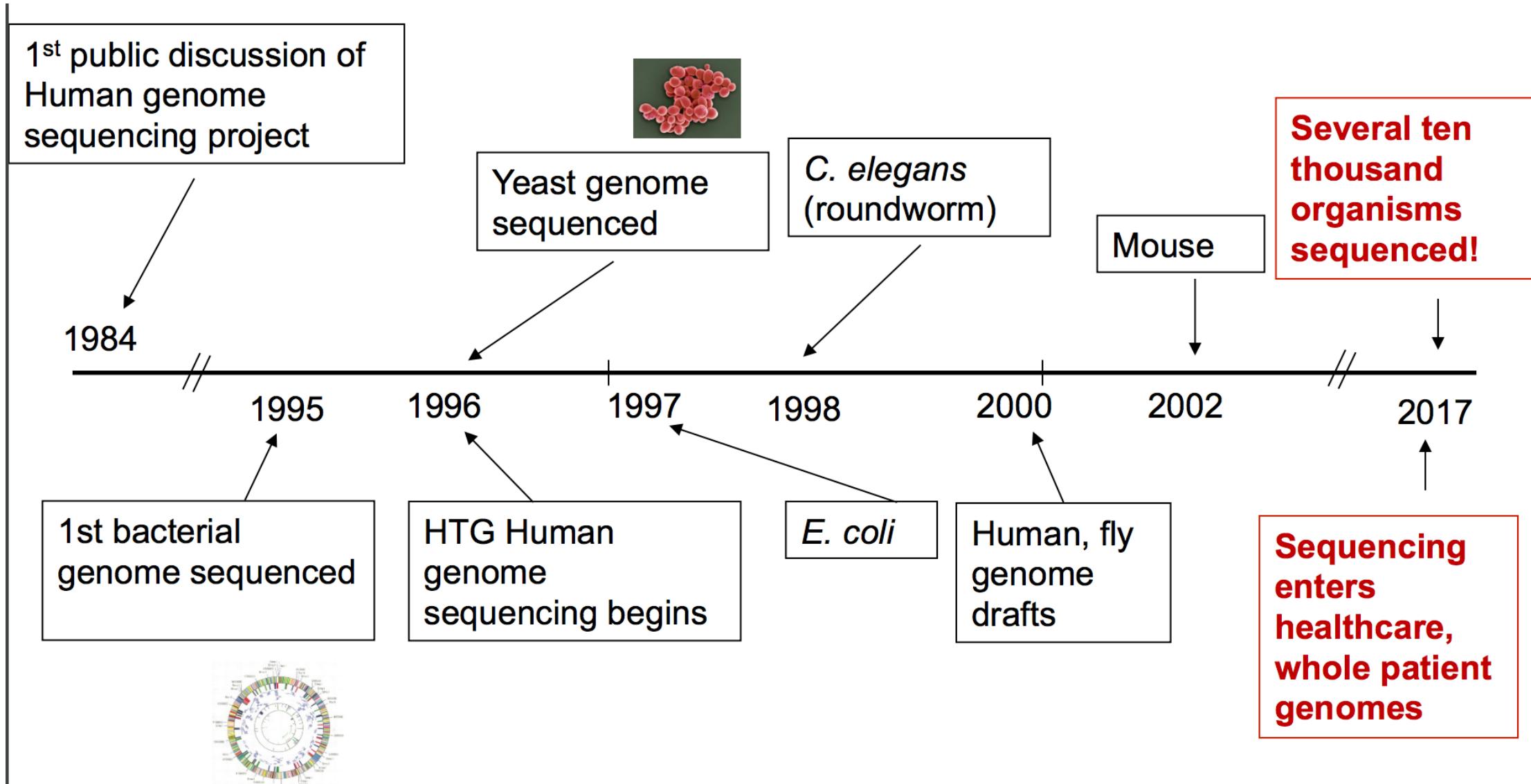




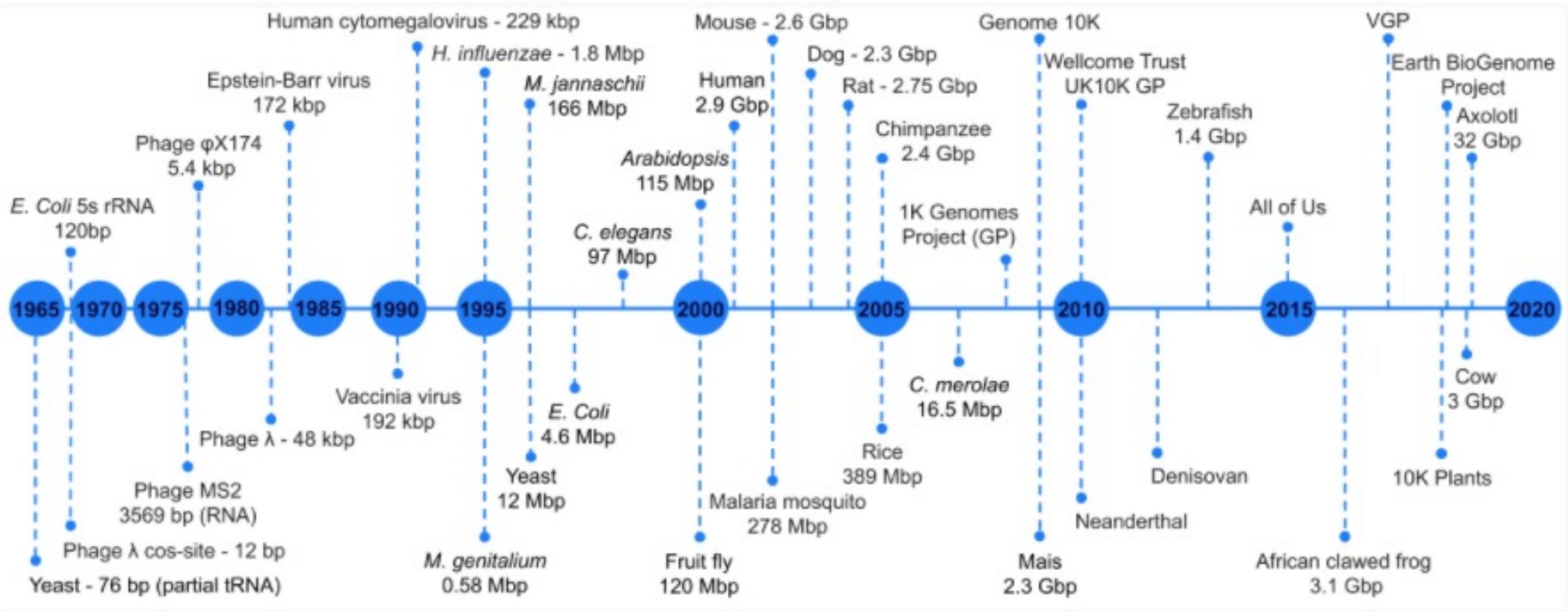
Small introduction to NGS and Genome assembly

Ricardo Leite
Instituto Gulbenkian de
Ciência

Brief story of sequencing projects



Brief story of sequencing projects





1986-2001

2005

2006

2010

2011

2014

ABI Sanger

Roche 454

Illumina

Life SOLiD

Life Ion
TorrentPacific
BiosciencesOxford
Nanopore

First Generation

Second Generation

Third Generation

ONI
FR



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

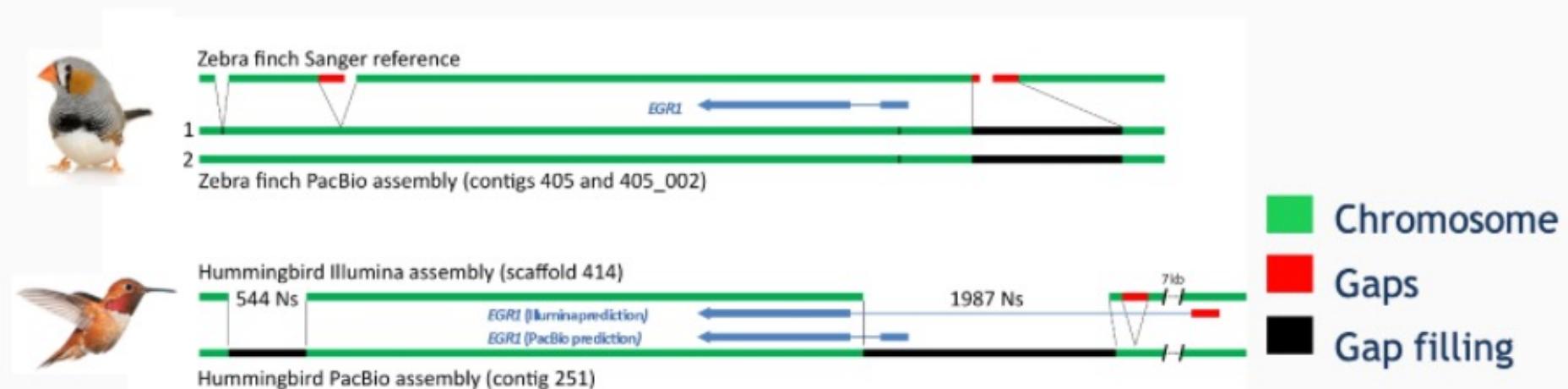


What type of sequencing instruments are in use?

- **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 long 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)



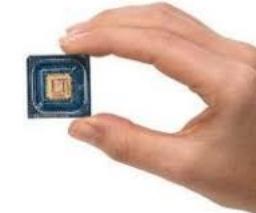
Korlach et al., 2017



New players
in town



6 TB/day



Genius semiconductor chip



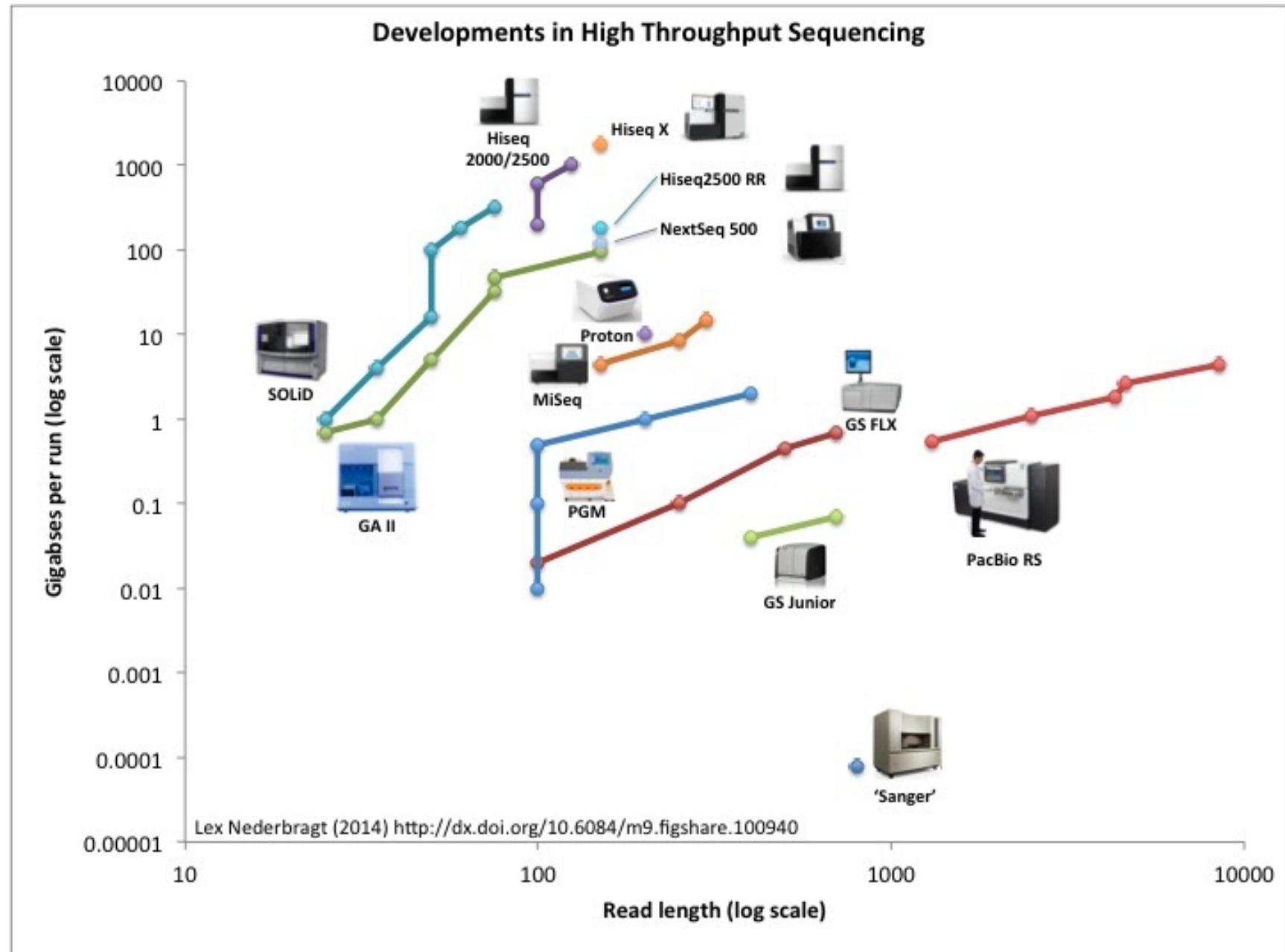
1,6 GB/run

Facts and numbers

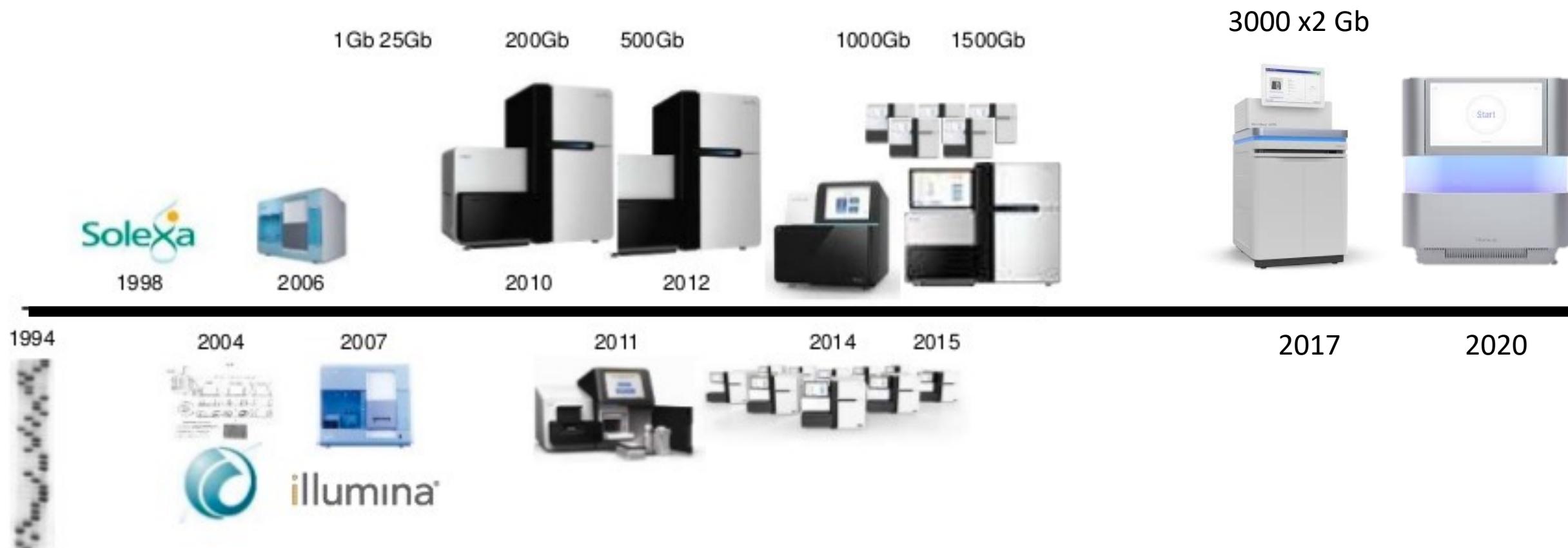
Instrument	Nanopore	Pacbio	Ion Torrent	Illumina	454	SOLID
Method	Single-molecule in real-time	Single-molecule in real-time	Ion semiconductor	synthesis	Pyrosequencing	Ligation
Amplification	none	none	emPCR	Bridge PCR	emPCR	emPCR
Read length	10kb average	12kb average	200 to 400 bp	50 to 300 bp	700 bp	50+35 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 length



Illumina History

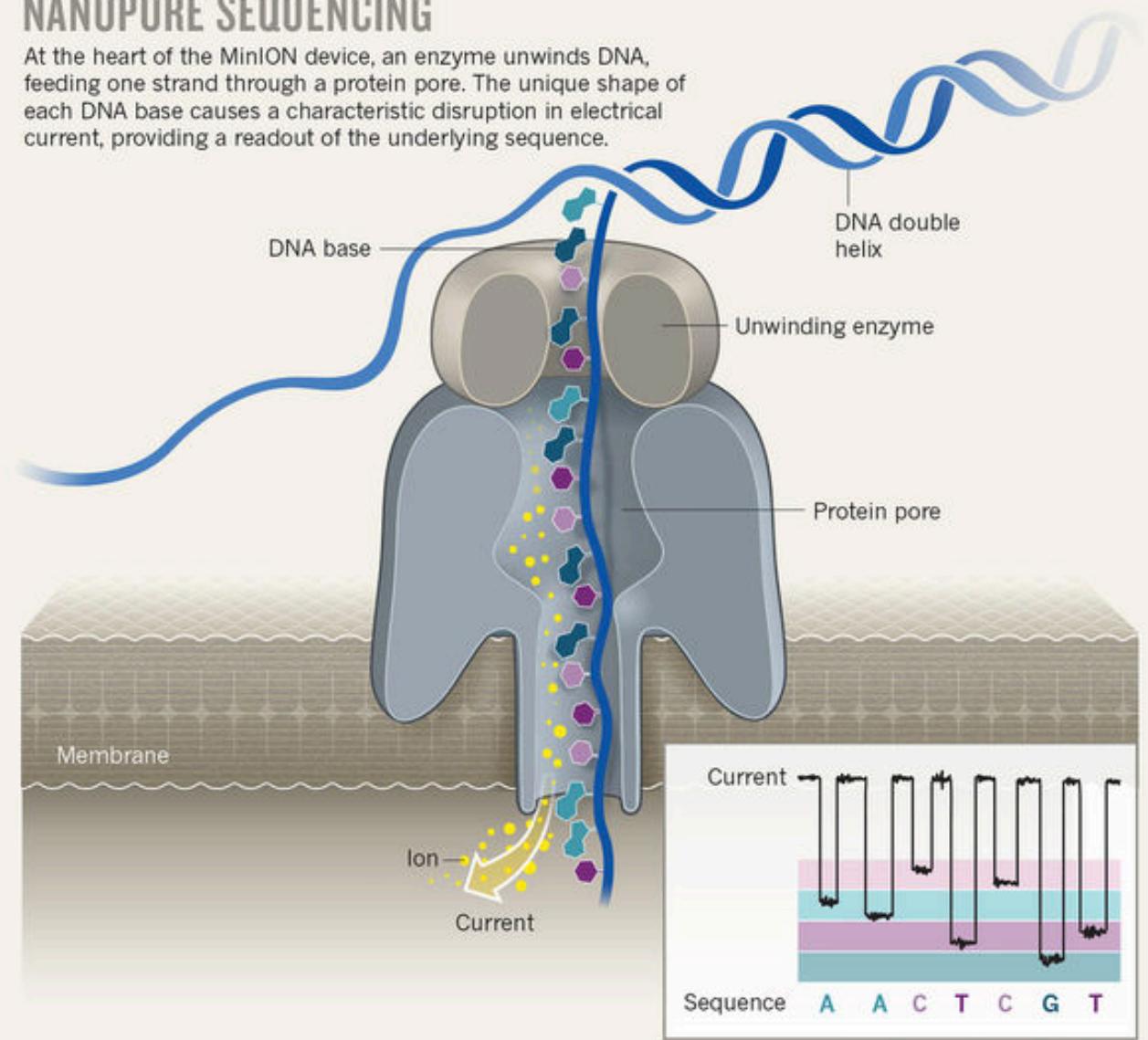


Adapted from [illumina.com](https://www.illumina.com)

Video Nanopore

NANOPORE SEQUENCING

At the heart of the MinION device, an enzyme unwinds DNA, feeding one strand through a protein pore. The unique shape of each DNA base causes a characteristic disruption in electrical current, providing a readout of the underlying sequence.



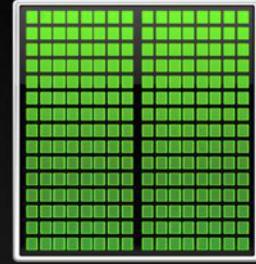
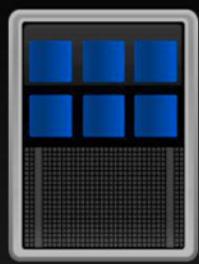


=



A GPU Accelerates Computing

The Right Processor for the Right Task

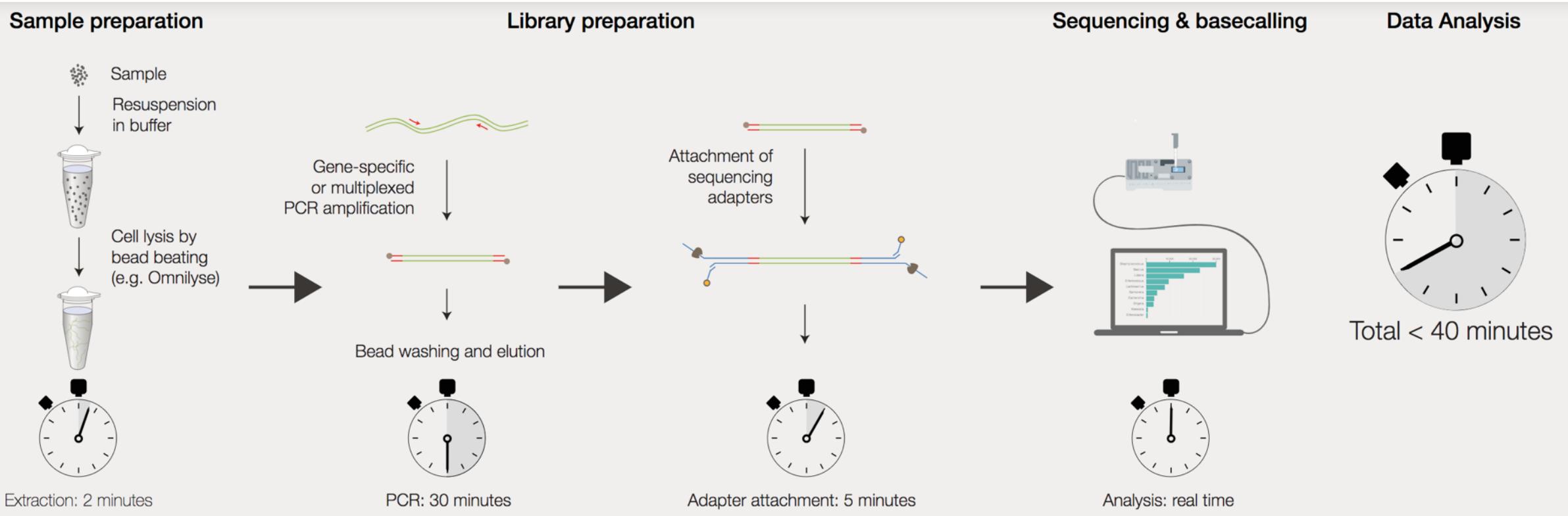


CPU
Several sequential cores

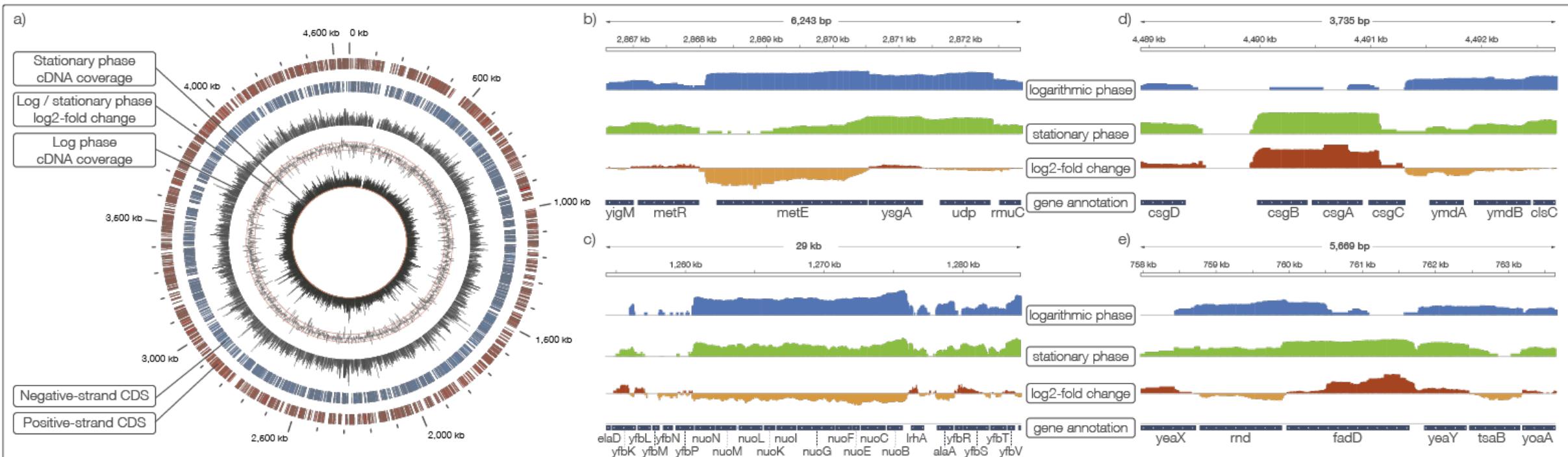
CUDA GPU
Hundreds of parallel cores

GPUs for Science

Fast solutions: 12 samples in 40 minutes



Bacterial transcriptome





No barcodes 12 barcodes 24 barcodes 24 barcodes (Fl)

Flow cell price	\$500	\$500	\$500	\$90 (Flongle flow cell)
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 deceptively 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 -.

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.

Fastq import facts

- The FASTQ format is a **multi-line format** just as the FASTA. In the early days of hightthroughput 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.

FASTQ facts

```
@SEQ_ID
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTTT
+
!^*((( (**+))%&++)(%&%).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

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

Is there more information in FASTQ headers?

@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG Contains 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

Bioinformatic Portal: usegalaxy.eu

The screenshot shows the usegalaxy.eu homepage. At the top, there is a large logo with the text "usegalaxy.eu". Below the logo, the page title is "Galaxy / Europe". The main navigation bar includes links for "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", "Login or Register", and a search icon. A banner at the top right announces "Workflows of the month" for HICExplorer 2.0, featuring two workflow buttons: "HICExplorer HCPuMatrix workflow" and "HICExplorer HCsumMatrix workflow". It also mentions that HICExplorer is available from <https://hicexplorer.usegalaxy.eu>. On the left, a sidebar lists various tool categories: Tools, FILE AND META TOOLS (Get Data, Convert Formats, Collection Operations), GENERAL TEXT TOOLS (Text Manipulation, Filter and Sort, Join, Subtract and Group), GENOMICS, NGS (Extract Features, Enrich Alignments, Operate on Genomic Intervals, Multiple Alignments, FASTA/FASTQ manipulation, Picard, Quality Control, Assembly, Mapping, Variant Calling, Genome editing), GATK Tools, Gemini Tools, RNA Analysis, Peak Calling, Epistatistics, Ethnoepistatistics, Phenotype Association, and Phylogenetics. The main content area has sections for "News" and "Events". The "News" section lists several recent articles: "Mar 16, 2018 Galaxy User Conference: second and last day", "Mar 15, 2018 Galaxy User Conference: a successful first day", "Mar 15, 2018 Freiburg Galaxy becomes useGalaxy.eu", "Mar 14, 2018 ELIXIR Galaxy community kick-off meeting", and "Mar 9, 2018 Propagating hashtags in a Galaxy history". The "Events" section lists upcoming events: "Apr 3, 2018 - Apr 5, 2018 Galaxy Africa", "Mar 15, 2018 - Mar 16, 2018 Galaxy User Conference", "Mar 14, 2018 ELIXIR Galaxy kick-off meeting in Freiburg", "Feb 26, 2018 - Mar 2, 2018 Galaxy HTS data analysis workshop 26.02.-02.03.2018 in Freiburg", and "Jan 8, 2018 - Jan 12, 2018 European galaxy administrator workshop in Norway". At the bottom, there is a chart titled "Currently Running and Queued Jobs" showing the number of jobs over time.

Data for this exercise:

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

A screenshot of a NextCloud interface showing a list of three files in a folder named "InspirarCiencia". The files are "Assemblag", "R1_50.fastq", and "R2_50.fastq". Each file has a checkmark icon and a small preview thumbnail. The total size of the selected files is 320,4 MB. A context menu is open over the files, with the following options: Mover ou copiar (Move or copy), Transferir (Transfer), Select file range (Select file range), and Eliminar (Delete). The menu is titled "... Ações".

	Ficheiro	Tamanho	Momento
✓	Assemblag	4,3 MB	há 2 anos
✓	R1_50.fastq	158 MB	há 2 anos
✓	R2_50.fastq	158,1 MB	há 2 anos



Galaxy Europe

[Analyze Data](#) [Workflow](#) [Visualize](#) [Shared Data](#) [Help](#) [Login or Register](#)

Using 0 bytes

Tools

[search tools](#) [Get Data](#)[Send Data](#)[Collection Operations](#)[GENERAL TEXT TOOLS](#)[Text Manipulation](#)[Filter and Sort](#)[Join, Subtract and Group](#)[GENOMIC FILE MANIPULATION](#)[Convert Formats](#)[FASTA/FASTQ](#)[FASTQ Quality Control](#)[Quality Control](#)[SAM/BAM](#)[BED](#)[VCF/BCF](#)[Nanopore](#)[COMMON GENOMICS TOOLS](#)[Operate on Genomic Intervals](#)[Fetch Sequences / Alignments](#)

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](#)

[Nov 24, 2020](#)

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](#)

Let's perform QC of the reads by using FastQC

The screenshot shows the Galaxy Europe web interface. On the left, a sidebar lists various tools and workflows. The 'Tools' section has a red circle around the 'fastqc' tool, which is highlighted in a dropdown menu. Below it are buttons for 'Upload Data' and 'Show Sections'. Other listed tools include 'FastQC Read Quality reports', 'Combine FASTA and QUAL into FASTQ', 'fastp', 'Map with PerM', 'Manipulate FASTQ', 'Create a model to recommend tools', and 'WORKFLOWS' (with 'All workflows' listed). The main content area is titled 'FastQC Read Quality reports (Galaxy Version 0.72+galaxy1)'. It contains sections for 'Short read data from your current history' (which says 'No fastq, fastq.gz, fastq.bz2, bam or sam dataset available.'), 'Contaminant list' (which says 'No tabular dataset available.'), 'Adapter list' (which says 'No tabular dataset available.'), 'Submodule and Limit specifying file' (which says 'No txt dataset available.'), and 'Disable grouping of bases for reads >50bp' (with a 'No' toggle switch). A note below explains that this option prevents the tool from crashing on long reads. At the bottom, there is a section for 'Lower limit on the length of the sequence to be shown in the report' and a note about setting a minimum length for sequence groups.

Should I be worried about the “stoplight” symbols?

NO

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per tile sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)
-  [Kmer Content](#)

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](#)

Tools   

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 (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

This dataset is large and only the first megabyte is shown below.
[Show all](#) | [Save](#)

```
>contig00001 len=132719 cov=21.6 corr=0 origname=Contig_113_21.5862 sw=shovill-skesa/1.1.0 date=20201209
CCCAACCTACTTTTGAGGAGACATCCCTGGAATAAGGGGGCTATTACGATAAATTAG
AGTCATAGATCAGTAGCCCATATTAACGTAGACTAATTCCGGCTACAGTTTATTATACC
GAGGCAGCTACTGTTTTACTTGCTGATATCGCTCATTCAATTCCGCTAACCGCGCT
TCGGTTACATCTTGCTTAGGATTAATCGCTGCTCAGTCCAACCGGGATACGTACCGCT
TTAAAAAAATCCATCAGACTCGCTGGTTATATTGTTCTGAGCCTCAACAATATGGAGT
TTGCTAAATCTGTGAGTTGCTCTCTCATTAGCATCTAATTGAGTACCGCTTTTGCT
ACTGCCGTTTAGAGGATTGTTGAACCTTAACCCCTTCCCAAATATTTTTGCT
TCTTCTATCGTTCAAAAATTGAACGTCTGACCTAATGAGTAATTACCAACTTGTCA
TAAACTAACTTATGAATGGAGAAACTAAGAAGAAGGTAATCTGCAAAACTATCACCT
ACTTCGTATATGCAGGACTAACATGGCGTACAATTTTCTATTTTATTCCTTATTTA
AAGATAAAAACCTATCCAAACTTGTATGCTGTTGACAACCTTGTGCTTATCCAAA
TAGCTGATTTCTGCTCTTGTACTCTACATTTCTATAGATCTTAAAGAATTACTGGC
```

History

search datasets

Unnamed history

9 shown

482.54 MB

9: Shovill on data 2 and data 6: Contigs

8: Shovill on data 2 and data 6: Contigs

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

The screenshot shows the NCBI BLAST homepage. At the top, there is a banner with COVID-19 information: "COVID-19 is an emerging, rapidly evolving situation. Get the latest public health information from CDC: <https://www.coronavirus.gov>. Get the latest research information from NIH: <https://www.nih.gov/coronavirus>. Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>". Below this, the main title is "Basic Local Alignment Search Tool". A description states: "BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance." There is a "Learn more" link. To the right, a "NEWS" sidebar mentions a webinar on December 9, 2020, and a Docker event on December 2, 2020. Below the main title, there are three search tool options: "Nucleotide BLAST" (nucleotide → nucleotide), "blastx" (translated nucleotide → protein), and "tblastn" (protein → translated nucleotide). To the right is "Protein BLAST" (protein → protein).

blast.ncbi.nlm.nih.gov/Blast.cgi

BLAST®

COVID-19 is an emerging, rapidly evolving situation.

Get the latest public health information from CDC: <https://www.coronavirus.gov>

Get the latest research information from NIH: <https://www.nih.gov/coronavirus>

Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

Using BLAST+ in Docker and on the cloud: [Webinar](#) on December 9, 2020.

In this webinar, the NCBI BLAST team will demonstrate containerized BLAST+ in Docker that is ready to use locally and in the cloud.

Wed, 02 Dec 2020 12:00:00 EST

More BLAST news...

Nucleotide BLAST

nucleotide → nucleotide

blastx

translated nucleotide → protein

tblastn

protein → translated nucleotide

Protein BLAST

protein → protein