Introduction à la Bio-informatique

Emmanuel Lokilo, Ir

Mars 2023

Plan

- Définition
- Compétences
- Tâches
- Séquençage

Définition

La bio-informatique est une science à l'interface des disciplines numériques (l'informatique et les mathématiques) et des sciences de la vie (biochimie, biologie, microbiologie, écologie, épidémiologie). Étant donné que les scientifiques de la vie génèrent une quantité croissante de nouvelles données portant sur les génomes, les biomolécules, les organismes, leurs interactions et leur évolution, il y a un besoin croissant d'approches informatiques pour la manipulation, le stockage, la visualisation et l'analyse de ces données souvent très complexes.

« Domaine interdisciplinaire, situé au carrefour de l'informatique, des mathématiques et de la biologie, qui traite de l'application de l'informatique aux sciences biologiques.

La bio-informatique est un vaste domaine qui recouvre l'ensemble des utilisations de l'informatique pour la gestion, l'entreposage, l'analyse, le traitement, l'organisation, la comparaison et la diffusion de données relatives à l'ensemble des sciences biologiques (physiologie, écologie, biochimie, biologie moléculaire et, dans une large mesure génétique et génomique). »(OLF, 2001)

Également, la bio-informatique joue un rôle important pour la recherche biomédicale. Les travaux sur les maladies génétiques et la génomique médicale sont en pleine croissance et l'avenir d'une médecine personnalisée dépend des approches de la bio-informatique.

Compétences

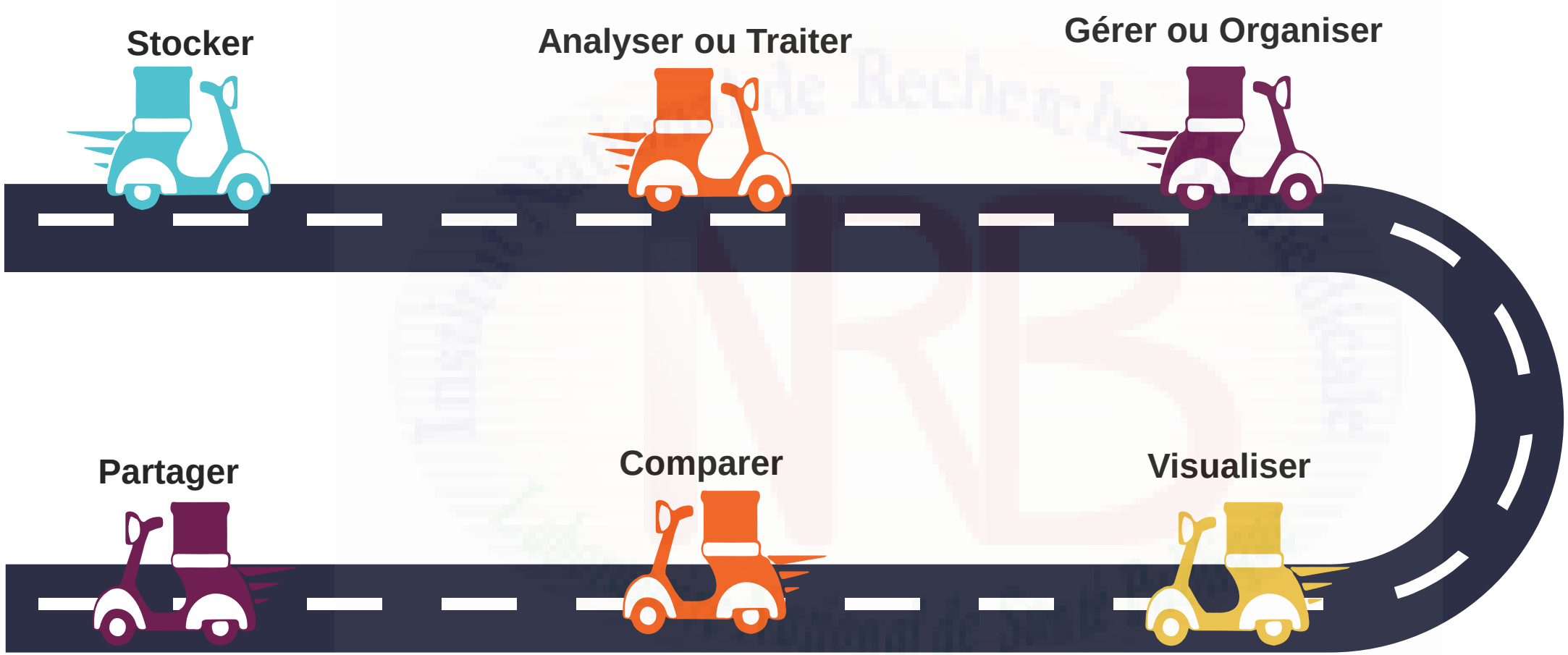
- Informatique

Web Services
Cloud Services (Nextcloud, OwnCloud, etc.)
SGBD (SQL, MySQL, etc.)
Files Sharing (FTP, etc.)
Programmation (Linux, perl, bash, java, javascript, etc.)
GIS and Remote Sensing

- Science de la vie

Biologie cellulaire et génétique
Biochimie et biologie moléculaire
Génomique
Protéomique
Phylogénétique

Tâches


Stocker
Analyser ou Traiter
Gérer ou Organiser
Partager
Comparer
Visualiser

Séquençage

Définition

En biochimie, le séquençage consiste à déterminer l'ordre linéaire des composants d'une macromolécule (les acides aminés d'une protéine, les nucléotides d'un acide nucléique comme l'ADN, les monosaccharides d'un polysaccharide, etc.).

En génétique, le séquençage concerne la détermination de la séquence des gènes voire des chromosomes, voire du génome complet, ce qui techniquement revient à effectuer le séquençage de l'ADN constituant ces gènes ou ces chromosomes.

« Le séquençage de l'ADN, consiste à déterminer l'ordre d'enchaînement des nucléotides d'un fragment d'ADN donné. »

La séquence d'ADN contient l'information nécessaire aux êtres vivants pour survivre et se reproduire

https://www.techno-science.net/glossaire-definition/Sequencage.html#:~:text=En%20biochimie%2C%20le,et%20se%20reproduire

Matériels et logiciel

1. Séquenceur

For MinION / GridION

Flongle

Adapter to enable small, rapid nanopore sequencing tests, for mobile or desktop sequencers

MinION Mk1B

Your personal nanopore sequencer, putting you in control

MinION Mk1C

Your personal nanopore sequencer including compute and screen, putting you in control

GridION Mk1

Higher-throughput, on demand nanopore sequencing at the desktop, for you or as a service

PromethION 24/48

Ultra-high throughput, on-demand nanopore sequencing, for you or as a service

https://nanoporetech.com/products/comparison

https://nanoporetech.com/applications/dna-nanopore-Sequencing#gns[searchValue]=sequencing

Matériels et logiciel

2. Ordinateur

Le sequence ADN avec Oxford Nanopore MinION (ONT) exige une puissance de calcul considerable, souvent dû à l'intensif processus d'appel des bases. En plus, réaliser une analyse des données en temps reel demande une grande capacité de stockage pour une recherche rapide à travers une large base de données. Le minimum requis se retrouve dans le processeur Intel i7, 8Gb de RAM, et 512Gb de stockage. Améliorer les specifications à > 1Tb de disque de stockage, 16Gb de mémoire avec un GPU (Graphical Processing Unit) est très recommandé. Chaque séquenceur peut générer jusqu'à plus de 150Gb de données (rélatif à la durée du séquençage), et le minimum de 512Gb de stockage sera rempli dans 3-4 runs de séquençage tout au plus

Matériels et logiciel

2. Ordinateur

Laptop: Dell XPS 15

Processor: Intel i7-10750H (5Ghz x 6 cores)

Memory: 32Gb RAM

Storage: 1Tb solid state drive

GPU: Nvidia GTX 1650 Ti

Total cost: ~$2,400

Lenovo Legion 5

Processor: 2.6 GHz Intel Core i7-10750H Six-Core

Memory: 16GB DDR4 RAM

Storage: 1TB M.2 NVMe SSD

GPU: NVIDIA GeForce RTX 2060 (6GB GDDR6)

Total cost: ~$1,300

N.B:la plupart des logiciels d'analyses de sequences d'ADN sont libre source et marchent sur les systems UNIX.

Matériels et logiciel

2’. SSD

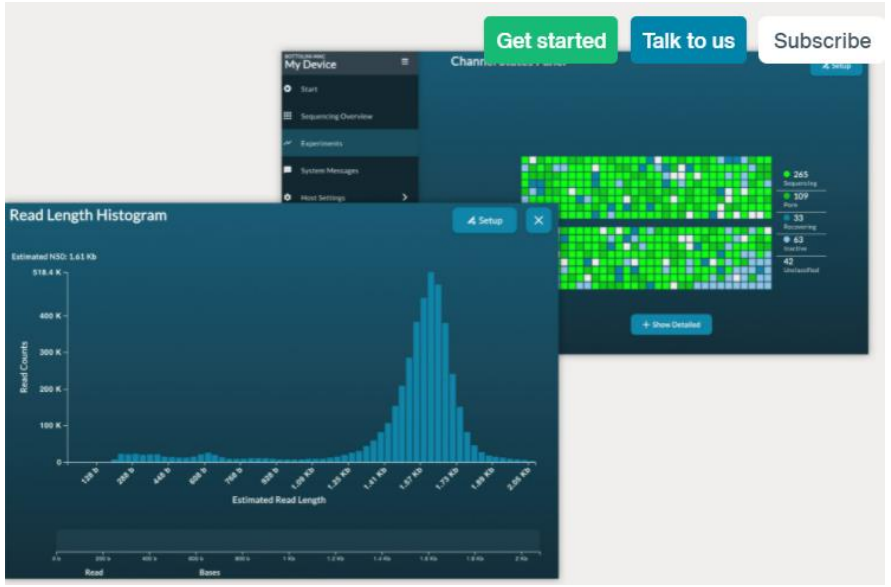
Matériels et logiciel

3. Wi-fi

https://fast.com/fr/

Matériels et logiciel

MinKNOW



Le séquençage Nanopore présente un certain nombre d'avantages significatifs qui permettent d'adapter le processus de séquençage à vos besoins :

- Basecalling en temps réel, permettant un accès immédiat aux résultats;
- Arrêtez le séquençage dès que des données suffisantes ont été obtenues;
- Arrêter, laver et réutiliser une Flow Cell ;
- Les appels de base à bord avec Guppy signifient que ni une infrastructure locale ni une connexion Internet stable ne sont nécessaires.

Le flux de travail de l'analyse du séquençage des nanopores est simple et facile à suivre : avec cinq étapes, de l'acquisition des données brutes à l'achèvement de l'analyse et à l'interprétation expérimentale. Dès le début de l'acquisition des données, l'analyse peut être effectuée en temps réel.

Matériels et logiciel

MinKNOW

Acquisition de données primaires avec MinKNOW

MinKNOW est le logiciel d'exploitation qui pilote les dispositifs de séquençage nanopore, effectue plusieurs tâches principales, notamment l'acquisition de données, l'analyse en temps réel et le retour d'expérience, l'appel de base local et la diffusion de données.

MinKNOW produit des fichiers FAST5 (HDF5) et/ou des fichiers FASTQ, selon vos préférences. Les fichiers FAST5 contiennent des données de signal brutes qui peuvent être utilisées pour l'appel de base.

Matériels et logiciel

MinKNOW

Appel de base

L'appel de base peut être défini comme le processus de conversion des signaux électriques générés par un brin d'ADN ou d'ARN traversant le nanopore en une séquence de bases correspondante du brin.

Un choix d'outils d'appel de base est disponible, dont certains sont entièrement pris en charge et d'autres en cours de développement. Guppy, un exemple du premier, est une boîte à outils de traitement de données qui contient les algorithmes d'appel de base d'Oxford Nanopore et plusieurs fonctionnalités de post-traitement bioinformatique, telles que le codage à barres/le démultiplexage, le rognage d'adaptateur et l'alignement.

La boîte à outils Guppy effectue également un appel de base modifié (5mC, 6mA et CpG) à partir des données de signal brutes, produisant un fichier FAST5 supplémentaire de probabilités de base modifiées.

Guppy est intégré à MinKNOW et est également disponible en version autonome.

Analyses bioinformatiques

1. Command lines

```
[root@localhost ~]# ping -q fa.wikipedia.org
PING text.pmtpa.wikimedia.org (208.80.152.2) 56(84) bytes of data.
^C
--- text.pmtpa.wikimedia.org ping statistics ---
1 packets transmitted, 1 received, 0% packet loss, time 0ms
rtt min/avg/max/mdev = 540.528/540.528/540.528/0.000 ms
[root@localhost ~]# pwd
/root
[root@localhost ~]# cd /var
[root@localhost var]# ls -la
total 72
drwxr-xr-x. 18 root root 4096 Jul 30 22:43 .
drwxr-xr-x. 23 root root 4096 Sep 14 20:42 ..
drwxr-xr-x.  2 root root 4096 May 14 00:15 account
drwxr-xr-x. 11 root root 4096 Jul 31 22:26 cache
drwxr-xr-x.  3 root root 4096 May 18 16:03 db
drwxr-xr-x.  3 root root 4096 May 18 16:03 empty
drwxr-xr-x.  2 root root 4096 May 18 16:03 games
drwxrwx--T.  2 root gdm  4096 Jun  2 18:39 gdm
drwxr-xr-x. 38 root root 4096 May 18 16:03 lib
drwxr-xr-x.  2 root root 4096 May 18 16:03 local
lrwxrwxrwx.  1 root root   11 May 14 00:12 lock -> ../run/lock
drwxr-xr-x. 14 root root 4096 Sep 14 20:42 log
lrwxrwxrwx.  1 root root   10 Jul 30 22:43 mail -> spool/mail
drwxr-xr-x.  2 root root 4096 May 18 16:03 nis
drwxr-xr-x.  2 root root 4096 May 18 16:03 opt
drwxr-xr-x.  2 root root 4096 May 18 16:03 preserve
drwxr-xr-x.  2 root root 4096 Jul  1 22:11 report
lrwxrwxrwx.  1 root root    6 May 14 00:12 run -> ../run
drwxr-xr-x. 14 root root 4096 May 18 16:03 spool
drwxrwxrwt.  4 root root 4096 Sep 12 23:50 tmp
drwxr-xr-x.  2 root root 4096 May 18 16:03 yp
[root@localhost var]# yum search wiki
Loaded plugins: langpacks, presto, refresh-packagekit, remove-with-leaves
rpmfusion-free-updates                                          | 2.7 kB     00:00
rpmfusion-free-updates/primary_db                               | 206 kB     00:04
rpmfusion-nonfree-updates                                       | 2.7 kB     00:00
updates/metalink                                                | 5.9 kB     00:00
updates                                                         | 4.7 kB     00:00
updates/primary_db                    73% [=======================         ]  62 kB/s | 2.6 MB     00:15 ETA
```

Analyses bioinformatiques

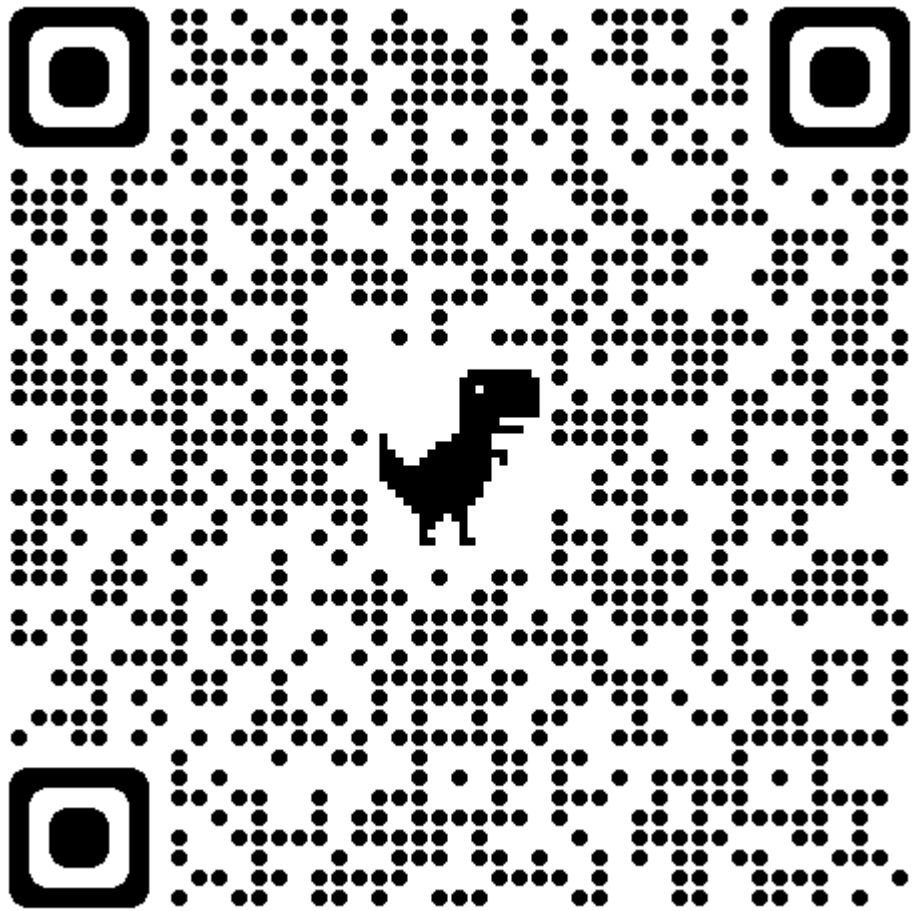
2. EPI2ME

Une gamme d'approches est disponible pour l'analyse en aval des données de séquençage des nanopores, afin de répondre à toutes les exigences et à tous les niveaux d'expertise en bioinformatique.

Oxford Nanopore propose EPI2ME, une plate-forme d'analyse basée sur le cloud fournissant des flux de travail d'analyse en temps réel, sans aucune expérience en ligne de commande requise. Les ordinateurs portables EPI2ME Labs fournissent des flux de travail d'analyse post-exécution dans un format de didacticiel conçu pour développer les compétences et la confiance dans l'analyse et l'exploration des données de séquence de nanopores. Les flux de travail d'EPI2ME Labs automatisent le flux de données du didacticiel, pour permettre une analyse de séquence à haut débit et plus "sans intervention". La sortie de données standard des dispositifs de séquençage nanopore peut également être utilisée dans une variété de logiciels de recherche qui sont continuellement développés et publiés par les équipes d'Oxford Nanopore.

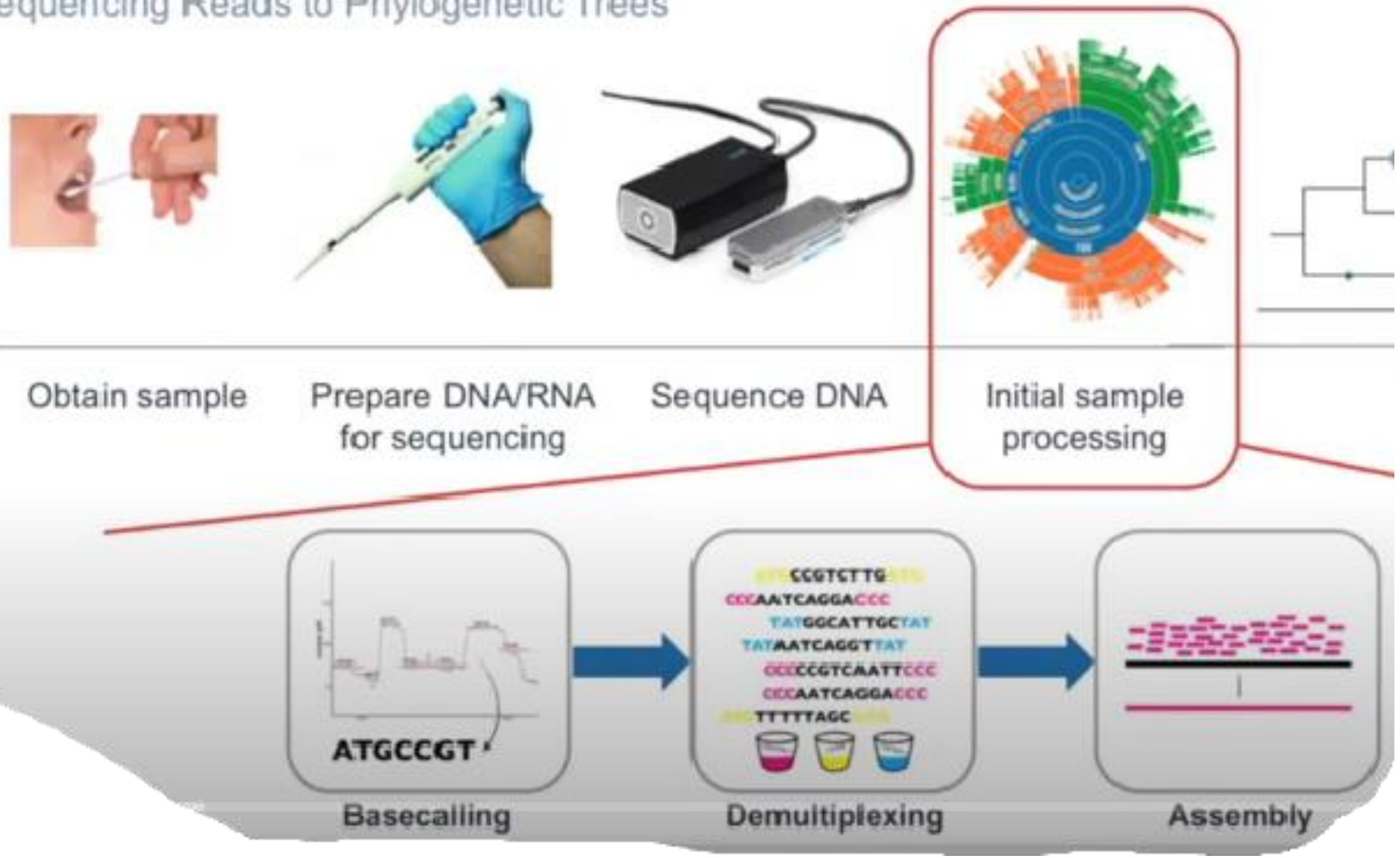
Enfin, une gamme d'outils développés par la Communauté sont disponibles, qui ont été développés par la communauté des utilisateurs pour une grande variété d'applications de recherche.

Principe

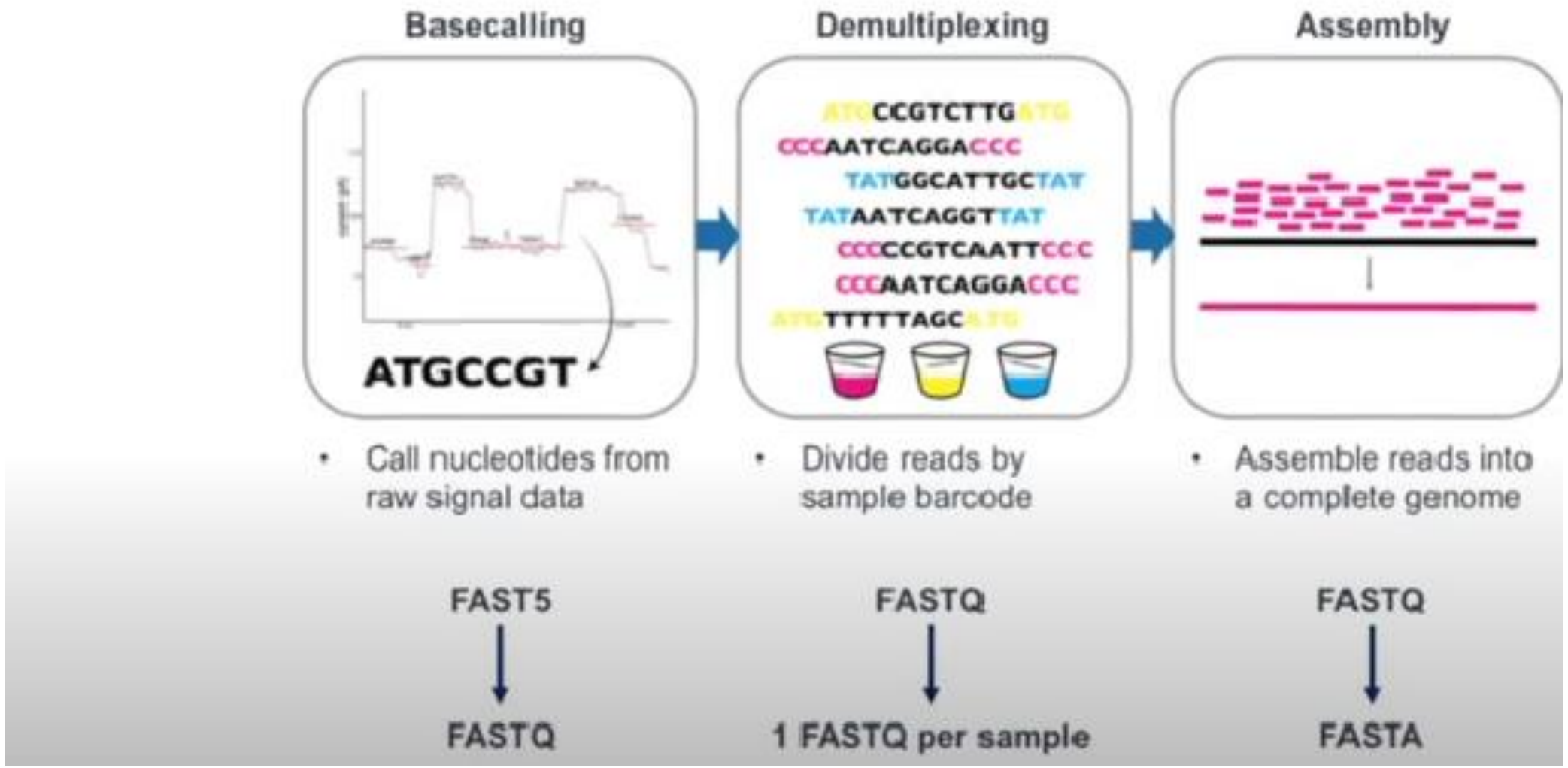


Overview

Sequencing Reads to Phylogenetic Trees

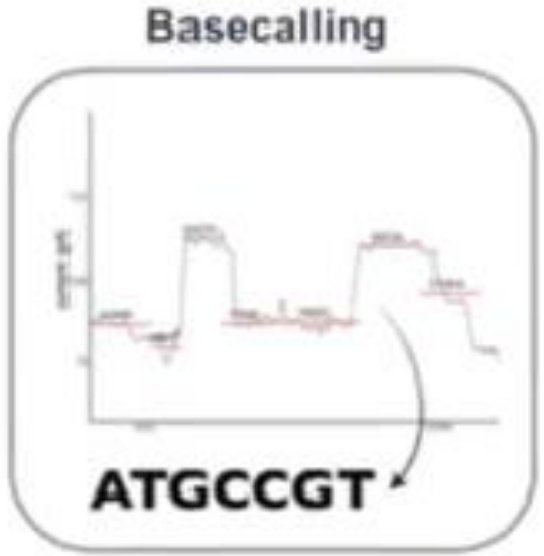


Initial Sample Processing



Basecalling

Understanding the FAST5 format



- Call nucleotides from raw signal data

FAST5

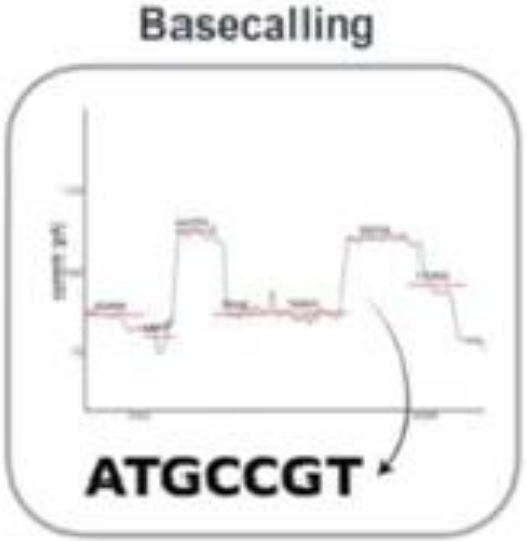
↓

FASTQ

The MinION sequencer produces files that end with **.fast5** – these contain raw signal data

Basecalling

Understanding the FAST5 format



- Call nucleotides from raw signal data

File name: FAL60025_86646155727949c32fbd860c1840059122e78671_128.**fast5**

```
<89>HDF
^Z
^@^@^@^@^@^H^H^@^D^@^P^@^@^@^@^@^@^@^@^@^@^@^@^@<FF><FF><F
F><FF><FF><FF><FF><FF><85><92><E2>^^^@^@^@^@<FF><FF><FF>
<FF><FF><FF><FF><FF>^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^A^@^@^
@^@^@^@^@<88>^@^@^@^@^@^@^@<A8>^B^@^@^@^@^@^@^@^A^@   ^@^A^@
^@^@^X^@^@^@^@^@^@^@^P^@^P^@^@^@^@^@^@<E0>^E^@^@^@^@^@^@P^@^
@^@^@^@^@^@TREE^@^@^@^@<FF><FF><FF><FF><FF><FF><FF><FF><FF>
<FF><FF><FF><FF><FF><FF><FF><FF><FF
```

These files are not human-readable — they look like gibberish!

FAST5

↓

FASTQ

The MinION sequencer produces files that end with **.fast5** – these contain raw signal data

Basecalling

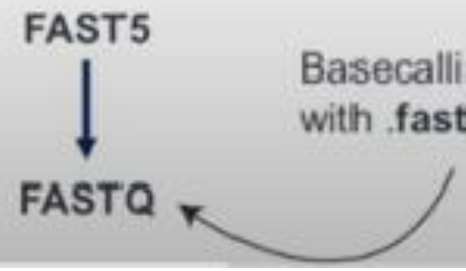
Understanding the FASTQ format

File name: sample1.**fastq**

Information for 1 read

```
@475b2a54-105d-438d-896d-ec3b37dadf2a runid=8664615 sampleid=Pool1 read=12341 ch=7
GCACCTCTTTCTTACTGCAGCACCGATTTCCCTCAGAAGAGGAGCAGACTCAACAATATCGGCATGTGCTGAACGCACTTAC
+
/:<CC<3AHEDLE2$&'*''&&)45357BCG)7B:8=5-4**%+,*4,/-45,5D:?@>1())>=<>=;<;22)&$$#$$49
@27bf0fae-6ea7-4a2f-a846-b5064ee98b25 runid=8664615 sampleid=Pool1 read=10231 ch=1
GCACCTTTCATGCTCAAGCTTGGCGCGTTAGCCACCGGATTGACCTGAATCGGTAGGTTTTGCTCTCTTCCAATGCCATACC
+
,49<749>95;@@3ACA0&&+5;26%'@A;0/.*,,-0-8AB47469;76<&+++-5&28<::5648;<%::7:7EH4332-
```

There are **4** lines of information for each read



Basecalling converts the .fast5 files into files that end with **.fastq** (or **.fastq.gz**) – these contain raw reads

Basecalling

Understanding the FASTQ format

Read identifier: starts with @

File name: sample1.**fastq**

```
@475b2a54-105d-438d-896d-ec3b37dadf2a runid=8664615 sampleid=Pool1 read=12341 ch=7
GCACCTCTTTCTTACTGCAGCACCGATTTCCCTCAGAAGAGGAGCAGACTCAACAATATCGGCATGTGCTGAACGCACTTAC
+
/:<CC<3AHEDLE2$&'*'*&&)45357BCG)7B:B=5-4**%+,*4,/-45,5D:?@>1())>=<>=;<;22)&$$#$$49
@27bf0fae-6ea7-4a2f-a846-b5064ee98b25 runid=8664615 sampleid=Pool1 read=10231 ch=1
GCACCTTTCATGCTCAAGCTTGGCGCGTTAGCCACCGGATTGACCTGAATCGGTAGGTTTTGCTCTCTTCCAATGCCATACC
+
,49<749>95;@@3ACAO&&+5;26%'@A;0/.*,,-0-8AB47469;76<&+++-5&28<::5648;<%::7:7EH4332-
```

There are 4 lines of information for each read

FAST5

↓

FASTQ

Basecalling converts the .fast5 files into files that end with .**fastq** (or .**fastq.gz**) – these contain raw reads

Basecalling

Understanding the FASTQ format

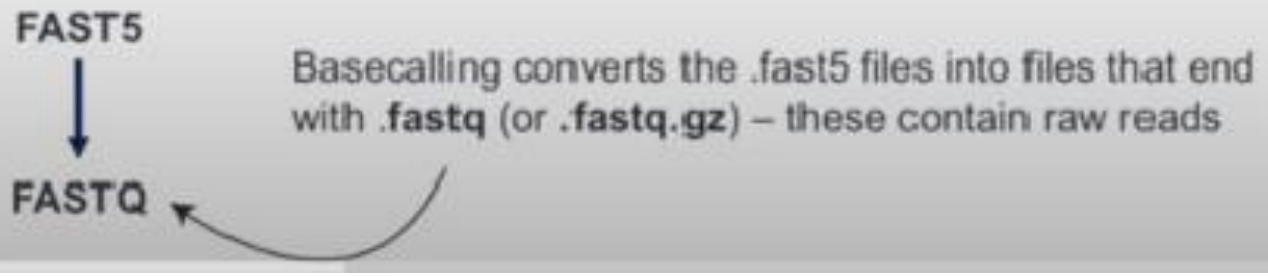
There are 4 lines of information for each read

Basecalling converts the .fast5 files into files that end with **.fastq** (or **.fastq.gz**) – these contain raw reads

Basecalling

Understanding the FASTQ format

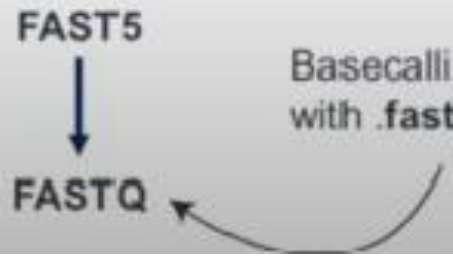
There are **4** lines of information for each read



Basecalling

Understanding the FASTQ format

There are 4 lines of information for each read

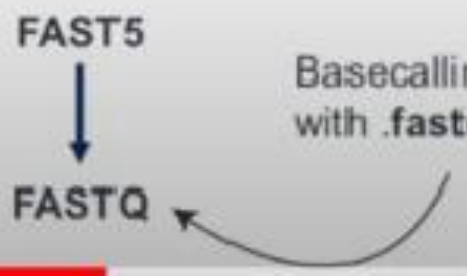


Basecalling converts the .fast5 files into files that end with **.fastq** (or **.fastq.gz**) – these contain raw reads

Basecalling

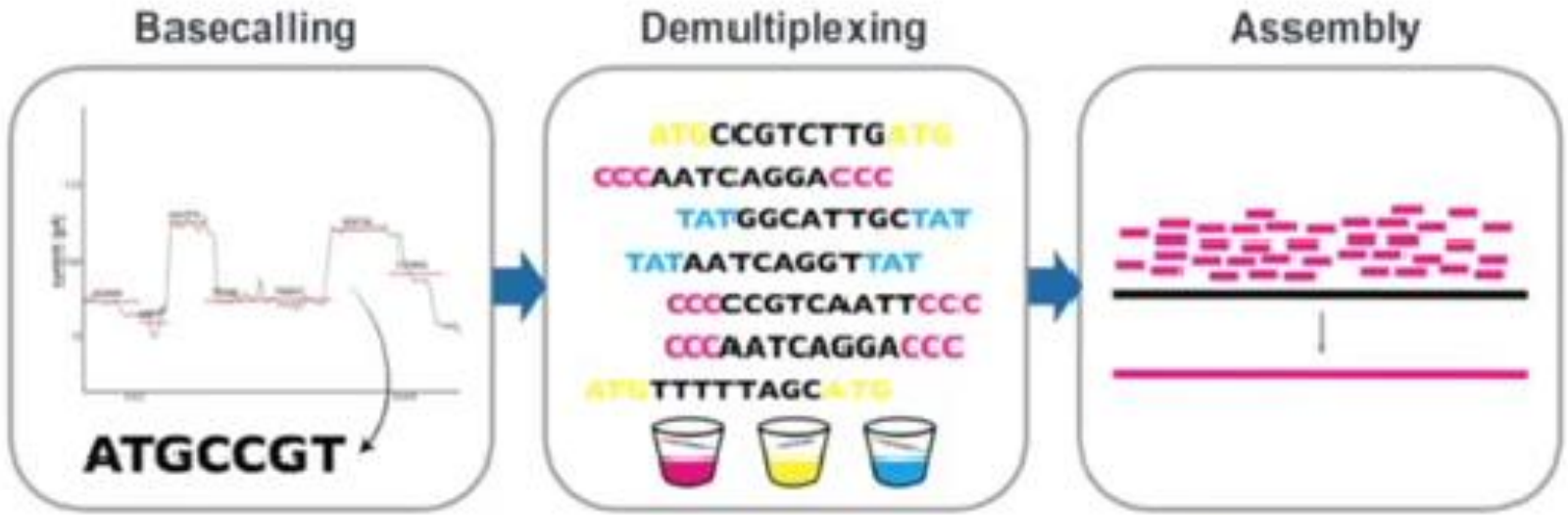
Requirements for calling nucleotides from raw signal data

- Basecalling software for Oxford Nanopore data: **Guppy**
- A computer with a GPU processor speeds up basecalling
  - Fast basecalling can also be done on the MinIT or GridION



Basecalling converts the .fast5 files into files that end with **.fastq** (or **.fastq.gz**) – these contain raw reads

Initial Sample Processing



- Call nucleotides from raw signal data
- Divide reads by sample barcode
- Assemble reads into a complete genome

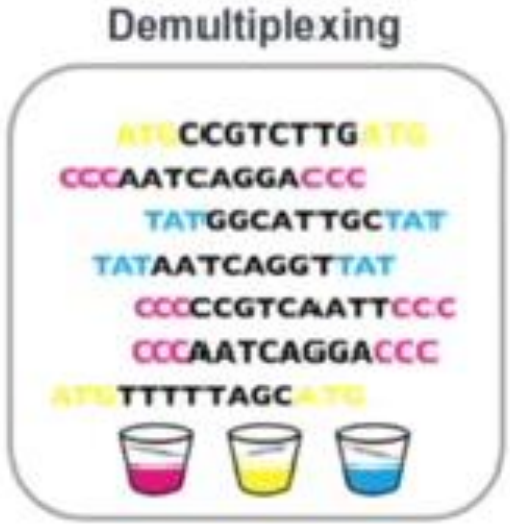
FAST5 → FASTQ

FASTQ → 1 FASTQ per sample

FASTQ → FASTA

Demultiplexing

Dividing reads by sample barcode



- Divide reads by sample barcode

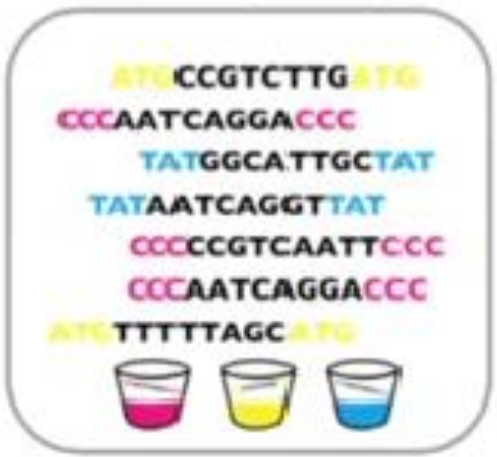
FASTQ

↓

1 FASTQ per sample

Demultiplexing

Dividing reads by sample barcode



- Divide reads by sample barcode

- Each read will contain some sequence corresponding to a barcode e.g. NB01
- All reads with the same barcode are put into one file

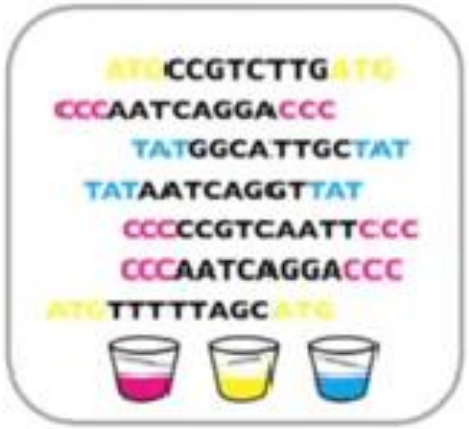
FASTQ

↓

1 FASTQ per sample

Demultiplexing

Dividing reads by sample barcode



- Divide reads by sample barcode

FASTQ

↓

1 FASTQ per sample

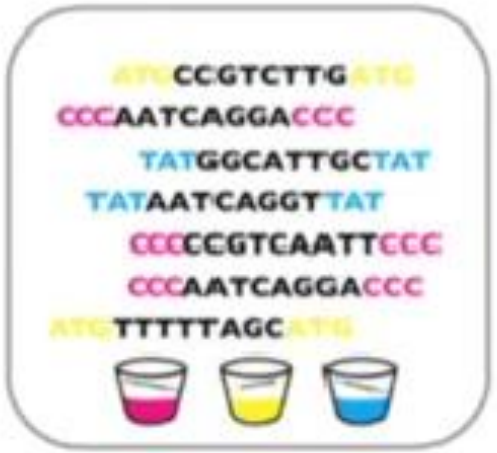
- Each read will contain some sequence corresponding to a barcode e.g. NB01
- All reads with the same barcode are put into one file

barcode01.fastq (yellow barcode)

barcode02.fastq (green b

Demultiplexing

Dividing reads by sample barcode



- Your **sample sheet** indicates which barcode was given to each sample:

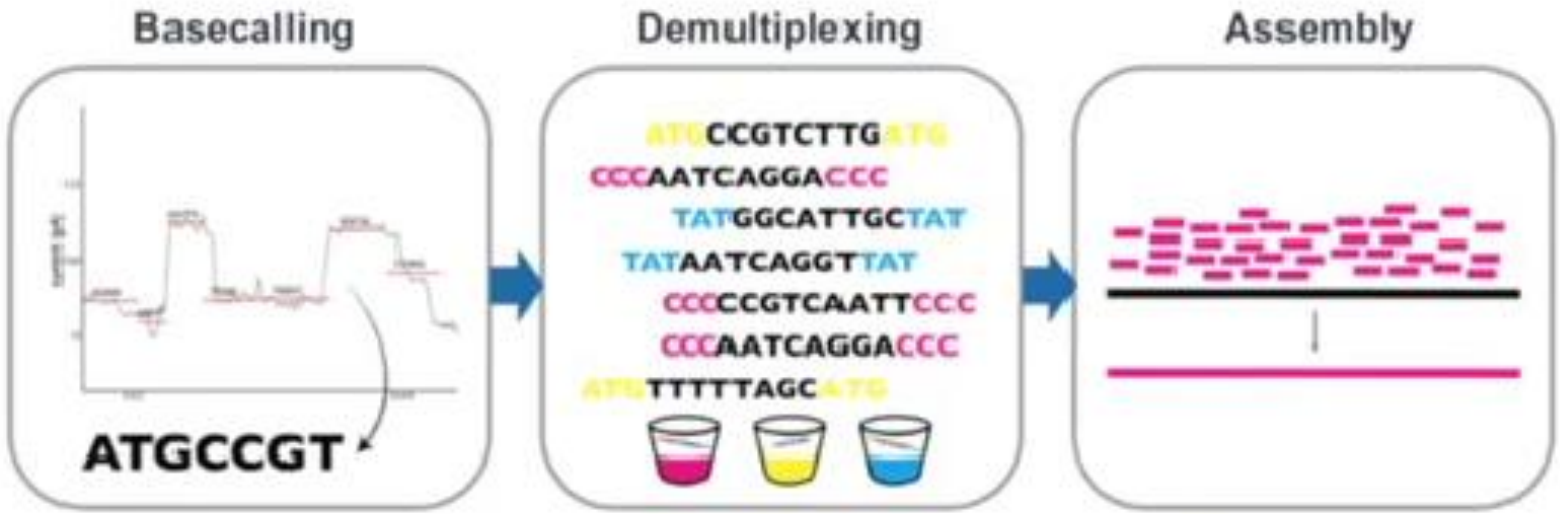
sample1	NB01
sample2	NB02
sample3	NB08
sample4	NB09
…	

← Text file containing this information

The demultiplexing software knows the specific sequence of each barcode: e.g., NB01 = GGCCAGTC

- After demultiplexing you should have 1 FASTQ file per barcode on your sample sheet:
  - barcode01.fastq
  - barcode02.fastq
  - barcode03.fastq
  - barcode04.fastq

Initial Sample Processing

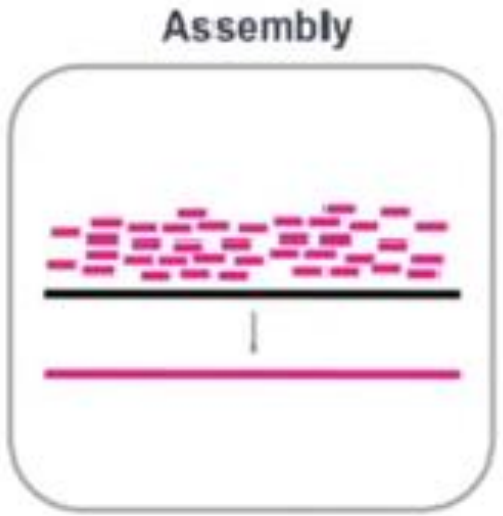


- Call nucleotides from raw signal data
- Divide reads by sample barcode
- Assemble reads into a complete genome

FAST5	FASTQ	FASTQ
↓	↓	↓
FASTQ	1 FASTQ per sample	FASTA

Assembly

Assembling reads into a complete genome



- Assemble reads into a complete genome

The first line of a sequence starts with >

File name: sample1.**fasta**

```
>sample1
AGGGTCATTAAATATATATAAAGATCTATATAGAGATCTTTTTATTAGATCTACTATTAA
GGAGCAGGATCTTTGTGGATAAGTGAAAAATGATCAACAAGATCATGCGATTCAGAAGGA
TCAGATCGTGTGATCAACCACTGATCTGTTCAAGGATTAGCTGGGATCAAAAACCTATGT
TATACACAGCCACCTTGGGATCTAAAACTTGTTATATGGATAACTATAGGAAGATCACCG
GATAATCGTATAGTTATCCACATGAGATTTGATTGAAAAAGCATCAATCAATTTTTCAC
TACCGTTAAATTTATCCACAATCCNAAAAAAAAGAGCGGCATTAAGCCGCTCTGCATGGAA
TAGGTCATTATTTAGAAGCGATTGATGACGCGTTTGAGCCAAGCTTCAGCGGCATCTTCA
GGCACTGGGTGCTCTTGTACATCGATGGTAAAGCAGTTGGCCAGAGGTTTAGCACCAATA
```

- FASTA files contain only the sequence name and nucleotide sequence
- All of the individual reads have been put together into a complete genome

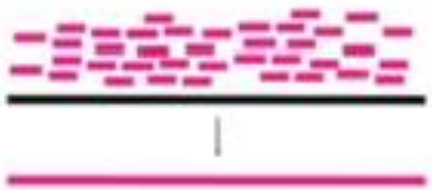
Assembly converts .fastq files into **.fasta** files

Assembly

Assembling reads into a complete genome

- Two types of genome assembly:

Reference-based assembly



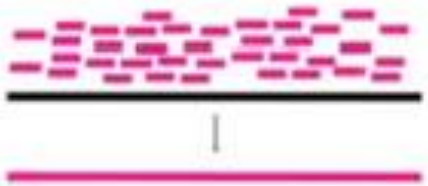
- Requires selection of an appropriate reference genome
- Each read is aligned to the correct position along the reference

Assembly

Assembling reads into a complete genome

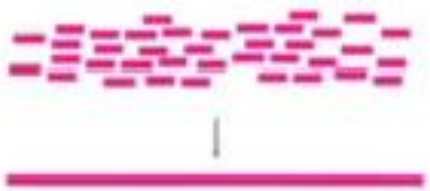
- Two types of genome assembly:

Reference-based assembly



- Requires selection of an appropriate reference genome
- Each read is aligned to the correct position along the reference

De novo assembly



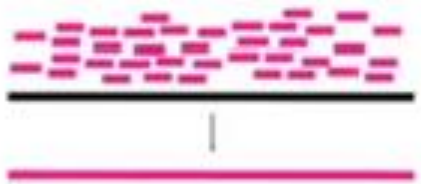
- Does not require a reference genome
- Useful when infectious pathogen is unknown
- Overlapping parts of reads are used to create genome

Assembly

Assembling reads into a complete genome

- Two types of genome assembly:

Reference-based assembly



- Requires selection of an appropriate reference genome
- Each read is aligned to the correct position along the reference
- SARS-CoV-2 reference genome: **MN908947.3**

De novo assembly

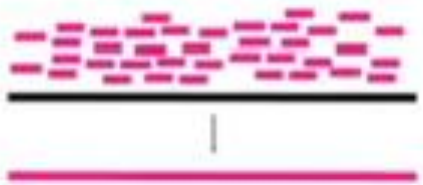
- Does not require a reference genome
- Useful when infectious pathogen is unknown
- Overlapping parts of reads are used to create genome

Assembly

Genome assembly from tiled amplicons

- Two types of genome assembly:

Reference-based assembly



- Requires selection of an appropriate reference genome
- Each read is aligned to the correct position along the reference
- SARS-CoV-2 reference genome: **MN908947.3**

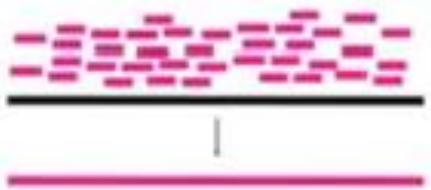
(1) Filter out reads that are not the appropriate length (SARS-CoV-2 amplicons are 400-700 bp)

Assembly

Genome assembly from tiled amplicons

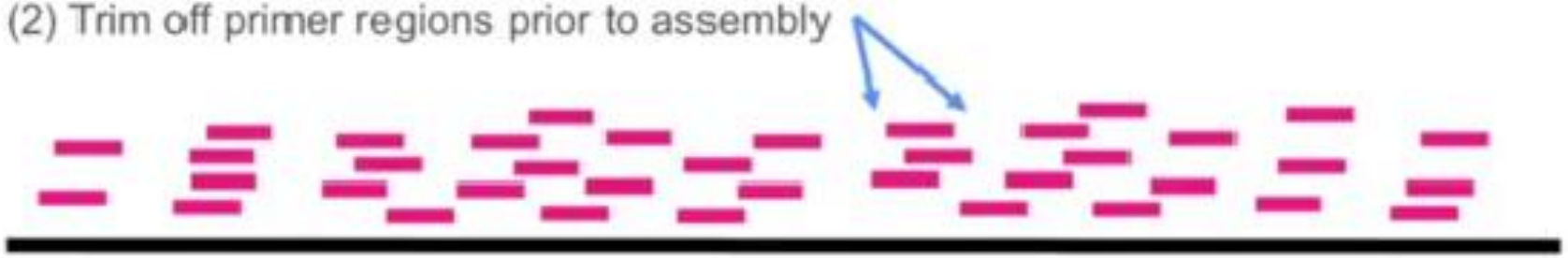
- Two types of genome assembly:

Reference-based assembly



- Requires selection of an appropriate reference genome
- Each read is aligned to the correct position along the reference
- SARS-CoV-2 reference genome: **MN908947.3**

(2) Trim off primer regions prior to assembly

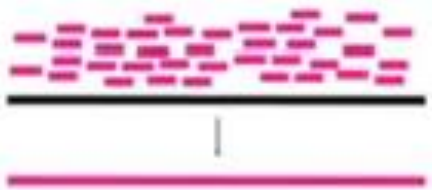


Assembly

Genome assembly from tiled amplicons

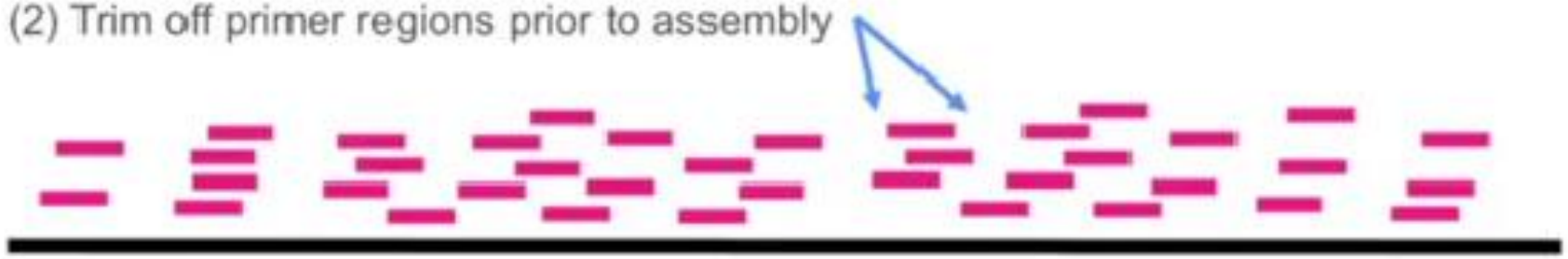
- Two types of genome assembly:

Reference-based assembly

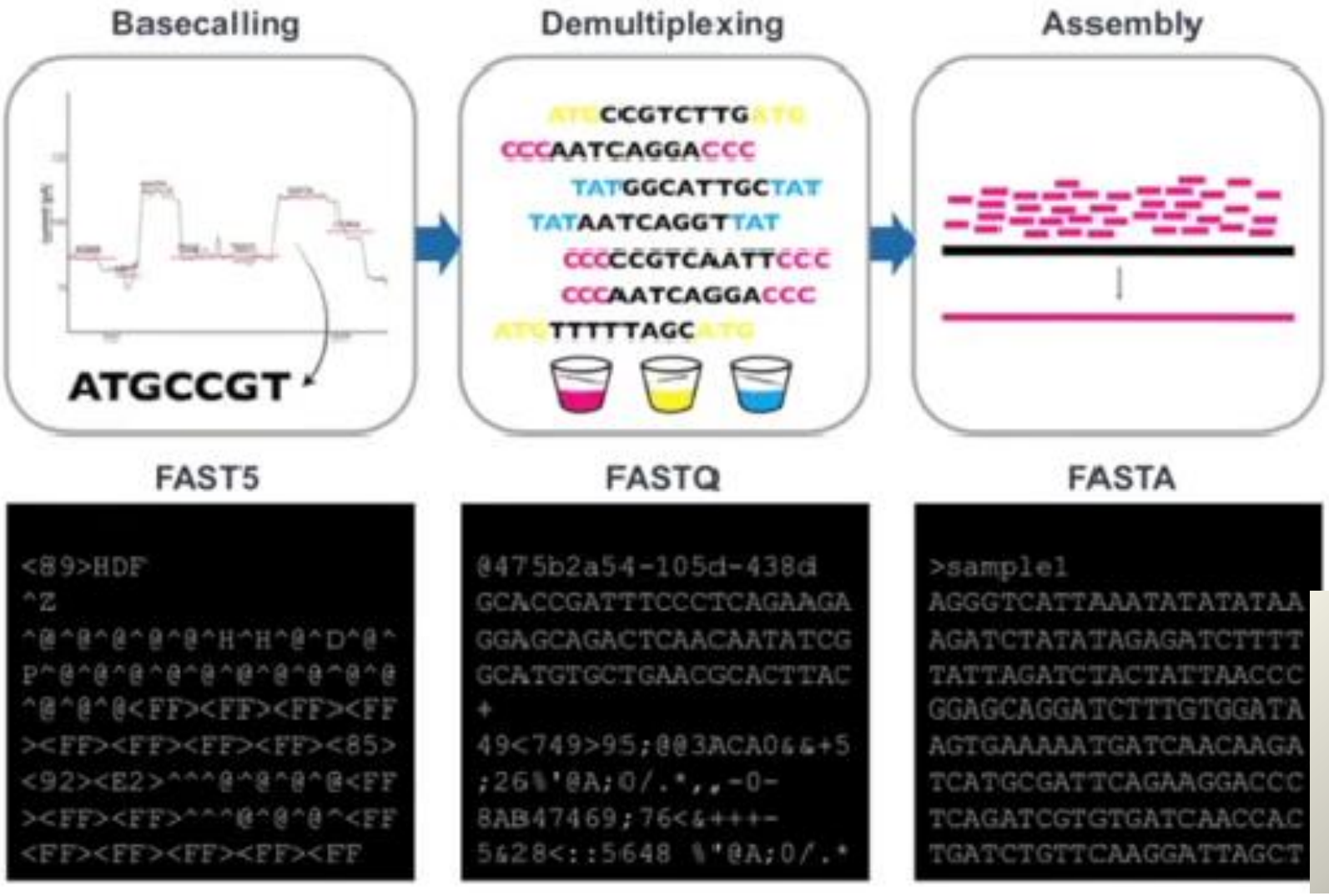


- Requires selection of an appropriate reference genome
- Each read is aligned to the correct position along the reference
- SARS-CoV-2 reference genome: **MN908947.3**

(2) Trim off primer regions prior to assembly



Recap: Initial Sample Processing



coverture

Merci