GENOMIC
PARIS CENTRE

One year of developments and collaborations around the MinION on the Genomic facility of the IBENS.

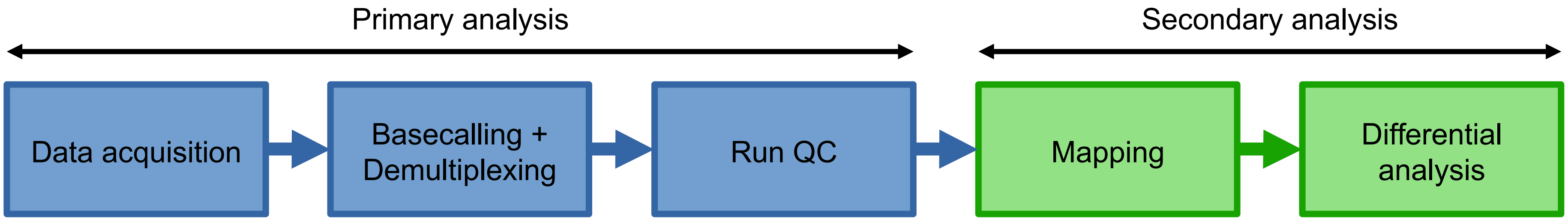
Laurent Jourdren (CNRS – IBENS)
Sophie Lemoine (CNRS – IBENS)
Bérengère Laffay (CNRS – IBENS)

Génoscope, Évry

December 13th, 2017

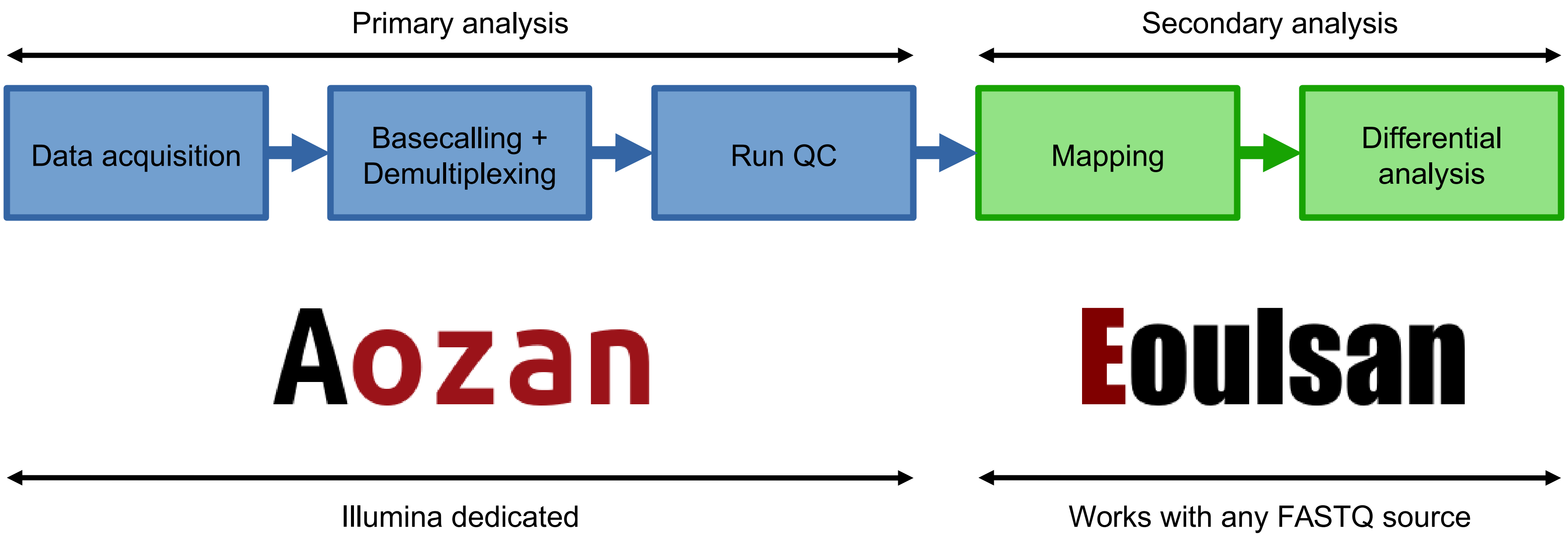
ONT analysis workflow

- Our aim is to develop a **RNA-Seq pipeline** from raw Nanopore data to differential analysis.



ONT analysis workflow

- Our aim is to develop a **RNA-Seq pipeline** from raw Nanopore data to differential analysis.
- Our current pipelines have been developed for Illumina data



Primary analysis

Secondary analysis

Data acquisition → Basecalling + Demultiplexing → Run QC → Mapping → Differential analysis

Aozan

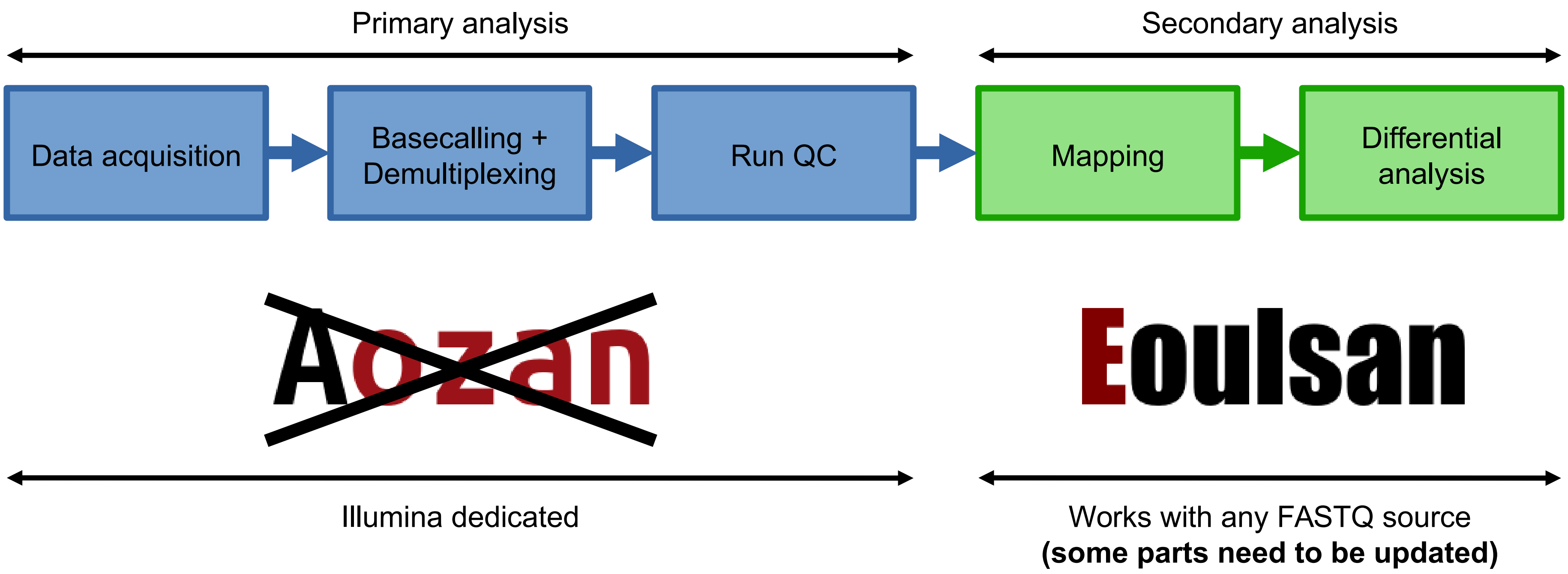
Eoulsan

Illumina dedicated

Works with any FASTQ source

ONT analysis workflow

- Our aim is to develop a **RNA-Seq pipeline** from raw Nanopore data to differential analysis.
- Our current pipelines have been developed for Illunina data



Primary analysis

Secondary analysis

Data acquisition → Basecalling + Demultiplexing → Run QC → Mapping → Differential analysis

~~Aozan~~

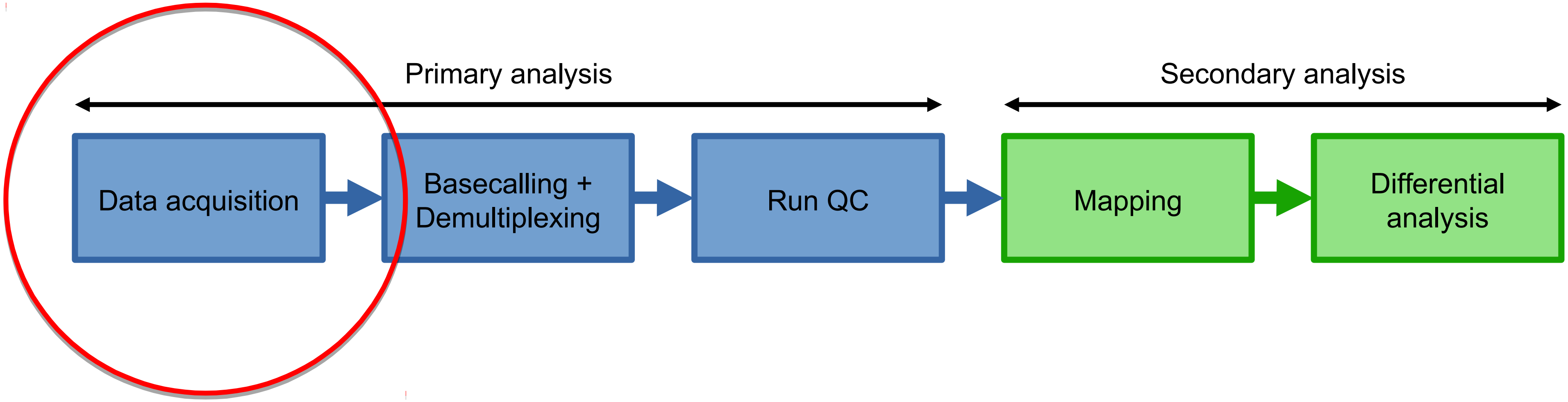
Eoulsan

Illumina dedicated

Works with any FASTQ source
(some parts need to be updated)

- We need to develop a **new post-sequencing pipeline** that will run on a **new dedicated infrastructure.**

Data acquisition



Data acquisition

- Data acquisition is performed using **MinKNOWN.**

- Use the **Linux** version of MinKNOW to avoid issues with anti-virus software that can stop runs.
- **Ubuntu 14.04 LTS** is the only Linux distribution officially supported by ONT.

- Our recommended hardware configuration:
  - **2 TB SSD** hard drive (ideally in RAID 1)
  - **32 GB** RAM (64GB for online basecalling)
- Create a large /var partition (where FAST5 files are stored)

- Connect your computer to a **UPS** to avoid power supply fail during the run.

MinKNOW updates

- New versions published **every 2 months.**

- New versions are often bugged especially the new major releases.

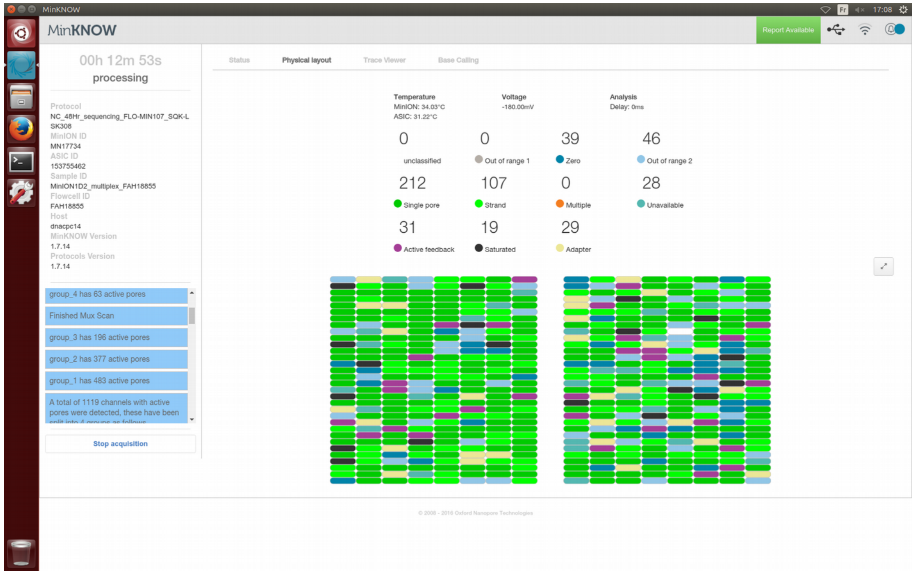
- **ONT do not provide access to previous versions**. “Customer shall install patches or new releases released by Oxford within one month after release”.

- We develop a script that **dump the ONT Ubuntu package repository** to be able to resinstall previous version of MinKNOWN.

- The script is not yet on GitHub but conctact us if you want it.

MinKNOW usage

- MinKNOW is a client/server software.
- Press F5 to refresh the client (a web browser interface).
- **Restart the computer before each new run** because it seems that the MinKNOW server part do not release all memory after a completed run.



00h 12m 53s
processing

Protocol
NC_48Hr_sequencing_FLO-MIN107_SQK-LSK308
MinION ID
MN17734
ASIC ID
153755462
Sample ID
MinION1D2_multiplex_FAH18855
Flowcell ID
FAH18855
Host
dnacpc14
MinKNOW Version
1.7.14
Protocols Version
1.7.14

group_4 has 63 active pores

Finished Mux Scan

group_3 has 196 active pores

group_2 has 377 active pores

group_1 has 483 active pores

A total of 1119 channels with active pores were detected, these have been split into 4 groups as follows

Temperature
MinION: 34.03°C
ASIC: 31.22°C

Voltage
-180.00mV

Analysis
Delay: 0ms

0	0	39	46
unclassified	● Out of range 1	● Zero	● Out of range 2
212	107	0	28
● Single pore	● Strand	● Multiple	● Unavailable
31	19	29	
● Active feedback	● Saturated	● Adapter	

MinKNOW data output transfer

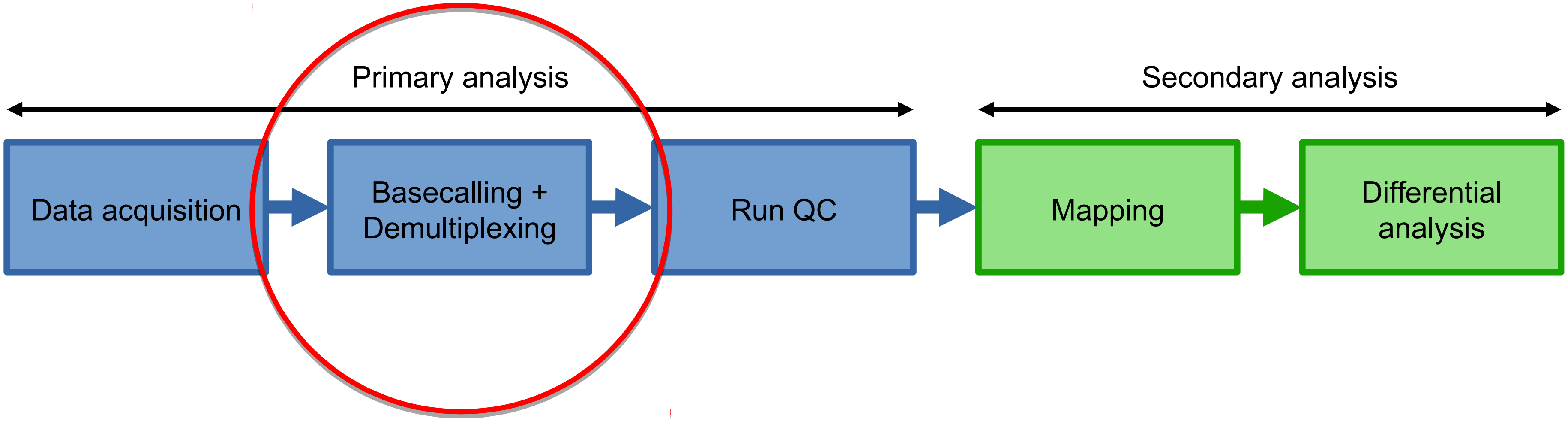
- MinKNOW creates **one FAST5 file for each read**.

- So for RNA-Seq up to **10,000,000 FAST5 files** are created for each run.

- The best solution to quickly copy/move your FAST5 files is to pack them in a **TAR archive**.

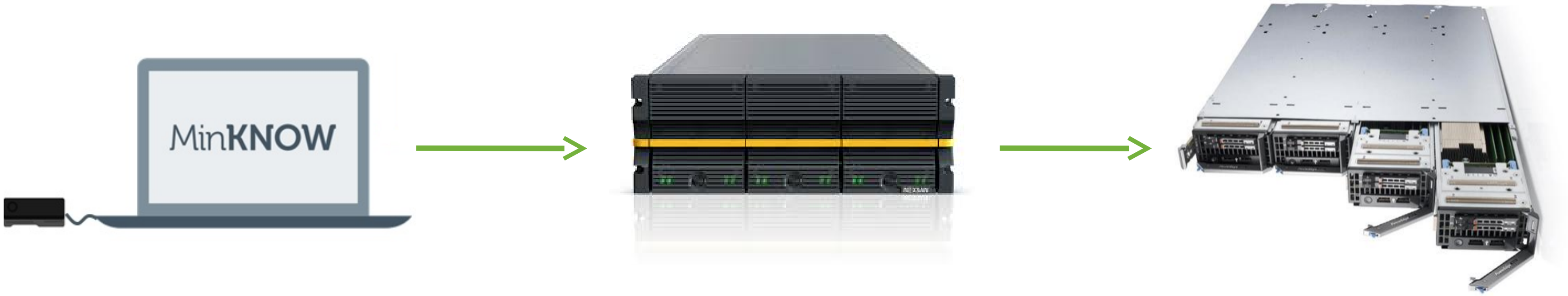
- You can also use Caltech's **bbcp** to use all the bandwidth of your WAN to transfert the data.

Basecalling and demultiplexing



Basecalling and demultiplexing hardware infrastructure

- Challenge: handle a **huge amount of small files and long computation time.**
- With the IBENS IT service, we built an **efficient and reliable infrastructure** to handle and process Nanopore Data.
- We developed a **tool to automatically launch data transfer and basecalling** once a run has finished.



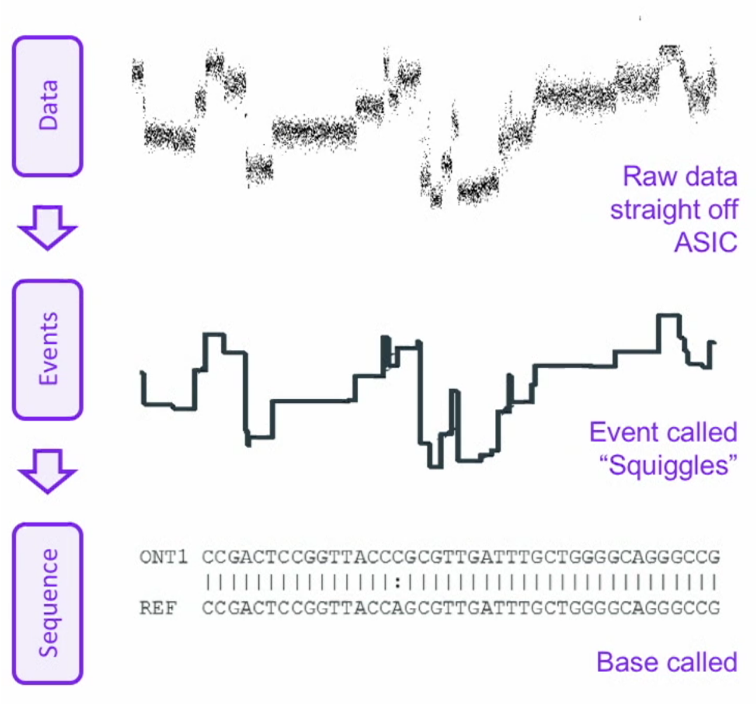
Acquisition
RAID 1 + UPS

Storage
85 TB

Processing
6x 16 cores - 196 GB

Raw data processing

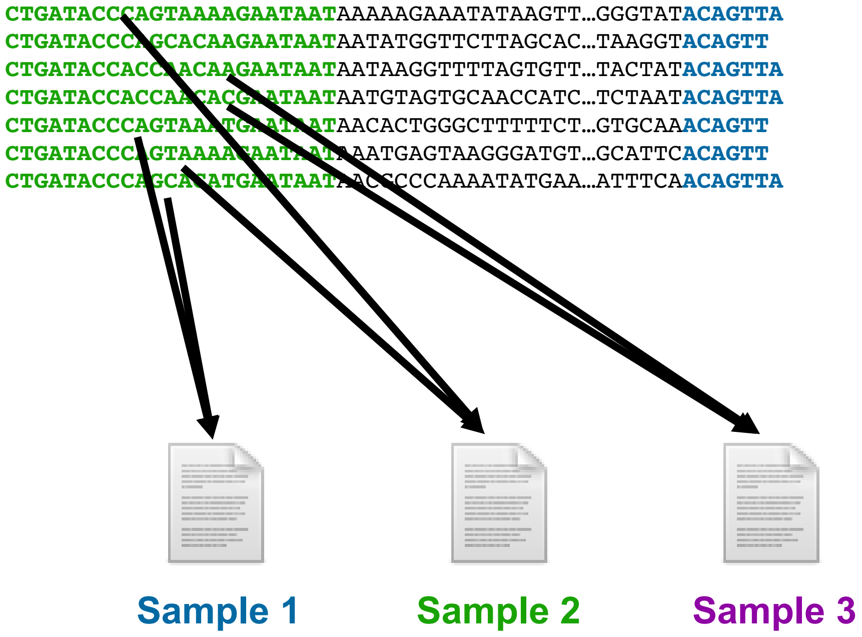
Basecalling



```
ONT1 CCGACTCCGGTTACCCGCGTTGATTTGCTGGGGCAGGGCCG
     |||||||||||||||:|||||||||||||||||||||||||
REF  CCGACTCCGGTTACCAGCGTTGATTTGCTGGGGCAGGGCCG
```

https://nanoporetech.com/

Demultiplexing



```
CTGATACCCAGTAAAAGAATAATAAAAAGAAATATAAGTT…GGGTATACAGTTA
CTGATACCCAGCACAAGAATAATAATATGGTTCTTAGCAC…TAAGGTACAGTT
CTGATACCACCAACAAGAATAATAATAAGGTTTTAGTGTT…TACTATACAGTTA
CTGATACCACCAACACGAATAATAATGTAGTGCAACCATC…TCTAATACAGTTA
CTGATACCCAGTAAATGAATAATAACACTGGGCTTTTTCT…GTGCAAACAGTT
CTGATACCCAGTAAAACAATAATAAATGAGTAAGGGATGT…GCATTCACAGTT
CTGATACCCAGCACATGAATAATAACCCCAAAATATGAA…ATTTCAACAGTTA
```

- ONT has 2 production basecallers / demultiplexers for production: **Metrichor** (deprecated since end of March) and **Albacore.**

Albacore

- Albacore is an **offline tool**.
- Produce FAST5 or **FASTQ** files (since 1.1, 5th May). Before that date, we used fast5tofastq (Aurélien Birer) to convert FAST5 to FASTQ.
- 23 versions of Albacore has been published since the beginning (including non-official). A **new major version** is published **every two months.**
- We provide Docker images.
- **Adaptors are not trimmed**.
- Always check the Albacore outputs for each new version.

https://hub.docker.com/r/genomicpariscentre/albacore/

https://github.com/GenomicParisCentre/toullig

Albacore: 1D performance

- **Never use a NFS share** to store/access FAST5 files (especially for basecalling) because there is a big performance issue.

- Perform a **benchmark** to find the optimal number of threads before starting to use Albacore in production.

- SSD hard drive is not mandatory to use Albacore for 1D data.

- 1D data is demultiplexed and basecalling in **one day**.

Albacore: $1D^2$ performance

- $1D^2$ basecalling requires the creation of **transitional FAST5 files**.

- Open/reading/writing FAST5/HDF5 files requires lot of I/O.

- SSD hard drive **is mandatory** to use Albacore for $1D^2$ data in reasonable amount of time.

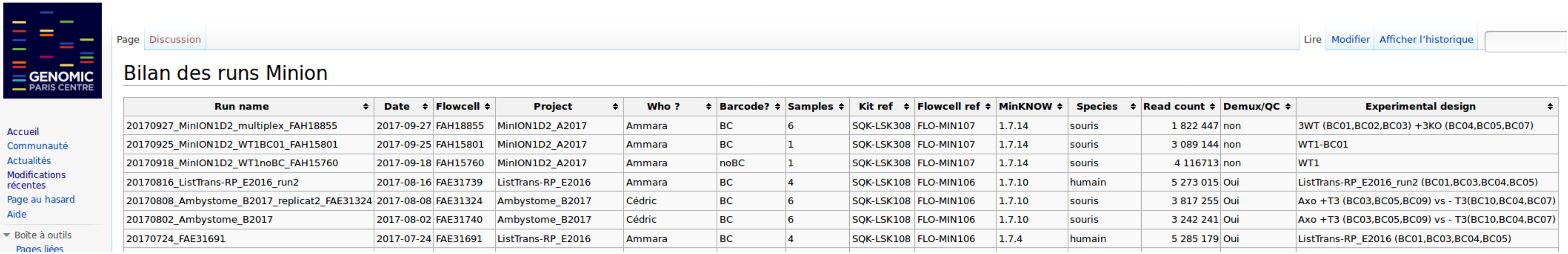
- For $1D^2$, 2 scripts are launched by `full_1dsquare_basecaller.py`. So we can save time by launching each scripts with different threads options.

- **One Month** computation time on a server with HD → **one week** on workstation with SSD.

Albacore: scripting

- We developed a **tool to automatically launch data transfer and basecalling** once a run has finished.
- We choose to not create a complex application like Aozan (Mix Python/Java) because ONT tools are still quickly evolving.
- We plan to create something better once we will buy a GridION.

- We currently use a wiki page to store kit reference, flowcell reference and experiment design for each run.



Bilan des runs Minion

Run name	Date	Flowcell	Project	Who ?	Barcode?	Samples	Kit ref	Flowcell ref	MinKNOW	Species	Read count	Demux/QC	Experimental design
20170927_MinION1D2_multiplex_FAH18855	2017-09-27	FAH18855	MinION1D2_A2017	Ammara	BC	6	SQK-LSK308	FLO-MIN107	1.7.14	souris	1 822 447	non	3WT (BC01,BC02,BC03) +3KO (BC04,BC05,BC07)
20170925_MinION1D2_WT1BC01_FAH15801	2017-09-25	FAH15801	MinION1D2_A2017	Ammara	BC	1	SQK-LSK308	FLO-MIN107	1.7.14	souris	3 089 144	non	WT1-BC01
20170918_MinION1D2_WT1noBC_FAH15760	2017-09-18	FAH15760	MinION1D2_A2017	Ammara	noBC	1	SQK-LSK308	FLO-MIN107	1.7.14	souris	4 116713	non	WT1
20170816_ListTrans-RP_E2016_run2	2017-08-16	FAE31739	ListTrans-RP_E2016	Ammara	BC	4	SQK-LSK108	FLO-MIN106	1.7.10	humain	5 273 015	Oui	ListTrans-RP_E2016_run2 (BC01,BC03,BC04,BC05)
20170808_Ambystome_B2017_replicat2_FAE31324	2017-08-08	FAE31324	Ambystome_B2017	Cédric	BC	6	SQK-LSK108	FLO-MIN106	1.7.10	souris	3 817 255	Oui	Axo +T3 (BC03,BC05,BC09) vs - T3(BC10,BC04,BC07)
20170802_Ambystome_B2017	2017-08-02	FAE31740	Ambystome_B2017	Cédric	BC	6	SQK-LSK108	FLO-MIN106	1.7.10	souris	3 242 241	Oui	Axo +T3 (BC03,BC05,BC09) vs - T3(BC10,BC04,BC07)
20170724_FAE31691	2017-07-24	FAE31691	ListTrans-RP_E2016	Ammara	BC	4	SQK-LSK108	FLO-MIN106	1.7.4	humain	5 285 179	Oui	ListTrans-RP_E2016 (BC01,BC03,BC04,BC05)

My List for Santa Albacore

Dear Santa,

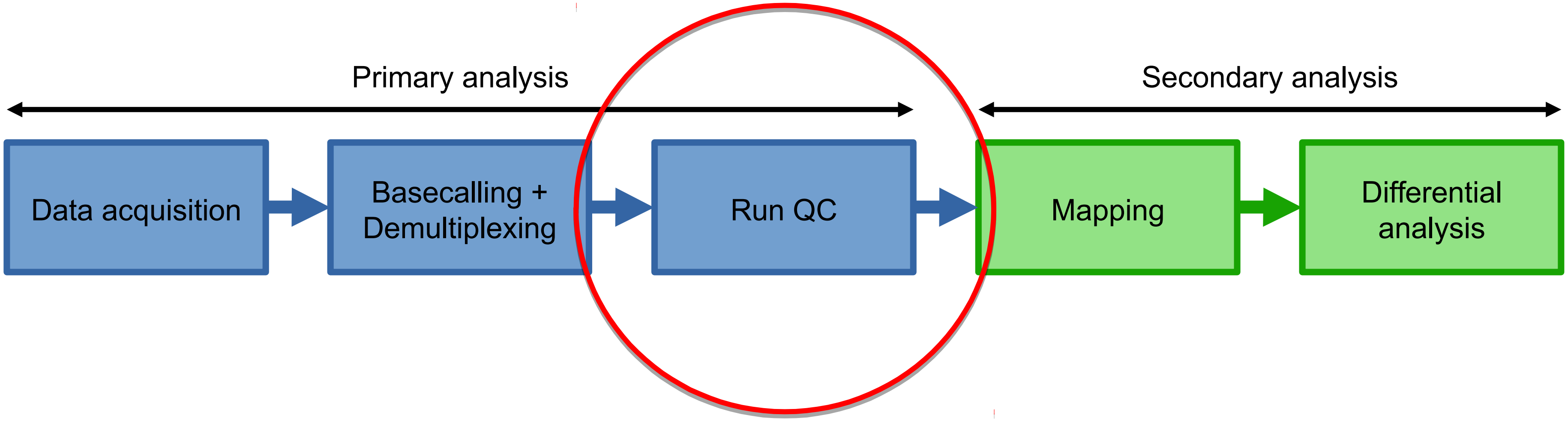
I have been very good this year so I hope you will bring me the presents listed below. Thank you!

Love, Laurent

- A **sample sheet** (like for bcl2fastq) for Albacore to avoid demultiplexing unnecessary barcodes.
- FASTQ entries with the **Pass/Fail flag** in each entry header.
- More Efficient file format to store raw data than the slow FAST5.
- No transitional FAST5 files creation for 1D^2 demultiplexing.
- **Adapters removing.**

Take this list with you when you visit Santa or send it to Santa at the North Pole.

Quality control

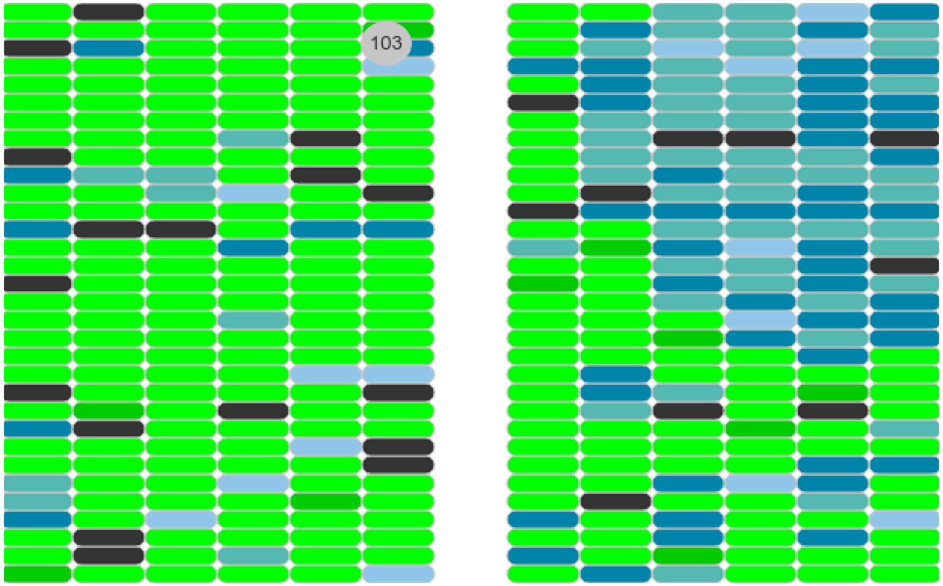


What do we have to evaluate a MinION Run?

- MinKNOW produces graphs and statistics during the run.

- The MinKNOW **report lacks information** and **is not adapted to RNASeq.**

- **Several tools are already available** (poretools , minotour, pore, ioniser...)
  - They produce interesting graphs and statistics;
  - But they are **not adapted to 1D runs** producing a lot of sequences and **using barcoded samples**.



We developed ToulligQC for better MinION run evaluation

- ToulligQC **gather all information in a single tool** adding graphs and statistics.
- It **efficiently handles files** to quickly produce a run QC (<5 minutes).
- ToulligQC is **adapted to RNASeq** and takes **barcoding** into account.
- The tool will soon handle **$1D^2$ runs**.
- ToulligQC is available on **GitHub**.
- Our software is easily installable using a **PyPi** package or a **Docker** image.

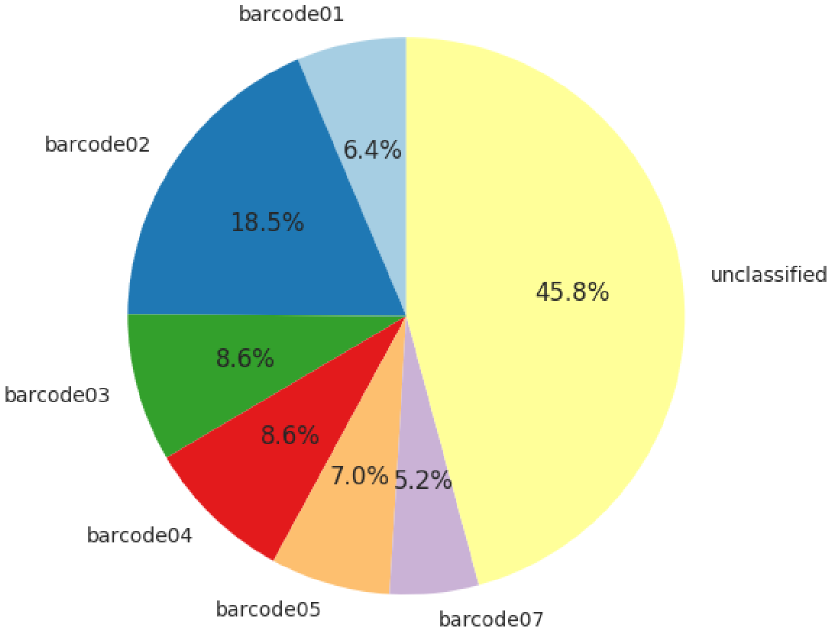
https://github.com/GenomicParisCentre/toulligQC

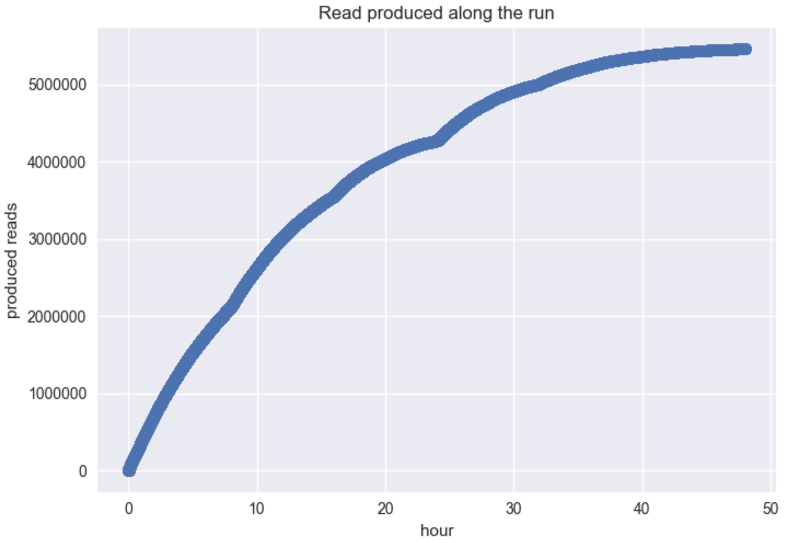
https://pypi.org/project/toulligqc/

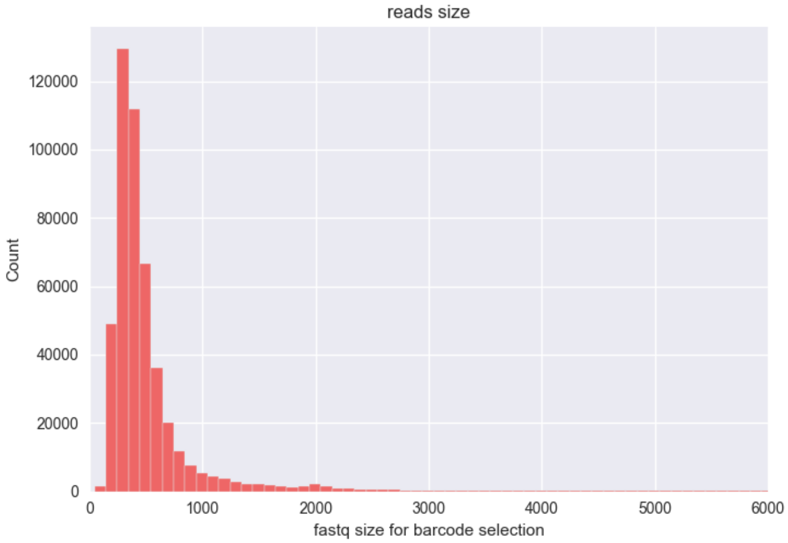
https://github.com/GenomicParisCentre/toulligQC

Examples of ToulligQC outputs

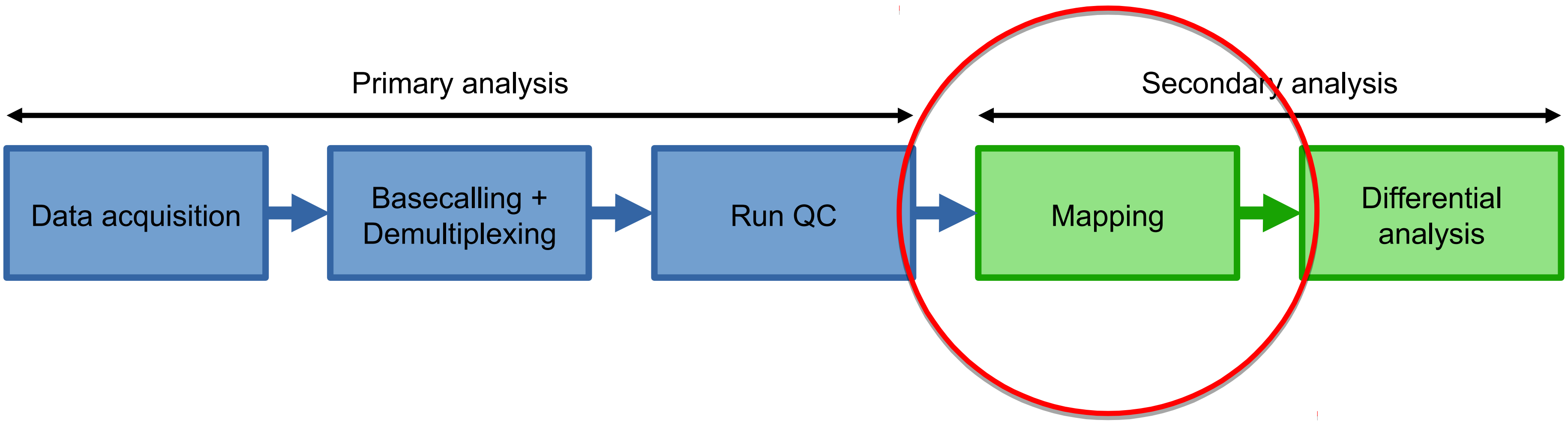
- Yield plot to **check homogeneous sequencing** along run time.
- **Transcript length** histogram.
- Easy access to **barcode proportion plot.**
- Flowcell map to **visualize spatial biases**.







Sequence alignment


Primary analysis
Secondary analysis
Data acquisition
Basecalling + Demultiplexing
Run QC
Mapping
Differential analysis