Module 1
Sequencing platforms and WGS terminology

The Fleming Fund
Regional Grants

15 February 2021

Jette S. Kjeldgaard, DTU FOOD

The Fleming Fund
Regional Grants

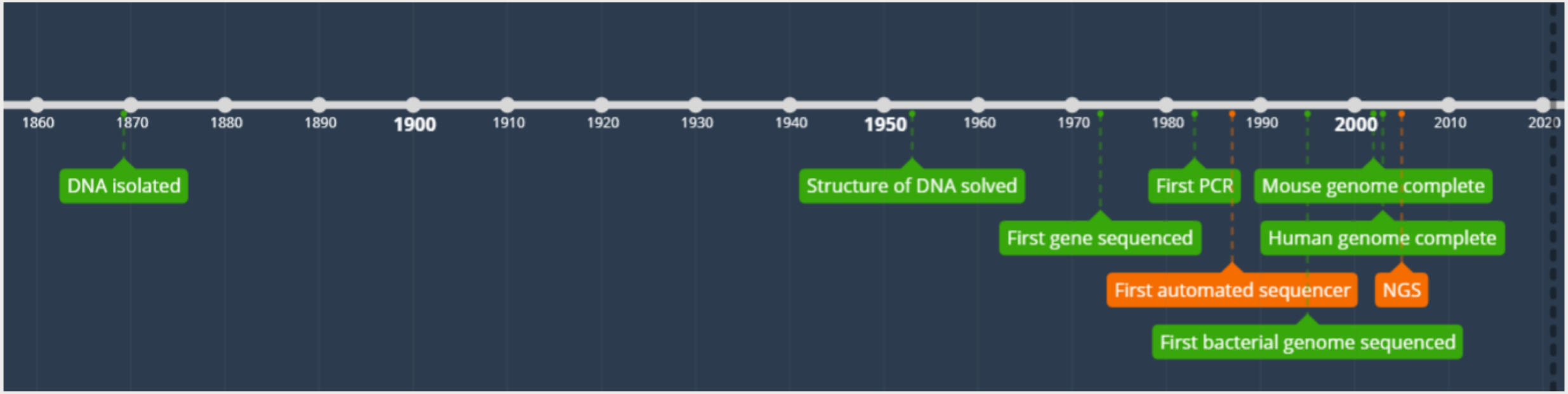
Overview

- Brief history of sequencing
- Sequencing platforms
  - Short reads
  - Long reads
- Pro's and con's
- Sequencing parameters for quality control

The Fleming Fund
Regional Grants

Brief history of sequencing

- First DNA isolation in 1869!
- The structure of DNA solved in 1953
- The first complete gene was sequenced in 1972



The Fleming Fund
Regional Grants

Second (next) generation sequencing

- Possible to sequence an entire genome at once
- General process is:
  - Fragmenting the genome into many, smaller pieces
  - Randomly sampling for a fragment
  - Sequencing the fragment using one of many possible technologies

The Fleming Fund
Regional Grants

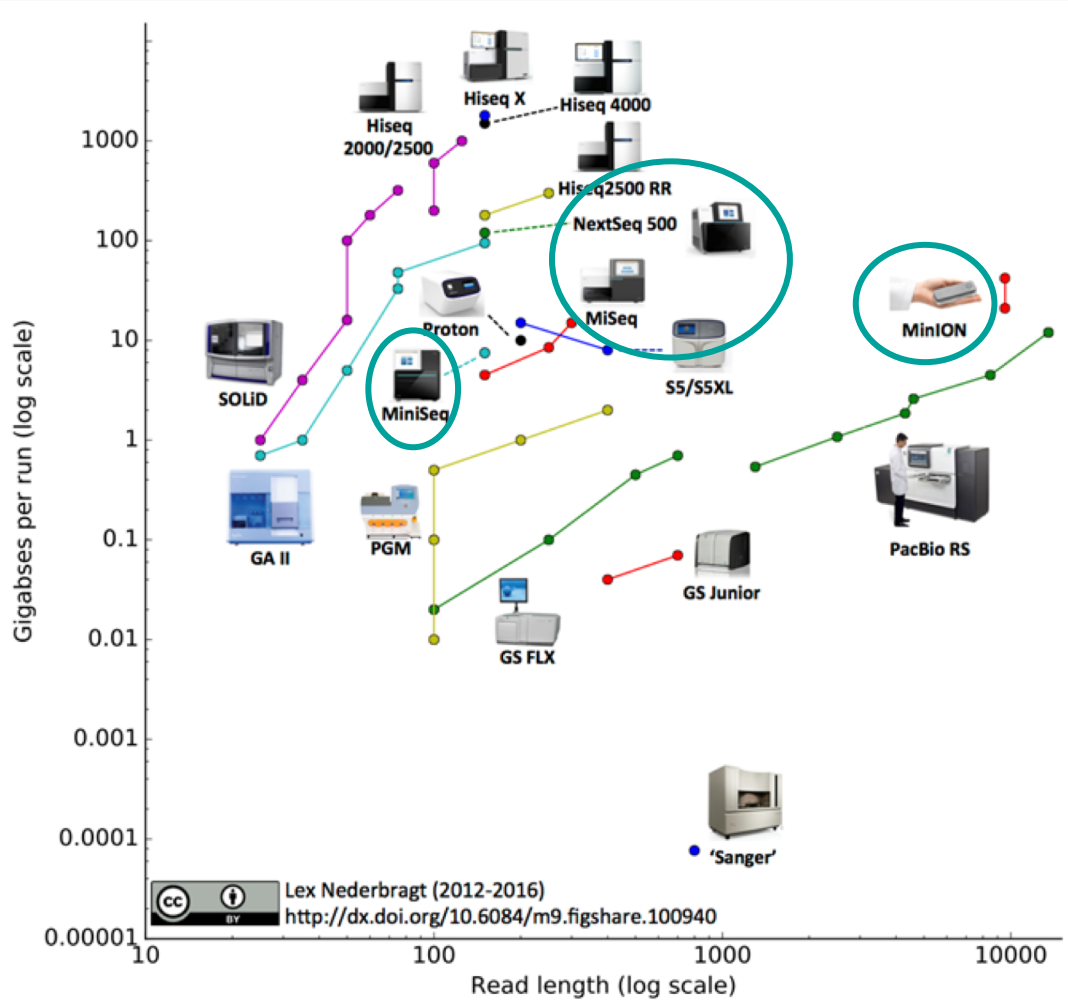
Sequencing platforms

- **Short read technologies** (50 - 300 bases)
  - Illumina (MiSeq, HiSeq etc.)
  - 454
  - Ion Torrent

- **Long read technologies**
  - Pacific Biosciences (PacBio) (~20 kb)
  - Oxford Nanopore Technologies (MinION) (up to 200 kb)

The Fleming Fund
Regional Grants

Sequencing platforms II



Some characteristics to consider:

Throughput/capacity

Run time

Instrumentation and service costs

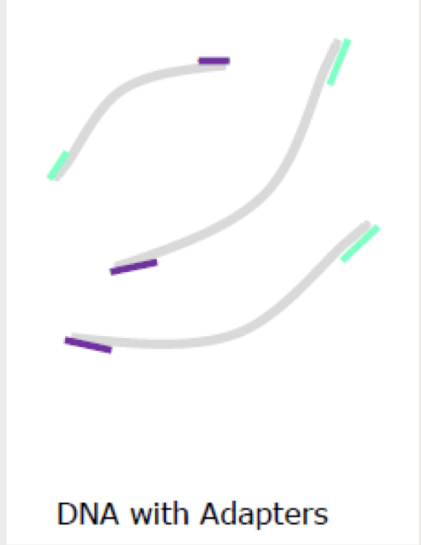
Reagent costs and availability

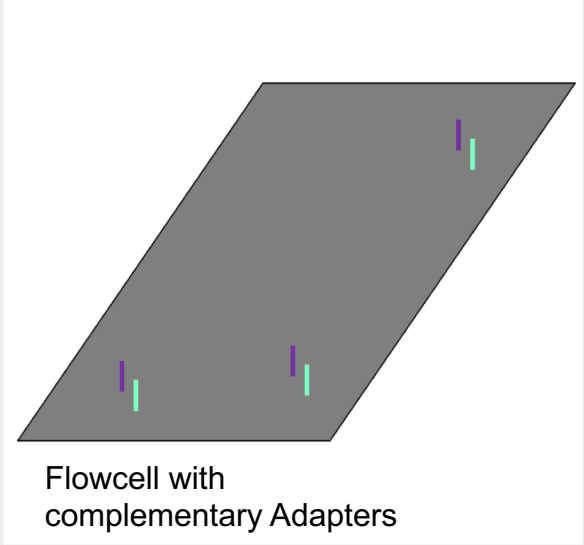
Infrastructure

Training of staff

Short read sequencers (Illumina)

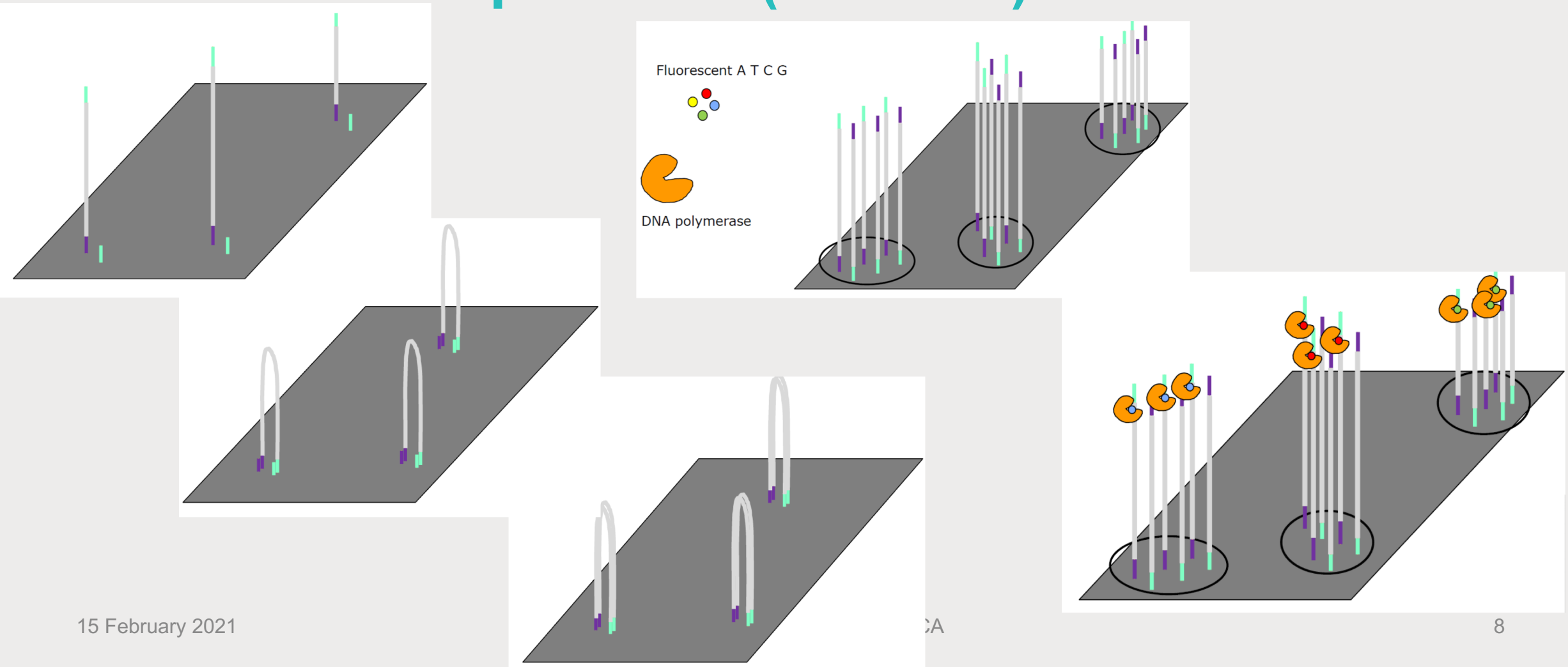
- How does Illumina sequencing work?
  - Sequencing by synthesis





The Fleming Fund
Regional Grants

Short read sequencers (Illumina)



Different Illumina's

- HiSeq, NextSeq, MiSeq, …
- General chemistry is the same
- HiSeq gives more reads, takes more time and costs more
- MiSeq is faster, cheaper but gives less reads
- NextSeq uses two-dye system (faster cycle times and less expensive platform (than HiSeq))

Comparison of short reads technologies

Illumina
- Good accuracy
- Error rate ~0.1%
- Some underrepresentation in AT and GC rich regions
- High throughput

454
- Longer reads (up to 700 bp.)
- Longer insert size
- High cost per mega base
- Error rate ~1%
- Issues with homopolymers
- Discontinued

Ion Torrent
- Fastest runtime and work-flow
- More hands on time
- Error rate ~1%
- Issues with homopolymers

Third Generation Sequencing

- Long reads
- Single-molecule
- Real-time
- Platforms:
  - Pacific Biosciences (PacBio)
  - Oxford Nanopore Technologies (MinIon)

Comparison

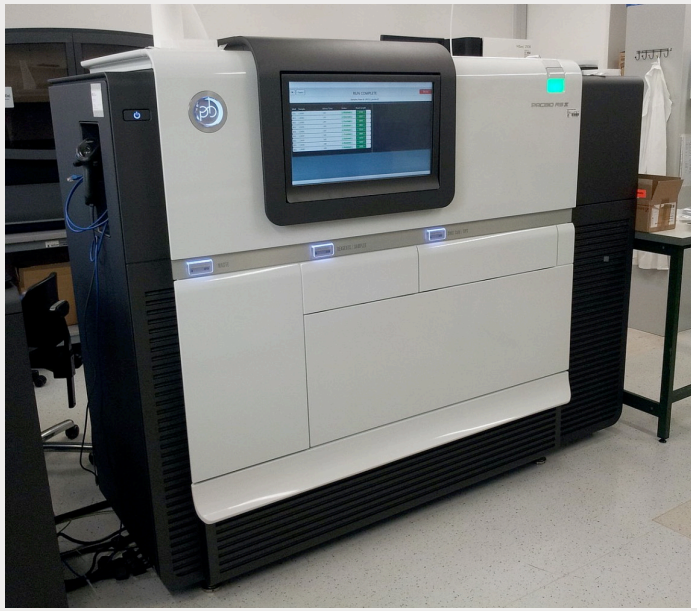
- Illumina
  - Short reads
  - High accuracy

Price:

- MiSeq: low instrument cost, higher cost per Gb data
- NextSeq: medium high instrument cost, lower cost per Gb
- HiSeq: high instrument cost, low cost per Gb data

Comparison (II)

- PacBio
- Long reads
- High error rates
- Price:
  - medium high instrument cost
  - very high cost per Gb data



The Fleming Fund
Regional Grants

Comparison (III)

- Nanopore
  - Very long reads
  - High error rates
  - Portable
- Price:
  - Very low instrument cost
  - High cost per Gb data

MinIon

GridIon

Comparison of long reads technologies

Pacific Biosciences

- Long reads. (Max: 50 kb. Avg.: 10-15 kb)
- Error rate ~15% (single pass)
- Precision can be improved to 99.999%
- Low throughput
- Expensive (acquisition cost)

Oxford Nanopore

- Very long reads (up to 900 kb.)
- Large error rates (3-8%)
- Portable

Still some limitations in sequence processing by tools available

The Fleming Fund
Regional Grants

Sequencing output

What is fastq?
Fasta + quality scores

Raw data – fastq files 1 read: 4 lines

```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACNGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCG
+
 BP`ccceggcegihiiighiifhihfddgfhi^efgfhhhhhegiiiiiiiiihiihihggeeccdddcccacWTT^acc[ab `]`[ b`^BBBBBBB
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCG
+
bb_eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[_X]]a[aacX
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGA
+
bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh_fhggdgegeaceeacbdcbcc\^aa]``_^bb]bcccccbac_a^b
```

The Fleming Fund
Regional Grants

Sequencing output

What is fastq?
Fasta + quality scores

Raw data – fastq files 1 read: 4 lines

```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACNGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCG
+
_BP`cccegggegihiiighiifhihfddgfhi^efgfhhhhhegiiiiiiiihiihihggeeccdddcccacWTT^acc[ab_`]`[_b`^BBBBBBB
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCG
+
bb_eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[_X]]a[aacX
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGA
+
bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh_fhggdgegeaceeacbdcbcc\^aa]``_^bb]bcccccbac_a^b
```

Header/ID

The Fleming Fund
Regional Grants

Sequencing output

What is fastq?
Fasta + quality scores

Raw data – fastq files 1 read: 4 lines

DNA sequence

```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACNGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCG
+
_BP`ccceggcegihiiighiifhihfddgfhi^efgfhhhhhegiiiiiiiihiihihggeeccdddcccacWTT^acc[ab_`]`[_b`^BBBBBBB
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCG
+
bb_eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[_X]]a[aacX
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGA
+
bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh_fhggdgegeaceeacbdcbcc\^aa]``_^bb]bcccccbac_a^b
```

The Fleming Fund
Regional Grants

Sequencing output

What is fastq?
Fasta + quality scores

Raw data – fastq files 1 read: 4 lines

Name field (optional)

```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACNGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCG
+
_BP`ccceggcegihiiighiifhihfddgfhi^efgfhhhhhegiiiiiiiihiihihggeeccdddcccacWTT^acc[ab_`]`[_b`^BBBBBBB
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCG
+
bb_eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[_X]]a[aacX
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGA
+
bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh_fhggdgegeaceeacbdcbcc\^aa]``_^bb]bcccccbac_a^b
```

The Fleming Fund
Regional Grants

Sequencing output

What is fastq?
Fasta + quality scores

Raw data – fastq files 1 read: 4 lines

Quality scores

```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACNGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCG
+
_BP`ccceggcegihiiighiifhihfddgfhi^efgfhhhhhegiiiiiiiihiihihggeeccdddcccacWTT^acc[ab_`]`[_b`^BBBBBBB
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCG
+
bb_eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[_X]]a[aacX
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGA
+
bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh_fhggdgegeaceeacbdcbcc\^aa]``_^bb]bcccccbac_a^b
```

The Fleming Fund
Regional Grants

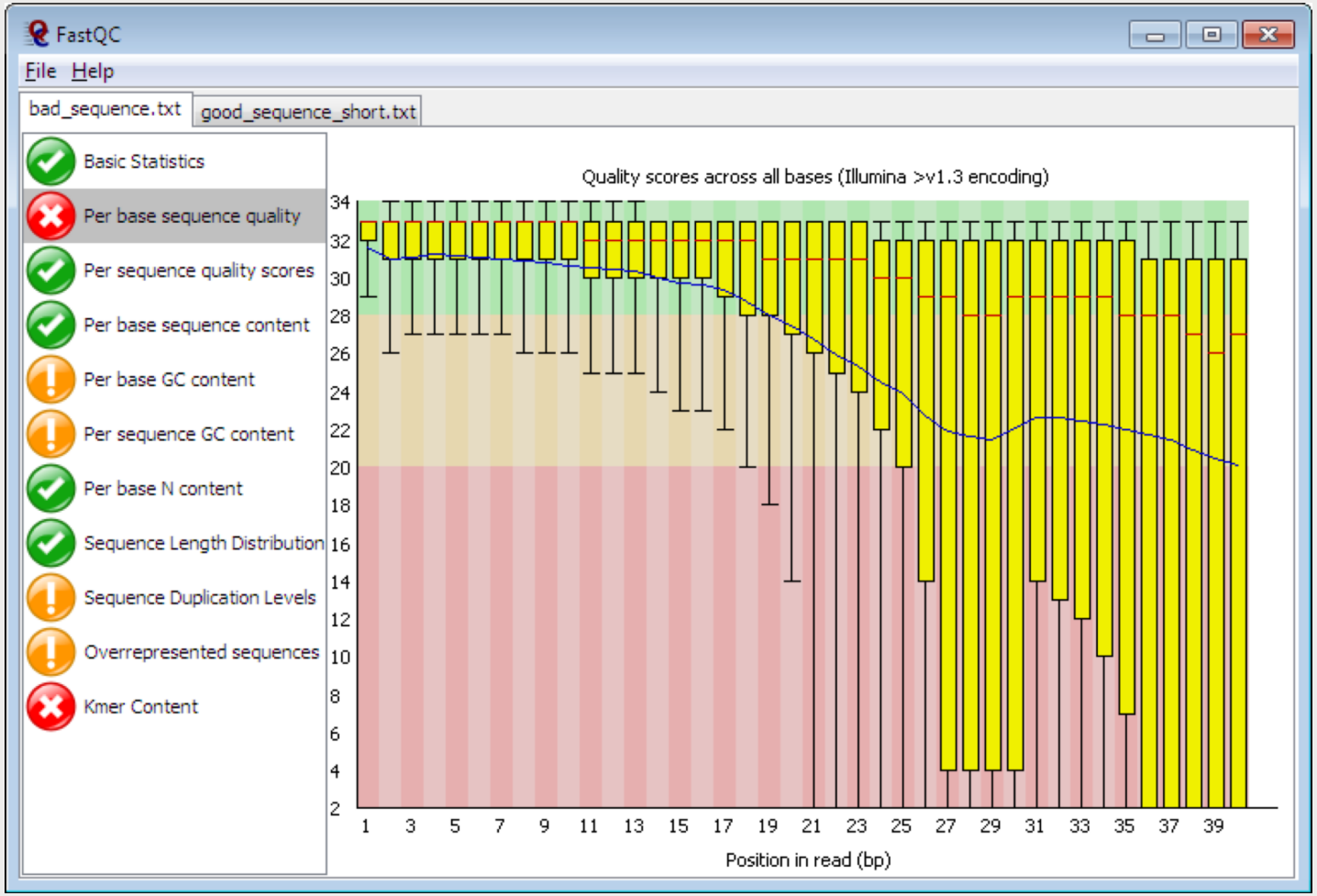
Post-sequencing steps

- Quality control
  - Trimming of adaptors and low quality reads
  - Error correction
- Assembly
- Validation
- Data analysis

The Fleming Fund
Regional Grants

Quality control (QC)

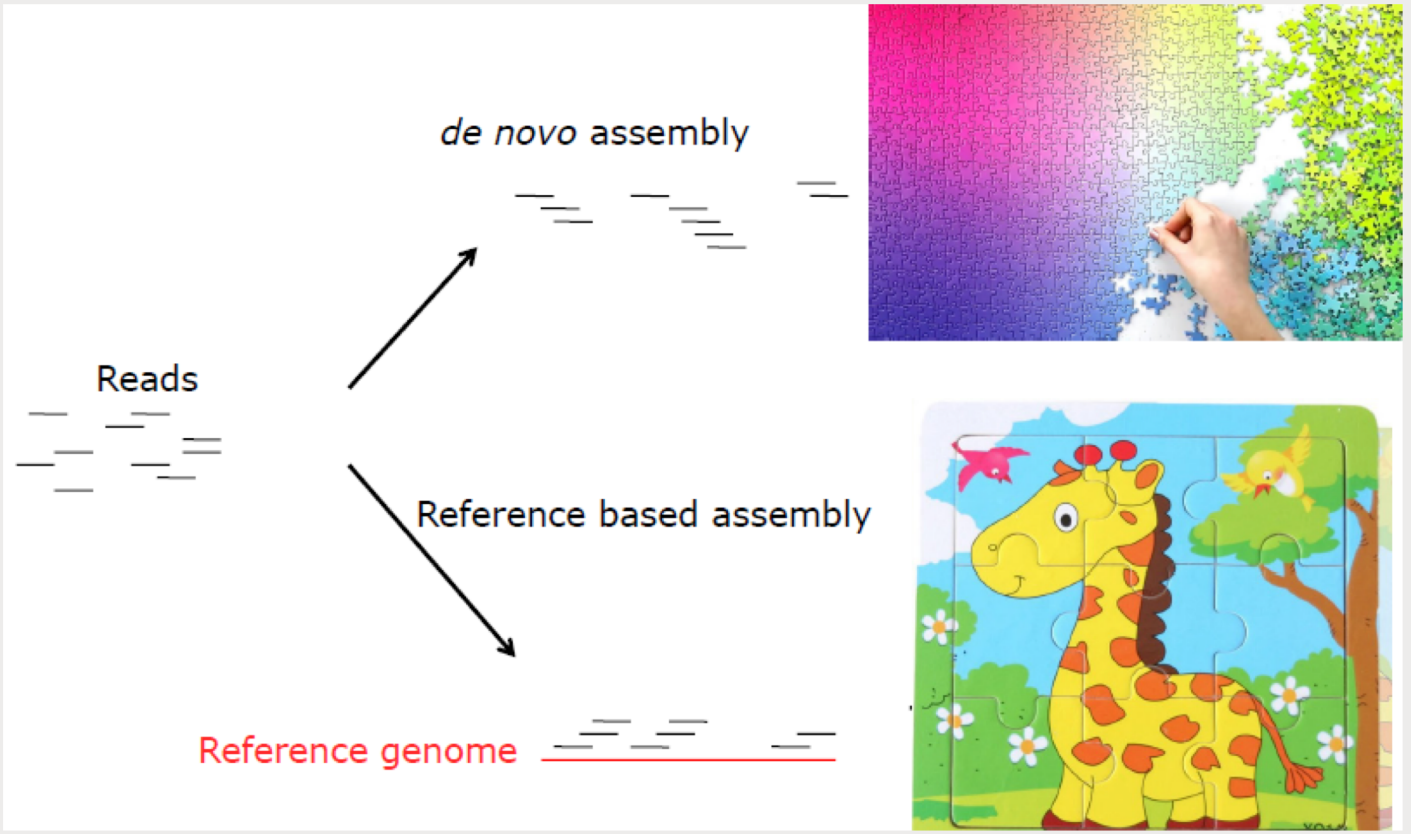
- Different tools available
- QC on raw reads
  - % of quality bases
  - # of reads
- QC on assembly
  - QC parameters
    - Genome size
    - # of contigs
    - N50
    - Coverage



The Fleming Fund
Regional Grants

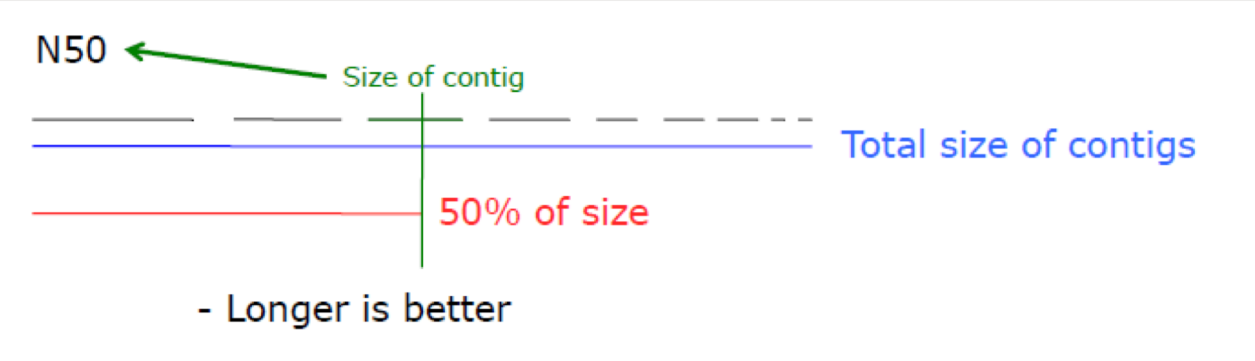
What is assembly?

- Assemble raw reads into larger stretches of DNA: contigs
  - Mapping to reference
  - *De novo assembly*



QC parameters- guidelines

- Number of contigs
  - < 500 contigs (the fewer, the better)
- Assembled genome size (Total size of contigs)
  - Close to expected size (+/- 20 %)
- N50 – higher is better (> 30.000)



QC parameters - guidelines

- Coverage
  - The number of times the genome is covered by the read data
  - Preferably > 20 X

$$C = N \cdot \frac{L}{G}$$

- N: Number of read
- L: Read length
- G: Genome size (target **or** assembly)

Example:
N = 5 mill
L = 100 bp
G = 5 Mbp

C = 5*100/5 = 100X

On average, 100 reads covers each position in the genome.

- Depth
  - The number of reads that covers a particular nucleotide in each position in the genome

$$\frac{\text{reads}}{\text{site}} = \text{depth}$$

The sequence equipment depend on your needs

- High variety in equipment
  - Throughput
  - Availability of reagents
  - Convenience in sequence analysis

Quality control is important!

- Ensure validity of your sequences
- Ensure adequate quality for further genome analysis
- Ensure comparability of results

QC
Genome
analysis

The Fleming Fund
Regional Grants

Thank you

NOGUCHI MEMORIAL INSTITUTE FOR MEDICAL RESEARCH
UNIVERSITY OF GHANA , LEGON

UNIVERSITY OF IBADAN
RECTE SAPERE FONS

NATIONAL INSTITUTE FOR COMMUNICABLE DISEASES
Division of the National Health Laboratory Service

KCRI Kilimanjaro Clinical Research Institute
Academic Centre for Evidence Based Health Interventions
An Institution of the Good Samaritan Foundation

UKaid
from the British people

MOTT MACDONALD

SEQAFRICA

DTU

This programme is being funded by the UK Department of Health and Social Care.
The views expressed do not necessarily reflect the UK Government's official policies.