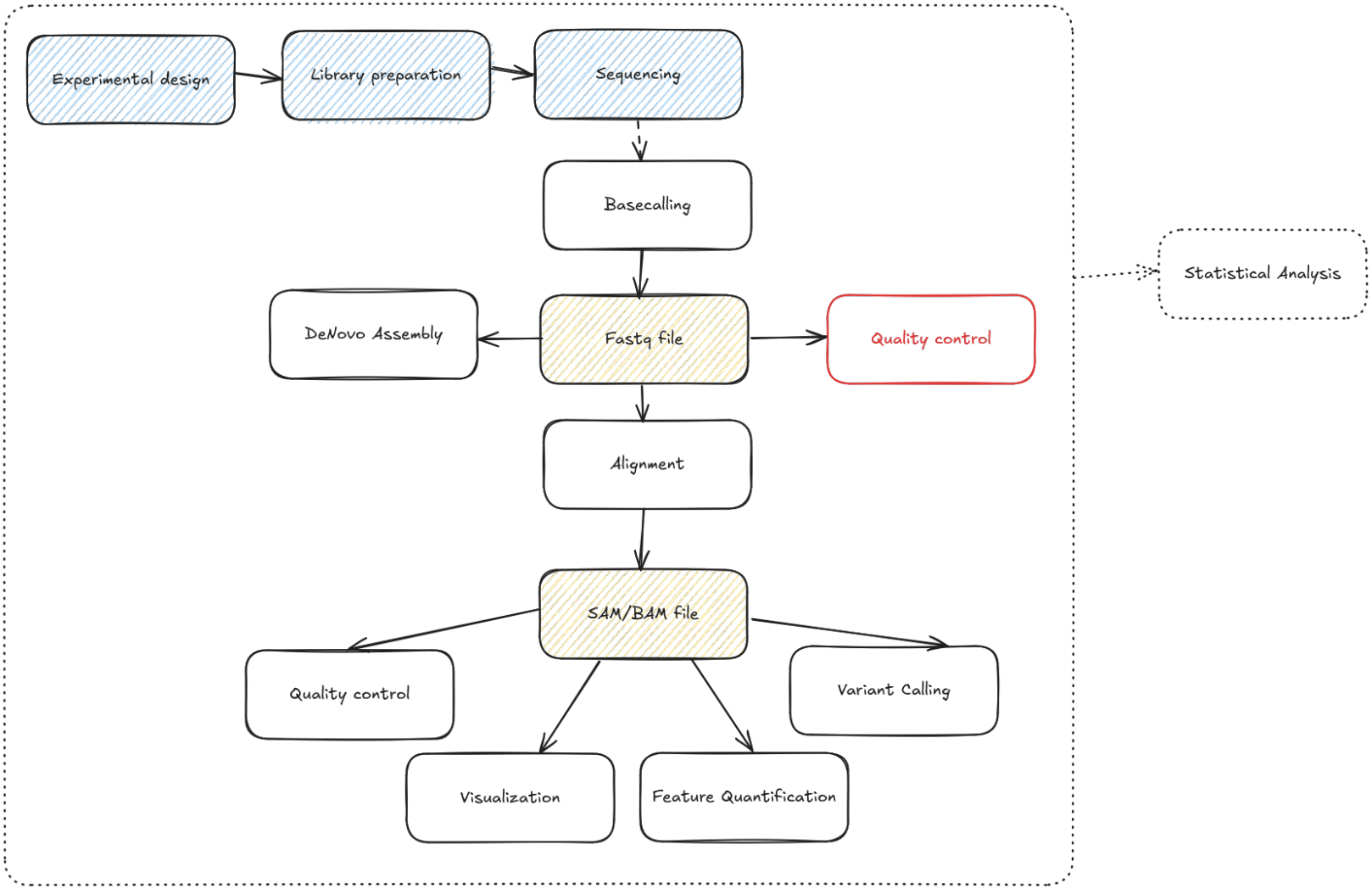
Examples



PHRED Score

The PHRED score is a quality score that indicates the accuracy of a nucleotide base call in DNA sequencing, with higher scores representing higher confidence and lower error probabilities.

Quality control



Quality control

- Describe the quality of sequencing data
- Set parameters of preprocessing (Trimming & Filtering)

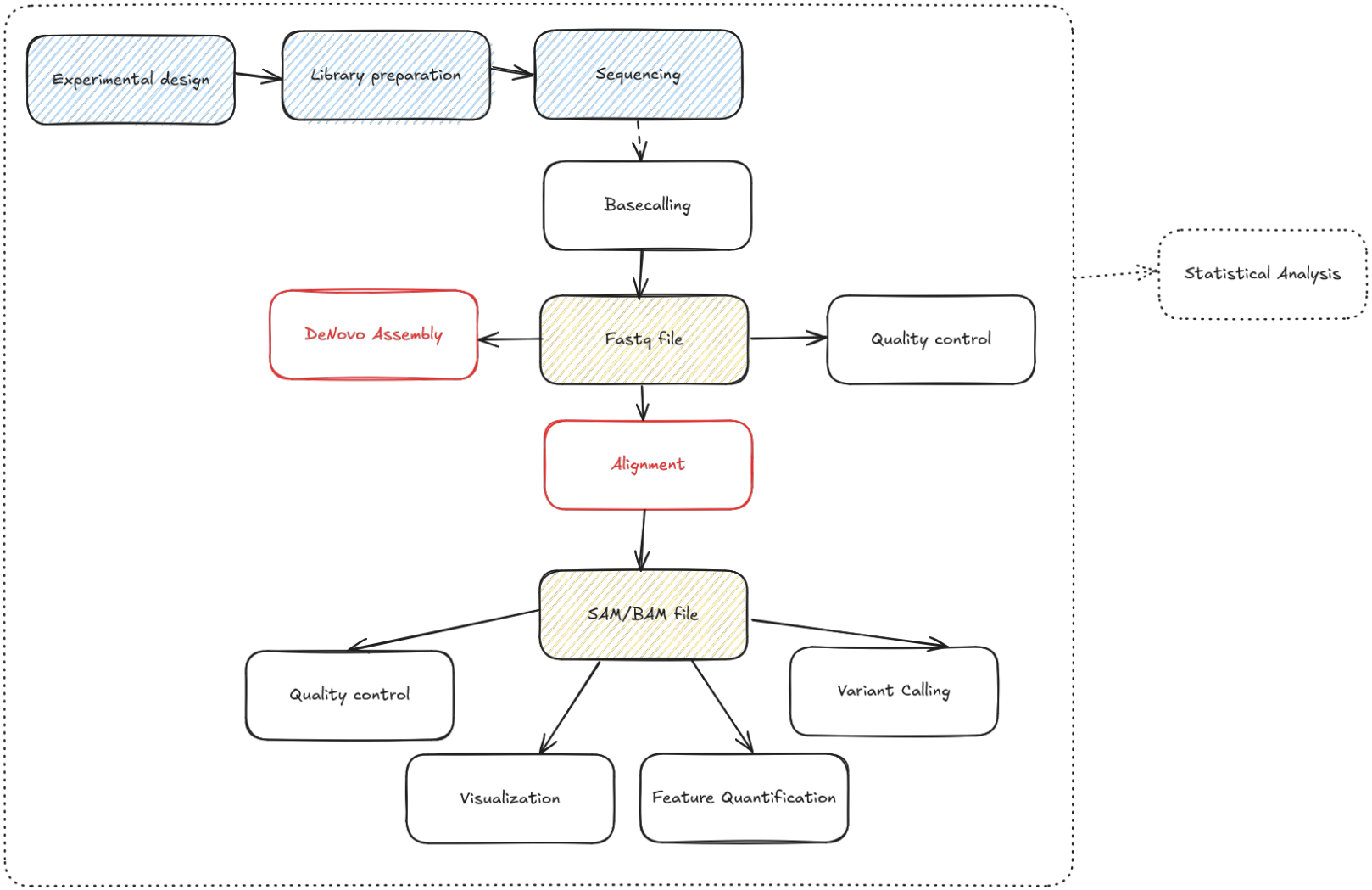
Examples

Question: What quality parameters to assess?

Examples

- Fastqc
- Nanoplot
- Fastp

Assembly & Alignment



Assembly & Alignment

DeNovo Assembly

De novo assembly is the process of constructing a genome sequence from short DNA fragments without the use of a reference genome, by assembling overlapping reads into longer contiguous sequences (**contigs**).

Alignment

Mapping is the process of aligning sequencing reads to a **reference genome** to determine the origin of each read and identify variations or similarities.

Examples

- DeNovo: SPADes
- Mappers: Bowtie2, BWA
- RNA Mappers: STAR (splice-aware mapping)

SAM/BAM format

SAM/BAM format

SAM (Sequence Alignment/Map) and BAM (Binary Alignment/Map) are file formats used to store aligned sequencing reads. Both include information about the read sequences, their alignment positions, mapping quality, and optional metadata.

Examples

```
HD VN:1.5 SD:coordinate
SD SN:ref LN:45
```

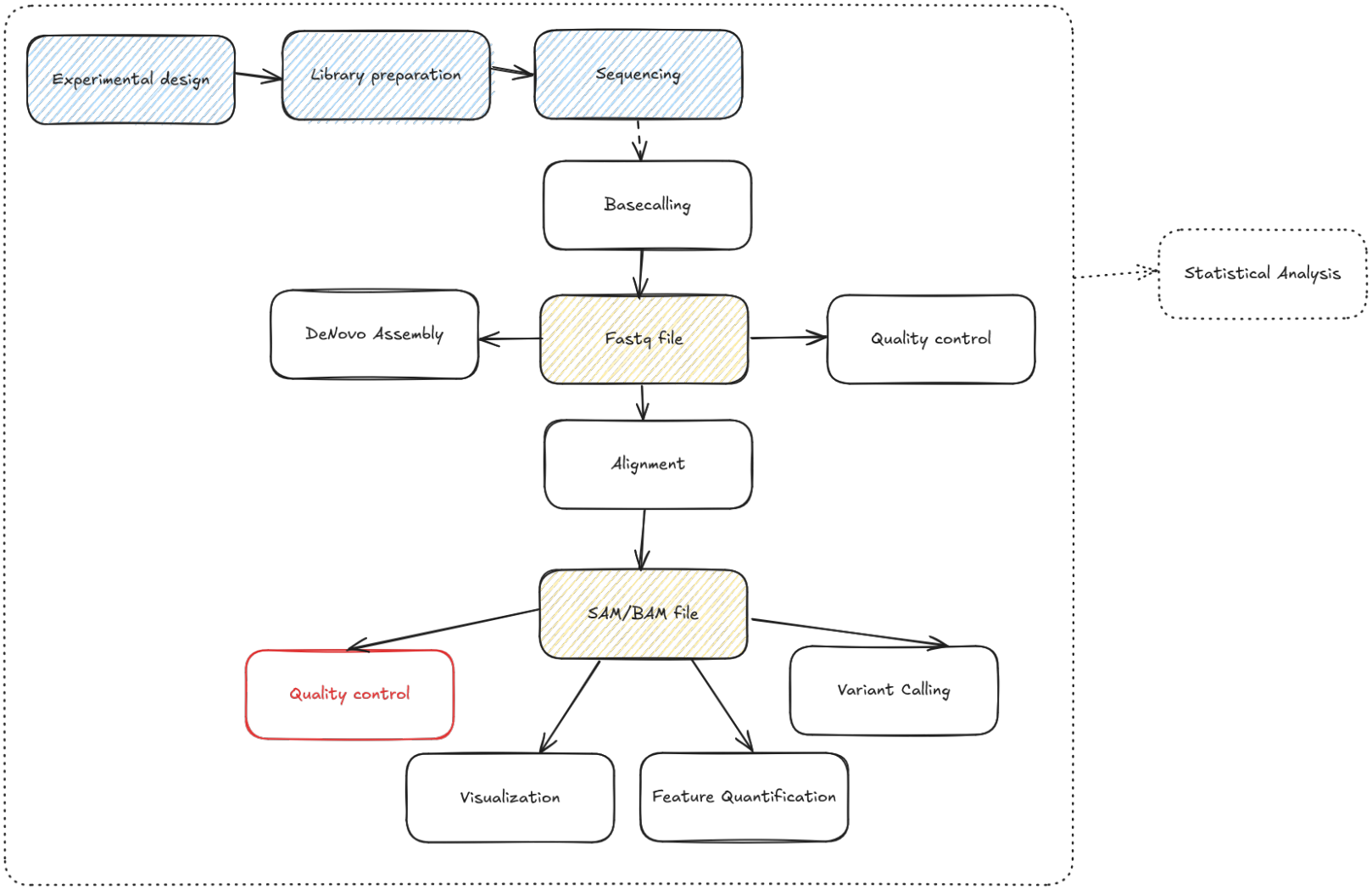
HEADER section

```
r001    99   ref    7   30   8M2I4M1D3M  =    37     39   TTAGATAAAGGATACTG *
r002     0   ref    9   30   3S6M1P1I4M  *     0      0   AAAAGATAAGGATA    *
r003     0   ref    9   30   5S6M        *     0      0   GCCTAAGCTAA       *   SA:Z:ref,29,-,6H5M,17,0;
r004     0   ref   16   30   6M14N5M     *     0      0   ATAGCTTCAGC       *
r003  2064   ref   29   17   6H5M        *     0      0   TAGGC             *   SA:Z:ref,9,+,5S6M,30,1;
r001   147   ref   37   30   9M          =     7    -39   CAGCGGCAT         *   NM:i:1
```

ALIGNMENT section

QNAME FLAG RNAME POS MAPQ CIGAR RNEXT PNEXT TLEN SEQ QUAL

Alignment QC



Alignment QC

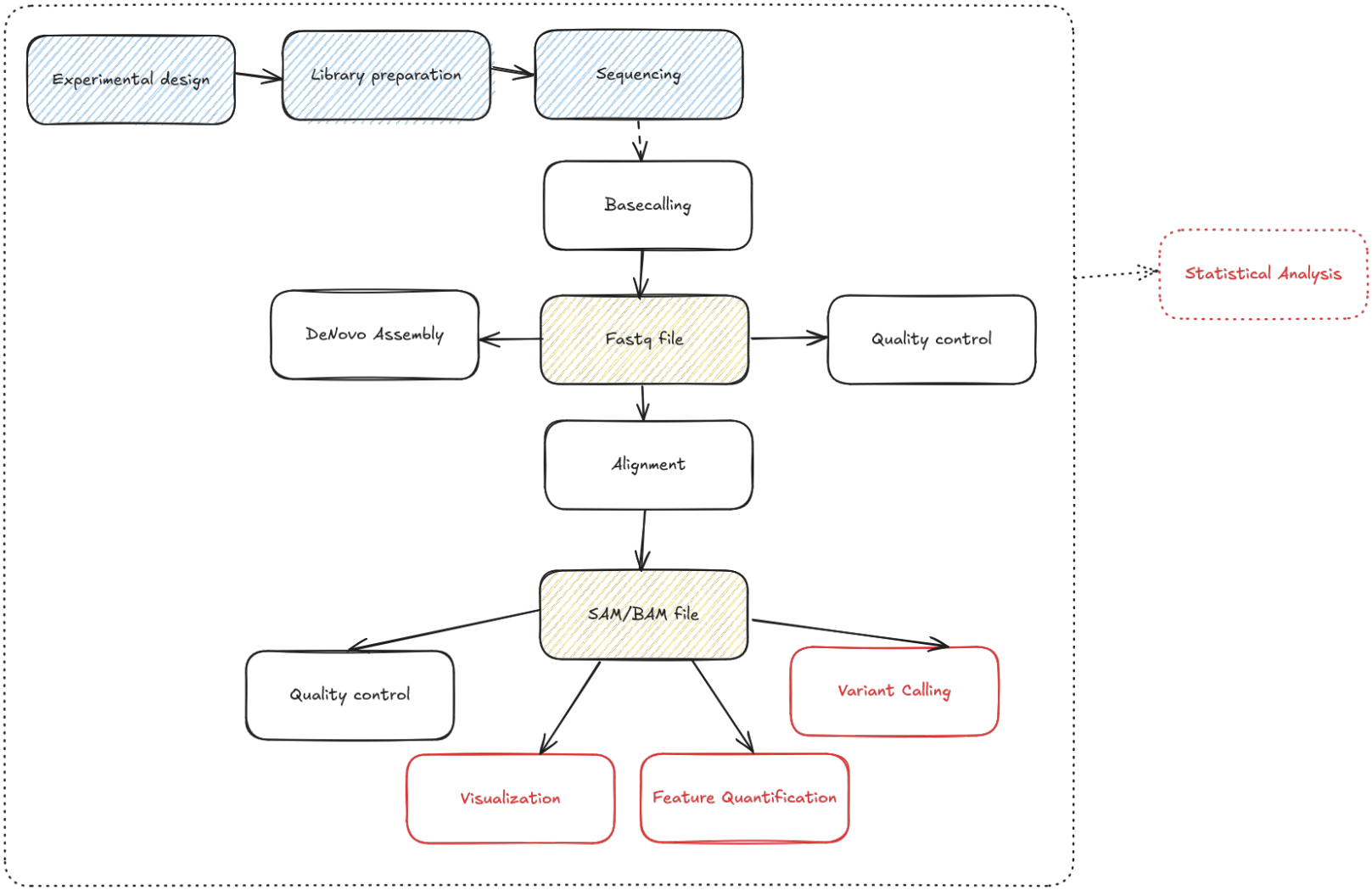
Examples

Question: What parameters to collect?

Examples

- Samtools
- QualiMap
- Picard tools

Postprocessing



Postprocessing

Depends on type of the experiment, quality of data, study design, hypotheses, ...

Visualization

Integrated Genome Browser (IGV)

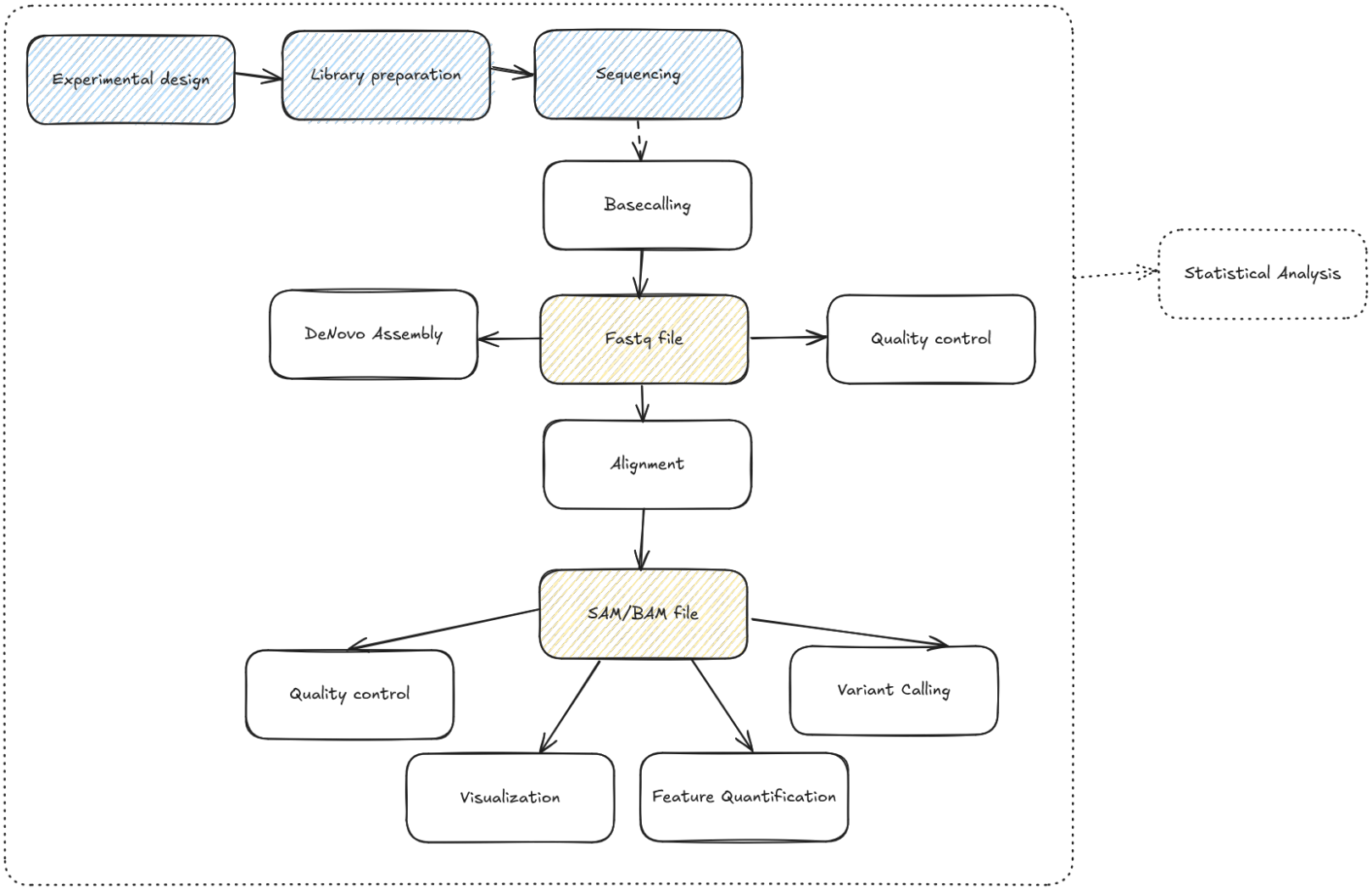
Feature Quantification

RNA-sequencing (genes), Metagenomics (bacteria)

Variant calling

Mutations, SNP, CNV

Summary


Experimental design
Library preparation
Sequencing
Basecalling
Statistical Analysis
DeNovo Assembly
Fastq file
Quality control
Alignment
SAM/BAM file
Quality control
Variant Calling
Visualization
Feature Quantification

To Remember

- Bioinformatics (and especially the sequencing bioinformatics) is a very new field
- No good books, no standards, nothing lasts forever, ... almost everything is old and outdated!
- Garbage in –> garbage out
- If you do not understand the whole process you don't know what the results mean

Keywords

Important terms

Sequencing, Illumina, Oxford Nanopore, Basecalling, Paired-end sequencing, PCR, bridge PCR, Adapters, Index, Pooling, Demultiplexing, Squiggle, Fasta, Fastq, SAM/BAM, DeNovo Assembly, Alignment, Mapping, Splice-aware, Quality control, Filtering, Trimming, Phred Score, SNP, Mutation, CNV, Workflow, ...

MASARYK
UNIVERSITY