EURGen-RefLabCap

Technical training workshop # 1

First day (virtual)

Tuesday, 29 November 2022

10:00 - 12:30 CET

Funded by the European Union

Virtual Housekeeping

Please **turn off your cameras and microphones** unless you're speaking – this will help with bandwidth and maximise audibility.

Do frequently **use the chat function** to share your views, comments and challenges. Keep the chat constructive, respectful and on topic!

If you wish to make a comment for e.g. the discussion, please use the **'Raise hand'** function.

Agenda

First day (virtual) – Tuesday 29 November 2022, 10:00 - 12:30 CET

10:00 - 10:15: Introduction and agenda for the day (Ana Rita Rebelo, DTU)

10:15 - 11:00: From isolate to WGS - biochemical principles (Ana Rita Rebelo, DTU)

11:00 - 11:15: Coffee break

11:15 - 12:00: From isolate to WGS - WGS data and bioinformatics (Jette Sejer Kjeldgaard, DTU)

12:00 - 12:30: Quality control of WGS data (Ana Rita Rebelo, DTU)

Funded by the European Union

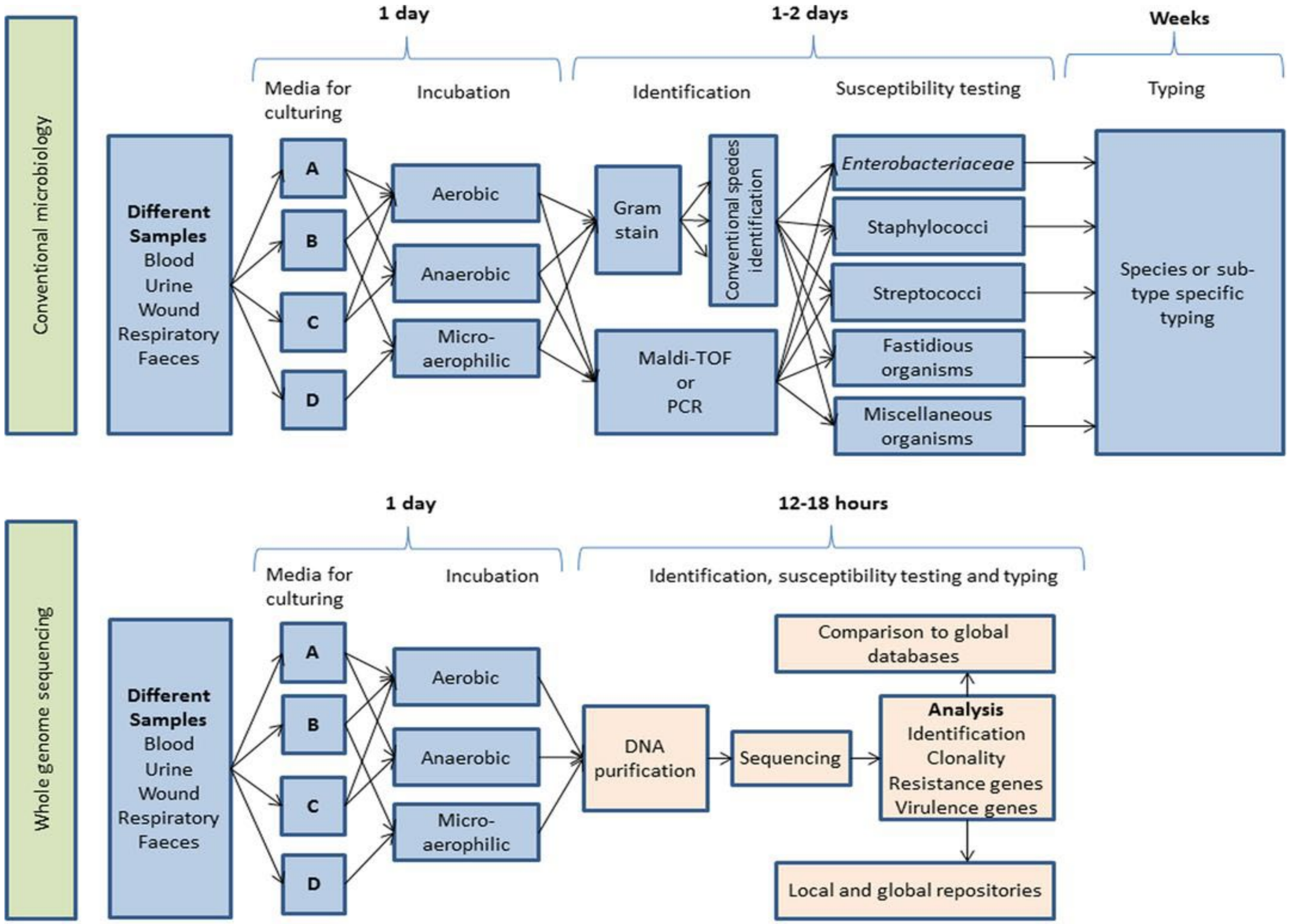
Ana Rita Rebelo

anrire@food.dtu.dk

From isolate to WGS – biochemical principles

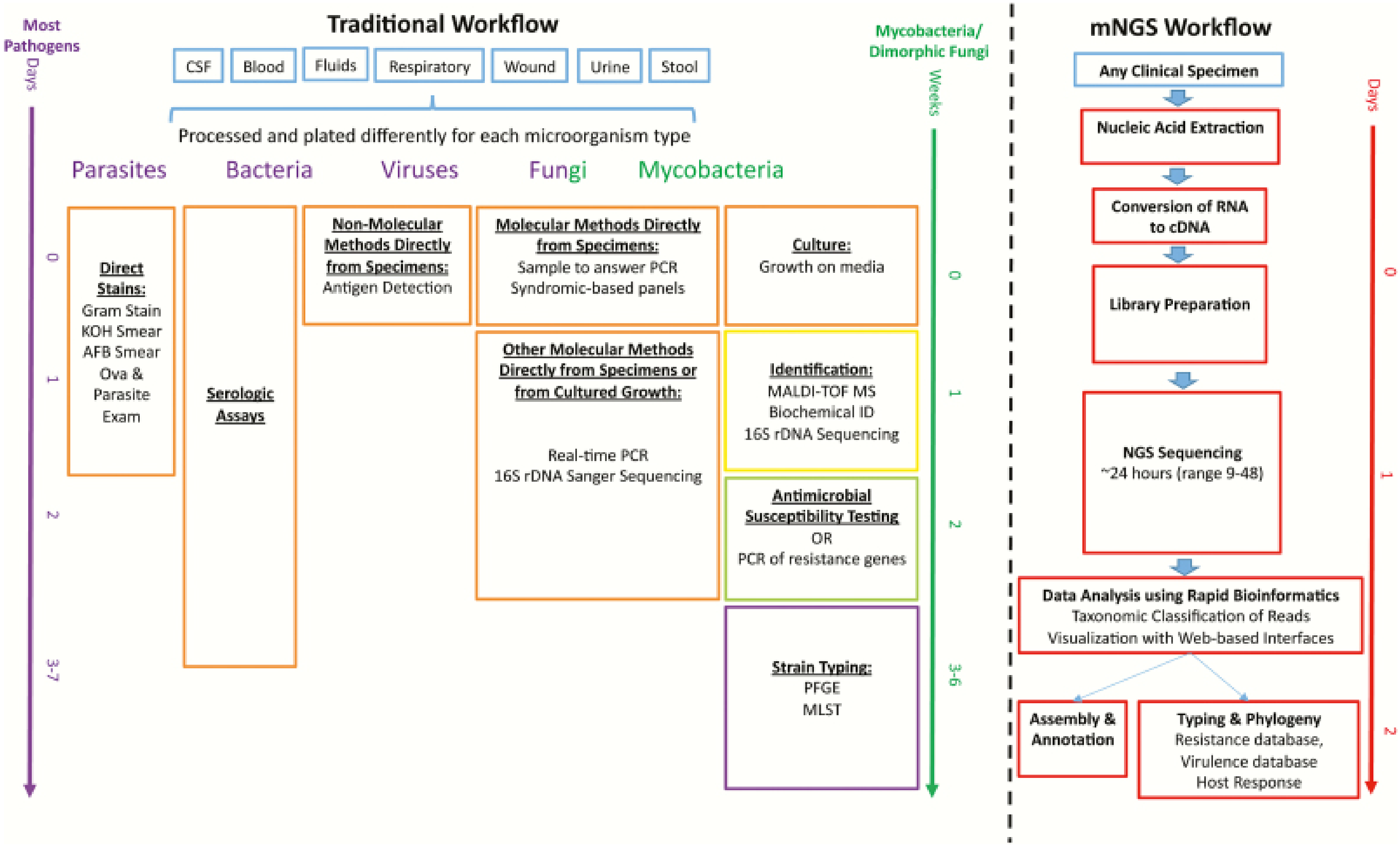
Funded by the European Union

WGS vs. classical methods



Hasman *et al*. 2013 (adapted)

WGS vs. classical methods



Traditional Workflow

CSF | Blood | Fluids | Respiratory | Wound | Urine | Stool

Processed and plated differently for each microorganism type

Parasites | Bacteria | Viruses | Fungi | Mycobacteria

Direct Stains: Gram Stain, KOH Smear, AFB Smear, Ova & Parasite Exam

Serologic Assays

Non-Molecular Methods Directly from Specimens: Antigen Detection

Molecular Methods Directly from Specimens: Sample to answer PCR, Syndromic-based panels

Other Molecular Methods Directly from Specimens or from Cultured Growth: Real-time PCR, 16S rDNA Sanger Sequencing

Culture: Growth on media

Identification: MALDI-TOF MS, Biochemical ID, 16S rDNA Sequencing

Antimicrobial Susceptibility Testing OR PCR of resistance genes

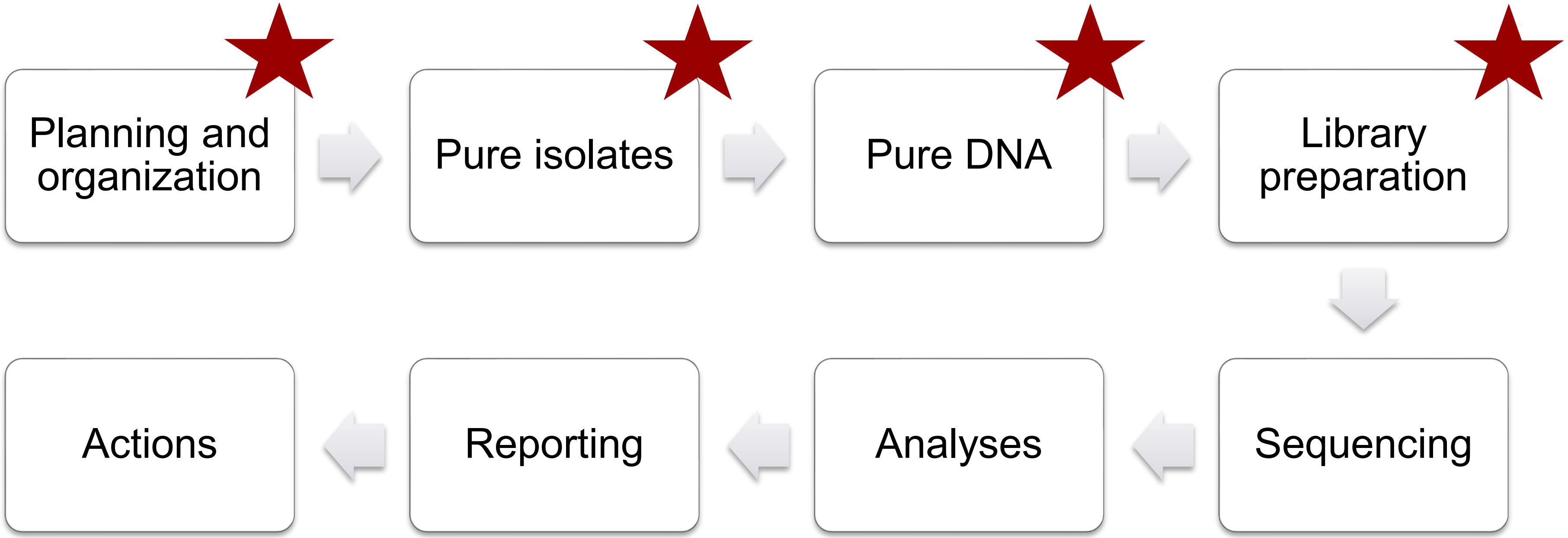
Strain Typing: PFGE, MLST

mNGS Workflow

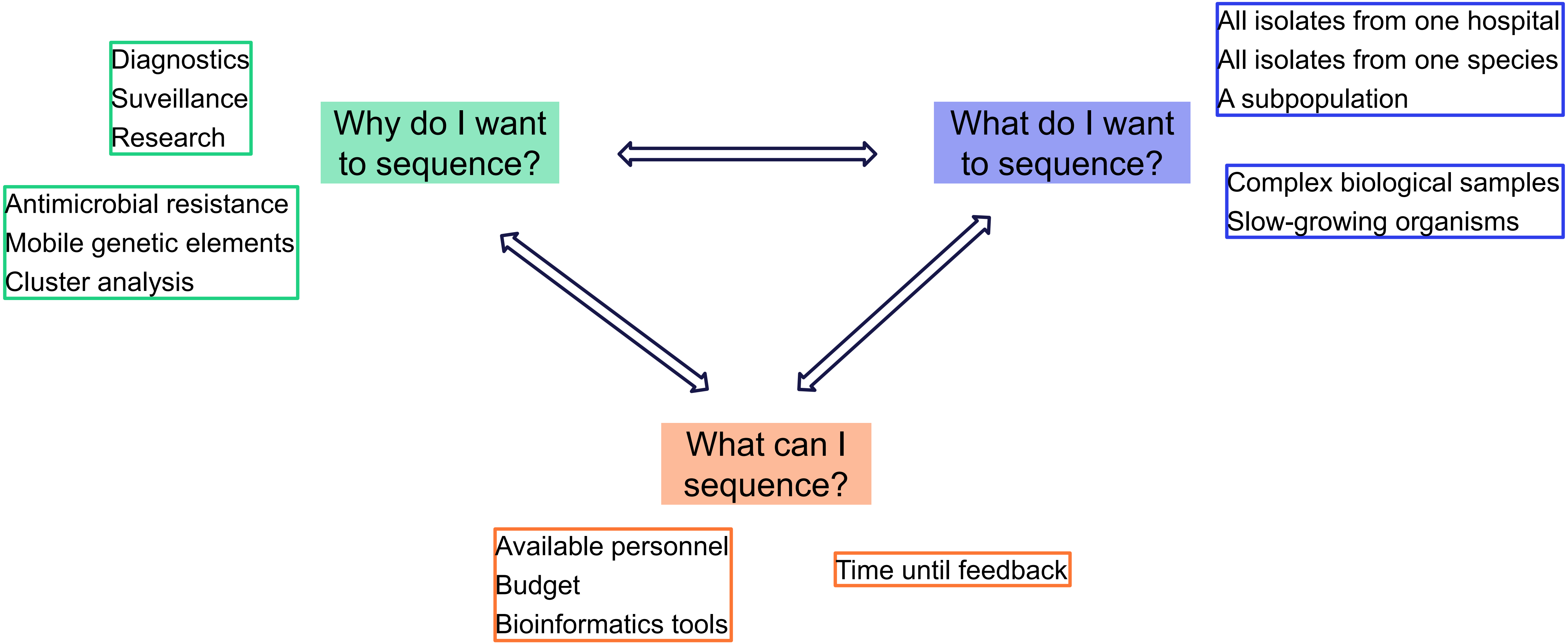
Any Clinical Specimen → Nucleic Acid Extraction → Conversion of RNA to cDNA → Library Preparation → NGS Sequencing ~24 hours (range 9-48) → Data Analysis using Rapid Bioinformatics: Taxonomic Classification of Reads, Visualization with Web-based Interfaces → Assembly & Annotation; Typing & Phylogeny: Resistance database, Virulence database, Host Response

Simner *et al*. 2018

The complete workflow



Decisions…



Why do I want to sequence?

Diagnostics
Suveillance
Research

Antimicrobial resistance
Mobile genetic elements
Cluster analysis

What do I want to sequence?

All isolates from one hospital
All isolates from one species
A subpopulation

Complex biological samples
Slow-growing organisms

What can I sequence?

Available personnel
Budget
Bioinformatics tools

Time until feedback

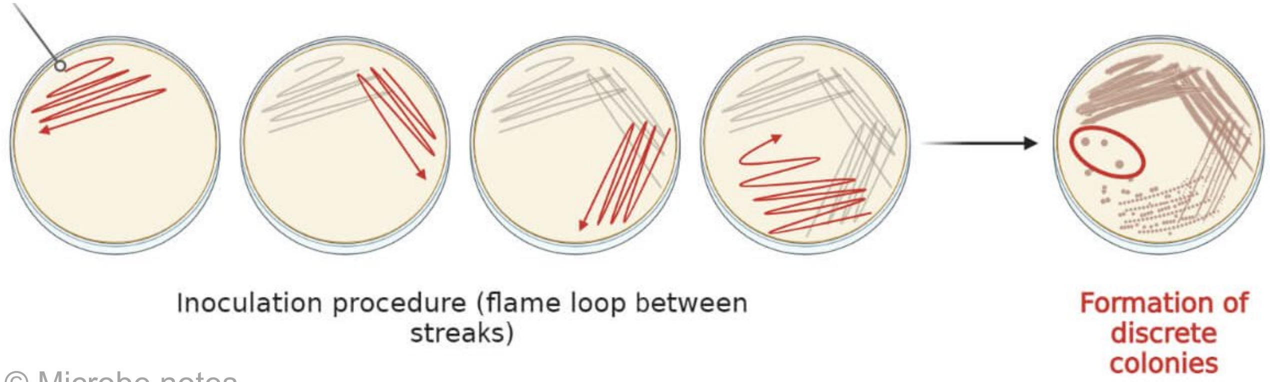
Biological samples and bacterial isolates

Cultures vs. complex samples

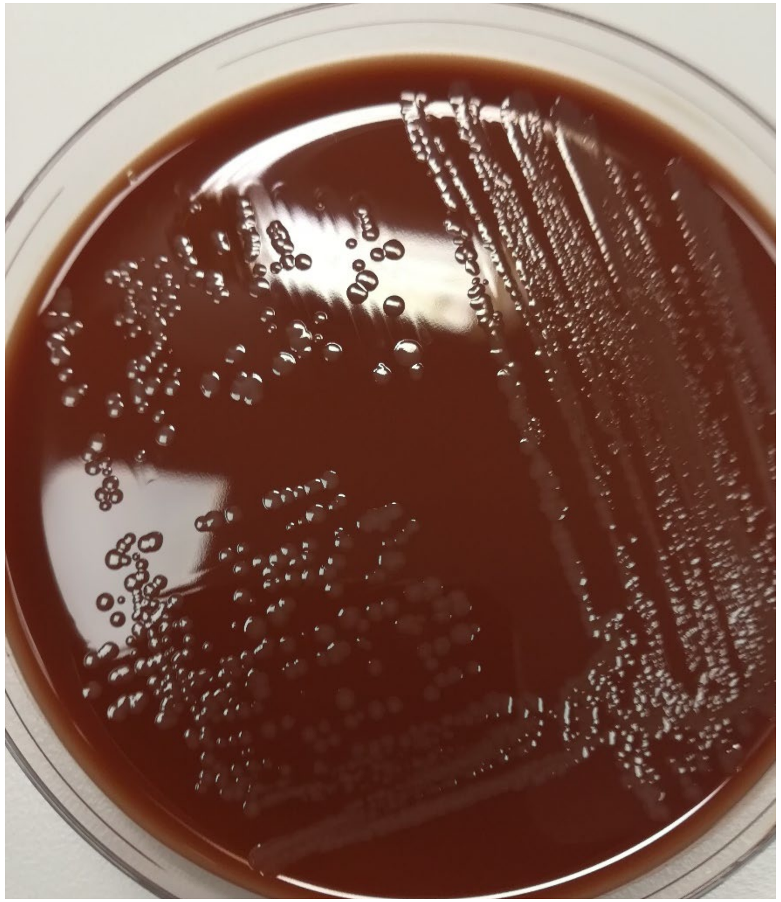
Selecting the adequate isolation methods

Selecting the correct isolates

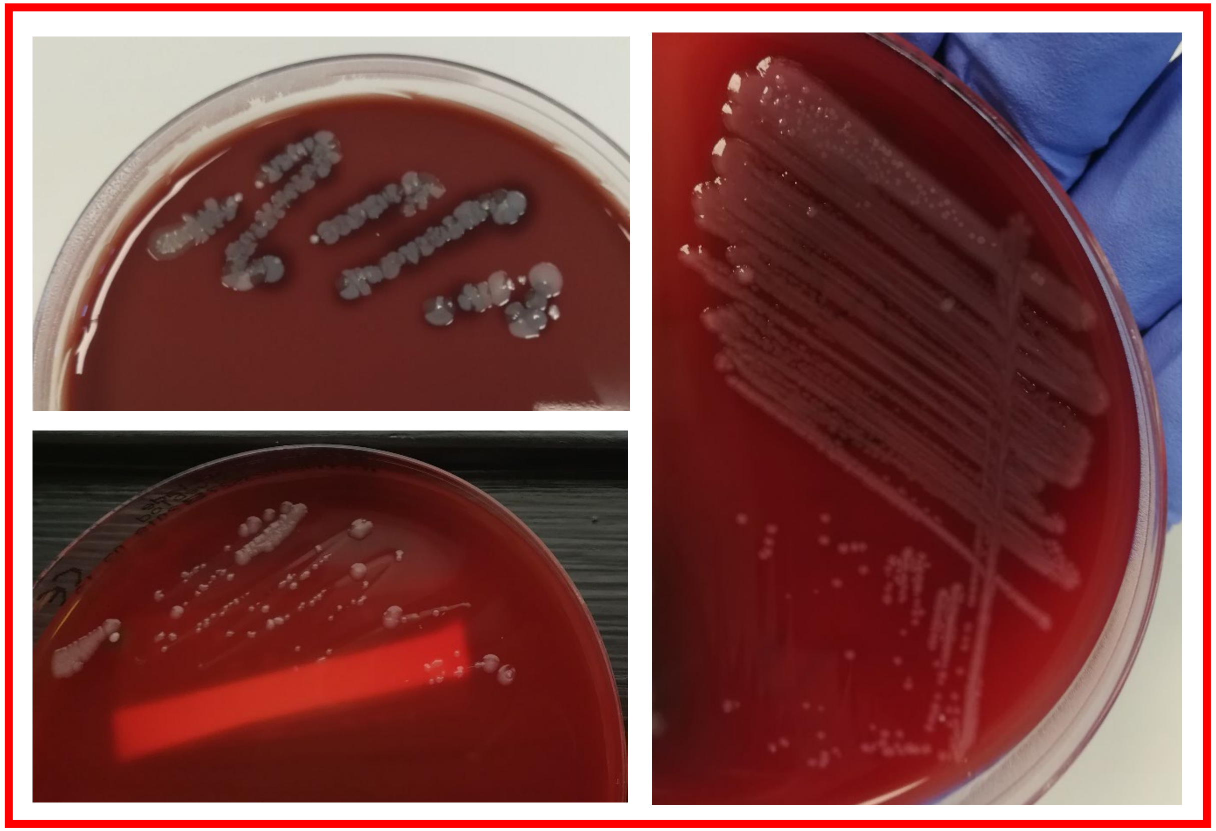
Bacterial cultures

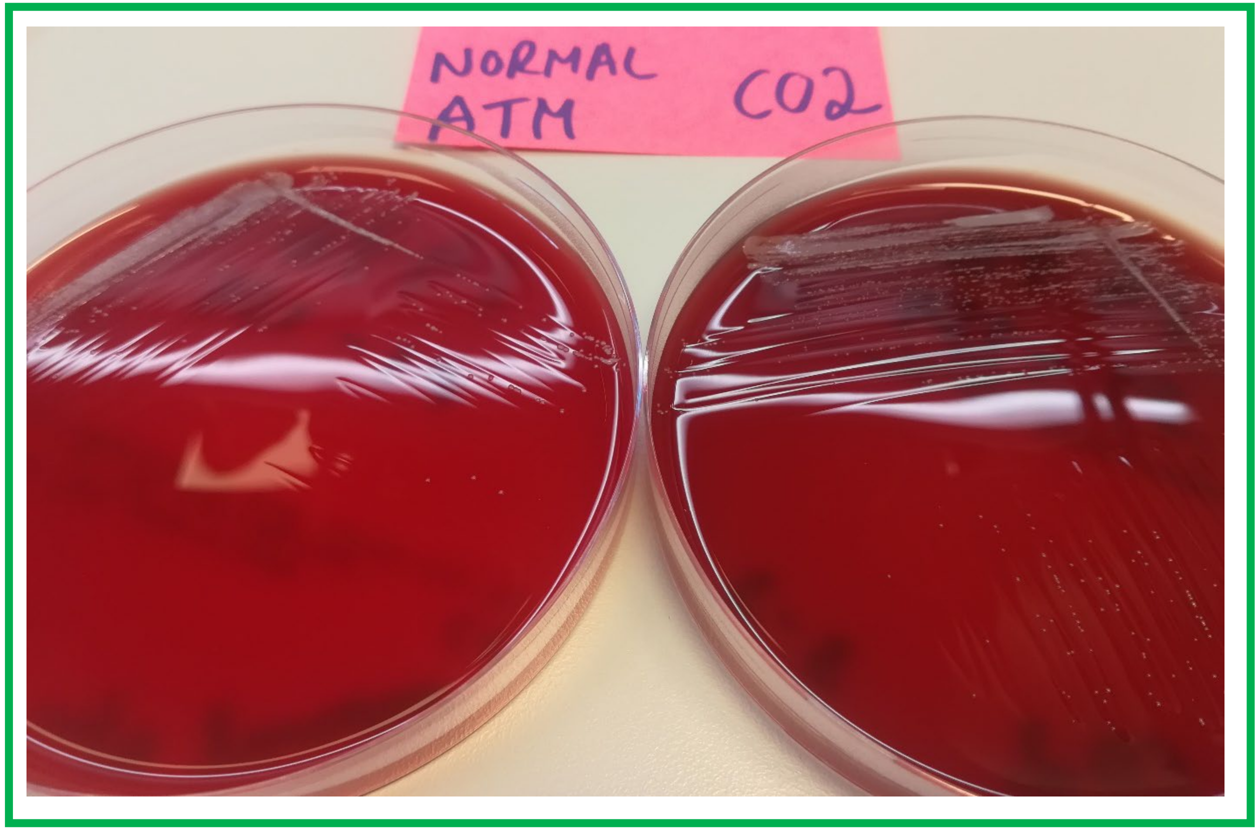

Inoculation procedure (flame loop between streaks)
Formation of discrete colonies

© Microbe notes



Bacterial cultures





DNA extraction

In house protocols or commercial kits

Cell lysis

-- Cell burst with release of intracellular components

-- Enzymes, temperature, mechanical lysis, detergents, etc.

Precipitation

-- Separation of the DNA and debris

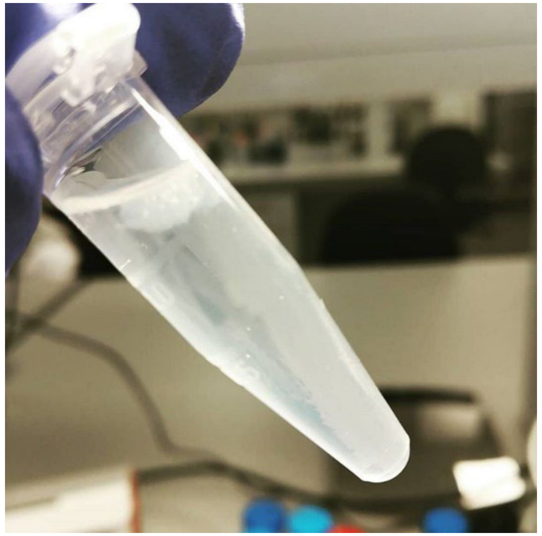
-- Organic solvents/alcohols and salts

Clean up

-- Recovery of DNA and removal of remaining salts and reagents

-- Organic solvents/alcohols

Resuspension



DNA dilution and quality control

UV 260/280 absorbance ratio values of the DNA samples should be in the interval 1.8 – 2.0

Bioanalyzer

(Agilent, Santa Clara, CA, USA)

© BioLabTech

Nanodrop spectrophotometers

(Thermo Scientific, Waltham, MA, USA)

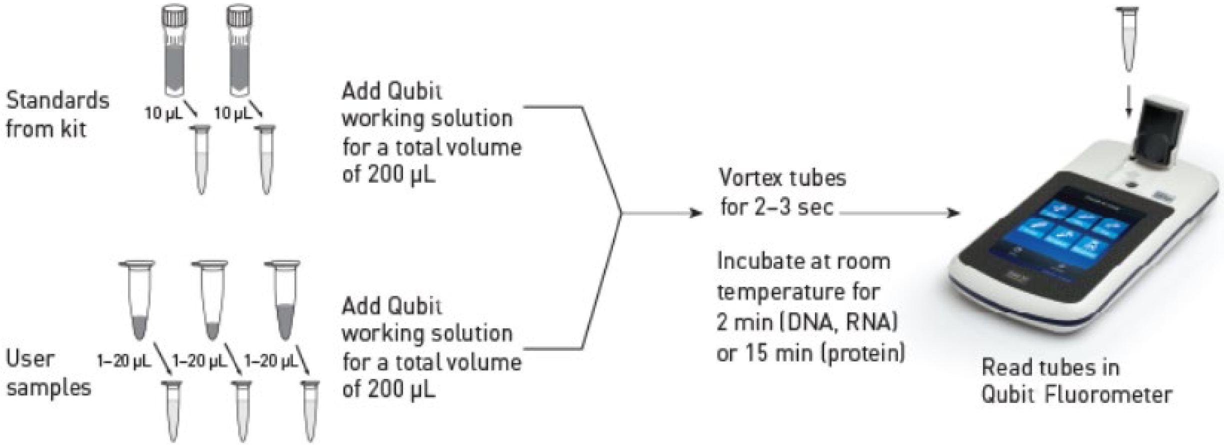
© Fischer Scientific

DNA dilution and quality control

Dilution of the extracted DNA

Qubit fluorometer
(Invitrogen, Carlsbad, CA, USA)

Accepted range: 0,18 – 0,28 ng/µl



Standards from kit

10 µL 10 µL

Add Qubit working solution for a total volume of 200 µL

User samples

1–20 µL 1–20 µL 1–20 µL

Add Qubit working solution for a total volume of 200 µL

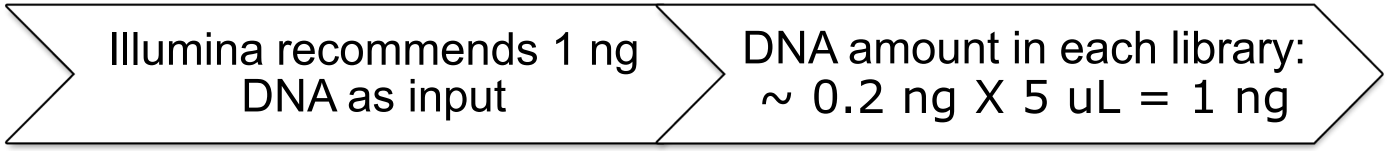
Vortex tubes for 2–3 sec

Incubate at room temperature for 2 min (DNA, RNA) or 15 min (protein)

Read tubes in Qubit Fluorometer

Qubit 4 Fluorometer, Thermo Fhisher Scientific, 2018

DNA dilution and quality control



Illumina recommends 1 ng DNA as input

DNA amount in each library: ~ 0.2 ng X 5 uL = 1 ng

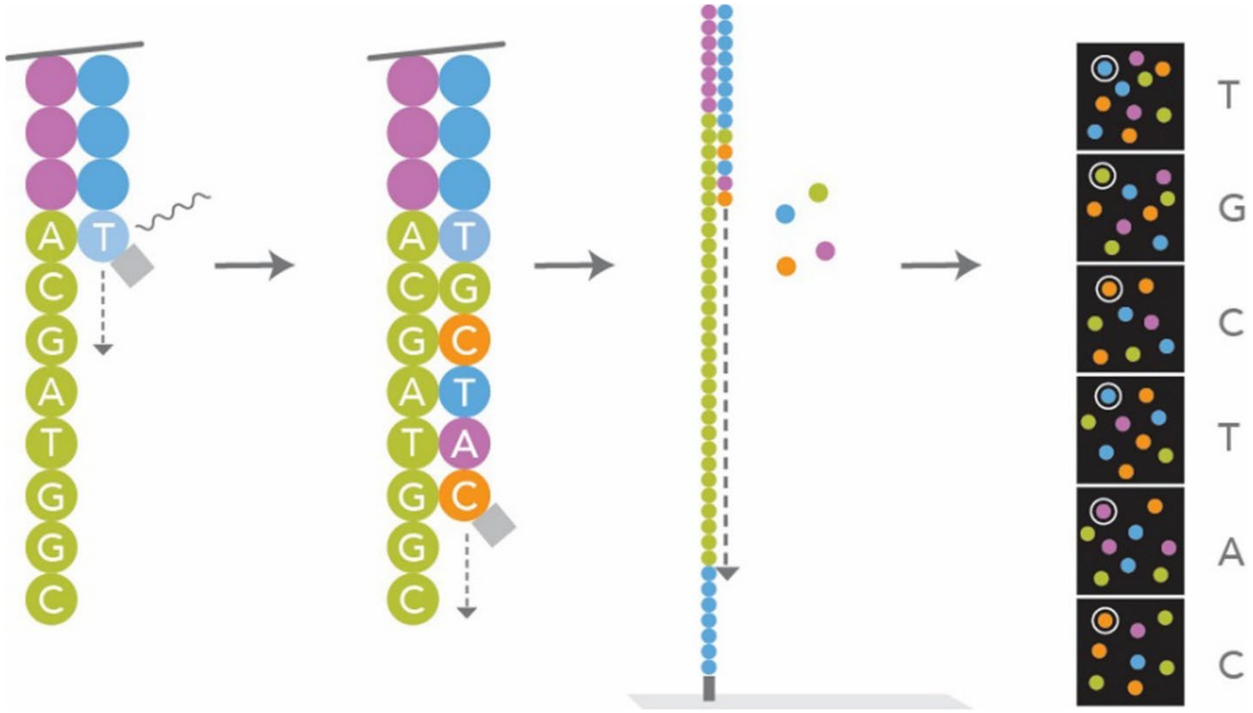
Dependant on species, extraction method, operator, …

DILUTION

Sample (examples)	DNA concentration before dilution (ng/uL)	DNA concentration after dilution (ng/uL)
Isolate A	46,5	0,277
Isolate B	103	0,258

Sequencing methods

Sequencing by synthesis, sequencing by ligation, the chain termination method, pyrosequencing, …



EFSA journal 2018;16(S1):e16086

Illumina platforms

(Illumina, Inc., San Diego, CA, USA)

Some concepts:

Short reads

The DNA fragments obtained by this process are shorter when compared to newer technologies

Sequencing by synthesis

The sequencing process works by using the DNA being analyzed – in the form of ssDNA - as a template to synthesize a complementary DNA strand with fluorescent nucleotides, which are then detected by the machine

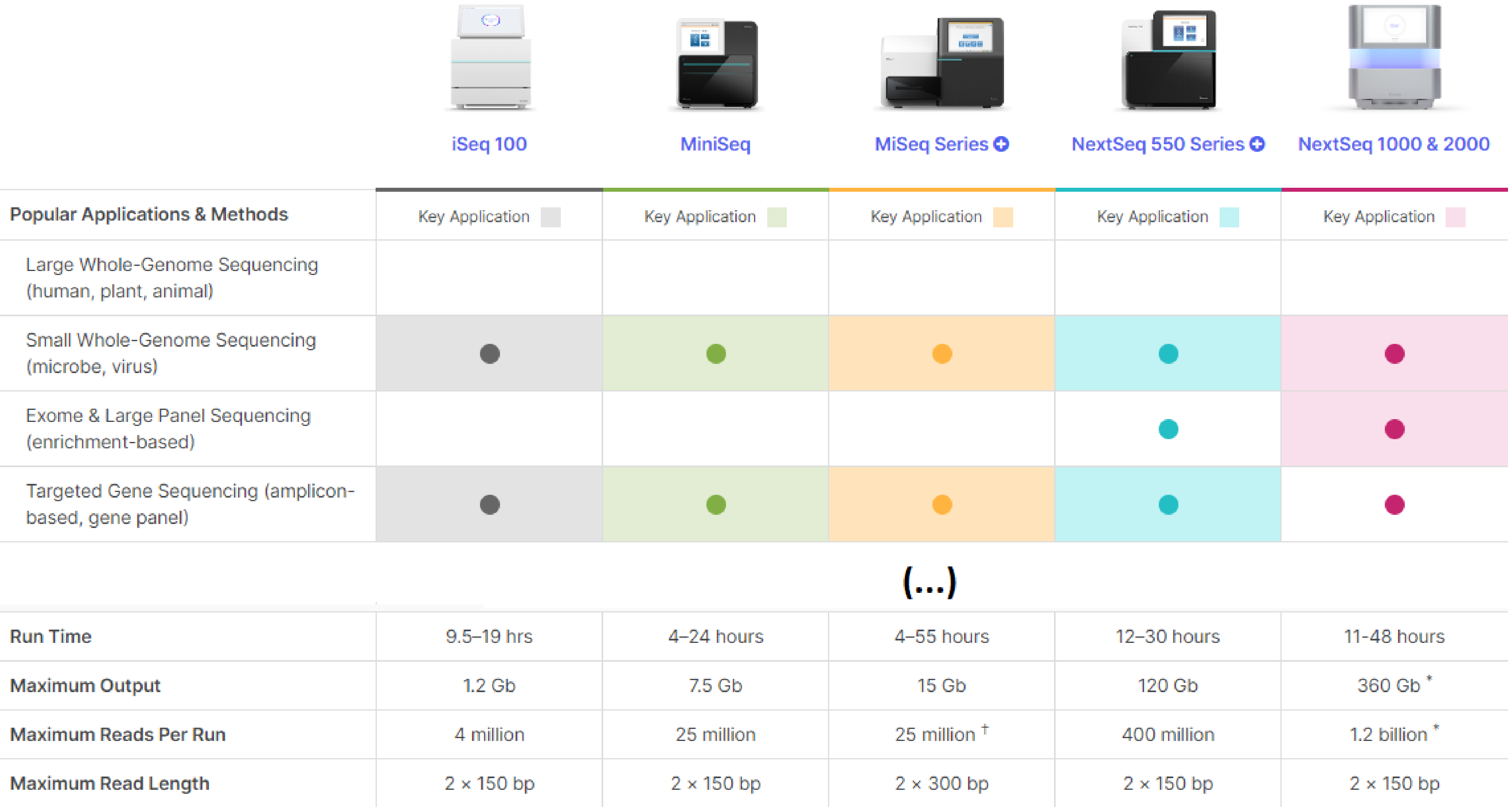
Paired end

The complementary DNA is synthesized from both ends to ensure accuracy

https://www.illumina.com/systems/sequencing-platforms.html

Illumina platforms

(Illumina, Inc., San Diego, CA, USA)



	iSeq 100	MiniSeq	MiSeq Series ⊕	NextSeq 550 Series ⊕	NextSeq 1000 & 2000
Popular Applications & Methods	Key Application ■	Key Application ■	Key Application ■	Key Application ■	Key Application ■
Large Whole-Genome Sequencing (human, plant, animal)					
Small Whole-Genome Sequencing (microbe, virus)	●	●	●	●	●
Exome & Large Panel Sequencing (enrichment-based)				●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●	●	●

(...)

Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	11-48 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	360 Gb *
Maximum Reads Per Run	4 million	25 million	25 million †	400 million	1.2 billion *
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

https://www.illumina.com/systems/sequencing-platforms.html

Illumina MiSeq

© Illumina

Illumina, Inc., 2017

Library preparation

Pre-performed steps

Processes that allow us to obtain pure, single isolates and to extract and dilute their DNA for sequencing.

In here is also included the preparation of the sequencing run in BaseSpace or other platforms

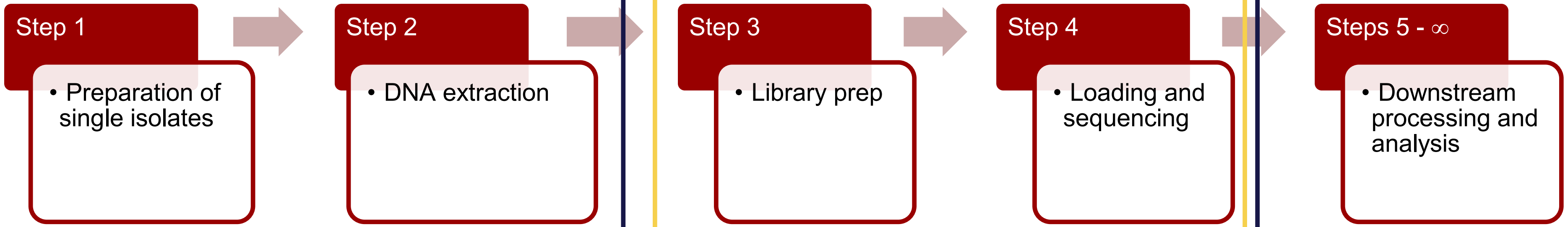
Library preparation

Laboratory procedures that allow for sequencing the extracted DNA.

They can be divided in several steps according to the biochemical processes taking place.

Post-sequencing steps

Processes that allow us control data quality and extract information from the DNA sequences



Step 1
- Preparation of single isolates

Step 2
- DNA extraction

Step 3
- Library prep

Step 4
- Loading and sequencing

Steps 5 - ∞
- Downstream processing and analysis

Tagmentation

Amplification

Clean-up

Normalization

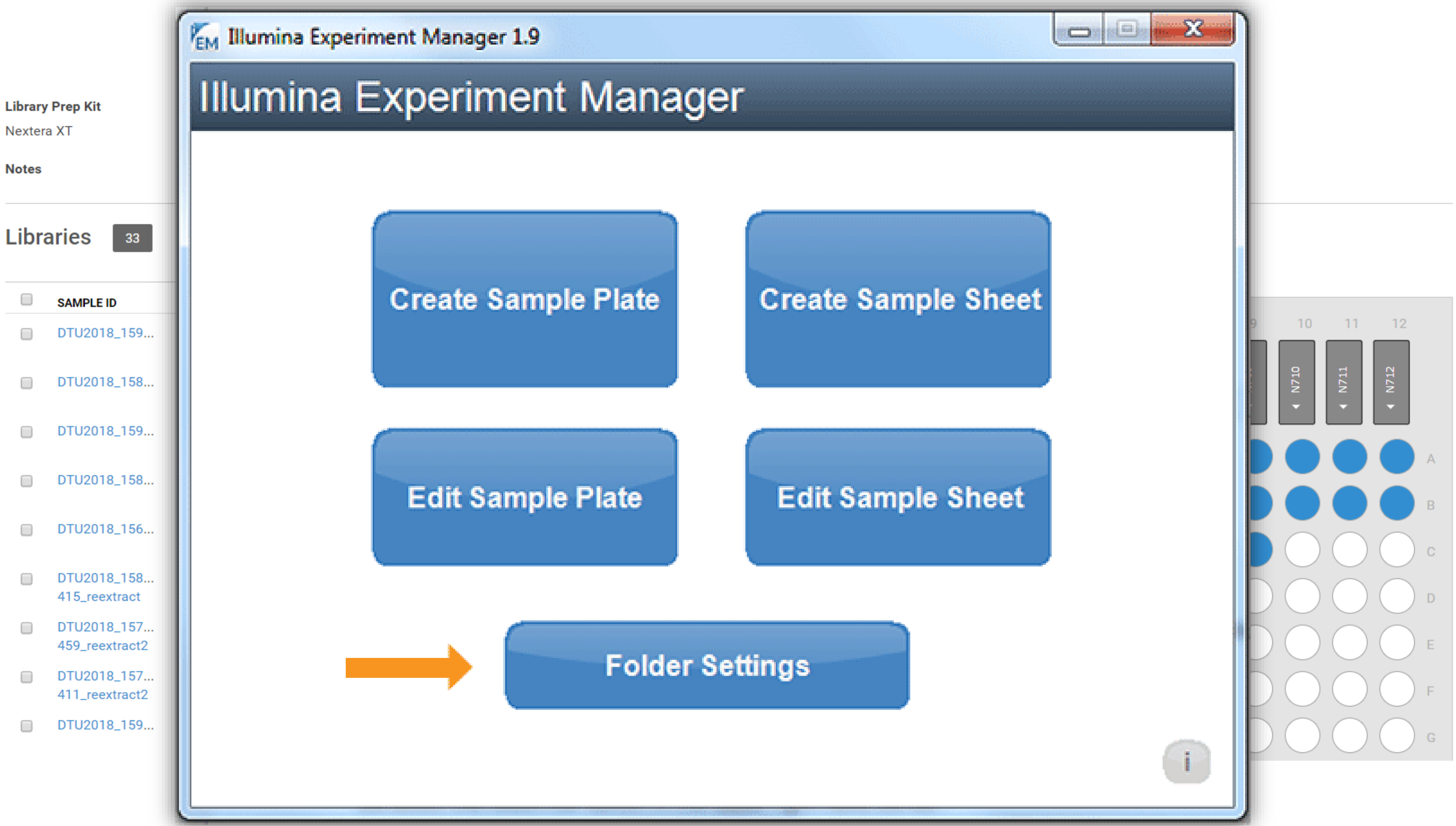
Pooling

Denaturing

Diluting

Preparation of sample sheets

Organization of samples for the run



Library Prep Kit

Nextera XT

Notes

Libraries 33

☐	SAMPLE ID
☐	DTU2018_159...
☐	DTU2018_158...
☐	DTU2018_159...
☐	DTU2018_158...
☐	DTU2018_156...
☐	DTU2018_158... 415_reextract
☐	DTU2018_157... 459_reextract2
☐	DTU2018_157... 411_reextract2
☐	DTU2018_159...

Library preparation – Tagmentation, indexing and amplification

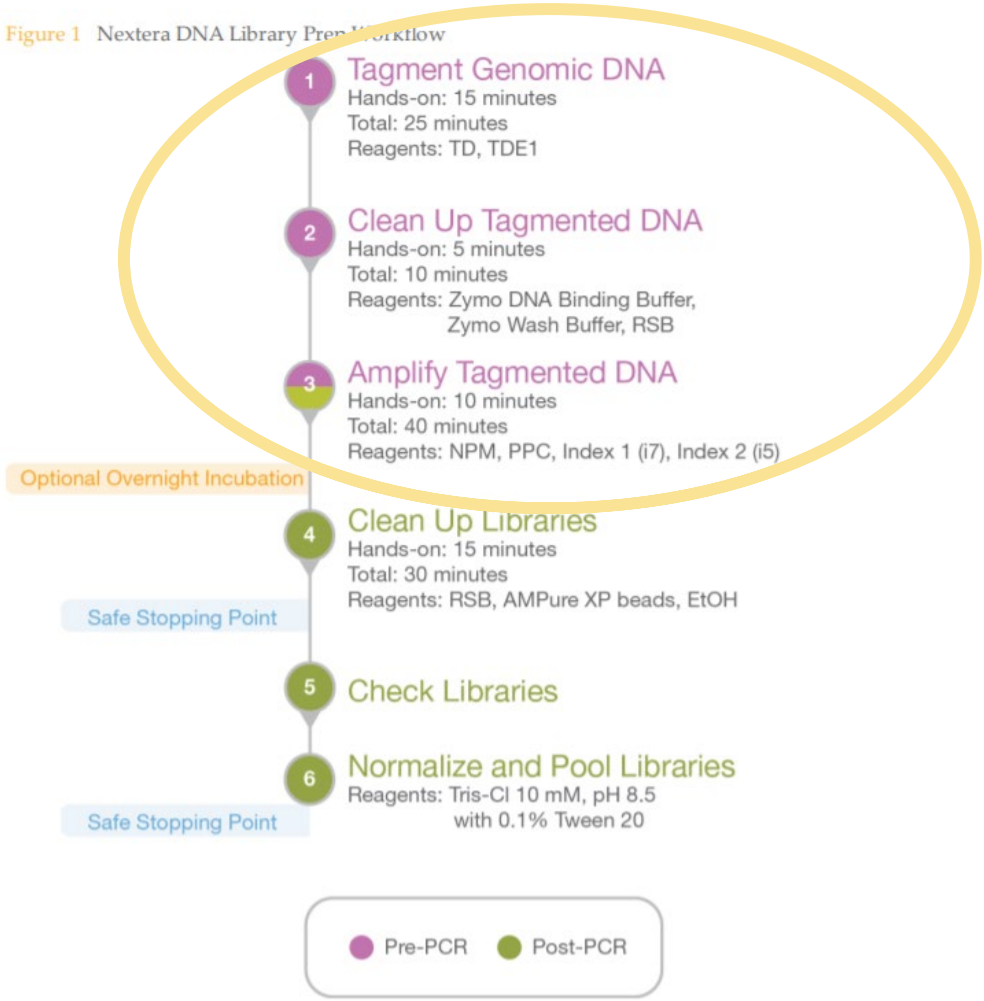


Figure 1 Nextera DNA Library Prep Workflow

1 Tagment Genomic DNA
Hands-on: 15 minutes
Total: 25 minutes
Reagents: TD, TDE1

2 Clean Up Tagmented DNA
Hands-on: 5 minutes
Total: 10 minutes
Reagents: Zymo DNA Binding Buffer,
Zymo Wash Buffer, RSB

3 Amplify Tagmented DNA
Hands-on: 10 minutes
Total: 40 minutes
Reagents: NPM, PPC, Index 1 (i7), Index 2 (i5)

Optional Overnight Incubation

4 Clean Up Libraries
Hands-on: 15 minutes
Total: 30 minutes
Reagents: RSB, AMPure XP beads, EtOH

Safe Stopping Point

5 Check Libraries

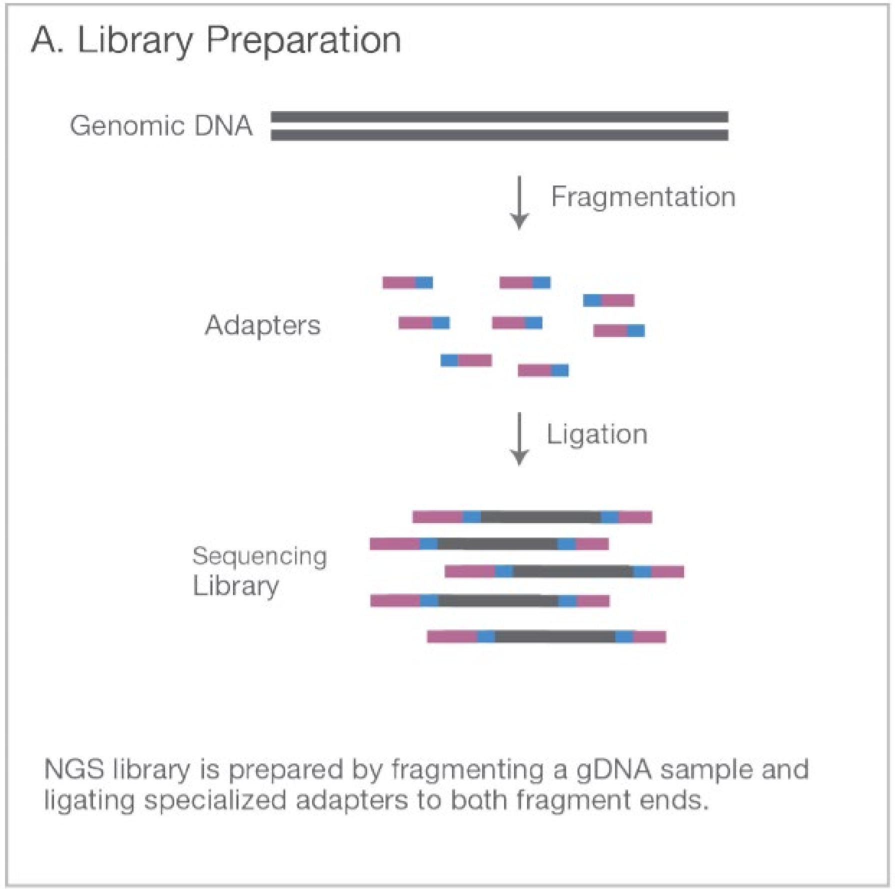
6 Normalize and Pool Libraries
Reagents: Tris-Cl 10 mM, pH 8.5
with 0.1% Tween 20

Safe Stopping Point

● Pre-PCR ● Post-PCR

© Illumina

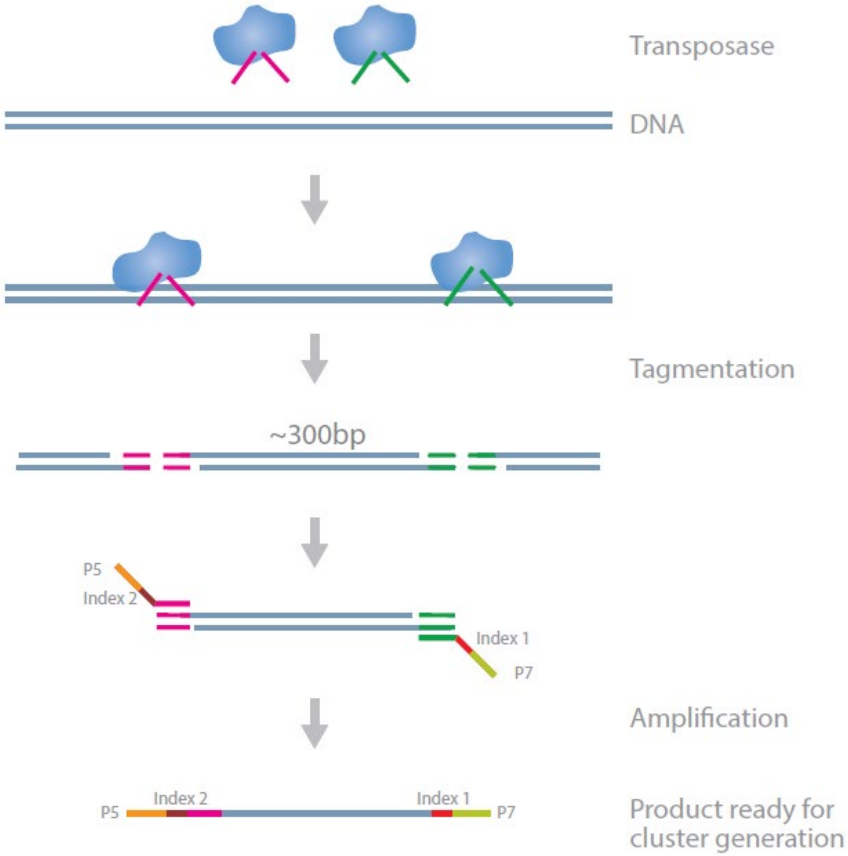
Library preparation – Tagmentation, indexing and amplification



A. Library Preparation

NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

An introduction to Next-Generation Sequencing Technology, Illumina, Inc., 2017



NGS libraries construction, Biofidal, 2018

Library preparation – Tagmentation, indexing and amplification

Video: https://www.youtube.com/watch?v=womKfikWlxM&ab_channel=Illumina

Illumina Sequencing Technology, Illumina, Inc., 2017 (adapted)

Library preparation – Tagmentation, indexing and amplification

How important are these steps?

The most important!

Fragmentation: If DNA is too fragmented it will be lost during clean-up

Tagmentation: Adapters provide binding sites for indexing. If tagmentation fails the DNA is useless

Indexing: Indexing marks each library independently of each other. If indexing fails the DNA is useless as it cannot be attributed to one specific library

Library preparation – Clean-up

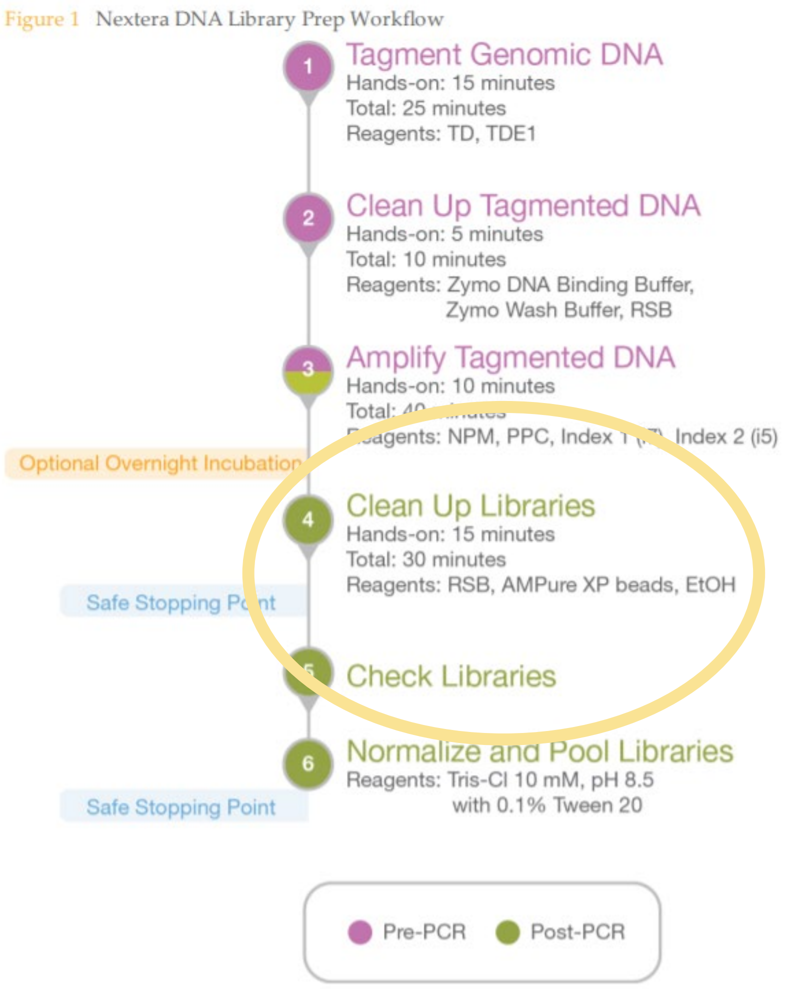


Figure 1 Nextera DNA Library Prep Workflow

1 Tagment Genomic DNA
Hands-on: 15 minutes
Total: 25 minutes
Reagents: TD, TDE1

2 Clean Up Tagmented DNA
Hands-on: 5 minutes
Total: 10 minutes
Reagents: Zymo DNA Binding Buffer, Zymo Wash Buffer, RSB

3 Amplify Tagmented DNA
Hands-on: 10 minutes
Total: 40 minutes
Reagents: NPM, PPC, Index 1 (i7), Index 2 (i5)

Optional Overnight Incubation

4 Clean Up Libraries
Hands-on: 15 minutes
Total: 30 minutes
Reagents: RSB, AMPure XP beads, EtOH

Safe Stopping Point

5 Check Libraries

6 Normalize and Pool Libraries
Reagents: Tris-Cl 10 mM, pH 8.5 with 0.1% Tween 20

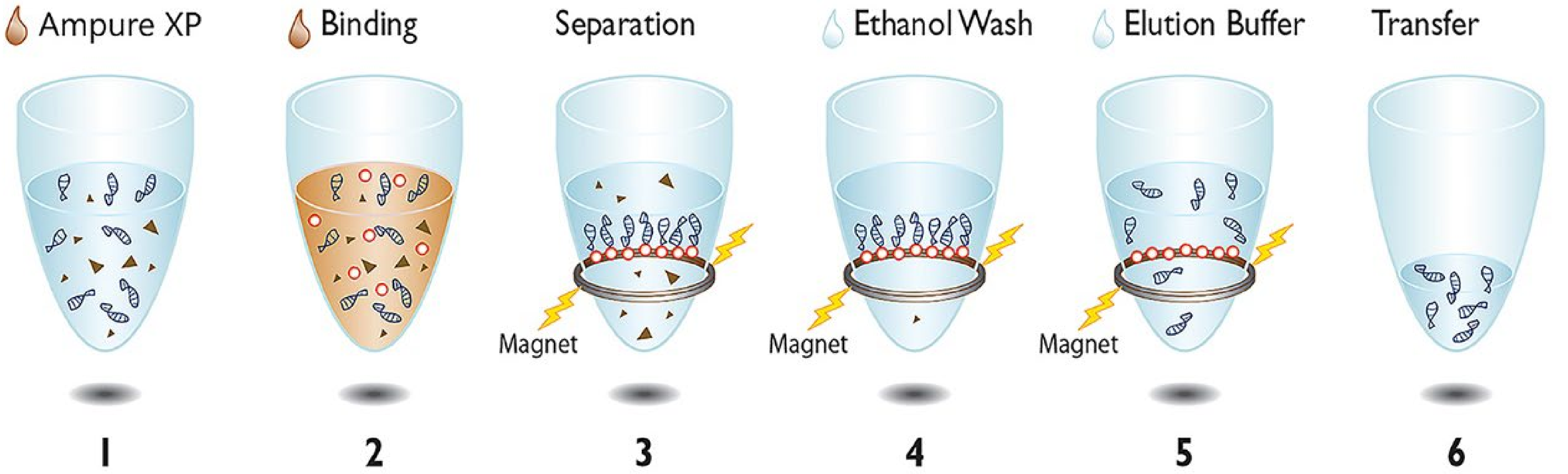
Safe Stopping Point

● Pre-PCR ● Post-PCR

© Illumina

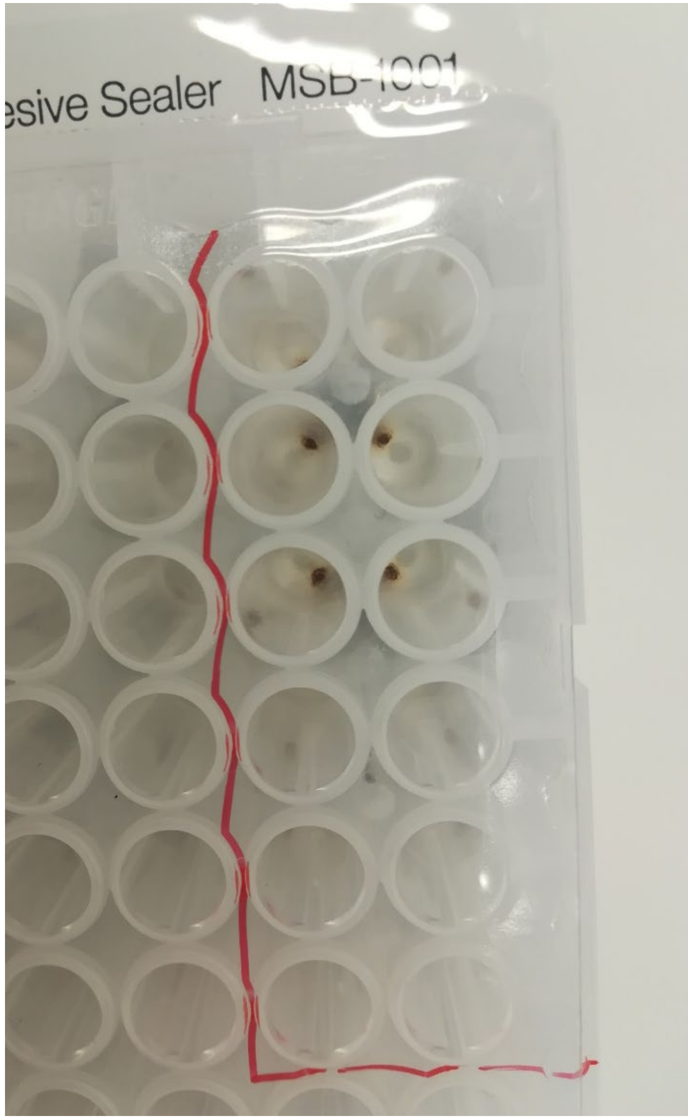
Library preparation – Clean-up

Size selection with AMPure XP beads



Ampure XP, Beckman Coulter Life Sciences, 2018

Library preparation – Clean-up



Library preparation - Normalization

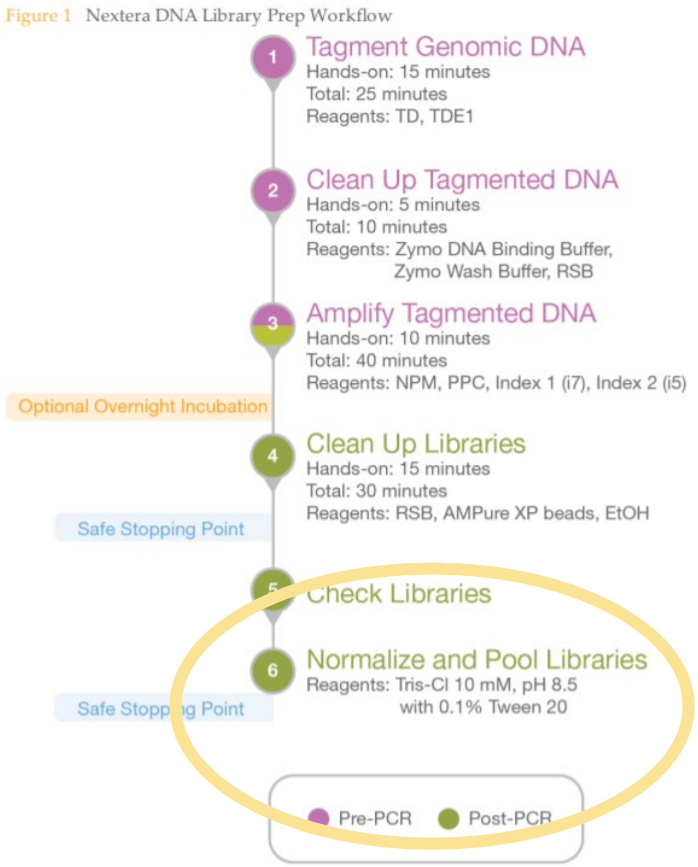


Figure 1 Nextera DNA Library Prep Workflow

1 Tagment Genomic DNA
Hands-on: 15 minutes
Total: 25 minutes
Reagents: TD, TDE1

2 Clean Up Tagmented DNA
Hands-on: 5 minutes
Total: 10 minutes
Reagents: Zymo DNA Binding Buffer, Zymo Wash Buffer, RSB

3 Amplify Tagmented DNA
Hands-on: 10 minutes
Total: 40 minutes
Reagents: NPM, PPC, Index 1 (i7), Index 2 (i5)

Optional Overnight Incubation

4 Clean Up Libraries
Hands-on: 15 minutes
Total: 30 minutes
Reagents: RSB, AMPure XP beads, EtOH

Safe Stopping Point

5 Check Libraries

6 Normalize and Pool Libraries
Reagents: Tris-Cl 10 mM, pH 8.5 with 0.1% Tween 20

Safe Stopping Point

● Pre-PCR ● Post-PCR

© Illumina

Library preparation - Normalization

Normalization: Adjusting library concentration to ensure a proper clustering and an even data distribution.

Recommended method in Nextera protocol: **bead normalization**

Somewhat similar to what happens in clean-up:

- DNA fragments are denatured and selected by using magnetic beads
- DNA fragments are bound to the surface of the beads in the same amount in all libraries
- Excess DNA is removed by washing
- The final concentration of DNA is the same in all libraries

Library preparation - Normalization

STANDARD NORMALIZATION

It's another way to adjust DNA input for the library - manually measure DNA and adjust DNA concentration using a spectrophotometeric method

Illumina recommends loading pooled libraries at 6-20 pM (for MiSeq) or 1.8 pM (for NextSeq)

Individual library concentration before pooling should be known, normalized and measured

Dependant on fragment size, clean-up success, …

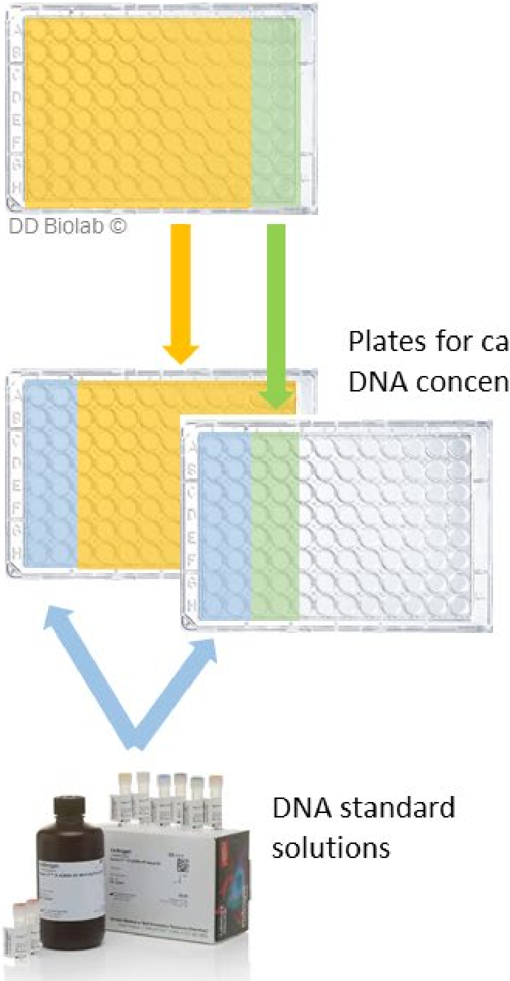
Fragment size has to be known to convert between units

Sample (examples)	Library concentration before normalization (ng/uL)	Library concentration before normalization (nM)
Isolate A	5,643	9,501
Isolate B	1,979	3,332
Final pooled library concentration: 0.71 nM		

NORMALIZATION AND POOLING

Library preparation - Normalization

Library prep plate


DD Biolab ©

Plates for calculating DNA concentration

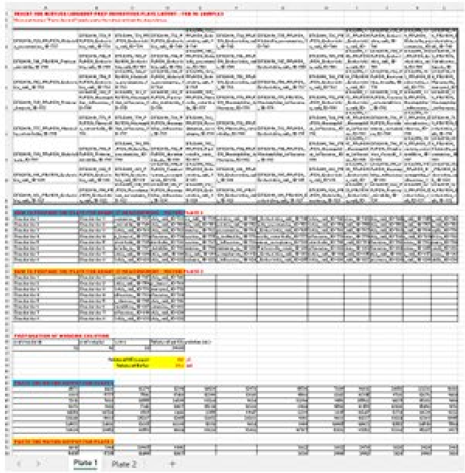
DNA standard solutions

ThermoFisher Scientific©

Fluorescence 485nm/535nm

Salford Scientific©

Automatic processing in Excel



Library preparation - pooling and diluting libraries

The normalization method will influence the last step of library prep: **pooling, denaturing and diluting libraries.**

Protocol B - Bead Normalization

It is the proper protocol to follow if the *library normalization* step was performed by bead normalization – but there are others

Dilution of the library followed by brief denaturation at 98 C

Protocol A - Standard Normalization (”manual”)

Denaturation with NaOH followed by dilution of the library

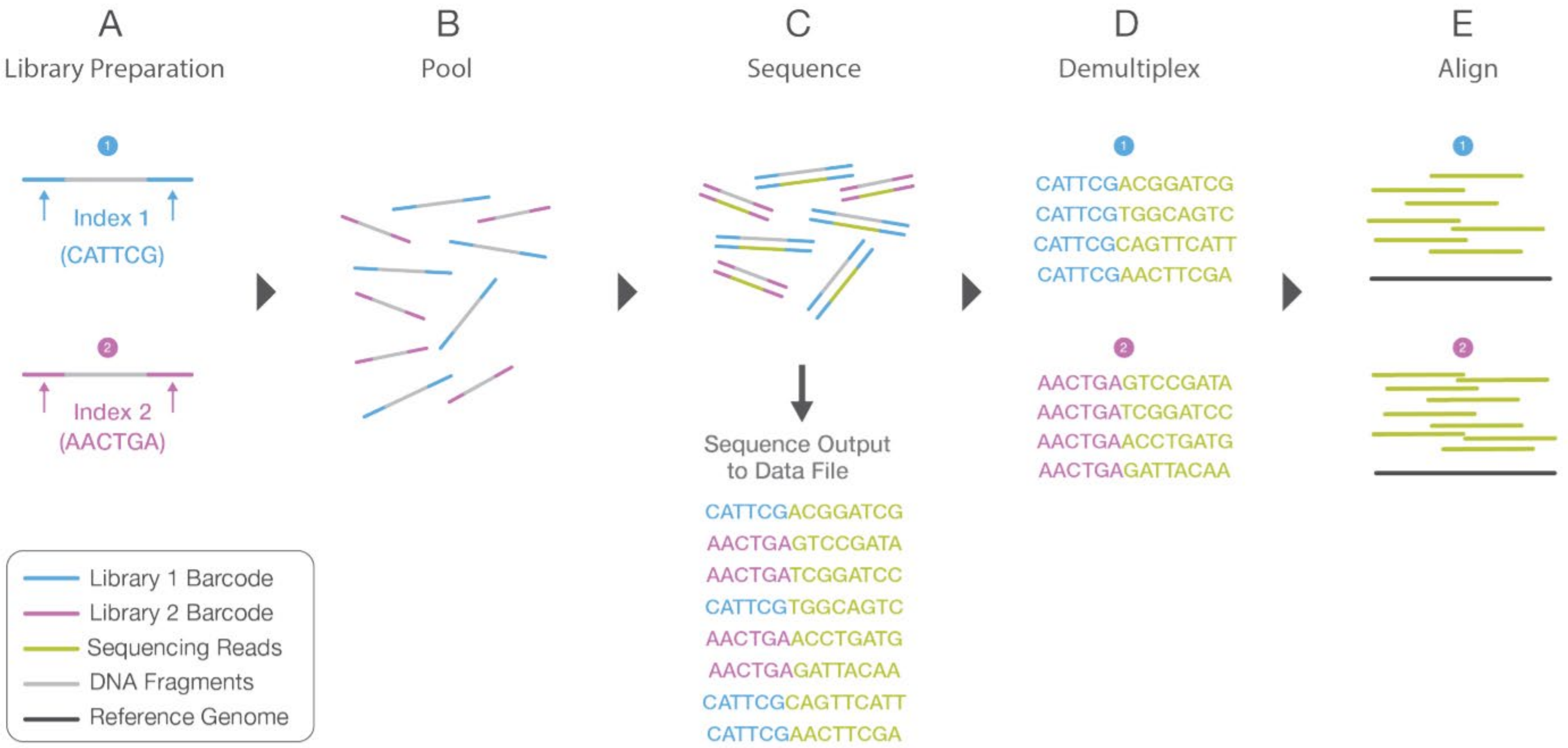
Library preparation - pooling and diluting libraries

MiSeq VS. NextSeq

Main difference – final loading volume and library concentration

NextSeq: > 1 ml at 1.8 pM MiSeq: < 1 ml at 6 – 20 pM

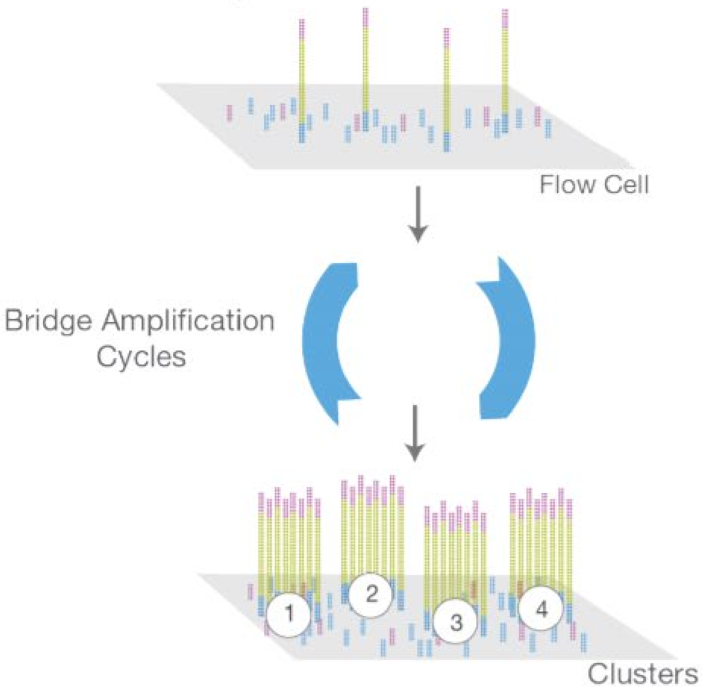
Library preparation - pooling and diluting libraries



An introduction to Next-Generation Sequencing Technology, Illumina, Inc., 2017 (adapted)

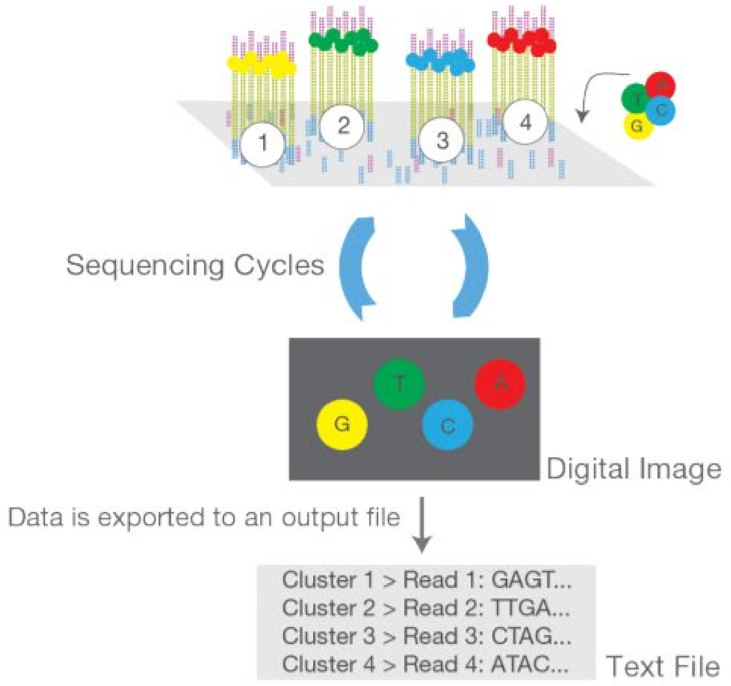
Loading and sequencing

B. Cluster Amplification



Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

C. Sequencing



Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...

Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases.

An introduction to Next-Generation Sequencing Technology, Illumina, Inc., 2017

Loading and sequencing

- Video: https://www.youtube.com/watch?v=womKfikWIxM&ab_channel=Illumina

Illumina Sequencing Technology, Illumina, Inc., 2017 (adapted)

Troubleshooting

Low DNA quality

Presence of biological contaminants (improper isolate purification)

Data can appear of good quality but results cannot be used for downstream analyis.

Presence of chemical contaminants (left-overs from extraction)

Can lead to undertagmentation and afterwards to underclustering.

Improper DNA dilution

Too much input DNA (>1ng)

Can lead to undertagmentation and afterwards to underclustering.

Too little input DNA (<1ng)

Can lead to overtagmentation and afterwards to overclustering.

Improper clean-up

Can lead to overclustering**.**

Underclustering: Lower data output / Overclustering: Lower data quality

To keep in mind

Do not re-use materials – especially if you are not confident while doing it

- ○ Can you use the same tips to distribute the indexes during amplification?
- ○ Can you use the same tips to distribute ethanol during library clean up?

Pay attention to storage and thawing conditions

- ○ Why?

Have a clear (even if basic) understanding of what is happening in each step

- ○ This is the only way you will be able to modify the protocols according to your needs and resources.

Downstream processing

- Downloading data from platform
- Quality control
- Assembly
- Analysis
  - AMR genes
  - Virulence factors
  - MLST
  - Serotype
  - …

Coffee break

Back at 11:20.

Funded by the European Union

Jette Sejer Kjeldgaard

jetk@food.dtu.dk

From isolate to WGS – WGS data and bionformatics

Funded by the European Union

From isolate to WGS

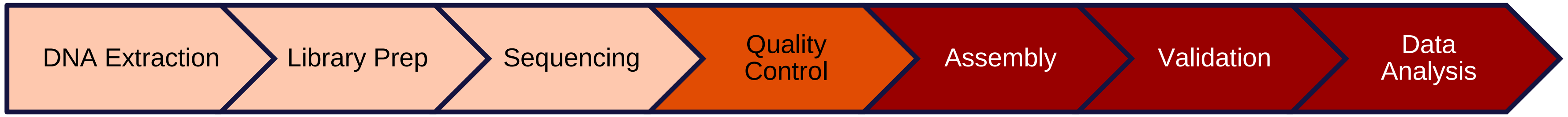
- WGS data and bioinformatics

EURGen-RefLabCap
Technical training workshop #1
29 November 2022

Jette S. Kjeldgaard
(jetk@food.dtu.dk)

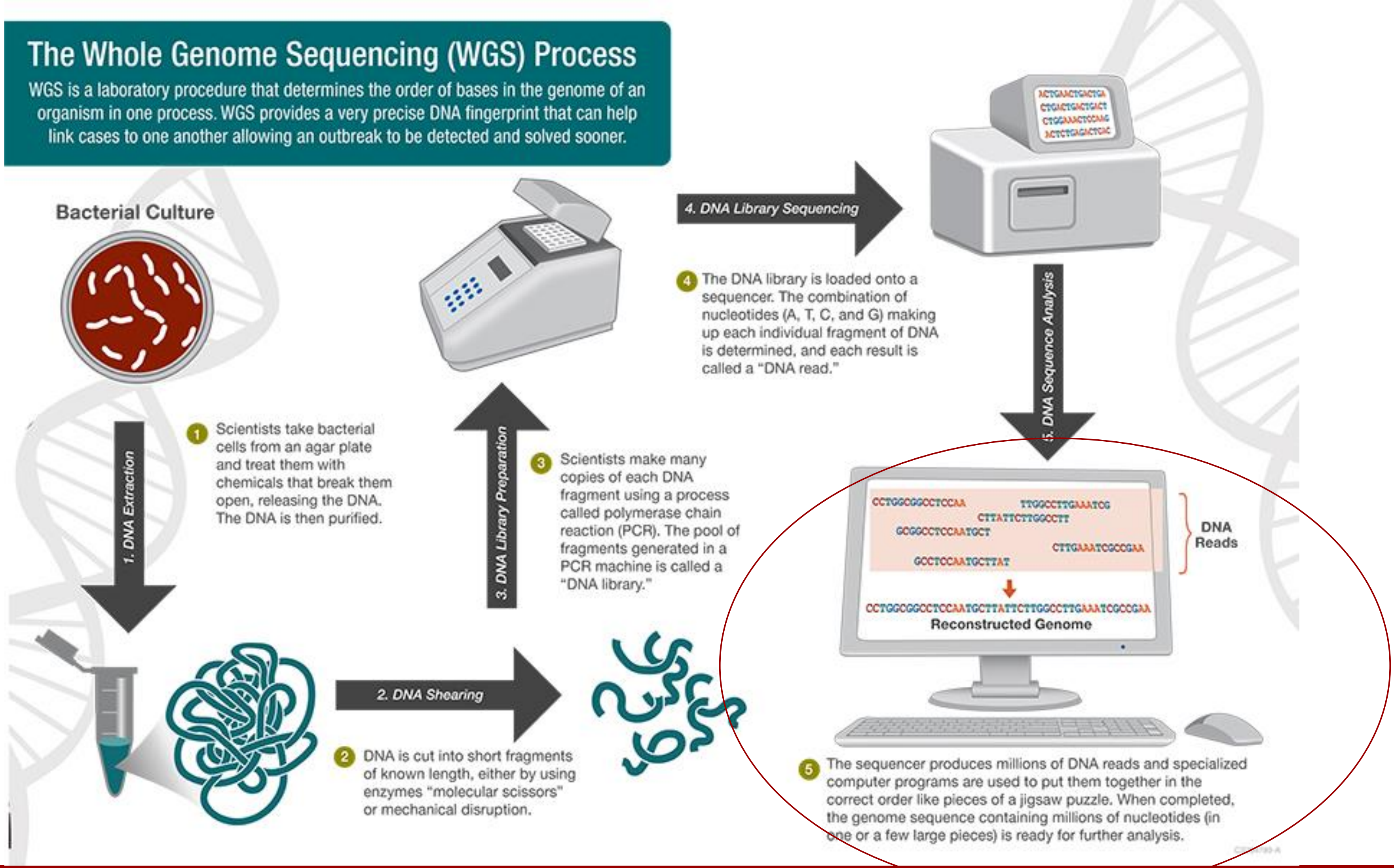
The many steps of sequencing

- Overview of workflow – bacterium to WGS result



DNA Extraction → Library Prep → Sequencing → Quality Control → Assembly → Validation → Data Analysis

WGS-based analysis of bacteria – Overview



The Whole Genome Sequencing (WGS) Process

WGS is a laboratory procedure that determines the order of bases in the genome of an organism in one process. WGS provides a very precise DNA fingerprint that can help link cases to one another allowing an outbreak to be detected and solved sooner.

1 Scientists take bacterial cells from an agar plate and treat them with chemicals that break them open, releasing the DNA. The DNA is then purified.

2 DNA is cut into short fragments of known length, either by using enzymes "molecular scissors" or mechanical disruption.

3 Scientists make many copies of each DNA fragment using a process called polymerase chain reaction (PCR). The pool of fragments generated in a PCR machine is called a "DNA library."

4 The DNA library is loaded onto a sequencer. The combination of nucleotides (A, T, C, and G) making up each individual fragment of DNA is determined, and each result is called a "DNA read."

5 The sequencer produces millions of DNA reads and specialized computer programs are used to put them together in the correct order like pieces of a jigsaw puzzle. When completed, the genome sequence containing millions of nucleotides (in one or a few large pieces) is ready for further analysis.

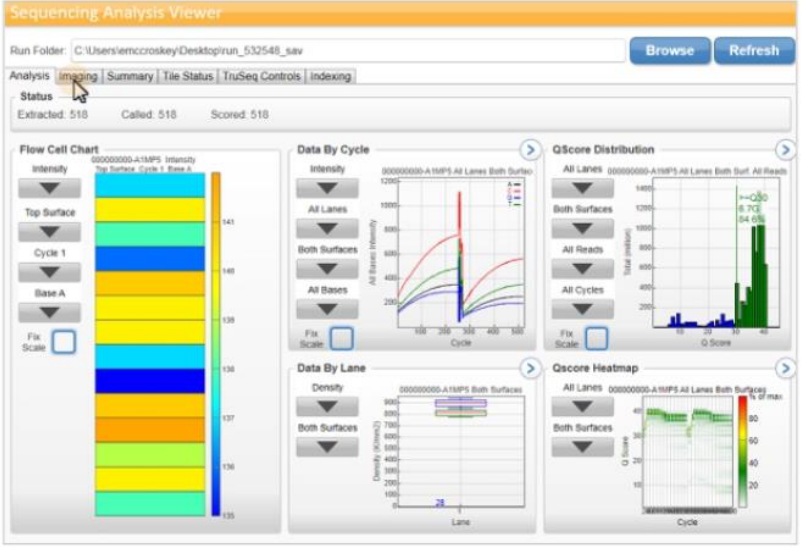
https://www.cdc.gov/pulsenet/pathogens/protocol-images.html#wgs

What to do when you have a sequence?

- Illumina sequencing

Analysis tools

Illumina sequence viewer



Sequencing Analysis Viewer

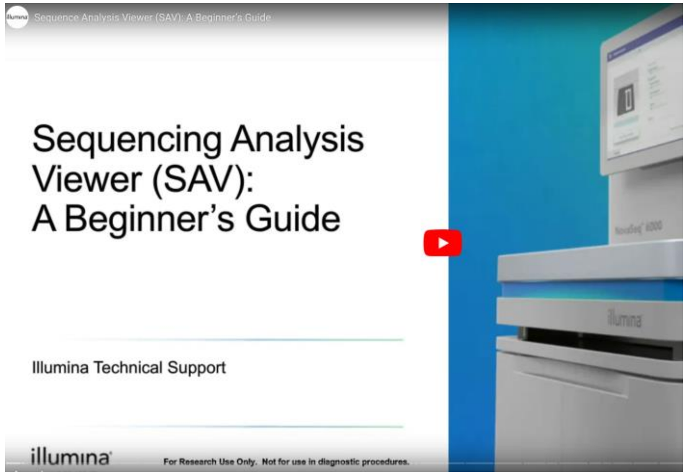
Run Folder: C:\Users\emccroskey\Desktop\run_532548_sav

Status

Extracted: 518 Called: 518 Scored: 518

Support.illumina.com

Illumina Featured Training



Sequence Analysis Viewer (SAV): A Beginner's Guide

Sequencing Analysis Viewer (SAV): A Beginner's Guide

Illumina Technical Support

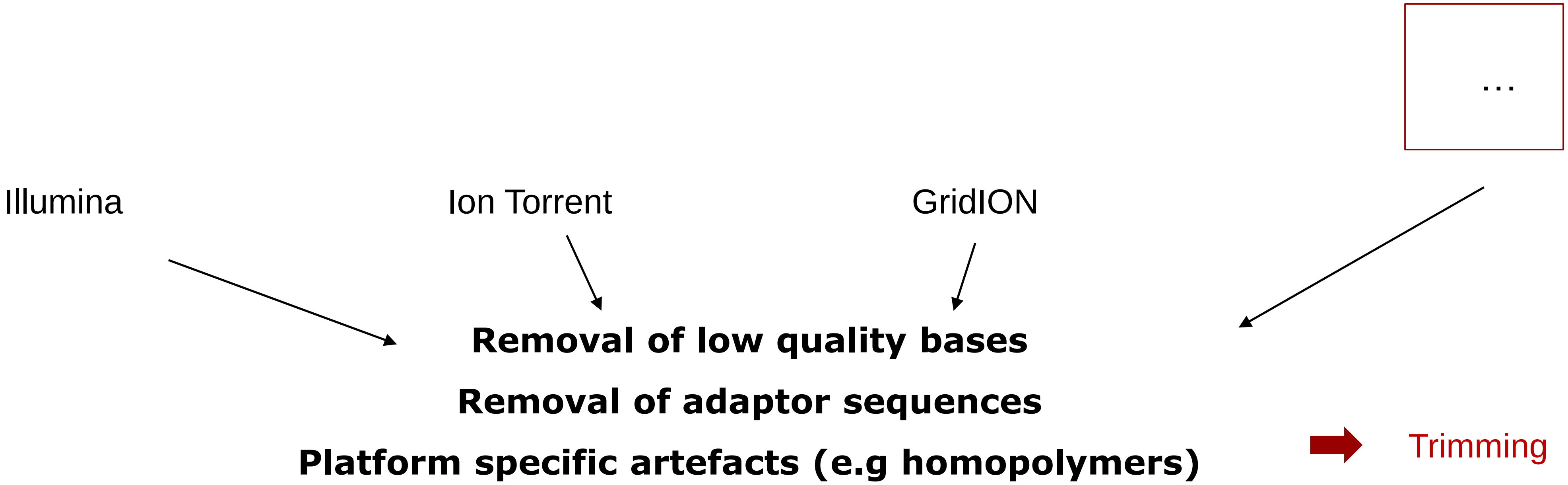
illumina

For Research Use Only. Not for use in diagnostic procedures.

https://support.illumina.com/sequencing/sequencing_software/sequencing_analysis_viewer_sav.html

Sequencing in-house or outsourced;

All platforms have errors and artefacts



Illumina

Ion Torrent

GridION

…

Removal of low quality bases

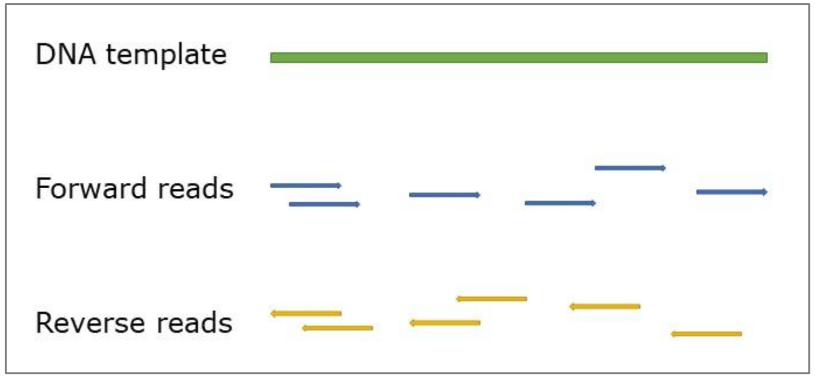
Removal of adaptor sequences

Platform specific artefacts (e.g homopolymers)

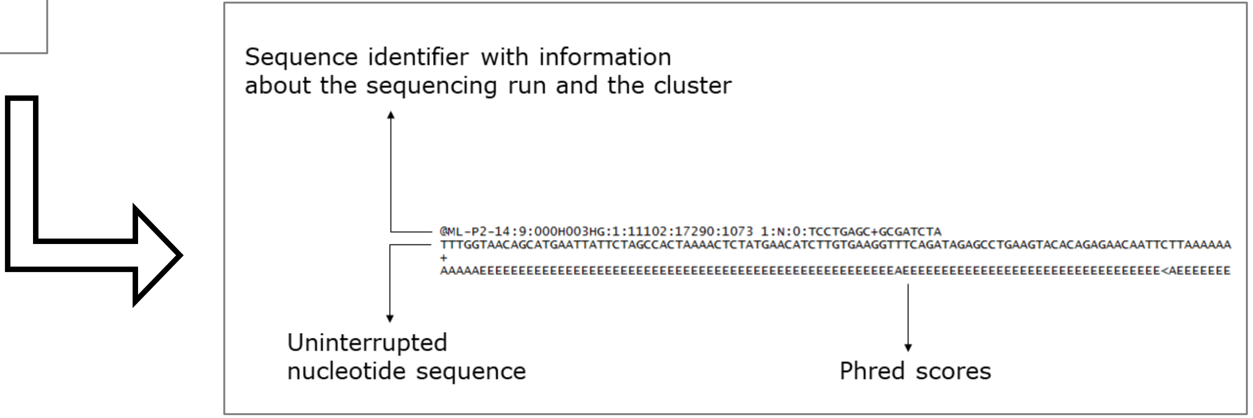
Trimming

(Illumina) WGS-based analysis of bacteria – How it works

- Illumina platforms sequence the DNA through a process called short-read paired-end massive parallel synthesis



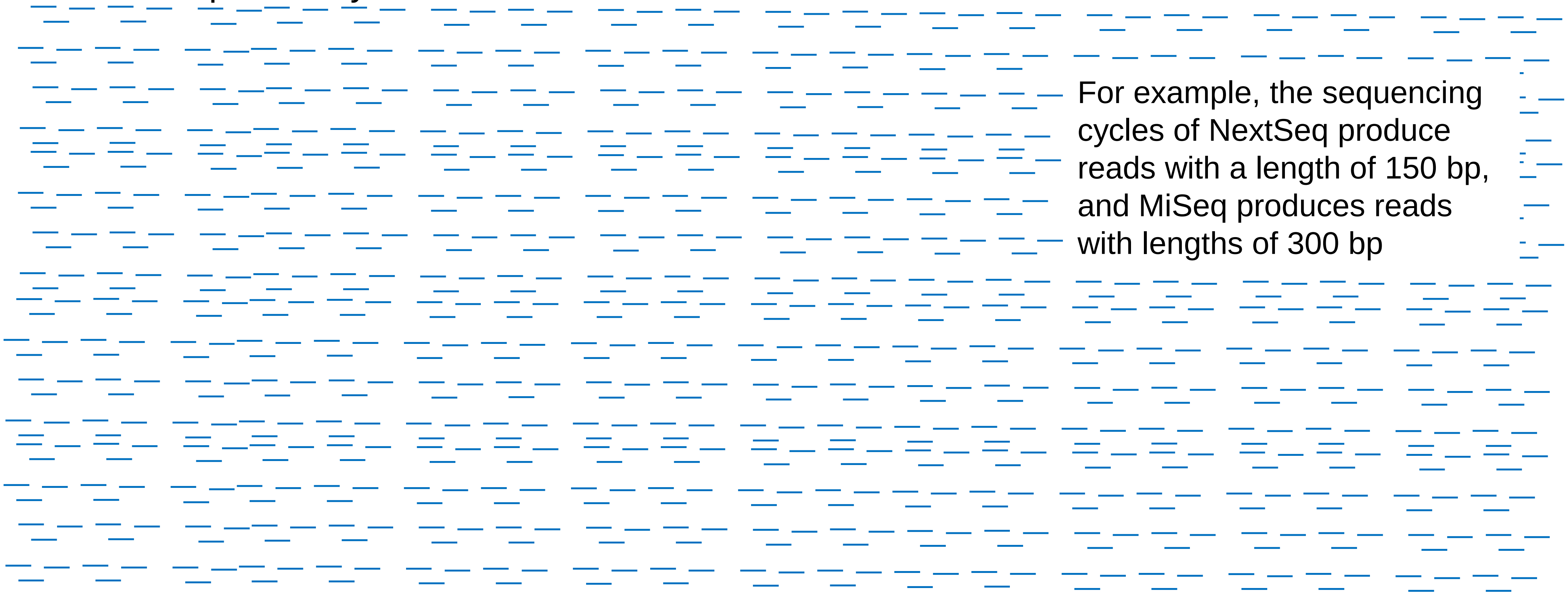
For example, the sequencing cycles of NextSeq produce reads with a length of 150 bp, and MiSeq produces reads with lengths of 300 bp


Sequence identifier with information about the sequencing run and the cluster
Uninterrupted nucleotide sequence
Phred scores

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

(Illumina) WGS-based analysis of bacteria – How it works

- Illumina platforms sequence the DNA through a process called short-read paired-end massive parallel synthesis



For example, the sequencing cycles of NextSeq produce reads with a length of 150 bp, and MiSeq produces reads with lengths of 300 bp

Paired-end short reads = fastq format

What is the data?

Fastq files

What is Fastq?

Fasta + quality scores

1 read, 4 lines

Fastq example:

```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACAGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCGCA
+
 BP`ccceggcegihiiighiifhihfddgfhi^efgfhhhhhegiiiiiiiihiihihggeeccdddcccacWTT^acc[ab `]`[ b`^BBBBBBBB
```

```
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCGAC
+
bb eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[ X]]a[aacXT
```

```
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGATT
+
bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh fhggdgegeaceeacbdcbcc\^aa]`` ^bb]bcccccbac a^bc
```

```
@FCC0CD5ACXX:1:1101:1239:2083#AGCGT/1
AGCGTCTGACTCACACAAAAACGGTAACACAGTTATCCACAGAATCAGGGGATAAGGCCGGAAAGAACATGTGAGCAAAAAGGCAAAGCCAGGACAAAAGG
+
bbbeeeeegggggiiiiiiiiiigifhhiiighiiihhiiiiiiihiiiiiiiiiiihiigcdbbdcdcccccdcccccccccaccccccccbcccaccccc
```

What is the data?

Fastq files

What is Fastq?

Fasta + quality scores

Fastq example:

Header/ID

```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACAGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCGCA
+
 BP`ccceggcegihiiighiifhihfddgfhi^efgfhhhhhegiiiiiiiiihiihihggeeccdddcccacWTT^acc[ab_`]`[_b`^BBBBBBBB
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCGAC
+
bb_eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[_X]]a[aacXT
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGATT
+
bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh_fhggdgegeaceeeacbdcbcc\^aa]``_^bb]bcccccbac_a^bc
@FCC0CD5ACXX:1:1101:1239:2083#AGCGT/1
AGCGTCTGACTCACACAAAAACGGTAACACAGTTATCCACAGAATCAGGGGATAAGGCCGGAAAGAACATGTGAGCAAAAAGGCAAAGCCAGGACAAAAGG
+
bbbeeeeegggggiiiiiiiiiigifhhiiighiiihhiiiiiiihiiiiiiiiiihiigcdbbdcdcccccdcccccccaccccccbcccaccccc
```

What is the data?

Fastq files

What is Fastq?

Fasta + quality scores

Fastq example:

DNA sequence

```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACAGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCGCA
+
_BP`ccceggcegihiiighiifhihfddgfhi^efgfhhhhhegiiiiiiiihiihihggeeccdddcccacWTT^acc[ab_`]`[_b`^BBBBBBBB
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCGAC
+
bb_eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[_X]]a[aacXT
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGATT
+
bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh_fhggdgegeaceeacbdcbcc\^aa]``_^bb]bcccccbac_a^bc
@FCC0CD5ACXX:1:1101:1239:2083#AGCGT/1
AGCGTCTGACTCACACAAAAACGGTAACACAGTTATCCACAGAATCAGGGGATAAGGCCGGAAAGAACATGTGAGCAAAAAGGCAAAGCCAGGACAAAAGG
+
bbbeeeeeggggiiiiiiiiiigifhhiiighiiihhiiiiiiihiiiiiiiiiihiigcdbbdcdcccccdcccccccaccccccbcccacccccc
```

What is the data?

Fastq files

What is Fastq?

Fasta + quality scores

Fastq example:

```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACAGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCGCA
+
_BP`ccceggcegihiiighiifhihfddgfhi^efgfhhhhegiiiiiiiihiihihggeeccdddcccacWTT^acc[ab_`]`[_b`^BBBBBBBB
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCGAC
+
bb_eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[_X]]a[aacXT
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGATT
+
bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh_fhggdgegeaceeacbdcbcc\^aa]``_^bb]bcccccbac_a^bc
@FCC0CD5ACXX:1:1101:1239:2083#AGCGT/1
AGCGTCTGACTCACACAAAAACGGTAACACAGTTATCCACAGAATCAGGGGATAAGGCCGGAAAGAACATGTGAGCAAAAAGGCAAAGCCAGGACAAAAGG
+
bbbeeeeeggggiiiiiiiiiigifhhiiighiiihhiiiiiiihiiiiiiiiiihiigcdbbdcdcccccdcccccccacccccccbcccaccccc
```

Name field (optional)

What is the data?

Fastq files

What is Fastq?

Fasta + quality scores

Fastq example:

Quality scores

```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACAGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCGCA
+
_BP`ccceggcegihiiighiifhihfddgfhi^efgfhhhhhegiiiiiiiihiihihggeeccdddcccacWTT^acc[ab_`]`[_b`^BBBBBBBB
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCGAC
+
bb_eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[_X]]a[aacXT
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGATT
+
bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh_fhggdgegeaceeacbdcbcc\^aa]``_^bb]bccccbac_a^bc
@FCC0CD5ACXX:1:1101:1239:2083#AGCGT/1
AGCGTCTGACTCACACAAAAACGGTAACACAGTTATCCACAGAATCAGGGGATAAGGCCGGAAAGAACATGTGAGCAAAAAGGCAAAGCCAGGACAAAAGG
+
bbbeeeeegggggiiiiiiiiiigifhhiiighiiihhiiiiiiihiiiiiiiiiihiigcdbbdcdcccccdcccccccaccccccccbcccacccccc
```

Phred scores

Phred scores:

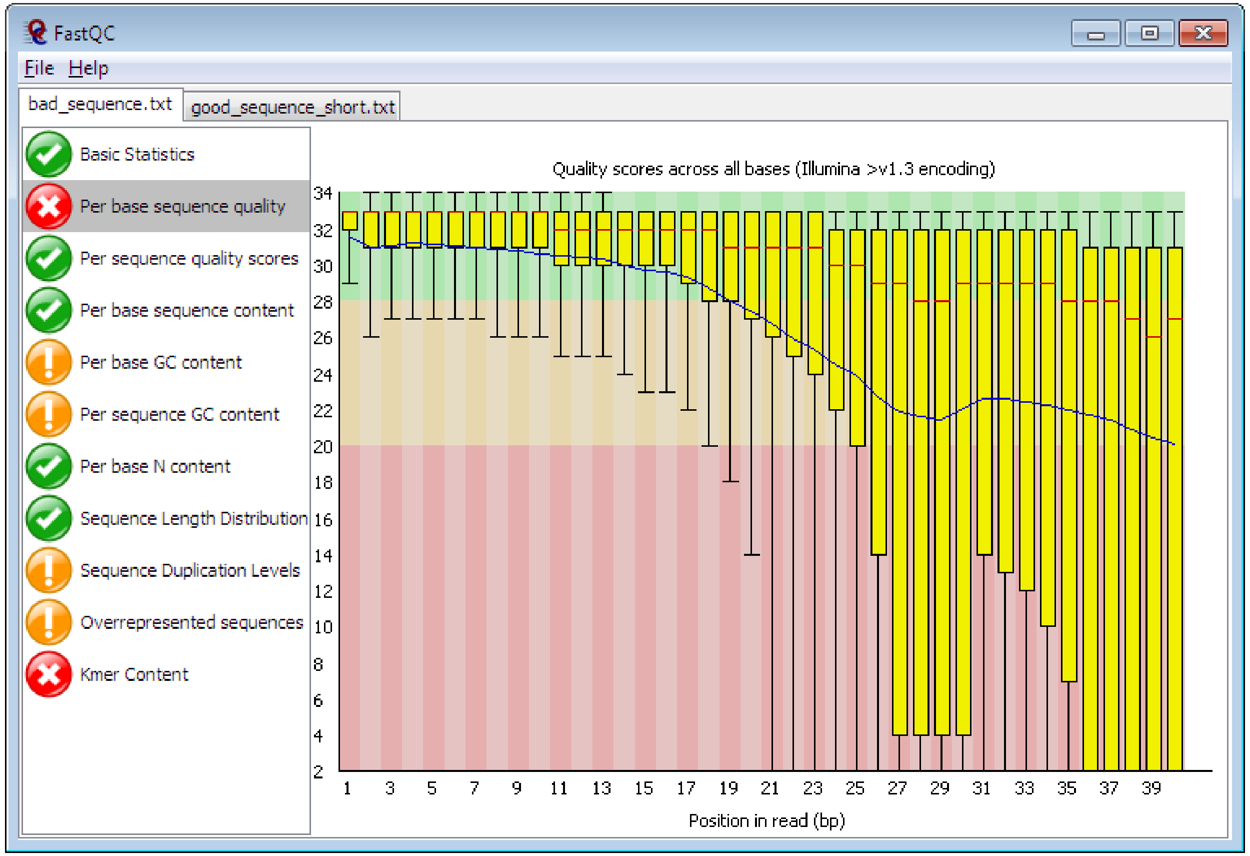
Show quality for a single nucleotides in ASCII codes. Should be ≥20

fastq format + quality scores

- Used for evaluation of quality of the sequences (QC tools)
- Used for trimming of poor quality reads
- Used for defining ‘true’ SNPs by SNP tools

- Trimming can be applied on raw reads
  - Also remove unpaired reads

- Trimmed reads = fastq format

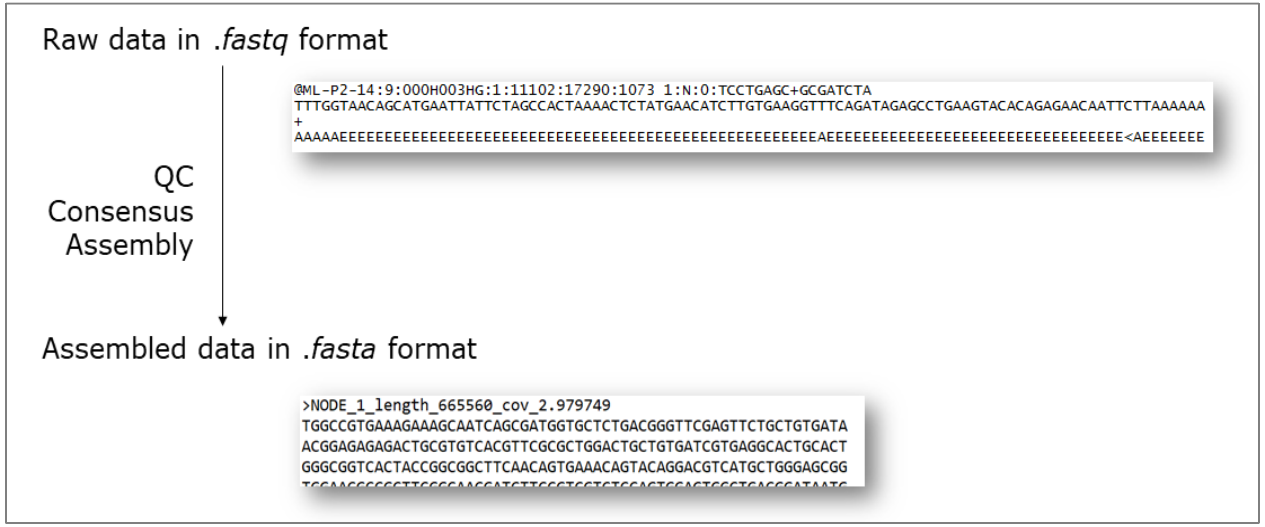


- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per base GC content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ✗ Kmer Content

FastQC- Quality control tool (online)

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

(Illumina) WGS-based analysis of bacteria
Fastq or fasta?



Raw data in *.fastq* format

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

QC
Consensus
Assembly

Assembled data in *.fasta* format

```
>NODE_1_length_665560_cov_2.979749
TGGCCGTGAAAGAAAGCAATCAGCGATGGTGCTCTGACGGGTTCGAGTTCTGCTGTGATA
ACGGAGAGAGACTGCGTGTCACGTTCGCGCTGGACTGCTGTGATCGTGAGGCACTGCACT
GGGCGGTCACTACCGGCGGCTTCAACAGTGAAACAGTACAGGACGTCATGCTGGGAGCGG
```

- **Compared with reference databases**
- *What genes from the database are present in this genome?*

File size?

Fastq:

E. coli – 100-500 MB

Two files: forward + reverse file

Fasta:

E. coli – 5 MB

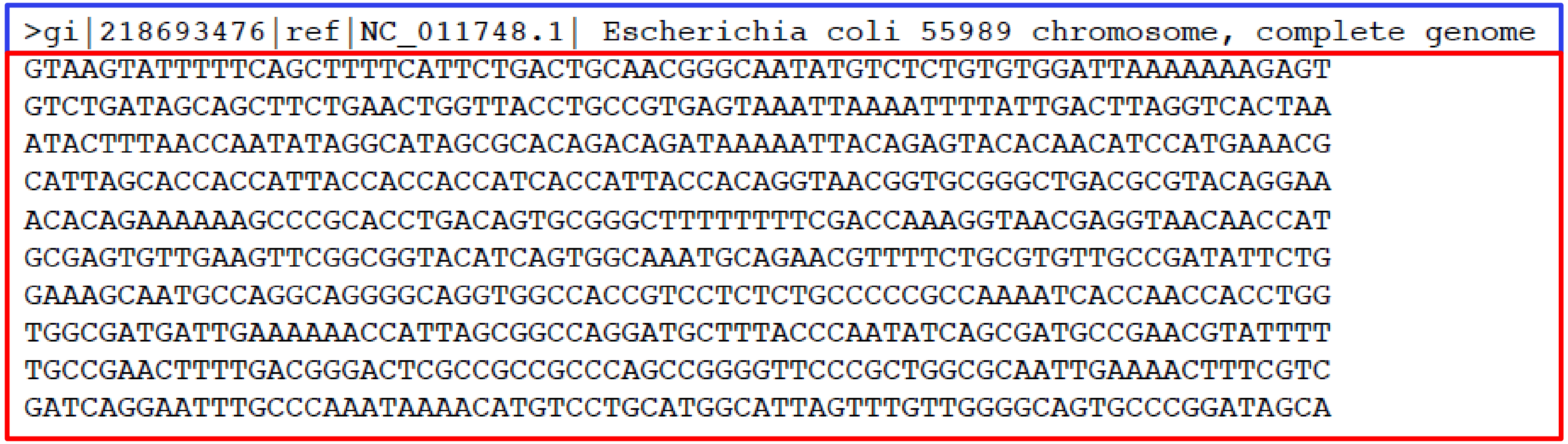
One file

=> different applications

Fasta files

- Sequence data (only) is stored in fasta files

Header



```
>gi|218693476|ref|NC_011748.1| Escherichia coli 55989 chromosome, complete genome
GTAAGTATTTTTCAGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGT
GTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAA
ATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACG
CATTAGCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAA
ACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCAT
GCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTG
GAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGG
TGGCGATGATTGAAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTT
TGCCGAACTTTTGACGGGACTCGCCGCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTC
GATCAGGAATTTGCCCAAATAAAACATGTCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCA
```

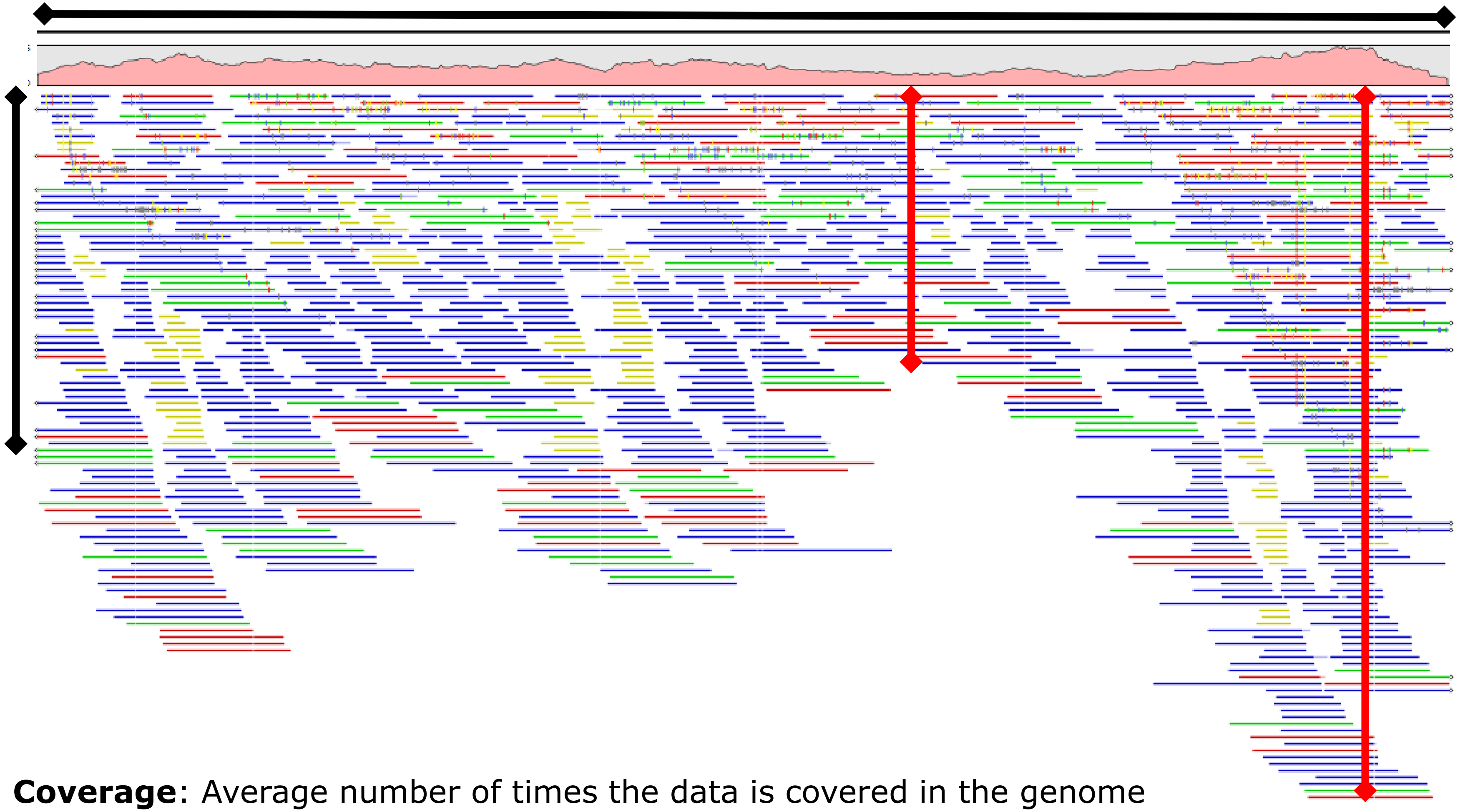
Sequence

Staph. aureus ~2.7 - 2.8 Mbp

E. coli ~ 4.5 - 5.5 Mbp

Human ~ 3.2 Gbp

Coverage vs. depth



The better coverage/higher depth

–> the better assembly

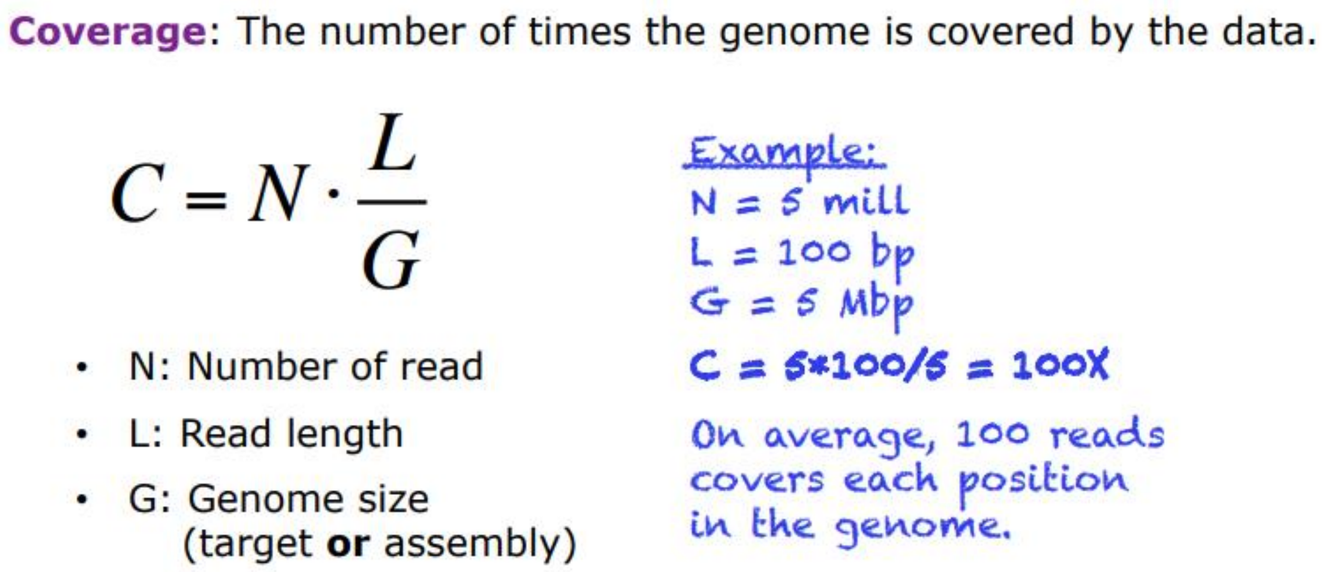
-> …in theory and to a certain limit

Coverage: Average number of times the data is covered in the genome
Depth: Number reads that covers a particular nucleotide in each position in the genome.

Coverage

- Good coverage is important to ensure all of the genome is covered
  - High variation in local coverage over the genome
  - Low copy plasmids can be hard to find
    - (trimmed off)

- QC tools output average coverage
- Can also be calculated:



Coverage: The number of times the genome is covered by the data.

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

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

Example:

N = 5 mill

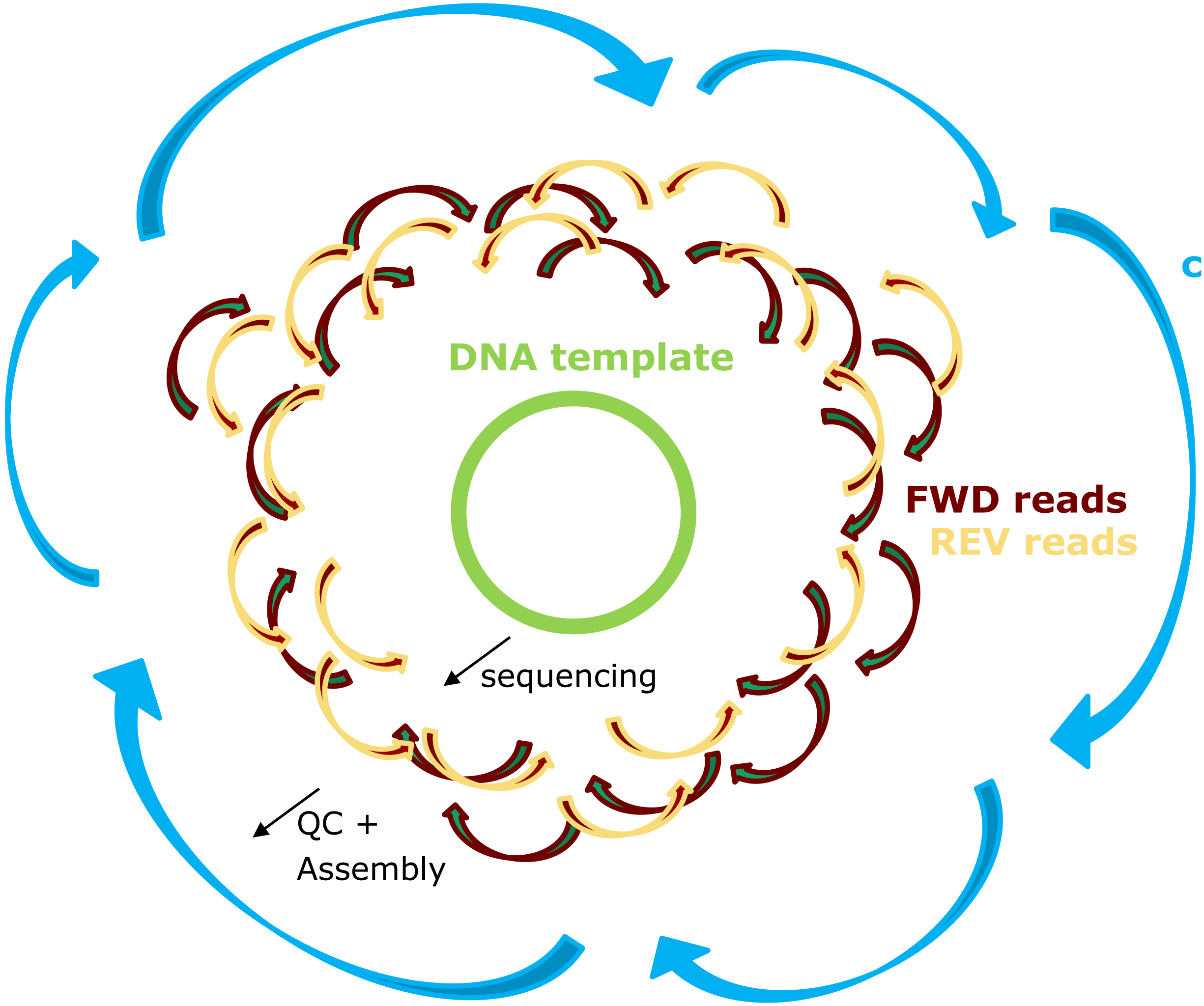
L = 100 bp

G = 5 Mbp

C = 5*100/5 = 100X

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

Assembly



Contig size?

- Preferably 100,000's of bases
- Small contigs can (often) be removed

Assembly methods

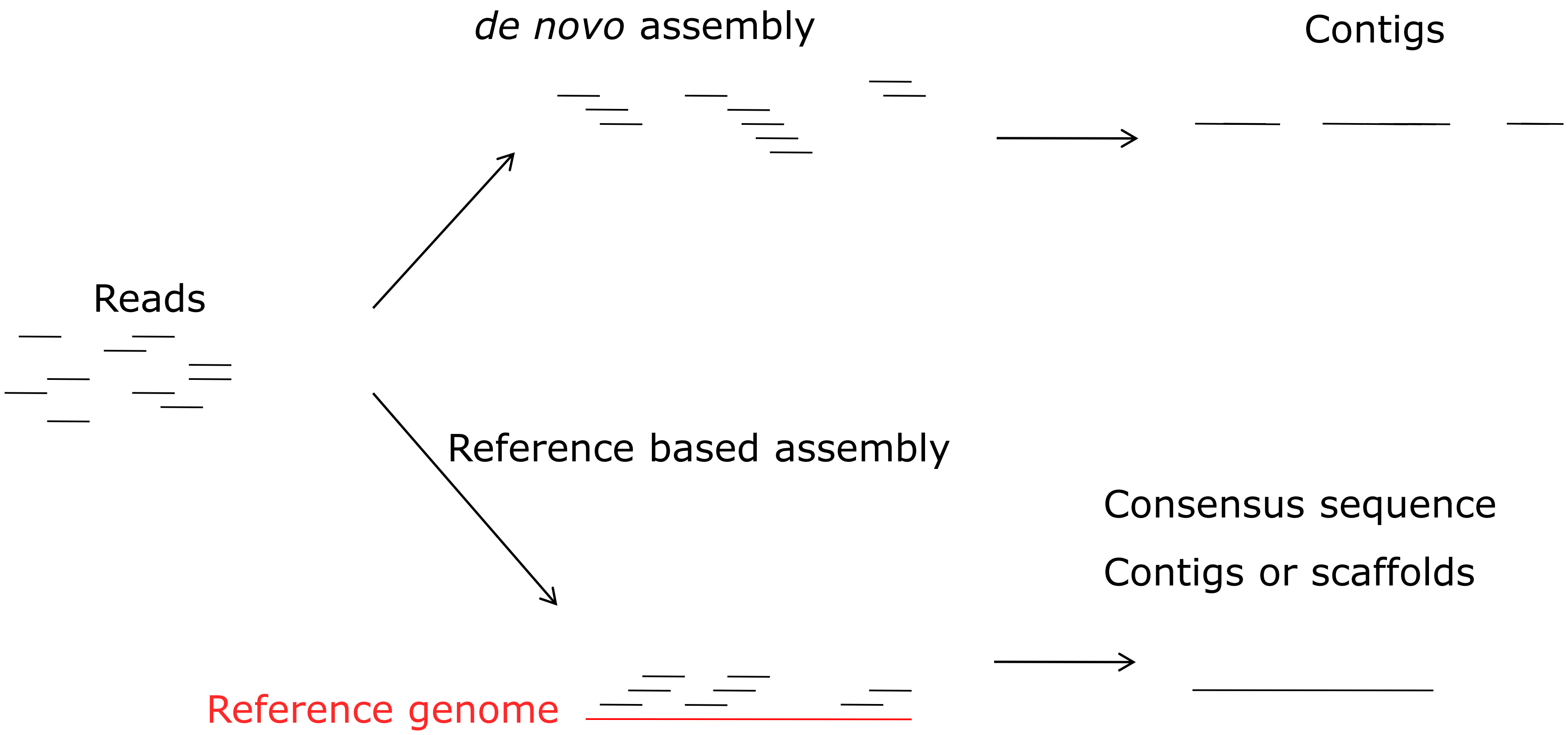
1. Mapping to reference

or

2. *de novo* assembly

Assembly methods

1



de novo assembly: short vs long reads

- You want as few and long contigs (or scaffolds) as possible
- Short reads are difficult to assemble to such long sequences
- Long reads have more errors
- Current state of the art is to use
  both short and long reads (hybrid assembly)
- Various software for different applications

Fasta vs fastq

- Fastq format
  - Much more data
  - Quality scores
  - Low depth reads included

- Better resolution for analyses
  - Quality parameters on SNPs

- Low quality reads can affect results
- Trimming is a benefit

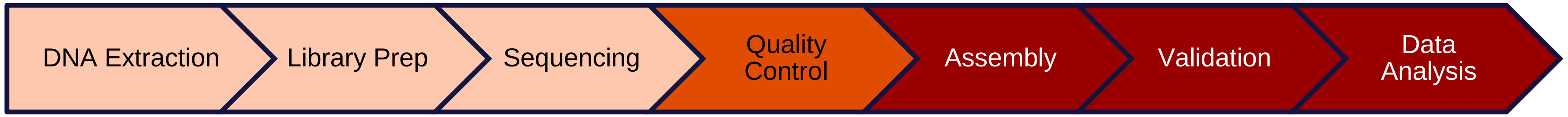
- Fasta format
  - Smaller files to handle
  - No quality scores
  - Only consensus sequence

- Context of genes and up/downstream bases

- Fasta might be adequate for most processes

The many steps of sequencing

- Overview



DNA Extraction

Library Prep

Sequencing

Quality Control

Assembly

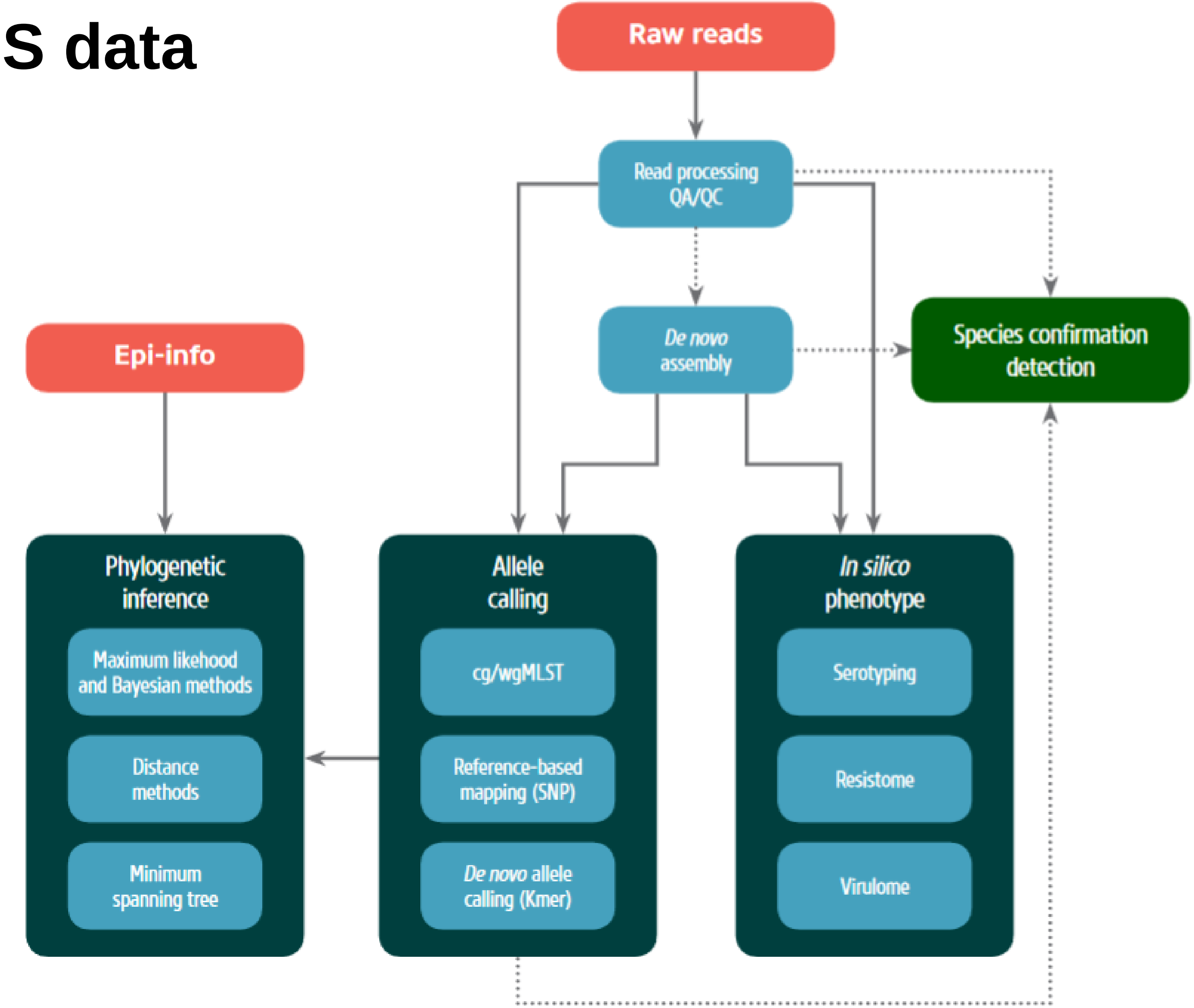
Validation

Data Analysis

Data analysis of WGS data

- Different approaches for different organisms (and subsets)
- Single-isolate analysis and/or phylogenetic analysis of all (relevant) isolates

Schematic representation of WGS pipeline

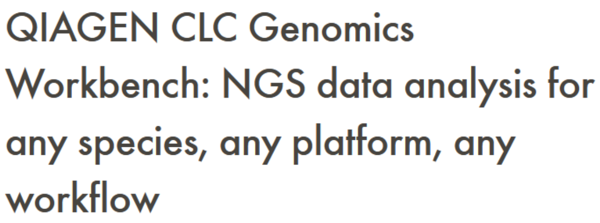


WHO landscape paper https://www.who.int/publications/i/item/789241513869

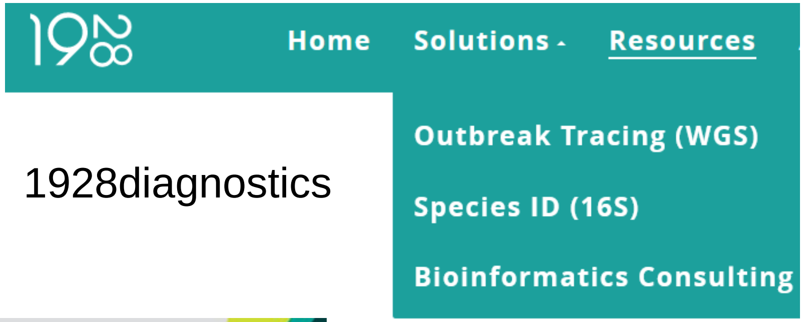
Commercial pipelines

- Various solutions available – including:

Bionumerics



QIAGEN CLC Genomics Workbench: NGS data analysis for any species, any platform, any workflow



Home Solutions Resources

Outbreak Tracing (WGS)

Species ID (16S)

Bioinformatics Consulting

1928diagnostics

Whole genome sequencing for foodborne disease surveillance

Landscape paper

World Health Organization

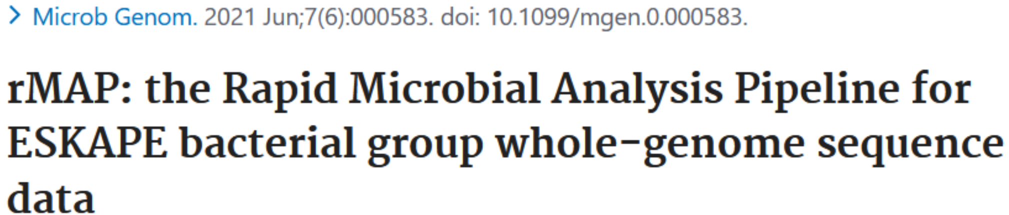
WHO Whole genome sequencing for foodborne disease surveillance: landscape paper

Great resource – also for clinical labs!

https://www.who.int/publications/i/item/789241513869

Published pipelines and perspectives

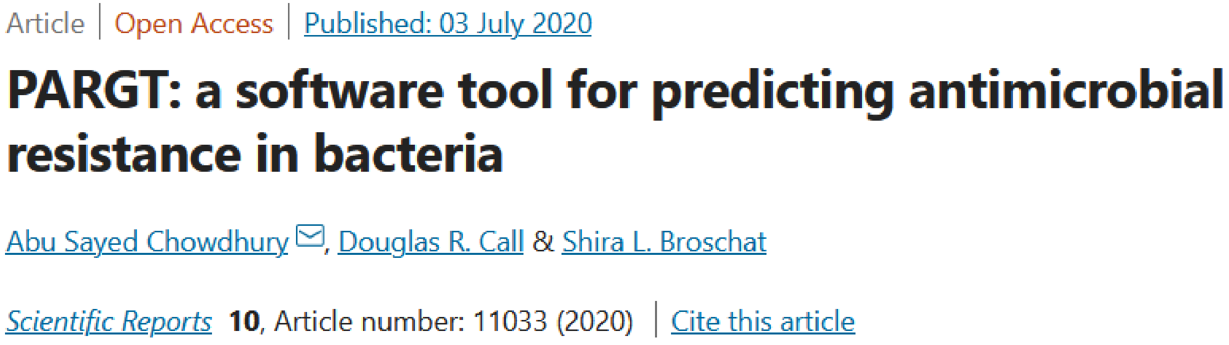
– almost random examples



› Microb Genom. 2021 Jun;7(6):000583. doi: 10.1099/mgen.0.000583.

rMAP: the Rapid Microbial Analysis Pipeline for ESKAPE bacterial group whole-genome sequence data

Ivan Sserwadda [1] [2], Gerald Mboowa [3] [2]

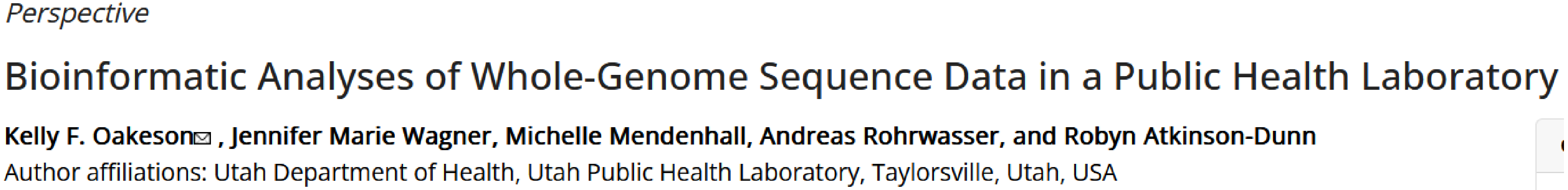


Article | Open Access | Published: 03 July 2020

PARGT: a software tool for predicting antimicrobial resistance in bacteria

Abu Sayed Chowdhury ✉, Douglas R. Call & Shira L. Broschat

Scientific Reports **10**, Article number: 11033 (2020) | Cite this article



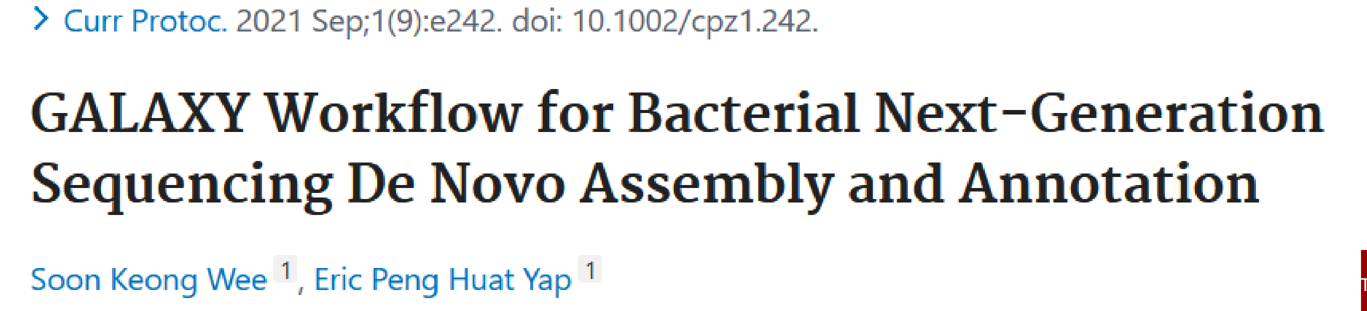
Volume 23, Number 9—September 2017

Perspective

Bioinformatic Analyses of Whole-Genome Sequence Data in a Public Health Laboratory

Kelly F. Oakeson✉ , Jennifer Marie Wagner, Michelle Mendenhall, Andreas Rohrwasser, and Robyn Atkinson-Dunn

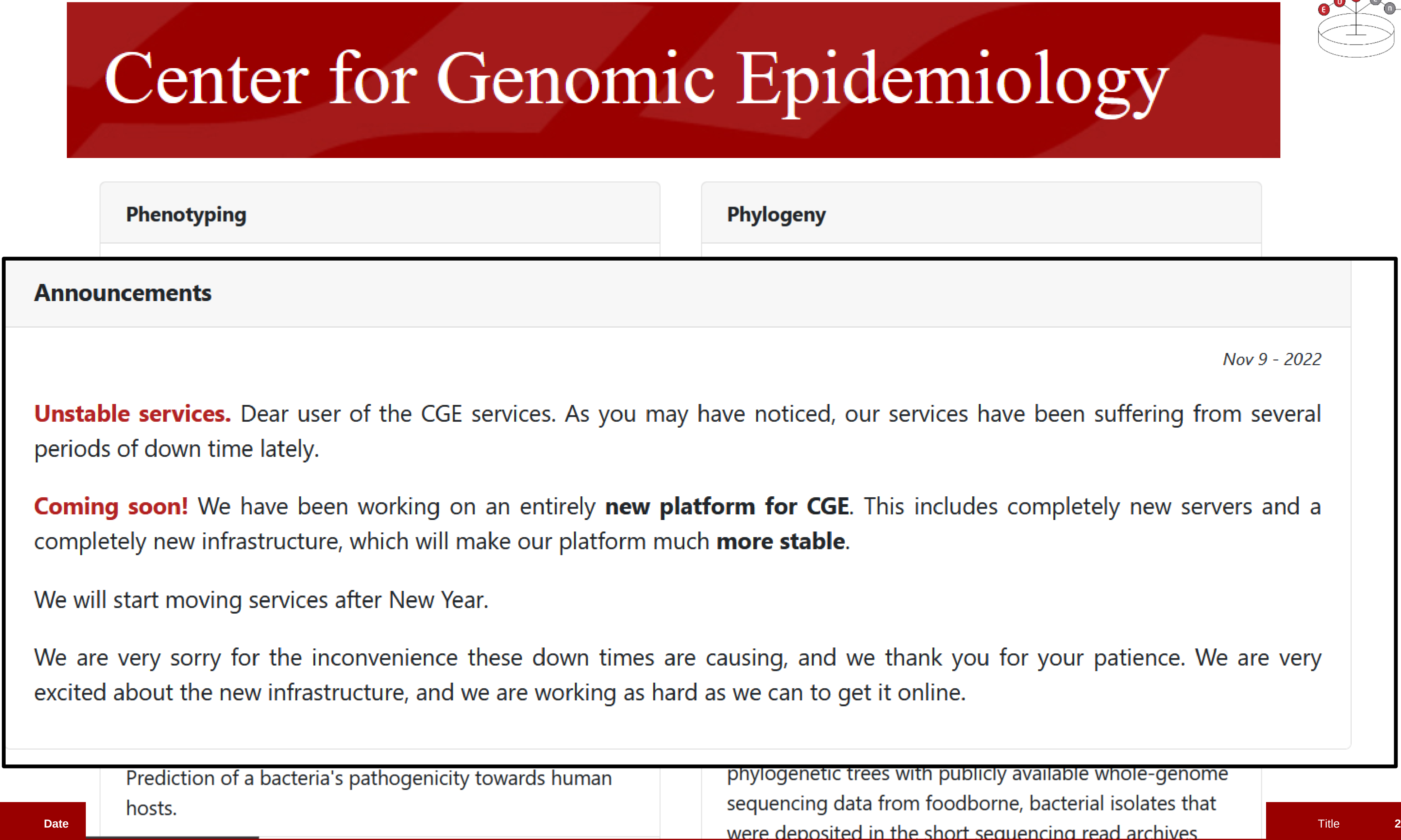
Author affiliations: Utah Department of Health, Utah Public Health Laboratory, Taylorsville, Utah, USA



› Curr Protoc. 2021 Sep;1(9):e242. doi: 10.1002/cpz1.242.

GALAXY Workflow for Bacterial Next-Generation Sequencing De Novo Assembly and Annotation

Soon Keong Wee [1], Eric Peng Huat Yap [1]



Center for Genomic Epidemiology

Phenotyping

Phylogeny

Announcements

Nov 9 - 2022

Unstable services. Dear user of the CGE services. As you may have noticed, our services have been suffering from several periods of down time lately.

Coming soon! We have been working on an entirely **new platform for CGE**. This includes completely new servers and a completely new infrastructure, which will make our platform much **more stable**.

We will start moving services after New Year.

We are very sorry for the inconvenience these down times are causing, and we thank you for your patience. We are very excited about the new infrastructure, and we are working as hard as we can to get it online.

Prediction of a bacteria's pathogenicity towards human hosts.

phylogenetic trees with publicly available whole-genome sequencing data from foodborne, bacterial isolates that were deposited in the short sequencing read archives

WGS-based analysis of bacteria – Requirements

- Expertise on DNA extraction methods
- Expertise on library preparation methods

Not too technically demanding
Ideally a dedicated room

- Access to sequencing platform
- Access and expertise on bioinformatics tools

Main challenges: cost, implementation

- Data management infrastructure

Main challenges: cost, compatibility

Personnel…

Collaboration

- Microbiologist/Molecular biologist
  - initial phenotypic/molecular identification and characterization of isolates, including culture purification and storage;
  - genomic DNA extraction and purification, library preparation with appropriate quality controls;
  - setting up of the sequencing run

- Bioinformatician
  - computational analysis of sequencing data,
  - Implementation and verification of tools/pipelines
  - variant detection and isolate clustering through construction of phylogenetic trees;
  - maintenance of accurate secure records of all procedures, including electronic databases of genome sequences and related quality control data;
  - quality assessment of original and processed sequencing data

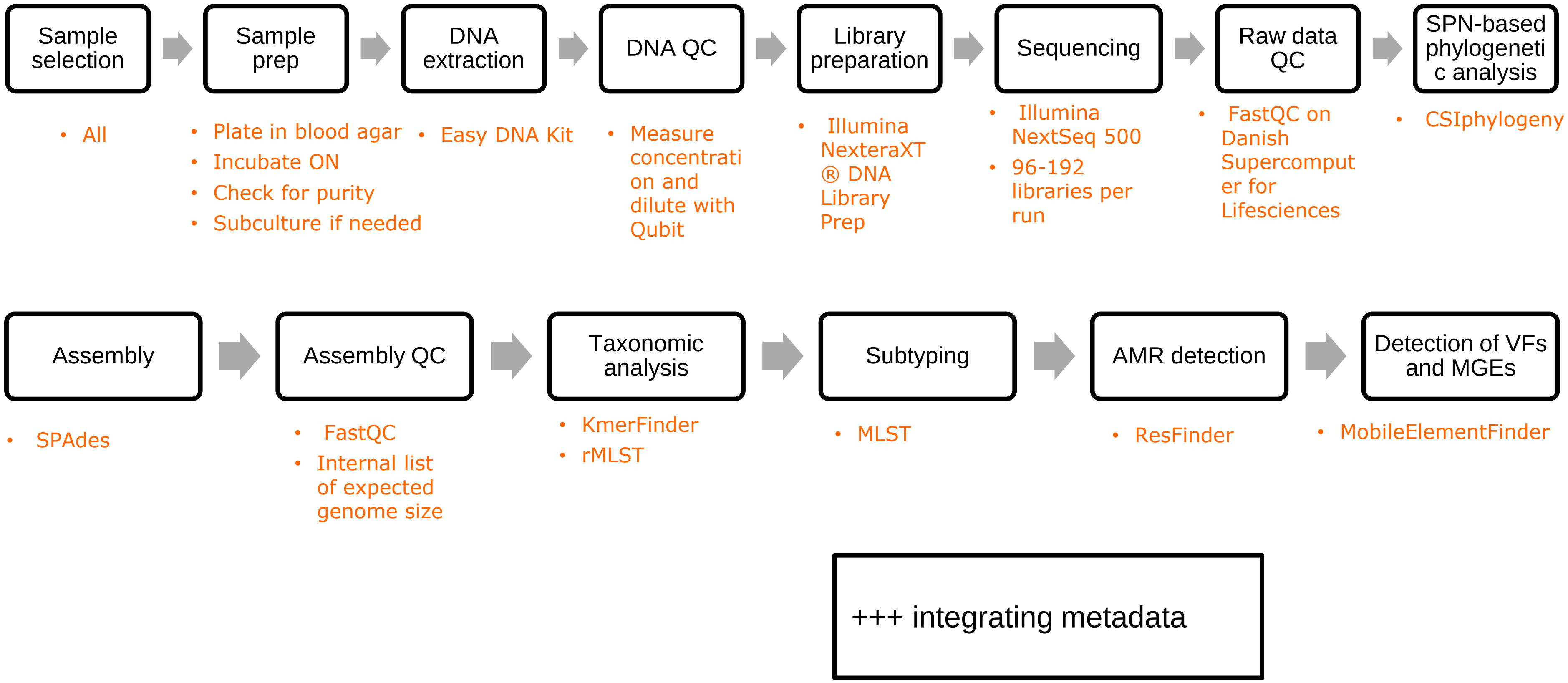
- Epidemiologist
  - collecting epidemiological information and integrating it with WGS data
  - setting of definitions for what constitutes a cluster to support epidemiological investigations
  - determination of which cases need to be followed up to collect epidemiological information, including determination of what isolates are part of a cluster

Adapted from WHO landscape paper https://www.who.int/publications/i/item/789241513869

- Questions – comments are welcome

  – Looking forward to the hands-on training next week!

Example: a complete WGS workflow



Sample selection
- All

Sample prep
- Plate in blood agar
- Incubate ON
- Check for purity
- Subculture if needed

DNA extraction
- Easy DNA Kit

DNA QC
- Measure concentration and dilute with Qubit

Library preparation
- Illumina NexteraXT® DNA Library Prep

Sequencing
- Illumina NextSeq 500
- 96-192 libraries per run

Raw data QC
- FastQC on Danish Supercomputer for Lifesciences

SPN-based phylogenetic analysis
- CSIphylogeny

Assembly
- SPAdes

Assembly QC
- FastQC
- Internal list of expected genome size

Taxonomic analysis
- KmerFinder
- rMLST

Subtyping
- MLST

AMR detection
- ResFinder

Detection of VFs and MGEs
- MobileElementFinder

+++ integrating metadata

Ana Rita Rebelo

anrire@food.dtu.dk

Quality control of WGS data

Funded by the European Union

Objectives

Many different:

- DNA extraction kits
- Sequencing platforms
- Bioinformatics approaches
- Bioinformatics tools

Well defined set of QC parameters

- For the raw data

  E.g. nr. and length of raw reads, depth of coverage
- For the assembled genomes

  E.g. N50, nr. of contigs, genome size
- For the performance of the tools

  E.g. accurately detect PMs and ARGs in sets of benchmarking data

Data QC

Raw data QC

Number of reads

Should be as high as possible. No assessed cut-off exist, but enough to obtain the desired coverage of the organism genome

Average read length

Should correspond to that expected from the sequencing platform and kit.

Illumina MiSeq avg read length = 300 bps
Illumina NextSeq avg read length = 150 bps

Coverage

Should as a minimum be 30x, and preferably even higher

Assembled data QC

Size of assembled genome

Enterobacterales: 4.5 Mb - 5.5 Mb
Deviation should not be higher than 10%

Total number of contigs

Should be less than 500

N50

Should be over 15.000 bp

$$Coverage = Number\ of\ reads\ x\ \frac{Read\ length}{Genome\ size}$$

Data QC

- **Number of contigs**

Is how many contigs (long sequences) were created during the assembly from good-quality raw reads. A low number of contigs means that the sequencing process was good enough to capture most of the genome and combine the raw data into long, uninterrupted sequences of nucleotides.

- **N50**

It’s a parameter that describes the length of all contigs that compose a genome.

- **Depth of coverage of sequenced genome**

Is how many times each bp present in the assembly was sequenced.

- **Genome size**

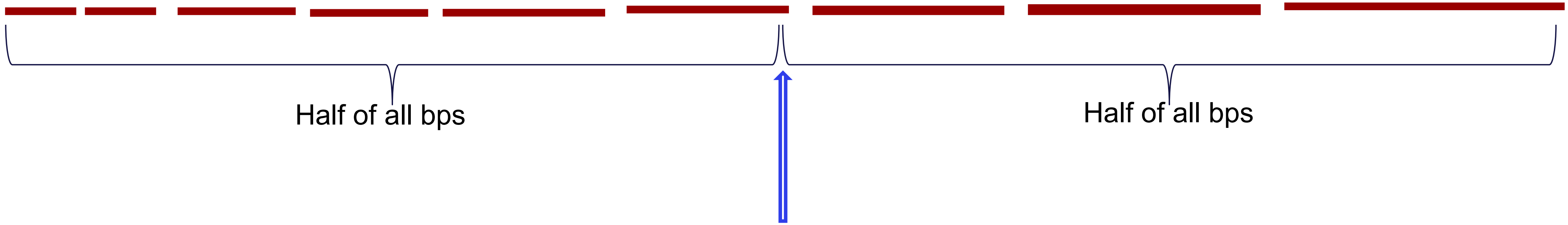
Is the number of individual bp that compose the assembled genome.

Data QC

- **N50**

It’s a parameter that describes the length of all contigs that compose a genome.

1) All contigs organized by size
2) Divide the total base-pairs in half
3) The contig that you “catch” has a certain length (bp) → that’s the N50



N50: the length of this contig (bps)

Genome sizes

Gram-positive expected genome size: 1.6 - 3 Mbp

Gram-negative expected genome size: 1.7 – 7 Mbp

Genera or group	Expected genome size (million bps)
Acinetobacter	5
Actinomyces	2.2 – 3
Aerococcus	1.6 – 2
Aeromonas	4.5
Anaerococcus	2
Bacteroides	5
Campylobacter	1.7
Clostridium	4.2 – 5
Corynebacterium	2.5 – 3
Enterobacterales (excluding *Proteus*)	4.5 – 5.5
Proteus	4
Enterococcus	3
Finegoldia	1.6 – 2
Fusobacterium	2
Haemophilus	1.8 – 2
Micrococcus	2.5
Moraxella	1.8 – 2
Neisseria	2
Pasteurella	2 – 2.2
Peptoniphilus	1.6 – 1.9
Prevotella	3 – 4
Propionibacterium	2 – 2.5
Pseudomonas	6.5 – 7
Rothia	2
Staphylococcus	2.5 – 2.8
Stenotrophomonas	4.5 – 5
Streptococcus	1.7 – 2.2

Troubleshooting

Usually poor **raw data** QC indicates:

- Inadequate DNA extraction
- Inadequate library preparation

Usually poor **assembly** QC indicates:

- Inadequate DNA extraction
- Contaminations

Re-sequence or re-extract?

Evaluation of QC becomes easier with experience + understanding the biochemical principles of the protocols.

QC of bioinformatics analysis

Thresholds for analysis and interpretation depend on the bioinformatics tools

In general:

- Be familiar with the recomended thresholds of each tool
- Use relevant control strategies
- Be critical when evaluating the results

Examples of recommended thresholds

Species identification with rMLST:

- at least 96% of support and absence of hits belonging to different species

Prediction of antimicrobial resistance with AMRFinder:

- minimum 90% identity and minimum 60% length

Thresholds for other tools and purposes:

○ EURGen-RefLabCap WGS protocol

○ Publication by the authors of the tool

○ Publications by other professionals using the tool

Examples of relevant control strategies

When analysing **isolates independently** (for example detecting AMR determinants in isolates):

○ Use **control strains** that harbour the same or similar genetic determinants you want to find

When analysing **isolates together** (for example performing cluster analysis):

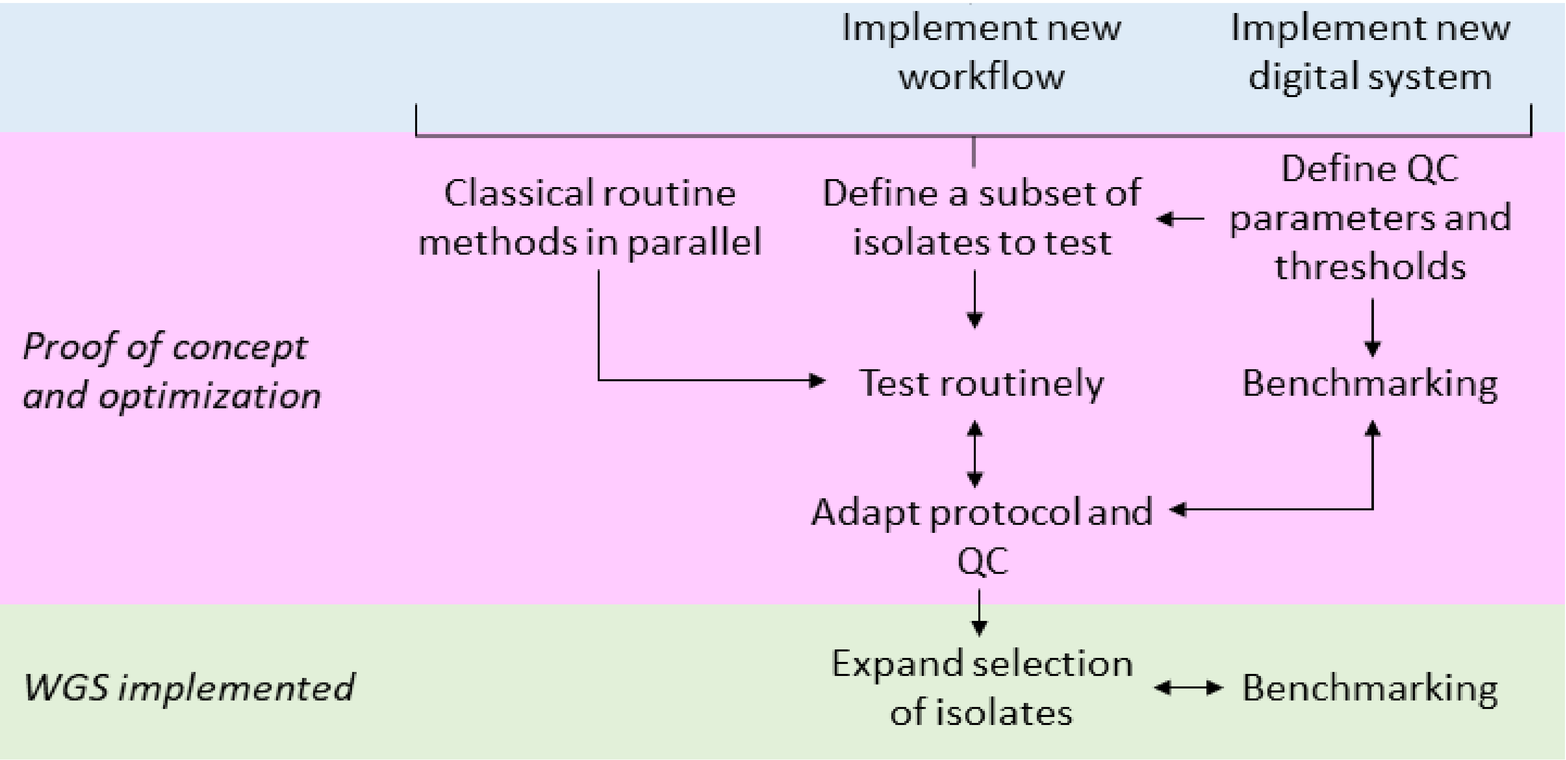
○ Use **groups of isolates** with well-established genetic relatedness

Examples of nonsense results

Why are these specific cases nonsense?

- Species identification tool detects a high proportion of hits belonging to different species
- Different *bla* genes in the same position of the genome with <100% coverage
- Lack of known AMR determinants when there is phenotypic resistance
- Zero SNPs between isolates very separated according to metadata (time/space)

Benchmarking in your settings


Implement new workflow
Implement new digital system
Classical routine methods in parallel
Define a subset of isolates to test
Define QC parameters and thresholds
Proof of concept and optimization
Test routinely
Benchmarking
Adapt protocol and QC
WGS implemented
Expand selection of isolates
Benchmarking

Ana Rita Rebelo

anrire@food.dtu.dk

Questions and wrapping up the day

Funded by the European Union

Updated agenda for next days

Second day (physical) – Wednesday 7 December 2022

9:00 - 9:30: Introduction and agenda for the day (Rene S. Hendriksen, DTU)

9:30 - 15:00: Laboratory work - Illumina MiSeq library preparation and sequencing (including *ad hoc* coffee, snacks and lunch)

15:00 - 15:30: Coffee break

15:30 - 16:15: Exercise about quality control of WGS data (Ana Rita Rebelo, DTU)

16:15 - 17:00: Exercise about bioinformatics tools for species identification and serotyping (Jette Sejer Kjeldgaard, DTU)

Funded by the European Union

DTU

STATENS SERUM INSTITUT

EURGen-RefLabCap@food.dtu.dk

Thank you on behalf of the EURGen-RefLabCap team

Funded by the European Union