# Small introduction to NGS and Genome assembly

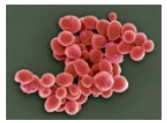
Ricardo Leite

Instituto Gulbenkian de Ciência

# Brief story of sequencing projects

1st public discussion of Human genome sequencing project

1984



Yeast genome sequenced

1996

*C. elegans* (roundworm)

1998

Mouse

2002

**Several ten thousand organisms sequenced!**

2017

1995

1st bacterial genome sequenced

HTG Human genome sequencing begins

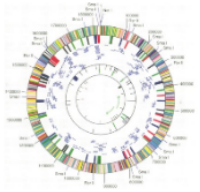
1997

*E. coli*

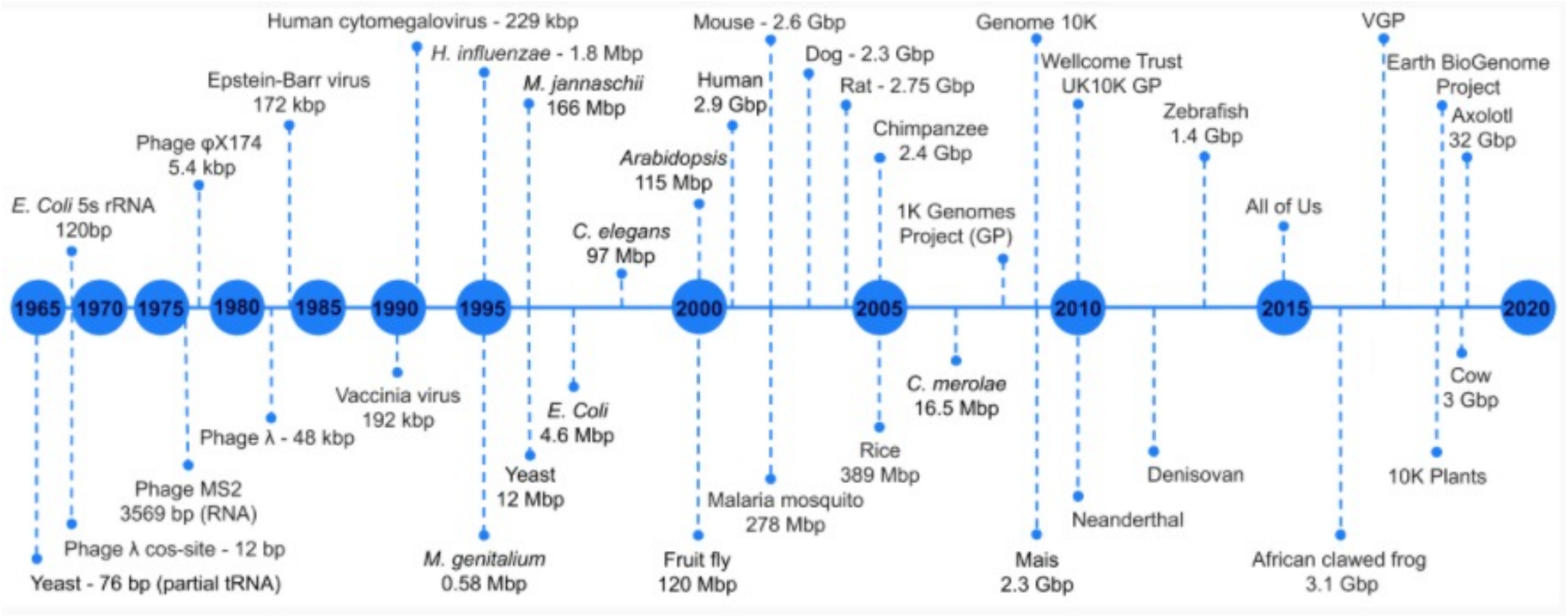
2000

Human, fly genome drafts

**Sequencing enters healthcare, whole patient genomes**



# Brief story of sequencing projects






1986-2001 2005 2006 2010 2011 2014

ABI Sanger

Roche 454

Illumina

Life SOLiD

Life Ion Torrent

Pacific Biosciences

Oxford Nanopore

First Generation

Second Generation

Third Generation



# What type of sequencing instruments are in use?

- **Illumina MiniSeq, MiSeq, NextSeq, HiSeq**
- Illumina is the current undisputed leader in the high-throughput sequencing market.
- Up to 300 million reads (HiSeq 2500)
- Up to 1500 GB per run (GB = 1 billion bases)
- **IonTorrent PGM, Proton**
- The IonTorrent platform targets more specialized clinical applications.
- Up to 400bp long reads
- Up to 12 GB per run

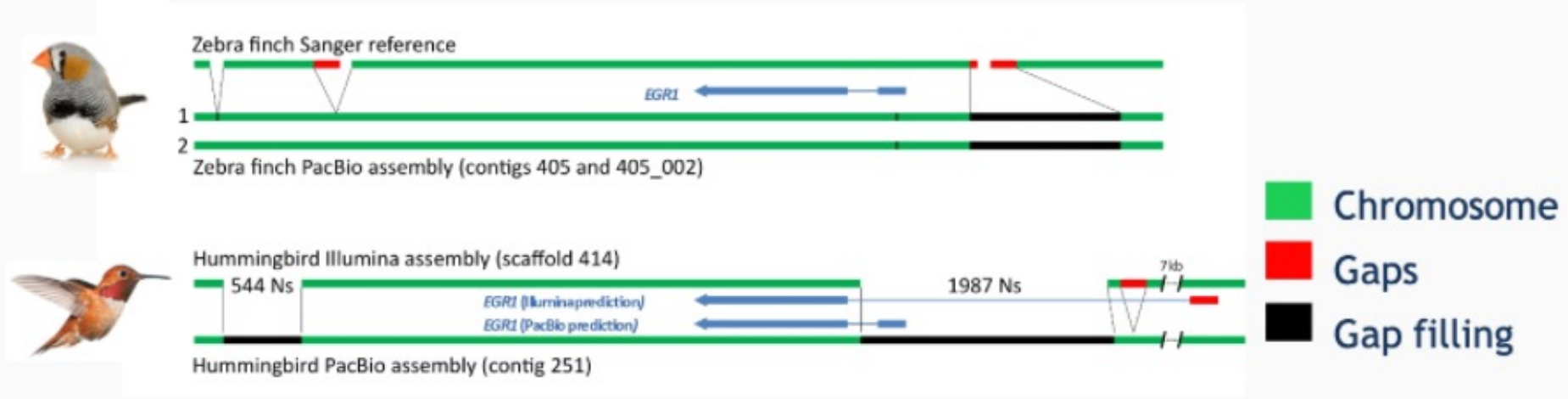
# What type of sequencing instruments are in use?

- **PacBio Sequel**
- This company is the leader in long read sequencing. ´
- Up to 12,000 bp long paired-end reads
- Up to 4 GB per run

- **MinION**
- A portable, miniaturized device that is not yet quite as robust and reliable as the other options. Up to 10,000 long reads
- Up to 240 MB per run (MB = 1 million bases)

# Importance of longe reads

High-quality error-free genome assemblies and annotations are necessary as current 1st and 2nd generation genome sequencing approaches generate numerous errors that cause a variety of problems in downstream analyses. **Parts of genes are missing, and some are incorrectly assembled, while others are completely missing from the assemblies despite pieces found in the raw sequence reads.** (Vertebrate Genomes Project)



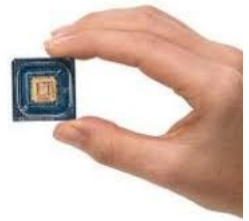

Korlach et al., 2017

MGI 华大智造

6 TB/day

New players in town

GenapSys



Genius semiconductor chip

1,6 GB/run

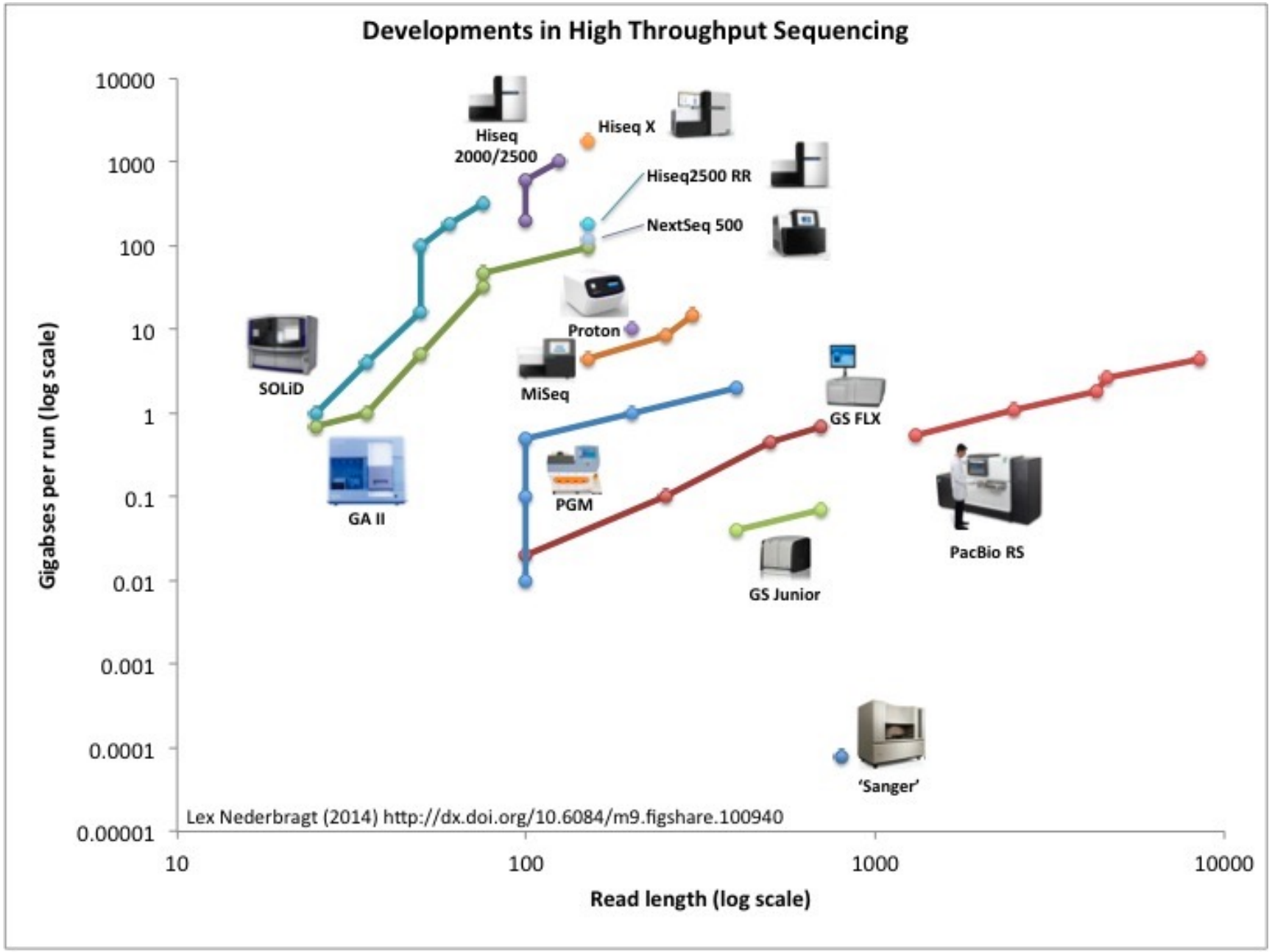
# Facts and numbers

| instrument | Nanopore | Pacbio | Ion Torrent | Illumina | 454 | SOLiD |
|---|---|---|---|---|---|---|
| Method | Single-molecule in real-time | Single-molecule in real-time | Ion semiconductor | synthesis | Pyrosequencing | Ligation |
| Amplification | none | none | emPCR | Bridge PCR | emPCR | emPCR |
| Read length | 10kb average | 12kb average | 200 to 400 bp | 50 to 300 bp | 700 bp | 50+35 or 50+50 bp |
| Error type | indel | indel | indel | substitution | indel | A-T bias |
| single-Pass Error rate % | 18 | 13 | ~1 | ~0.1 | ~0.1 | ~0.1 |
| Reads per run | variable | 75000–500000 | up to 4M | up to 3.2G | 1M | 1.2 to 1.4G |
| Time per run | 6h-72h | 30 min. to 4 hours | 2 to 7 hours | 1 to 10 days | 24 hours | 1 to 2 weeks |
| Cost per Gb (in US$) | $150 | $180 to $300 | $80 to $500 | $1 to $1'000 | $10'000 | $60 to $80 |

Adapted from http://www.molecularecologist.com/next-gen-fieldguide-2016/

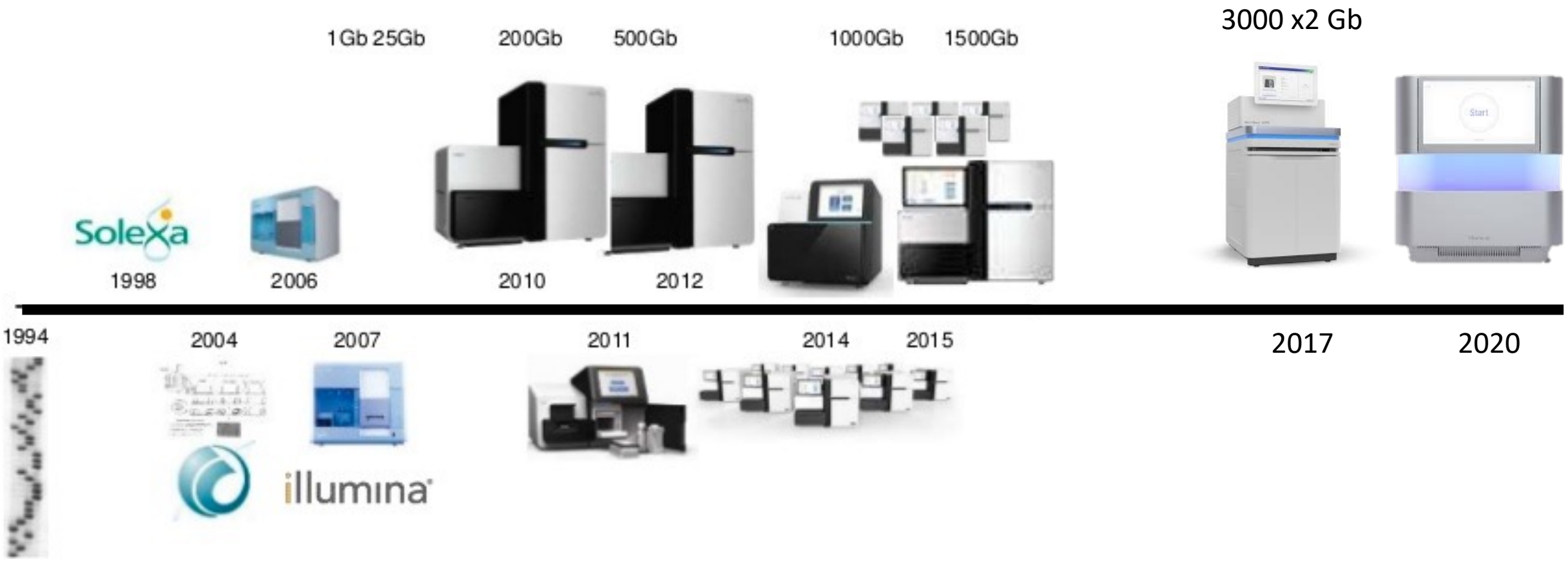
Output vs

Read legth



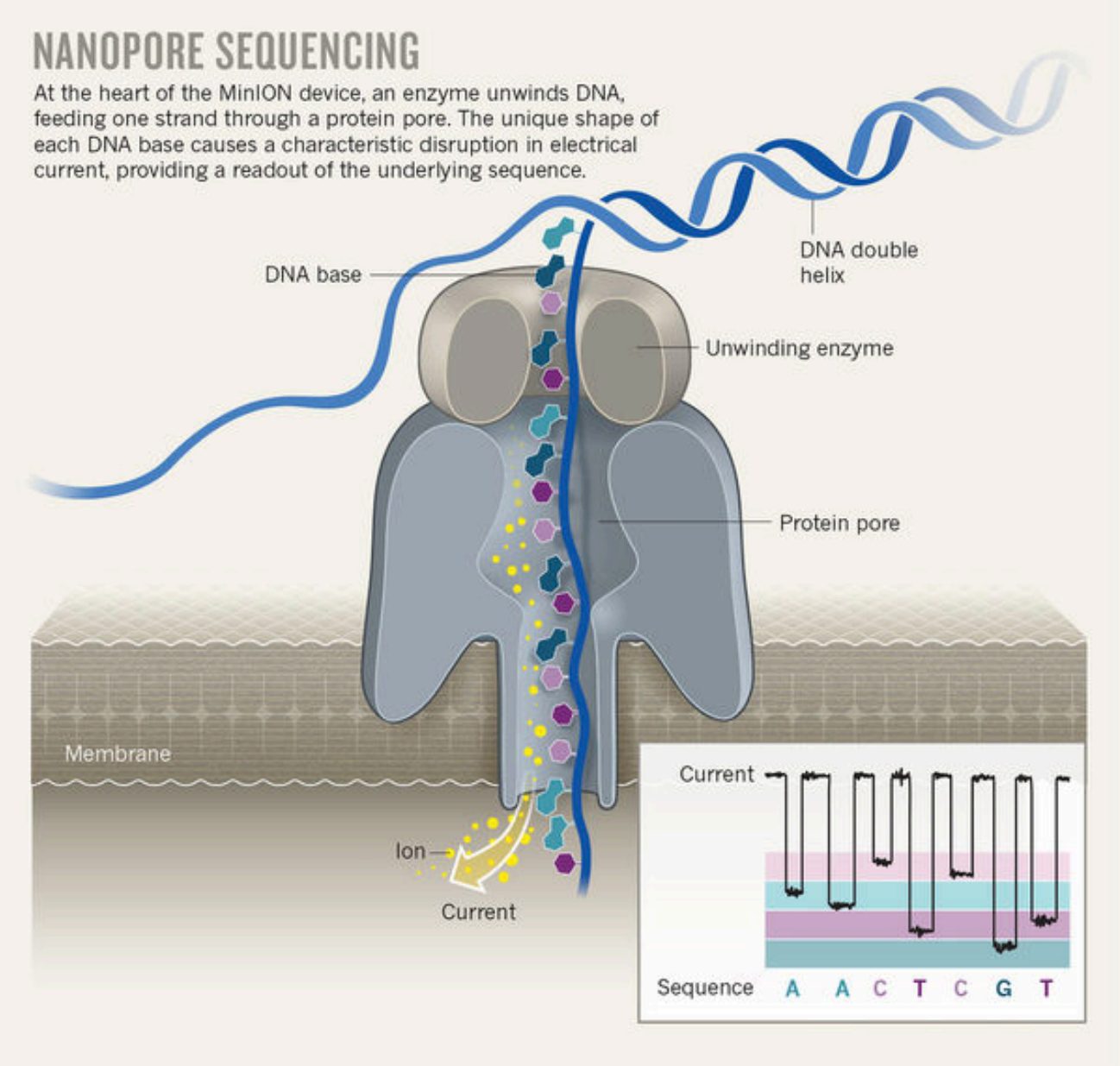

doi 10.6084/m9.figshare.100940.

# Illumina History

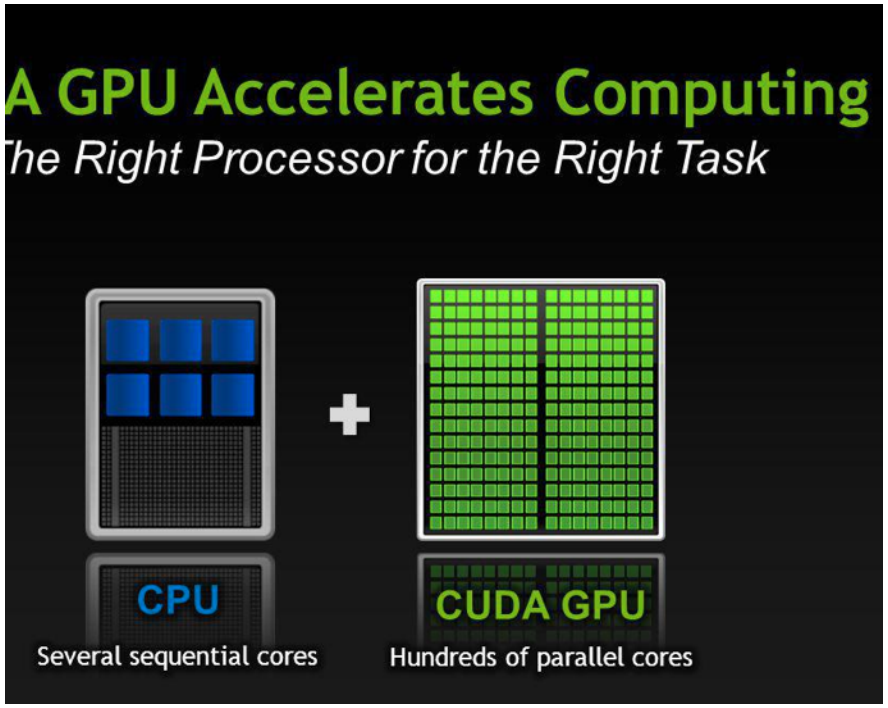



*Adapted from ilumina.com*

# Video Nanopore

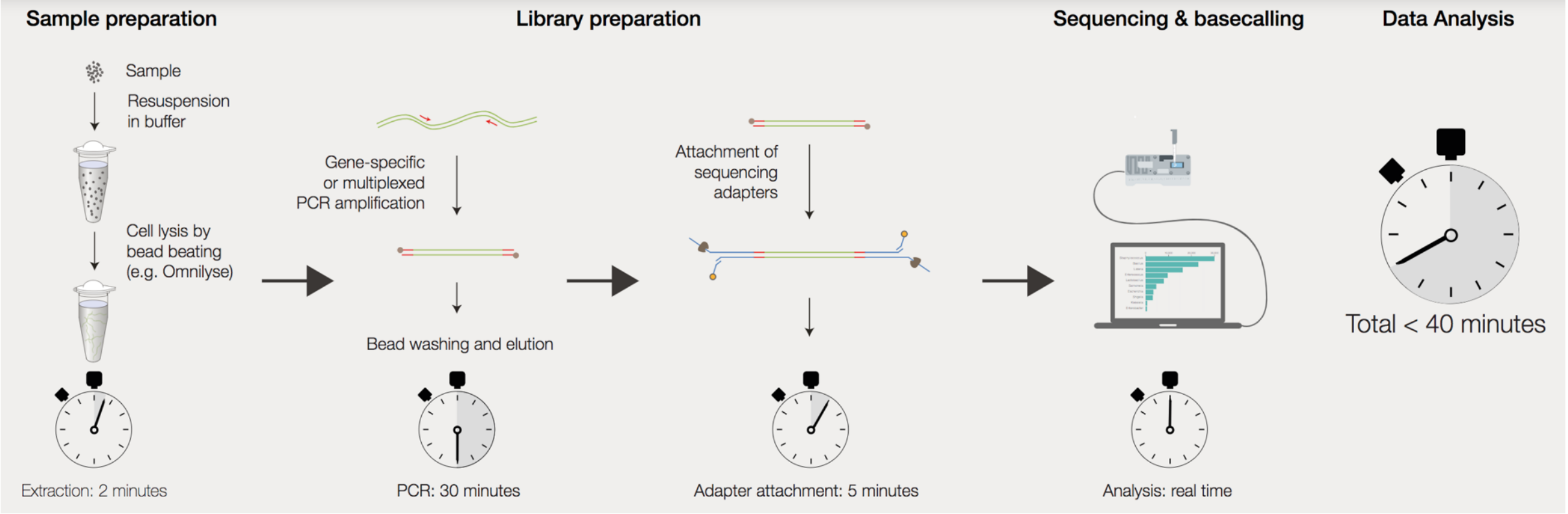



=

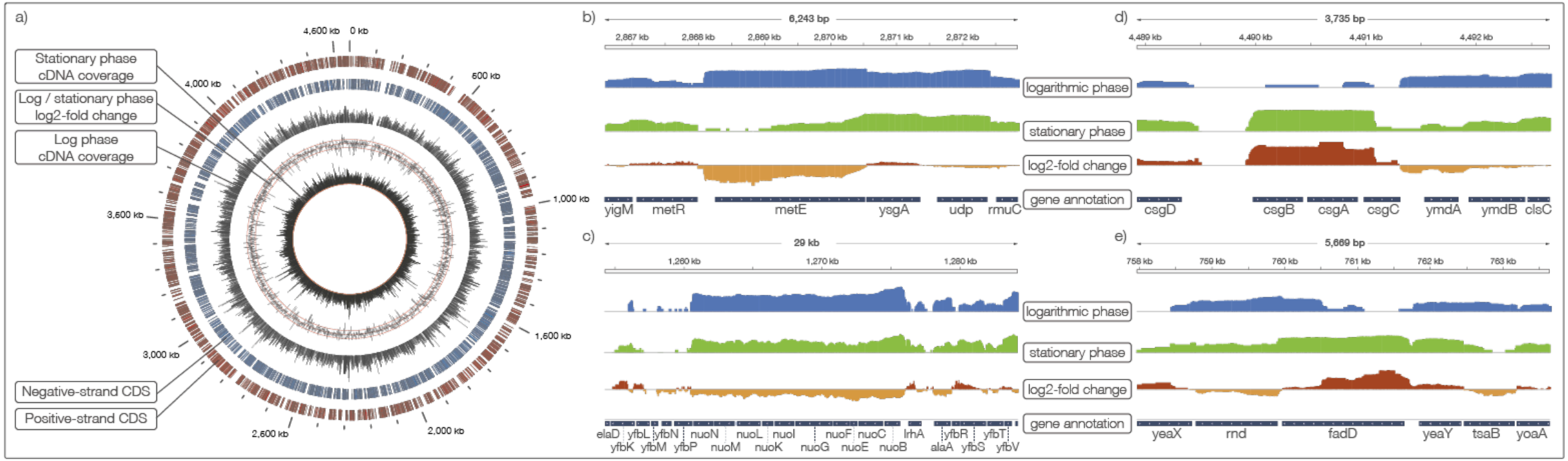




# GPUs for Science

# Fast solutions: 12 samples in 40 minutes



# Bacterial transcriptome




https://nanoporetech.com/resource-centre/bacterial-comparative-genomics-and-transcriptomics-using-long-nanopore-reads#image3&

| | No barcodes | 12 barcodes | 24 barcodes | 24 barcodes (Fl |
|---|---|---|---|---|
| Flow cell price | $500 | $500 | $500 | $90 (Flongle flov |
| Library price | $90 | $120 | $150 | $150 |
| **Price per sample** | **$590** | **$51.67** | **$27.08** | **$10** |

# Not that expensive

# Fasta Format - the "workhorse" of bioinformatics

- The format is deceivingly simple:
- A > symbol on the FASTA header line indicates a fasta record start.
- A string of letters called the sequence id may follow the > symbol. The header line may contain an arbitrary amount of text (including spaces) on the same line. Subsequent lines contain the sequence.

- The standard alphabet for nucleotides would contain: ATGC . An extended alphabet may also contain the N symbol to indicate a base that could be any of ATGCN . A different extension of a nucleotide alphabet may allow extended symbols such as W that corresponds to nucleotide that is either an A or T etc. Gaps may be represented via . or -.

Adapted from Biostar handbook

# FASTQ Format

- The FASTQ format is the de facto standard by which all **sequencing instruments** represent data.
- It may be thought of as a variant of the FASTA format that allows it to associate a quality measure to each sequence base, FASTA with **QUALITIES**.
- In simpler terms, it is a format where for **every base, we associate a reliability measure**: base is A and the probability that we are wrong is 1/1000 .
- Conceptually, the format is very similar to FASTA but suffers from even more flaws than the FASTA format.

Adapted from Biostar handbook

# Fastq import facts

- The FASTQ format is a **multi-line format** just as the FASTA. In the early days of highthroughput sequencing, instruments always produced the entire **FASTQ sequence on a single line (on account of them being so short 35-50bp)**

- **Flaw: @ sign is both a FASTQ record separator and a valid value of the quality string**. For that reason it is a little more difficult to design a correct FASTQ parsing program.

Adapted from Biostar handbook

# FASTQ facts

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

1. A FASTA-like header, but instead of the **> symbol it uses the @** symbol. This is followed by an ID and more optional text, similar to the FASTA headers.

2. The second section contains the **measured sequence** (typically on a single line), but it may be wrapped until the + sign starts the next section.

3. The third section is marked by the **+ sign** and may be optionally followed by the same sequence id and header as the first section

4. The last line encodes the **quality values** for the sequence and must be of the same length. It should (must?) also be wrapped the same way

Adapted from Biostar handbook

# Phred scores

The **weird characters** in the are the so called "**encoded" numerical values**, each character !''*(((( represents a **numerical value**: a so-called **Phred score** , encoded via a single letter encoding. This is a cautionary tale of what happens when scientists start to design data formats.

The idea behind the Phred score is to **map two digit numbers** to single characters so that the length of the quality string stays the same as the length of the sequence.

Each character has a numerical value, say ! will be mapped to 0 , + will be remapped to mean 10 and say 5 will be mapped to 30 and I will be mapped to 40

Adapted from Biostar handbook

# Is there more information in FASTQ headers?

@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACGContains the following information:

EAS139 the unique instrument name

136 the run id

FC706VJ the flowcell id

2 flowcell lane

2104 tile number within the flowcell lane

15343 ‘x’-coordinate of the cluster within the tile

197393 ‘y’-coordinate of the cluster within the tile

1 the member of a pair, 1 or 2 (paired-end or mate-pair reads only)

Y Y if the read is filtered, N otherwise

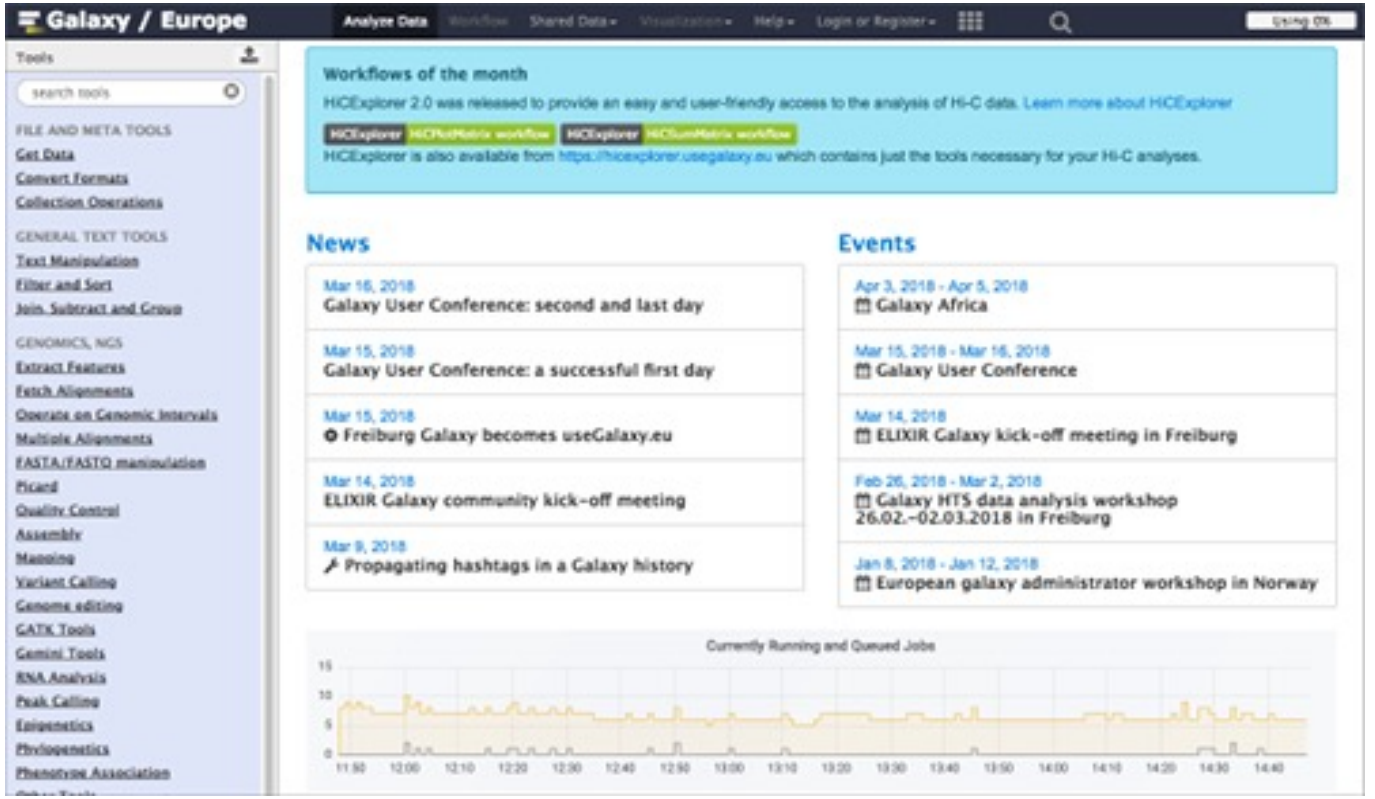
18 0 when none of the control bits are on, otherwise it is an even number

ATCACG index sequence

Adapted from Biostar handbook

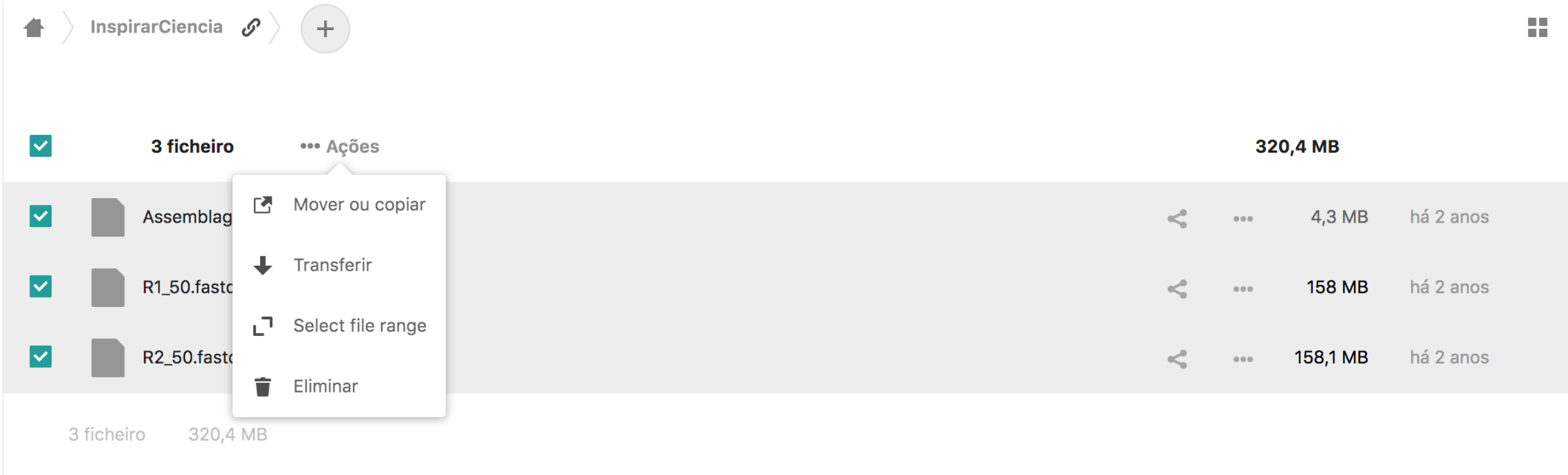
# Bioinformatic Portal: usegalaxy.eu

usegalaxy.eu



# Data for this exercise:

https://nextcloud.igc.gulbenkian.pt/index.php/s/KtRo2fWHP2GoSLc



# Galaxy Europe



## Tools

search tools



### COVID-19 research!

Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal at covid19.galaxyproject.org. We mirror **all public** SARS-CoV-2 data from ENA in a Galaxy data library for your convenience. The Galaxy community also created COVID-19 related trainings and we also maintain a running document with recent news. Our new preprint about **The landscape of SARS-CoV-2 RNA modifications** is out!

"Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." – Prof. Stephen Hawking

## News

Dec 8, 2020
**Training Infrastructure Feedback on the ELIXIR Belgium workshop "DDA and DIA proteomic analysis in Galaxy"**

Dec 3, 2020
**Galaxy Release 20.09 Video Summary**

Nov 28, 2020
**UseGalaxy.eu Tool Updates for 2020-11-28**

Nov 27, 2020
**Insights from the first cross-training between EOSC-Life and EOSC-Nordic**

Nov 27, 2020
**Training Infrastructure Feedback from Ambre Jousselin**

## Events

Feb 24, 2021
**DRS, long-read-sequencing, proteomics and more - An update to recent COVID-19 workflow developments**

Feb 17, 2021
**Viral Beacon and Galaxy variant workflows**

Feb 15, 2021 - Feb 19, 2021
**GTN Smörgåsbord: A Global Galaxy Course**

Feb 10, 2021
**Insights from selection analysis of complete genomes and read-level data**

Feb 3, 2021
**Supporting the COVID-19 Data portal: viral data cleaning from human reads and submission to ENA**

## History

search datasets

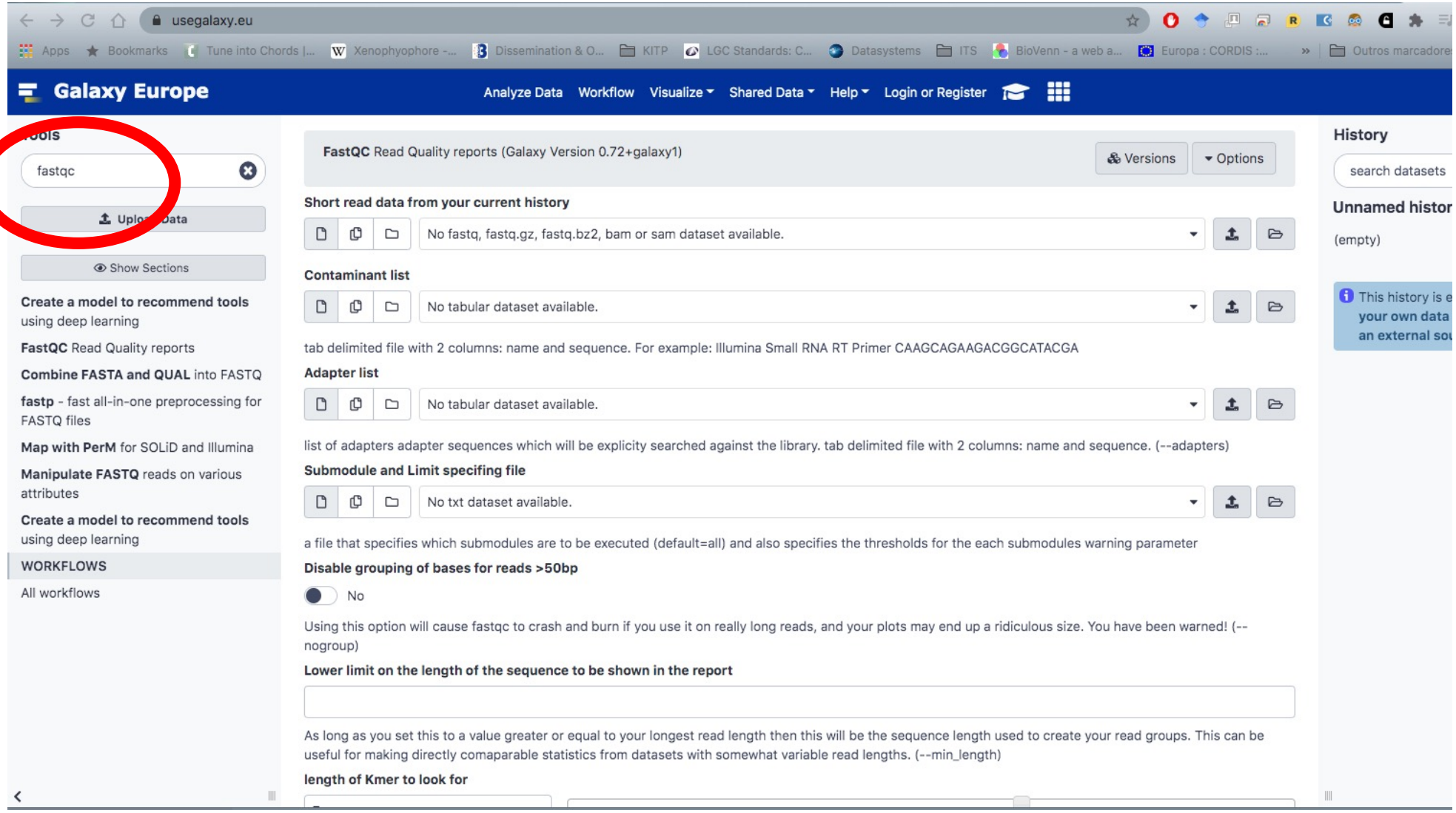
**Unnamed history**

(empty)

This history is empty. You can **load your own data** or **get data from an external source**

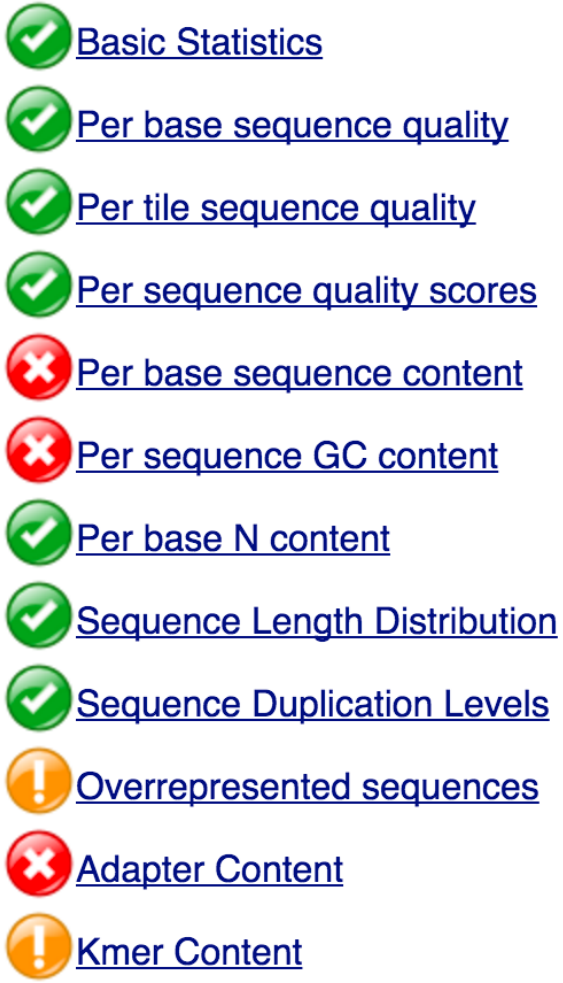
OPEN CHAT

# Let’s perform QC of the reads by using FastQC



# Should I be worried about the “stoplight” symbols?

NO

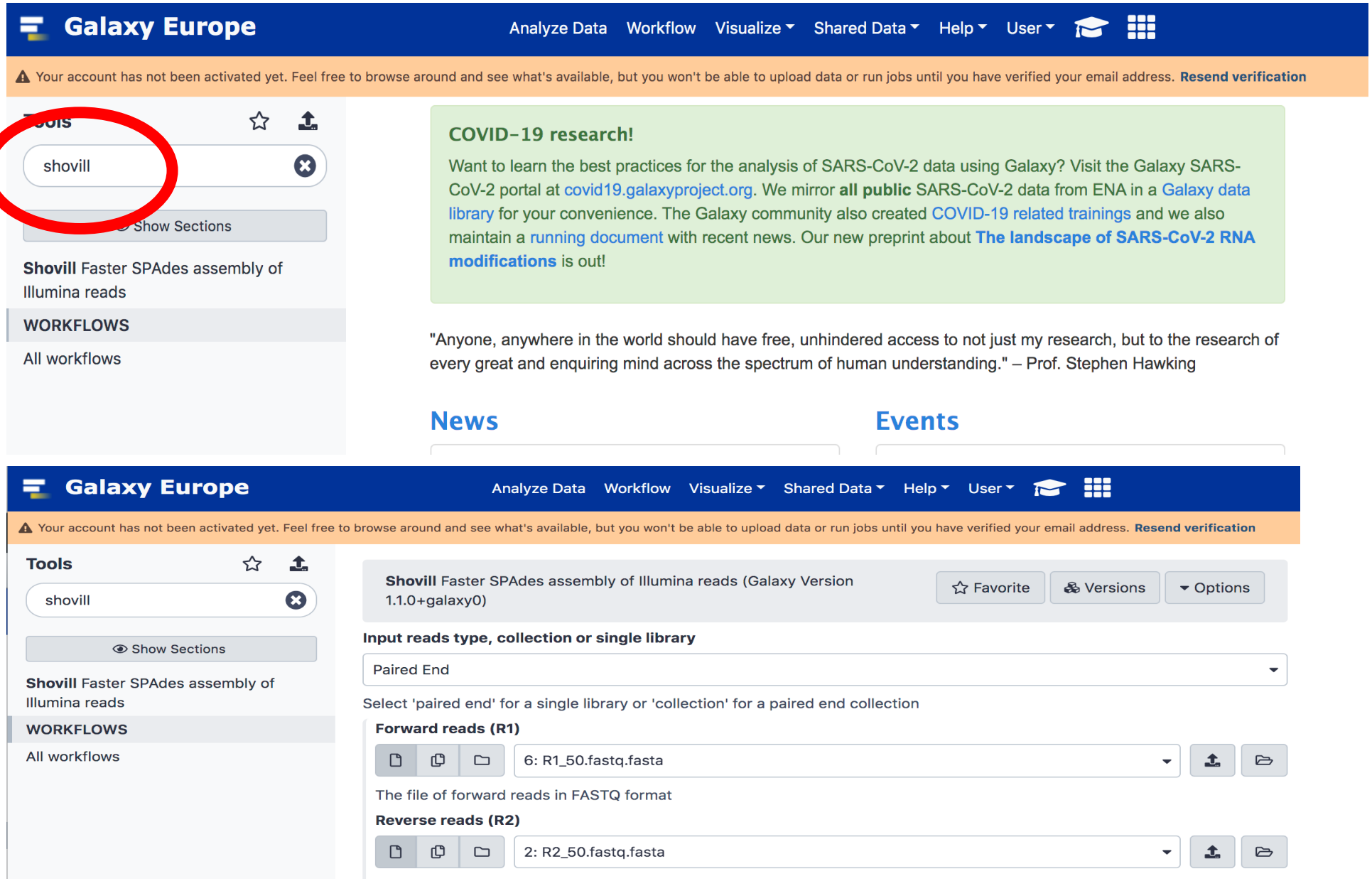


# Is there a list of QC tools?

A considerable number of QC tools have been published.

Recommend: Trimmomatic, BBDuk ,fastp and cut adapt. While all tools implement basic quality control methods, each often includes unique functionality specific to that tool.

Quality control tools often provide more utility, some are full sequence manipulation suites.


Galaxy Europe
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Your account has not been activated yet. Feel free to browse around and see what's available, but you won't be able to upload data or run jobs until you have verified your email address. Resend verification
shovill
Show Sections
Shovill Faster SPAdes assembly of Illumina reads
WORKFLOWS
All workflows
COVID-19 research!
Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal at covid19.galaxyproject.org. We mirror all public SARS-CoV-2 data from ENA in a Galaxy data library for your convenience. The Galaxy community also created COVID-19 related trainings and we also maintain a running document with recent news. Our new preprint about The landscape of SARS-CoV-2 RNA modifications is out!
"Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." – Prof. Stephen Hawking
News
Events
Galaxy Europe
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Your account has not been activated yet. Feel free to browse around and see what's available, but you won't be able to upload data or run jobs until you have verified your email address. Resend verification
Tools
shovill
Show Sections
Shovill Faster SPAdes assembly of Illumina reads
WORKFLOWS
All workflows
Shovill Faster SPAdes assembly of Illumina reads (Galaxy Version 1.1.0+galaxy0)
Favorite
Versions
Options
Input reads type, collection or single library
Paired End
Select 'paired end' for a single library or 'collection' for a paired end collection
Forward reads (R1)
6: R1_50.fastq.fasta
The file of forward reads in FASTQ format
Reverse reads (R2)
2: R2_50.fastq.fasta

# How to perform a genome assembly

**How do I run an assembly?**

Genome assembly is an iterative process where you will:

Apply quality control measures

Estimate initial parameters (k-mers-sizes, coverages)

Run the assembly

Evaluate assembly (then go to 2 if necessary)

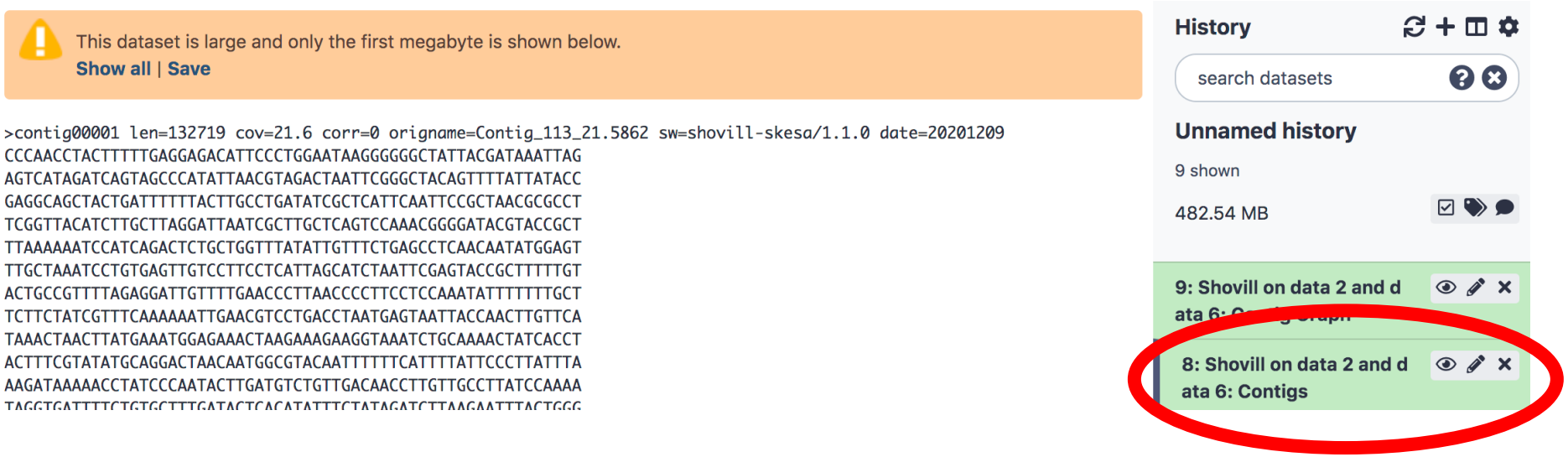
# How to evaluate a genome assembly

The term “assembly evaluation” has two distinct and unrelated meanings:

Characterize the assembled contigs by their observed properties.

Characterize the assembled contigs relative to the “reality,” typically a known, similar genome.

We can inspect and check our assembly (contigs) by pressing the visualization button on the history



# Rapid Genome annotation

Search for PROKKA Tool and run it with all options at default

The input are the contigs from our assembly from galaxy’s history

It produces 12 outputs:

- **13: Prokka on data 11: gff**
- **14: Prokka on data 11: gbk**
- **15: Prokka on data 11: fna**
- **16: Prokka on data 11: faa**
- **17: Prokka on data 11: ffn**
- **18: Prokka on data 11: sqn**
- **19: Prokka on data 11: fsa**
- **20: Prokka on data 11: tbl**
- **21: Prokka on data 11: tsv**
- **22: Prokka on data 11: err**
- **23: Prokka on data 11: txt**
- **24: Prokka on data 11: log**

- From the 12 outputs we want to inspect the faa extension that correspond to putative proteins annotate in aa format

Let’s copy (ctrl-v)10 random predicted proteins of this organism and perform a blast