Introduction on Sequence technologies, Common Bioinformatic file formats and BASH

Marcel van den Broek, Delft University of Technology (NL)
Marcel.vandenbroek@tudelft.nl

BiomeFun 2025, 16 September 2025

Sequencing Technology

DNA sequencing

- **DNA sequencing** is the process of determining the **nucleic acid sequence** – the order of nucleotides in DNA.

- It includes any method or technology that is used to determine the order of the four bases: **adenine, guanine, cytosine, and thymine**.

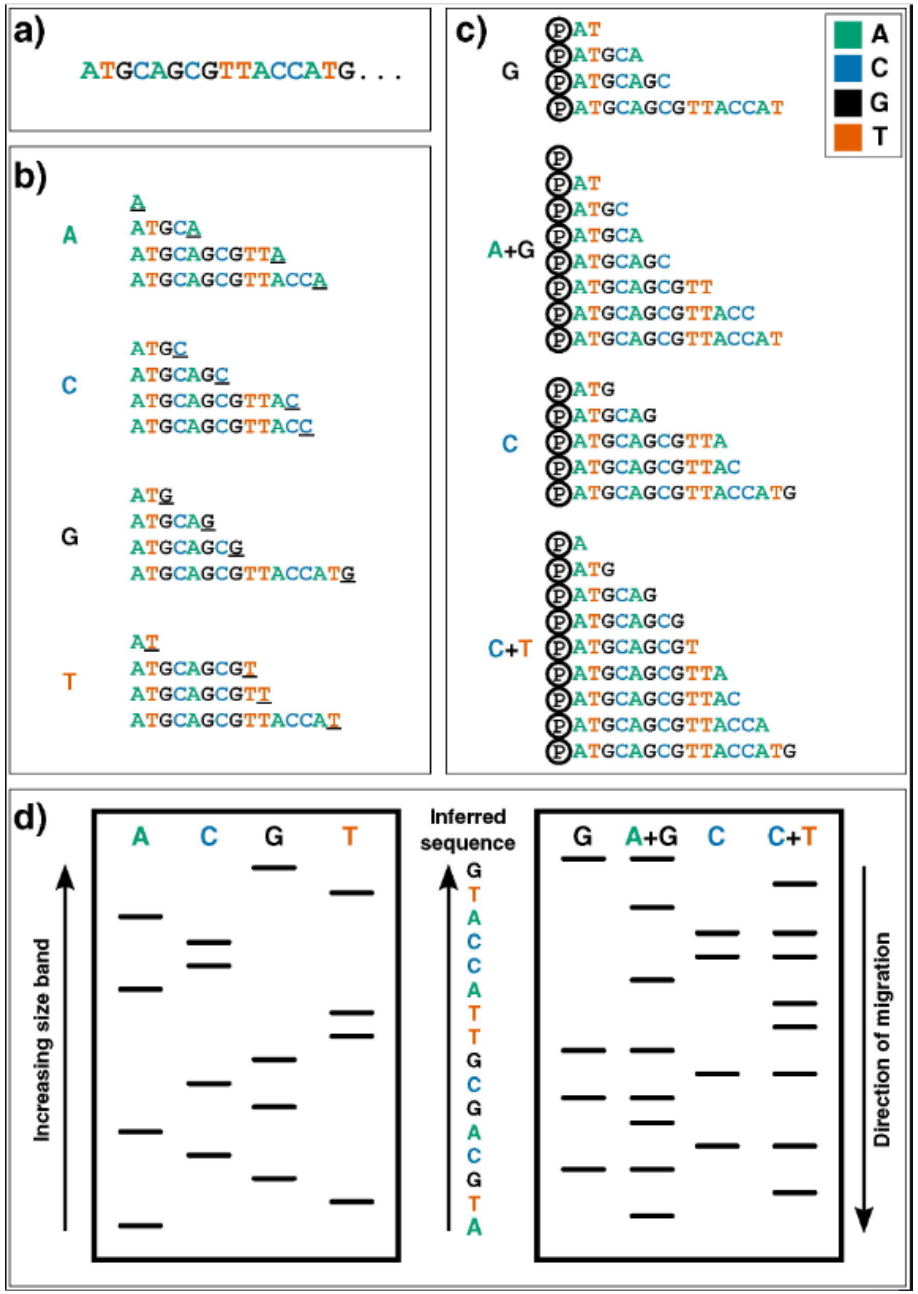
- The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery.

https://en.wikipedia.org/wiki/DNA_sequencing#

Applications

- Sequencing of:
  - Individual genes
  - Larger genetic regions (clusters of genes)
  - Full chromosomes
  - Entire genomes

History of sequencing



First generation:

1977: Frederick Sanger chain termination method.

(a) Example DNA to be sequenced.

(b) Sanger or **(c)** Maxam–Gilbert sequencing.

(b): Sanger. Radio- or fluorescently-labelled ddNTP nucleotides of a given type - which once incorporated, prevent further extension. Each of the four reactions, sequence fragments are generated with 3' truncations as a ddNTP is randomly incorporated at a particular instance of that base (underlined 3' terminal characters).

(d): Fragments visualized via electrophoresis on a high-resolution polyacrylamide gel: sequences are inferred by reading ‘up’ the gel. shorter DNA fragments migrate fastest.

Taken from Heather *et al.* The sequence of sequencers: The history of sequencing DNA
https://doi.org/10.1016/j.ygeno.2015.11.003

Sequencing Cost and Data Output

First generation

1977: Sanger chain termination method

1987: AB370 first automated instrument

1998: AB3730xl used in Human Genome Projects

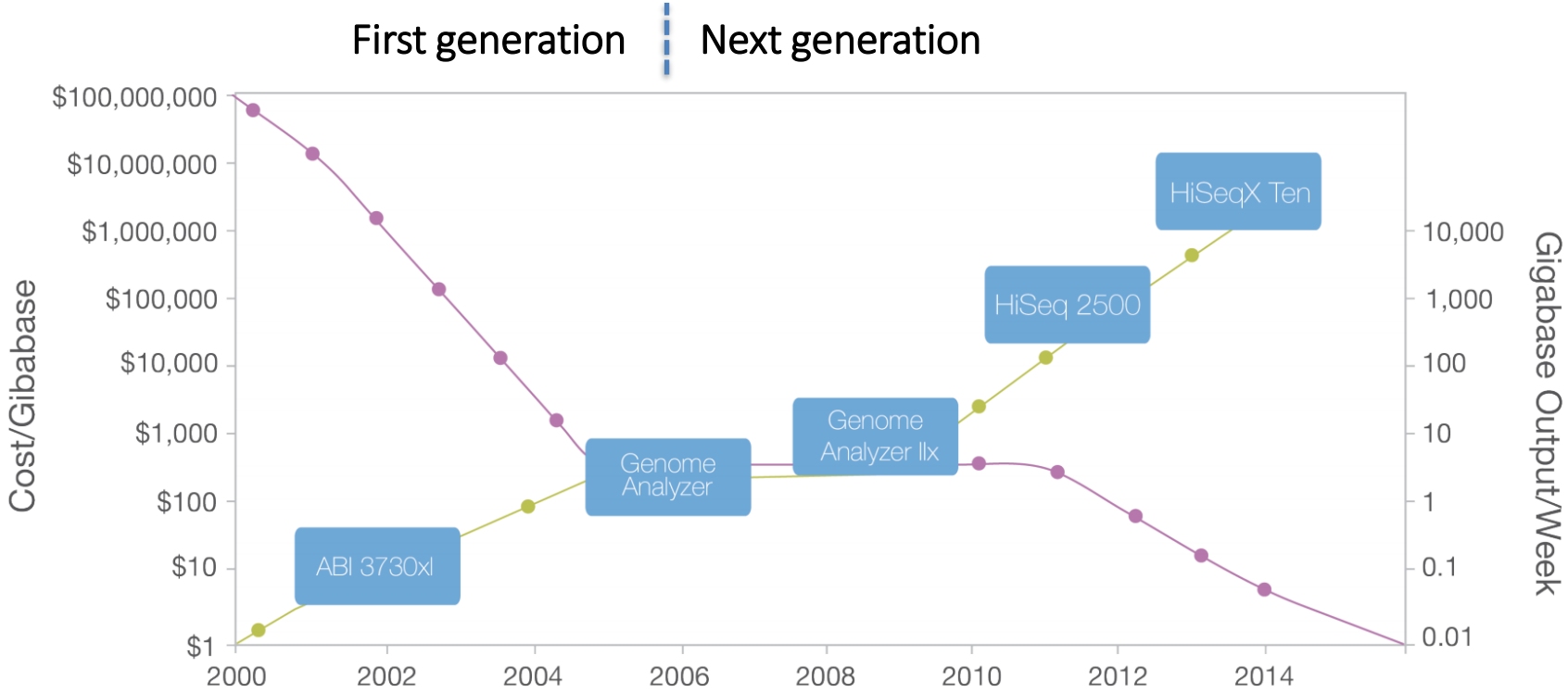


Figure 1: Sequencing Cost and Data Output Since 2000—The dramatic rise of data output and concurrent falling cost of sequencing since 2000. The Y-axes on both sides of the graph are logarithmic.

Next generation

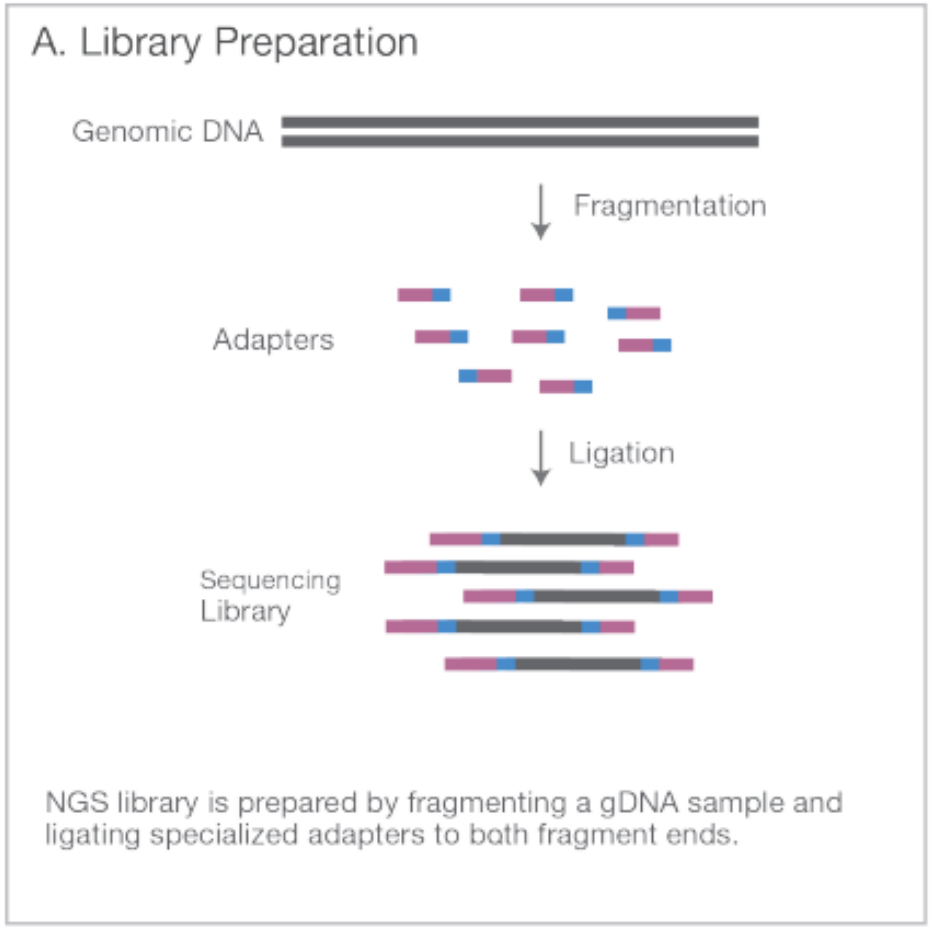
2005: Genome Analyzer: From 84 kilobase/run to 1 gigabase/run

2014: 1.8 terabases in a single run

Taken from Illumina sequencing introduction.pdf

Next or second-generation sequencing: 2005/6

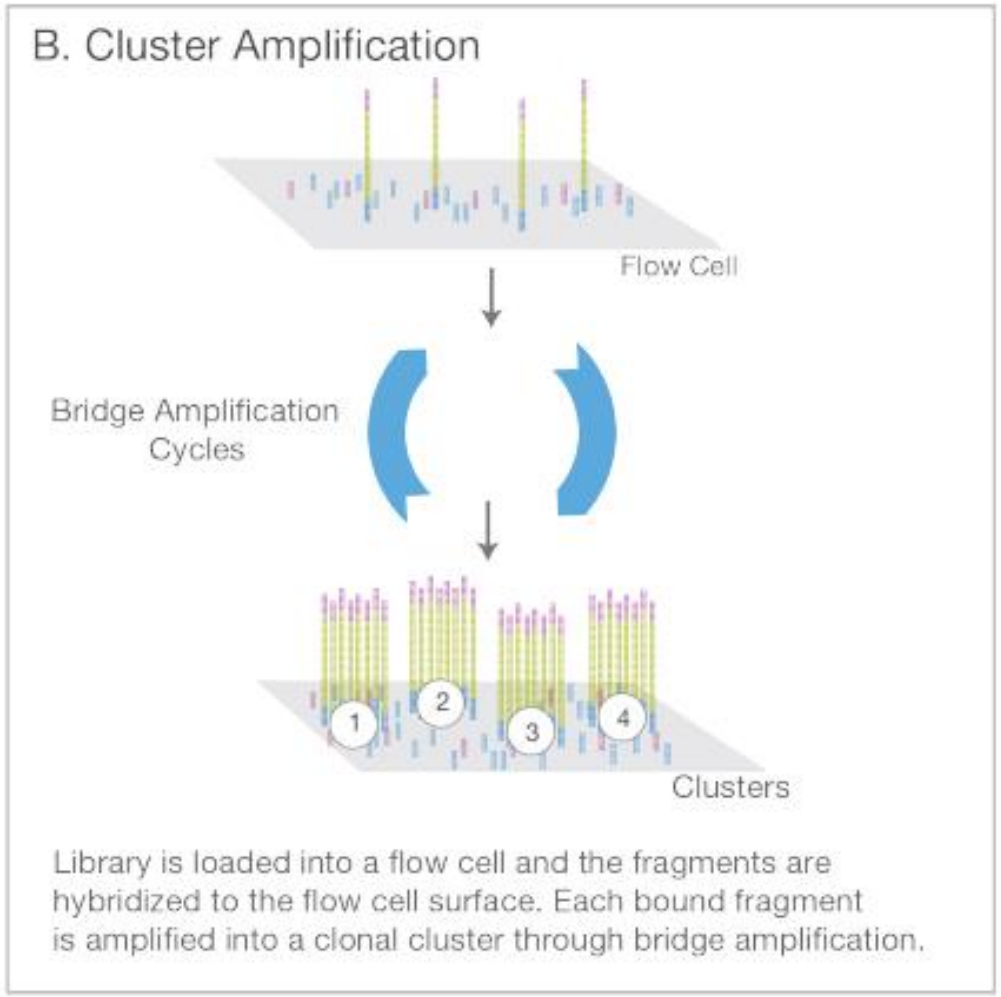
Illumina sequencing – Library Preparation



Taken from Illumina sequencing introduction.pdf

History of sequencing: 2005/6

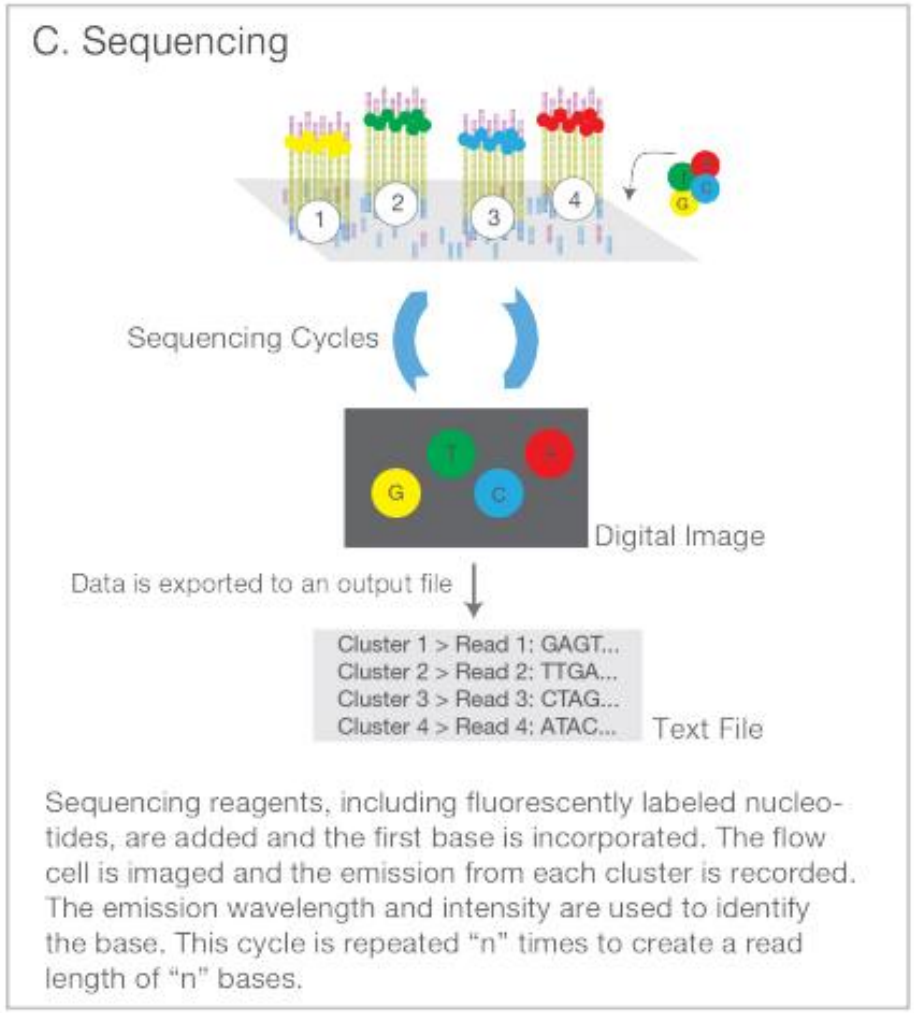
Illumina sequencing – Cluster Amplification



Taken from Illumina sequencing introduction.pdf

History of sequencing: 2005/6

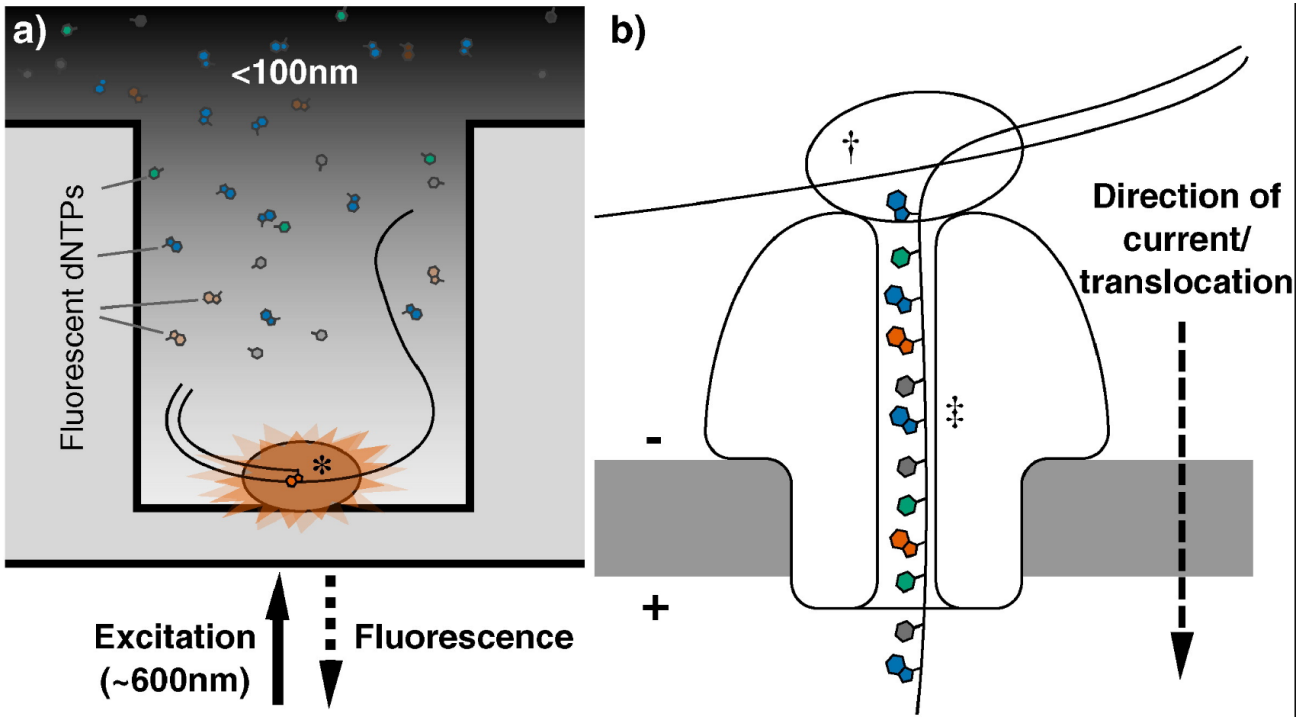
Illumina sequencing – Sequencing



Taken from Illumina sequencing introduction.pdf

Third generation sequencing: 2011/2012

Pacific Biosciences and Nanopore

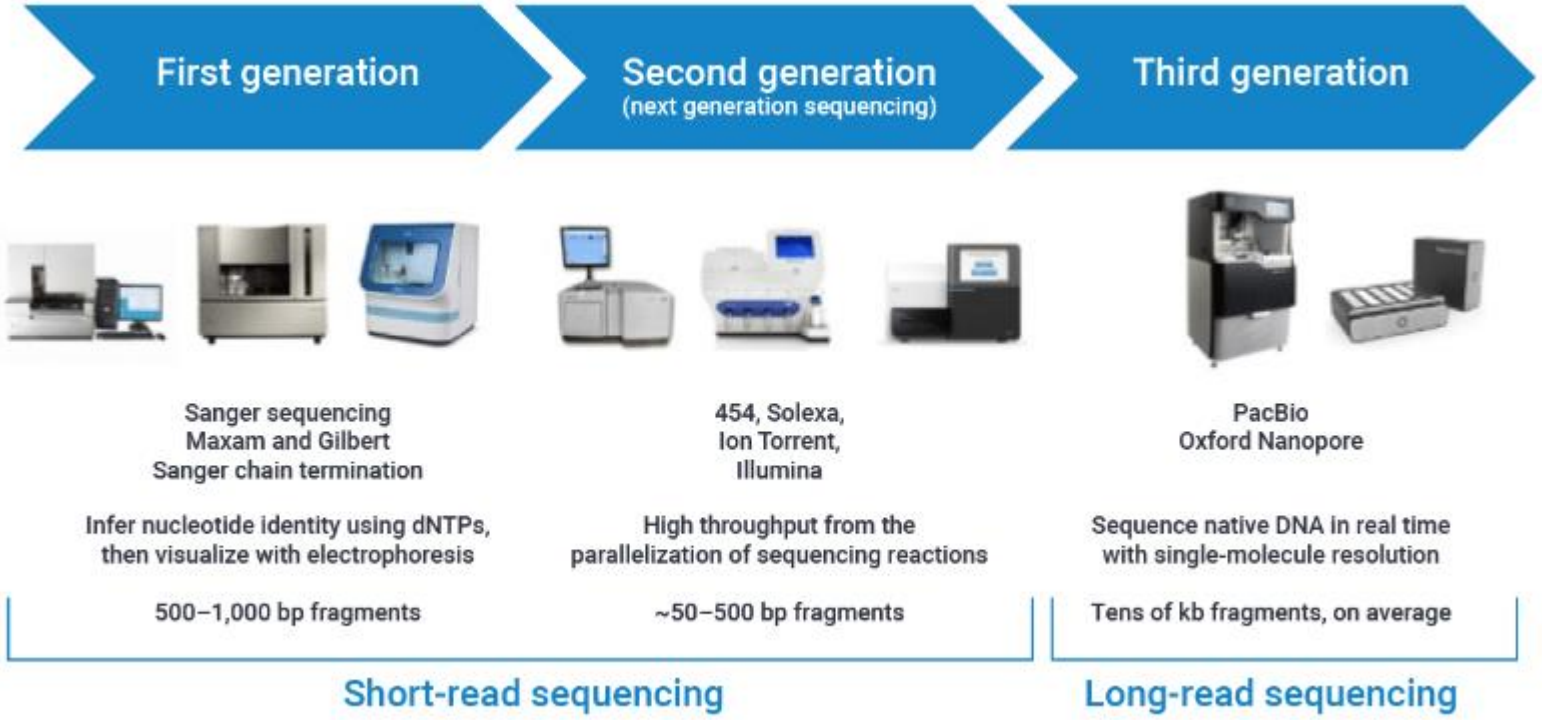


(A) **N**ucleotide detection in a zero-mode waveguide (ZMW), as featured in **PacBio sequencers**. **DNA polymerase molecules** are **attached to the bottom of each ZMW** (*), and target DNA and **fluorescent nucleotides are added**. As the diameter is narrower than the excitation light's wavelength, illumination rapidly decays travelling up the ZMW: nucleotides being incorporated during polymerisation at the base of the ZMW provide real-time bursts of fluorescent signal, without undue interference from other labelled dNTPs in solution.

(B) **Nanopore** DNA sequencing as employed in ONT's MinION sequencer. **Double stranded DNA** gets denatured by a processive enzyme (†) which ratchets one of the strands through a biological nanopore (‡) embedded in a synthetic membrane, across which a voltage is applied. As the **ssDNA passes through the nanopore** the different bases prevent ionic flow in a distinctive manner, allowing the sequence of the molecule to be inferred by monitoring the current at each channel.

Taken from van Dijk *et al.* The Third Revolution in Sequencing Technology
https://doi.org/10.1016/j.tig.2018.05.008

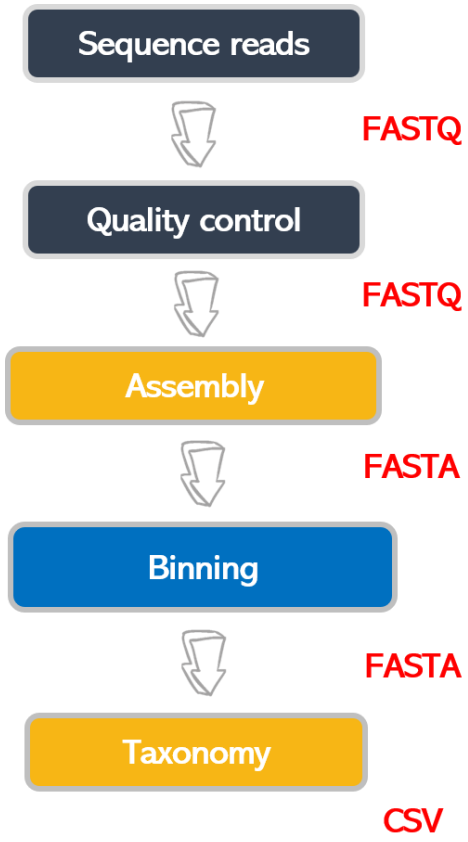
History of sequencing



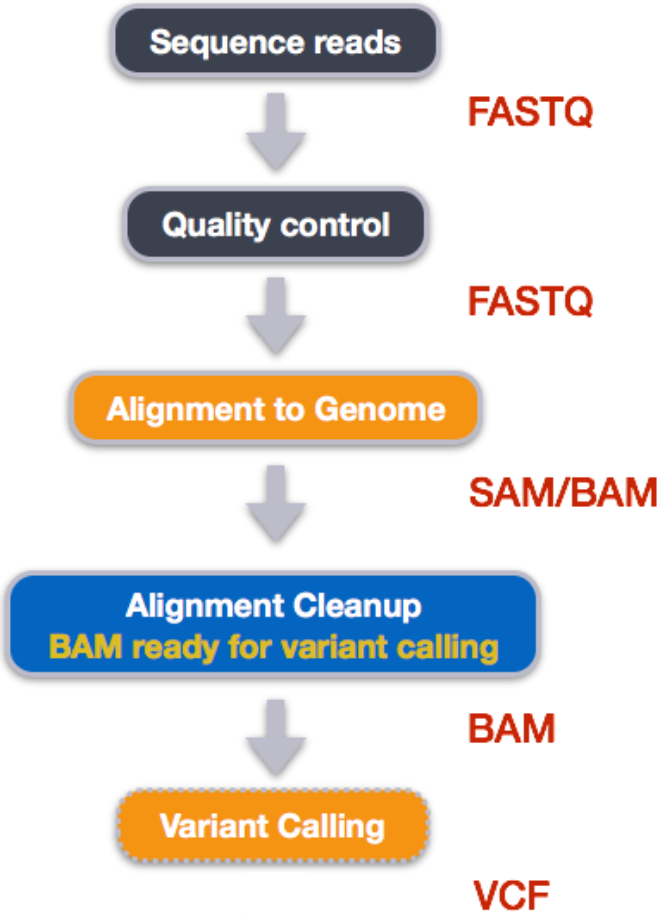
Taken from https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

genomics pipelines and file formats

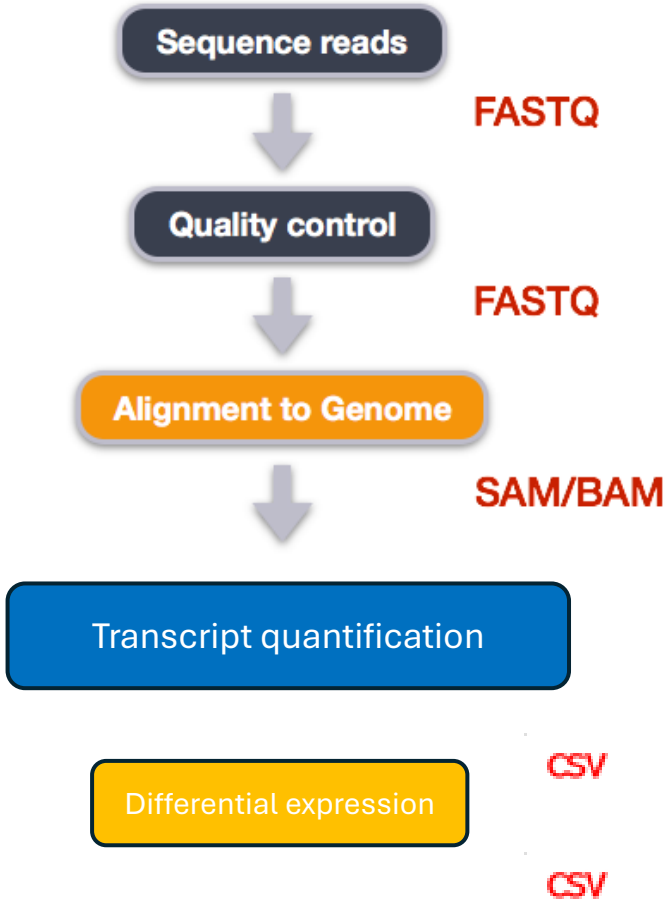
Meta-genomics



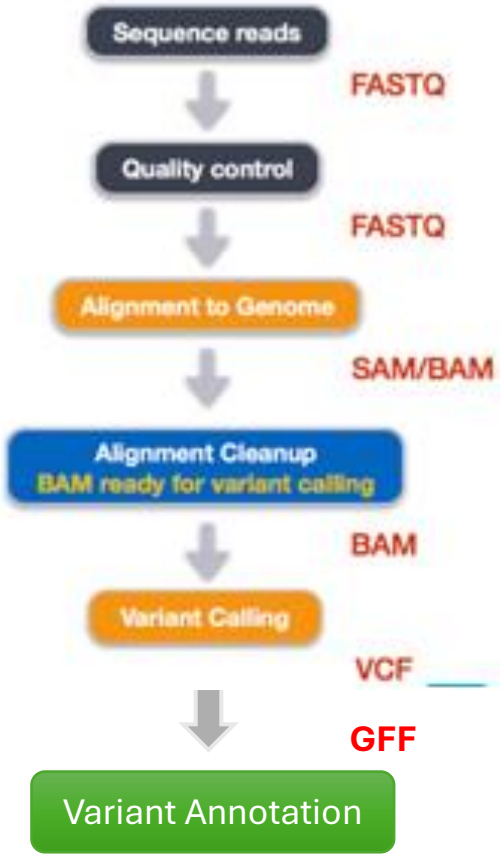
Variant calling

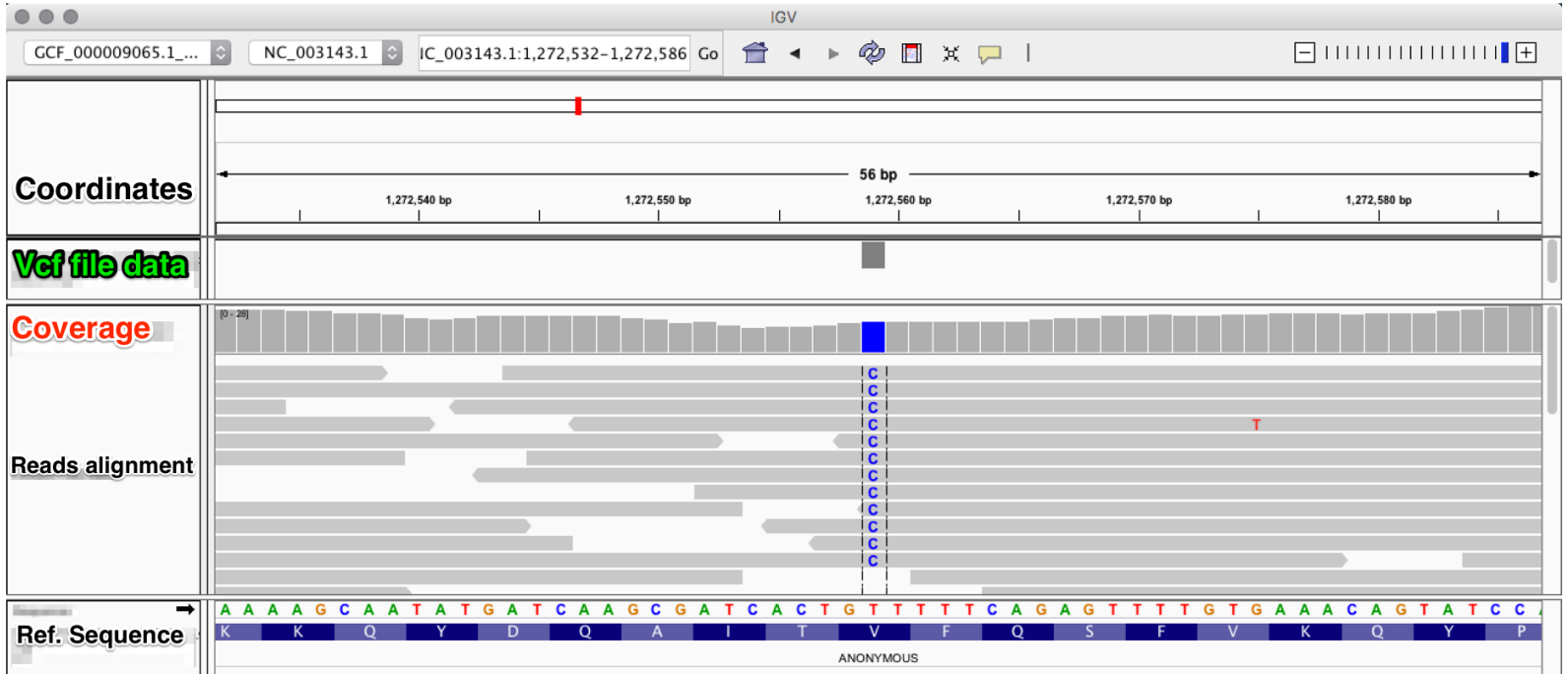


RNAseq



Variant Calling





Annotated VCF

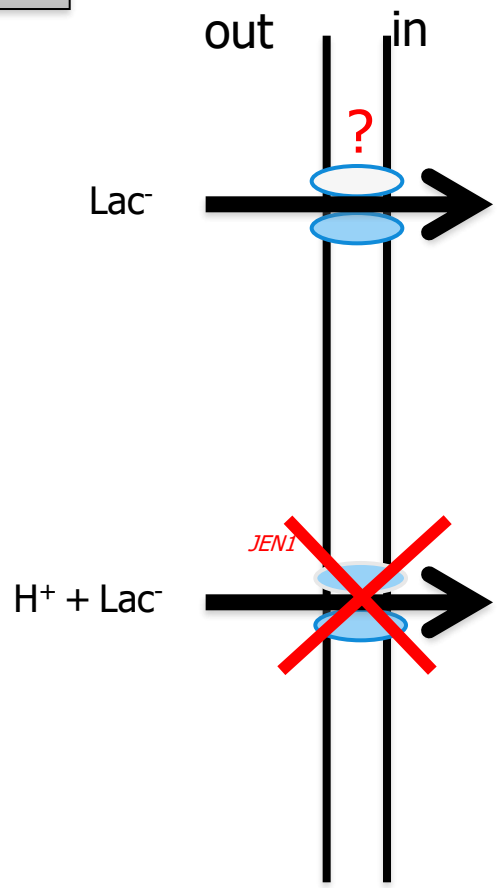
Visualisation with Integrated Genome Viewer (IGV)

Lactate transport in *S. cerevisiae*

Casal *et al.*, J.Bacteriol. 181, 2620-2623, 1999

- *JEN1*: only importer lactic acid
- *jen1Δ* is able to grow on lactate ($\mu<0.001\ h^{-1}$)

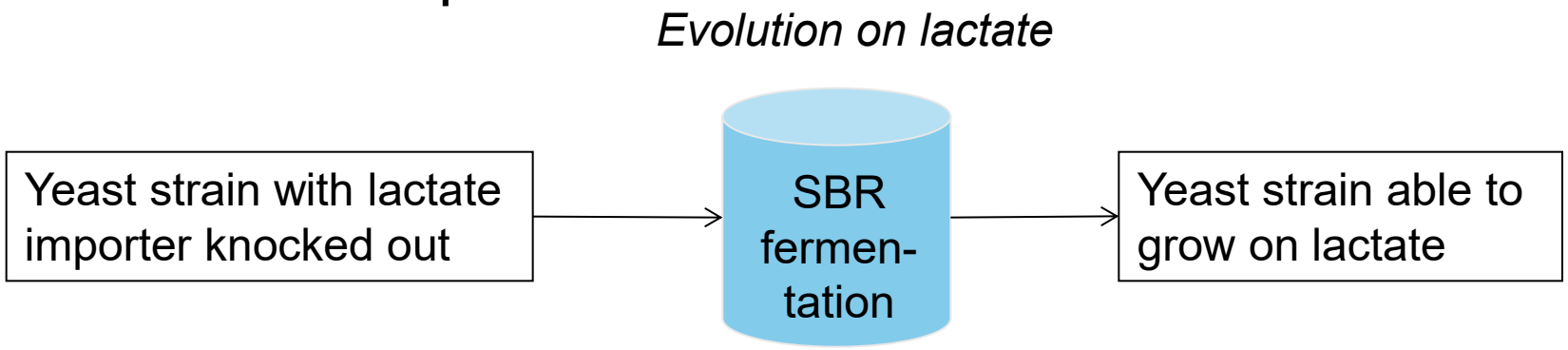
caused by a nuclear, monogenic mutation. The original mutant was named BLC 55, and a spore from a cross presenting the lactate-negative phenotype was termed BLC 142. In both cases, faint, residual growth on lactate was always observed. A strain mutated in the *PCK1* gene, encoding phosphoenolpyruvate carboxykinase, or a double mutant with alterations in the *CYB2* and *DLD* genes (12), encoding the D- and L-lactate



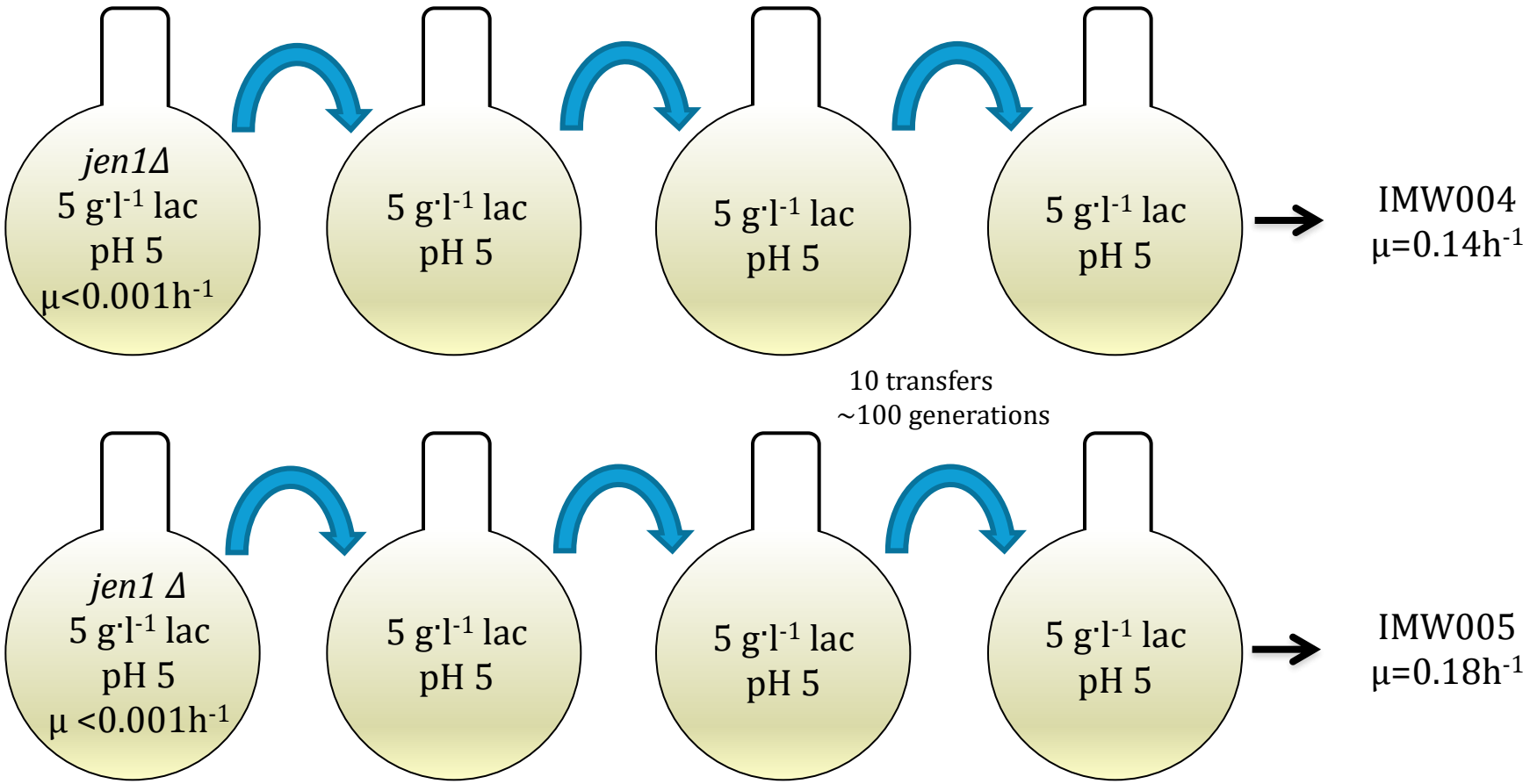
- Evolutionary engineering to identify additional lactate transporter.

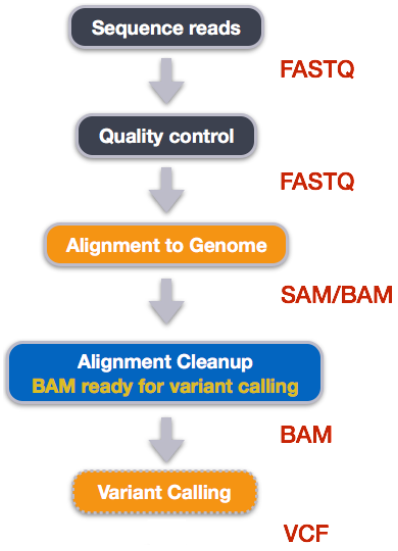
Experimental set-up

- The lactate importer *Jen1* has been knocked out
- Yeast has evolved with lactate as sole carbon source in serial batch reactors
- Result: a yeast strain able to grow on lactate *without* the lactate importer *Jen1*



Parallel *jen1Δ* evolution



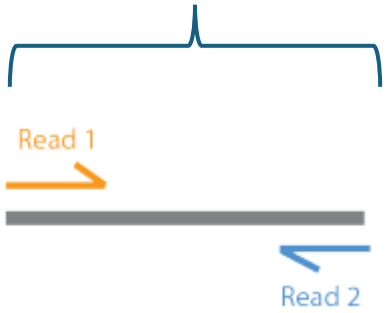


Whole genome sequencing

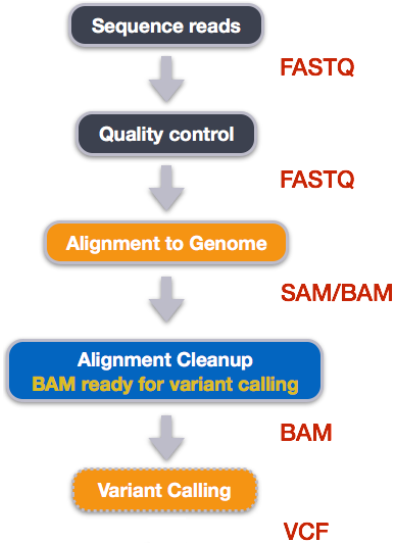
- Genomic DNA from IMW004 and IMW005 was isolated using the Qiagen 100/G kit (Qiagen, Hilden, Germany). A library of 200-bp genomic fragments was created and sequenced paired-end (50-bp reads) using an Illumina HiSeq 2000 sequencer by Baseclear BV (Baseclear).

Paired-end library

200 bp fragment length



50 bp read length

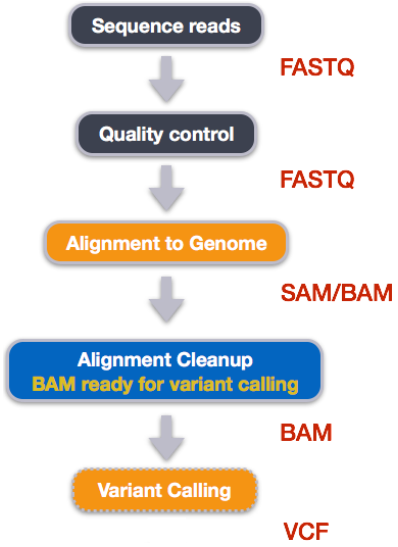


FASTQ Format

- Similar to FASTA format, but also contains quality information.
- Single record (sequence read) consists of four lines:

```
@HWI-ST330:304:H045HADXX:1:1101:1111:61397
CACTTGTAAGGGCAGGCCCCCTTCACCCTCCCGCTCCTGGGGGANNNNNNNNNNANNNCGAGGCCCTGGGGTAGAGGGNNNNNNNNNNNNNNGATCTTGG
+
@?@DDDDDDHHH?GH:?FCBGGB@C?DBEGIIIIAEF;FCGGI#########################################################
```

Line	Description
1	Always begins with '@' and then information about the read
2	The actual DNA sequence
3	Always begins with a '+' and sometimes the same info in line 1
4	Has a string of characters which represent the quality scores; must have same number of characters as line 2



FASTQ Format (Phred scores)

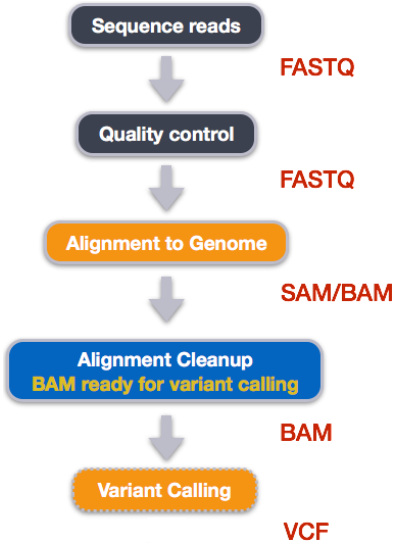
- Line 4 has characters encoding the quality of each nucleotide in the read.

```
@HWI-ST330:304:H045HADXX:1:1101:1111:61397
CACTTGTAAGGGCAGGCCCCCTTCACCCTCCCGCTCCTGGGGGANNNNNNNNNNANNNCGAGGCCCTGGGGTAGAGGGNNNNNNNNNNNNNNGATCTTGG
+
@?@DDDDDDHHH?GH:?FCBGGB@C?DBEGIIIIAEF;FCGGI#########################################################
```

- The legend below provides the mapping of quality scores (Phred-33) to the quality encoding characters.

```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                  |         |         |         |         |
   Quality score: 0........10........20........30........40
```

- The second nucleotide in the read (A) is called with a quality score of 30.
- (A->?->30)



FASTQ Format (accuracy)

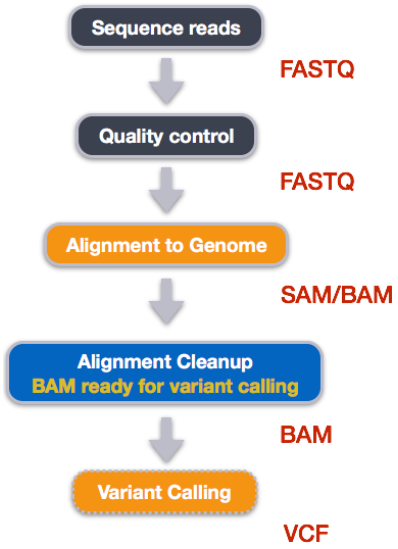
- Each quality score represents the probability that the corresponding nucleotide is incorrect. (A->?->30)

```
Q = -10 x log10(P), where P is the probability that a base call is erroneous
```

- The score values can be interpreted as:

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

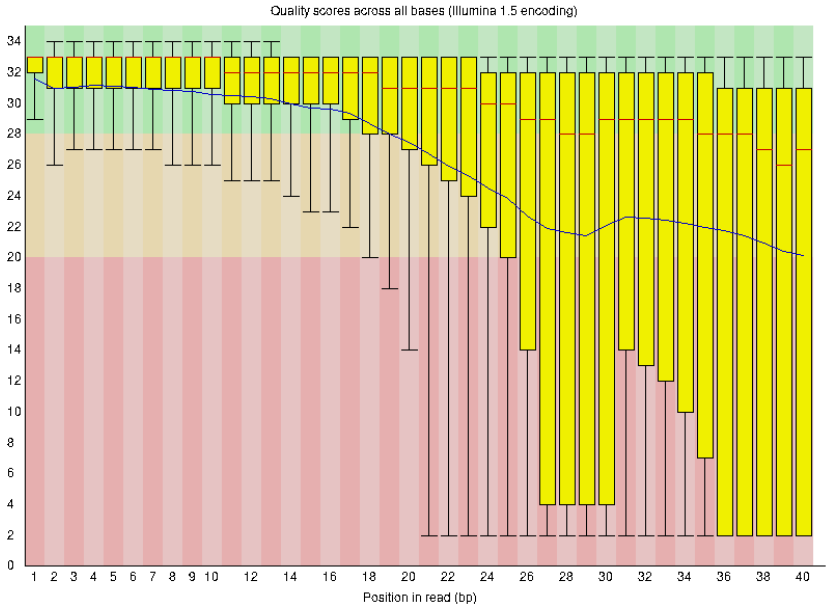
- The second nucleotide in the read (A) is less than a 1 in 1000 chance that the base was called incorrectly



Quality control and Trimming/filtering

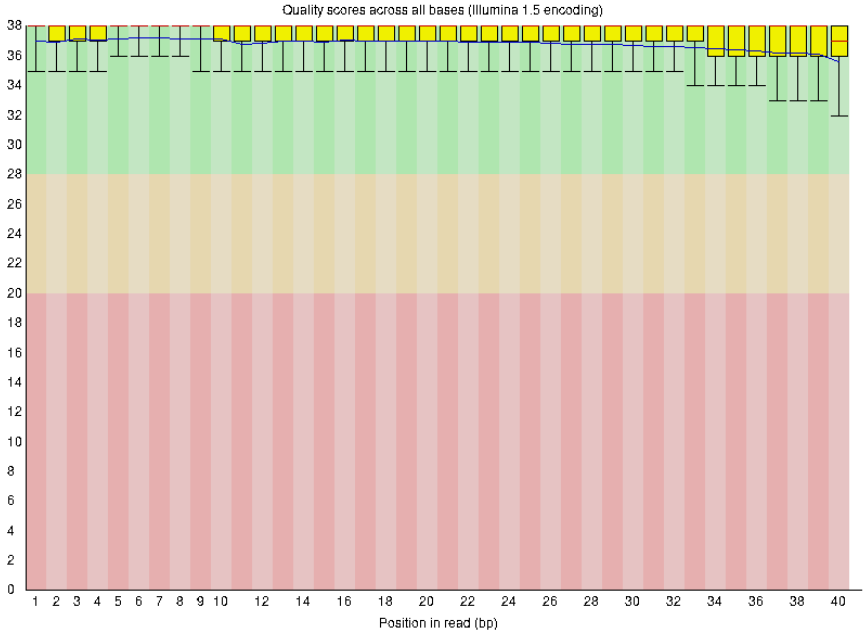
- FastQC: High throughput Quality control.
- Trimmomatic: Illumina quality trimming tool.

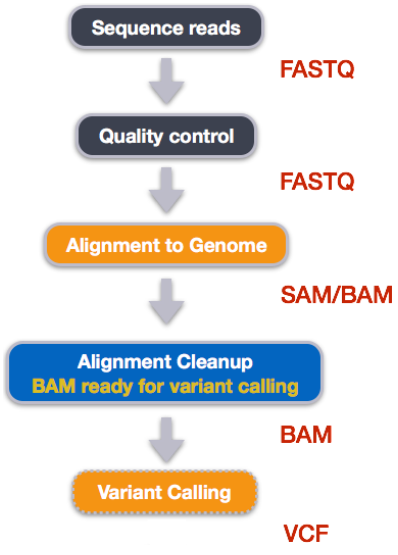
Raw sequencing reads



Trimmomatic

Trimmed/filtered sequencing reads

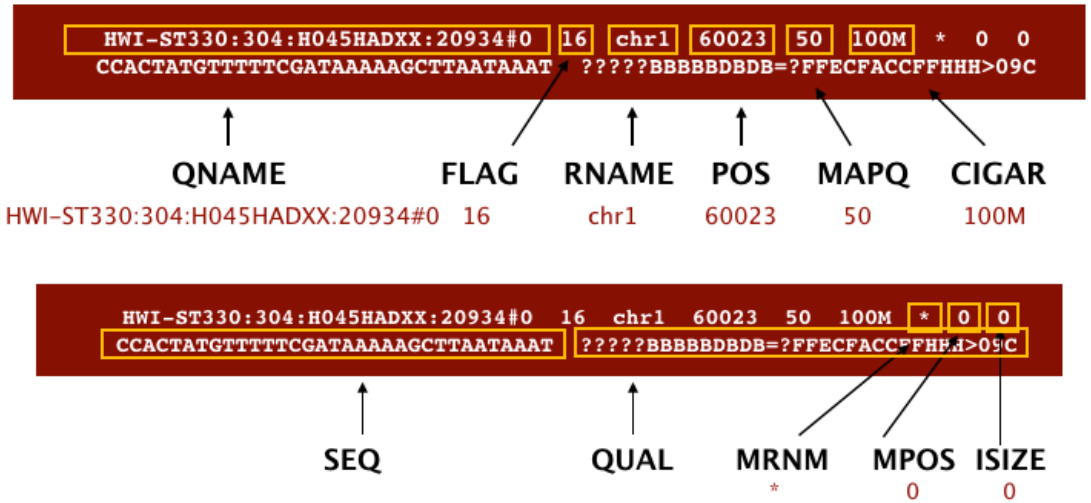


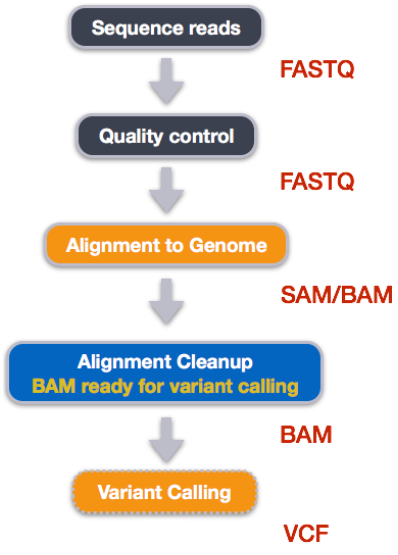


Sequence Alignment (Map): SAM

- To detect variants in the samples, reads are aligned to a close reference and stored in a SAM file.
- The SAM file is a tab-delimited text file that contains information for each individual read and its alignment to the genome.
- Each line corresponds to alignment of a single read.
- Each alignment has 11 mandatory fields.

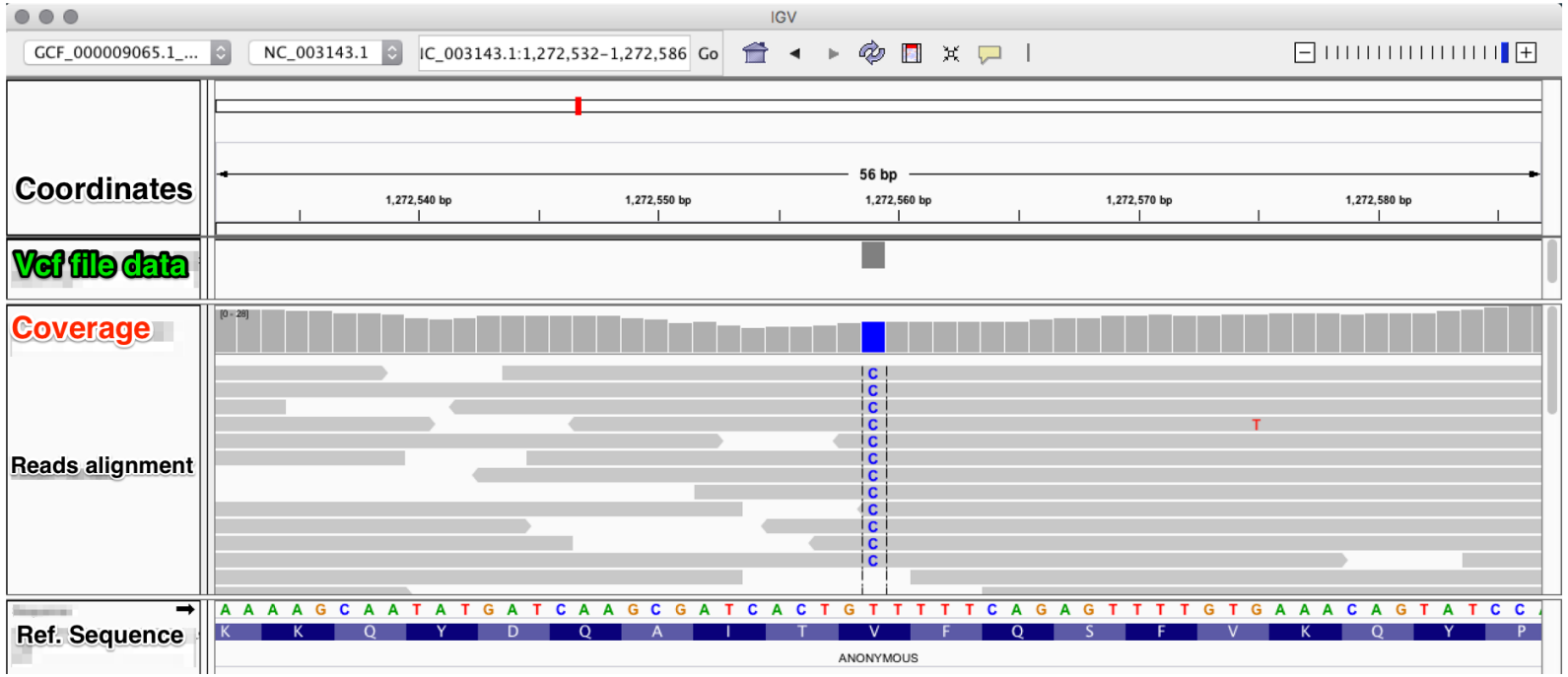
QNAME	Read identifier
RNAME	Ref sequence name
POS	Read mapping position
MAPQ	Mapping quality
SEQ	Read sequence
QUAL	Quality scores
MPOS	Read mapping position of mate
ISIZE	Insert size

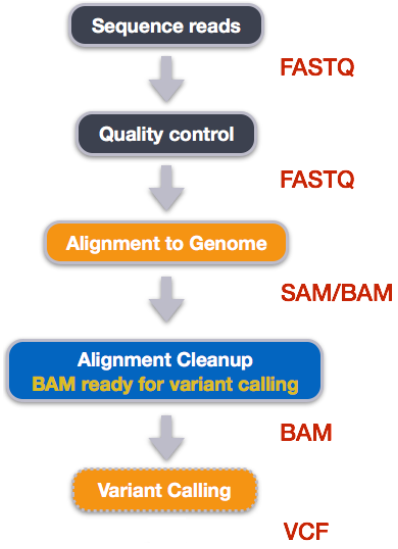




(Sorted) BAM

- The compressed binary version of SAM is called a BAM file.
- Has reduced file size and allows for indexing, which enables efficient random access of the data necessary for downstream analysis and visualisation.
- To be able to call variants BAM needs to be sorted on reference coordinates.





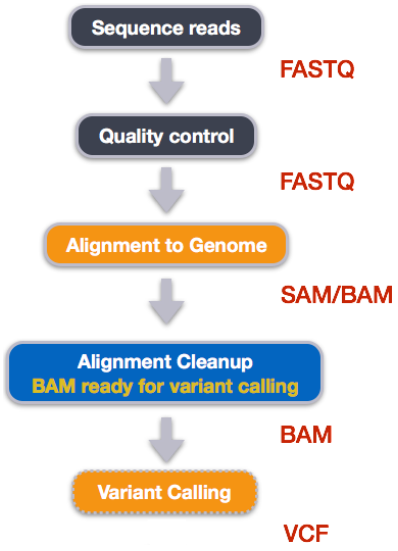
Variant Calling (File)

- A variant call is a conclusion that there is a nucleotide difference vs a reference at a given position.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr1	1521	.	C	G	207	.	DP=32;MQ=55;BC=0,0,32,0
chr2	10563	.	T	A	225	.	DP=40;MQ=60;BC=40,0,0,0

column	info
CHROM	contig location where the variation occurs
POS	position within the contig where the variation occurs
ID	a `.` until we add annotation information
REF	reference genotype (forward strand)
ALT	sample genotype (forward strand)
QUAL	Phred-scaled probability that the observed variant exists at this site (higher is better)
FILTER	a `.` if no quality filters have been applied, PASS if a filter is passed, or the name of the filters this variant failed

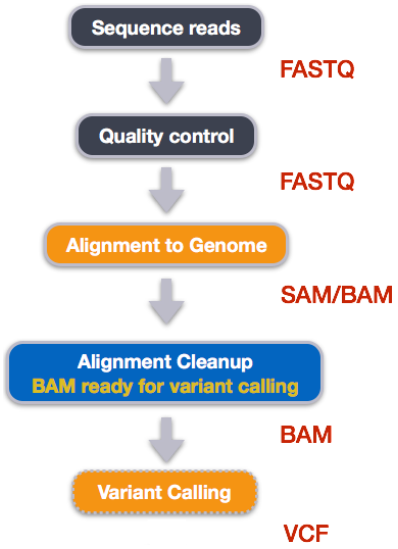
DP	Depth
MQ	Mapping Quality
BC=	Base Count



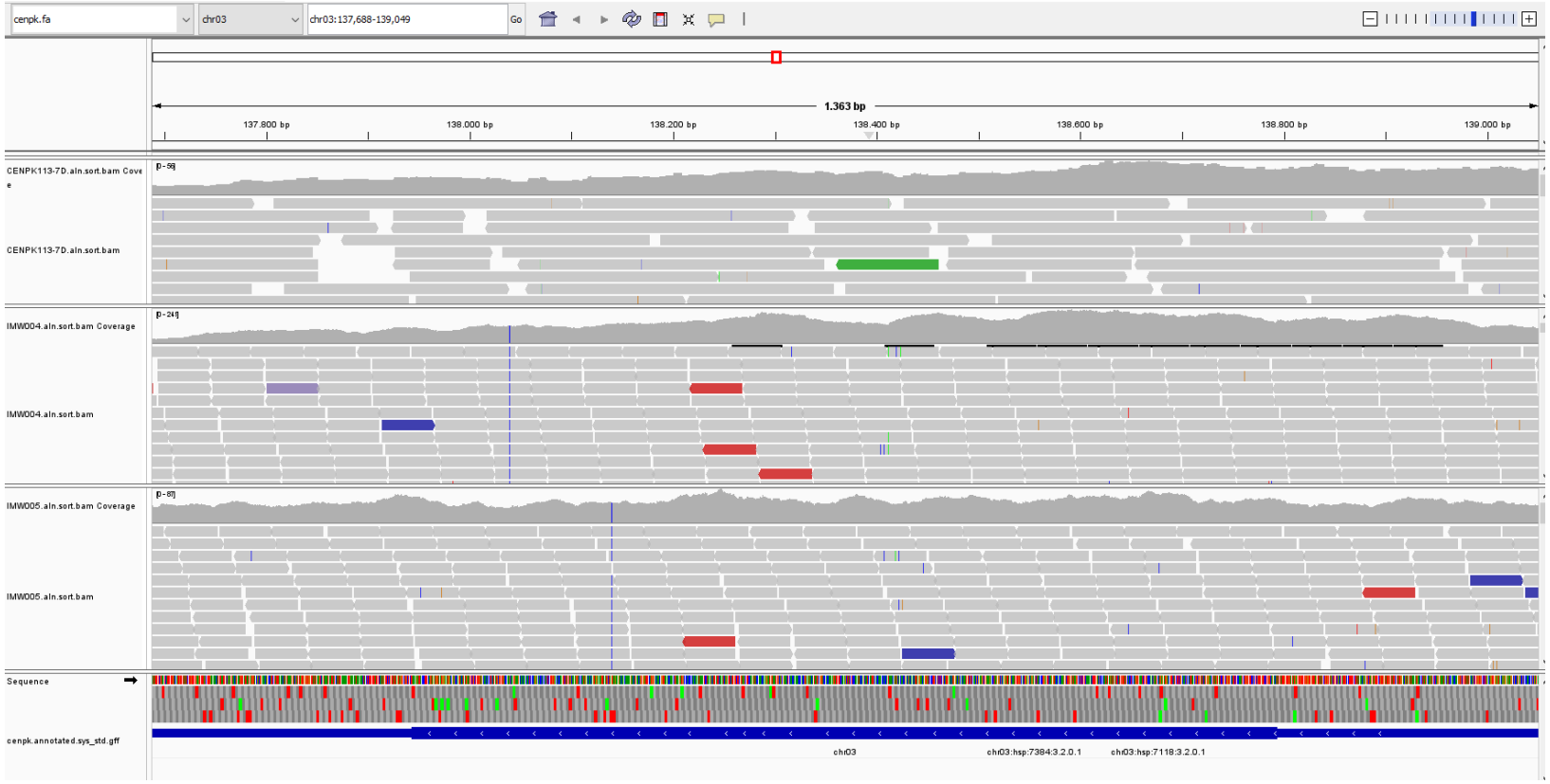
General Feature File (GFF)

- A GFF file is used to describing gene annotations and other features of DNA, RNA and protein sequences.

Seqid	source	type	Start	End	Score	Strand	phase	attributes
Chr03	maker	gene	137943	138794	.	-	.	ID=gene122;Name=ADY2
Chr03	maker	mRNA	137943	138794	.	-	.	ID=gene122-mRNA;Parent=gene122
Chr03	maker	CDS	137943	138794	.	-	.	ID=gene122-mRNA-cds;Parent=gene122-mRNA
Chr03	maker	exon	137943	138794	.	-	.	ID=gene122-mRNA-exon;Parent=gene122-mRNA



Annotated Variant Calls



WT

IMW004

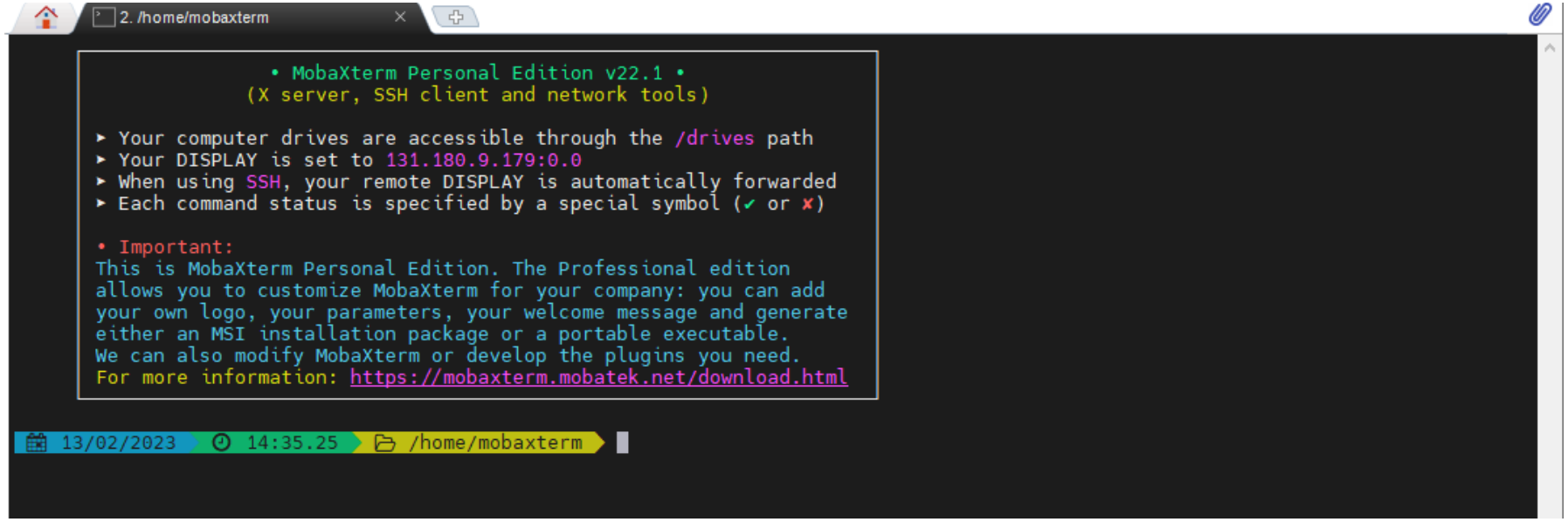
IMW005

Sample	Gene	Description	Nucleotide change	Amino acid change
IMW004	ADY2	Acetate transporter	C755G	Ala252Gly
IMW005	ADY2	Acetate transporter	C655G	Leu219Val

Command Line for Genomics

What is the shell?

A *shell* is a computer program that presents a command line interface which allows you to control your computer using commands entered with a keyboard instead of controlling graphical user interfaces (GUIs) with a mouse/keyboard combination.

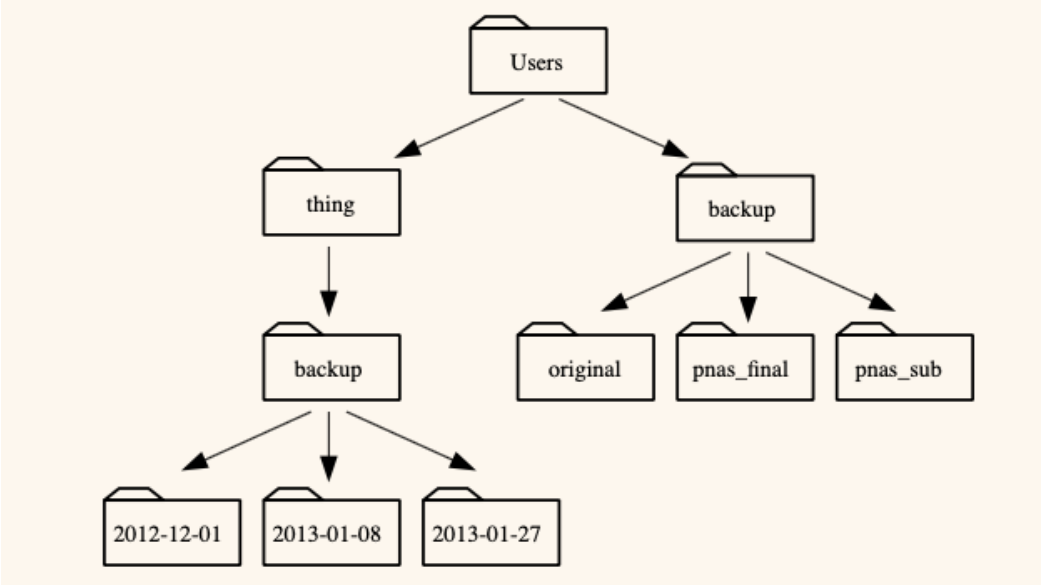


Why to use a shell

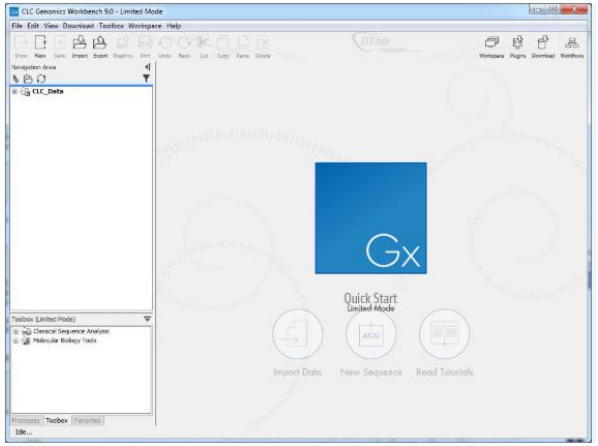
- **Automate repetitive tasks**
  - Makes your work less error-prone and more reproducible.
- **Computing power**
  - Many bioinformatic tasks can’t realistically be run on your own machine.
- **Tool availability**
  - Many tools can only be used through a command line interface.
  - Tools like BLAST, has advanced options only via the command line.

Shell commands, navigating file system

- ls listing folder content
- pwd print working directory
- cd change directory

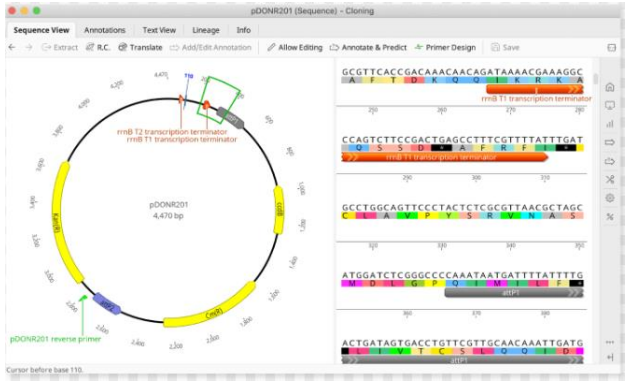


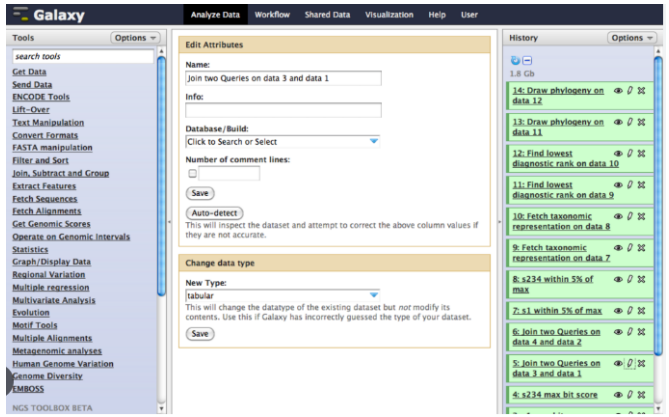
GUI Software?



- CLC Genomics Workbench (QIAGEN)
- Geneious (Dotmatics)

Expensive, Yearly fee





- Galaxy, Open-source web-based platform

Which Cloud?

Commercial Clouds

Azure — Microsoft Azure

aws — Amazon WebServices
EC2 – Amazon Elastic Compute Cloud

Google Cloud — Google Compute Engine

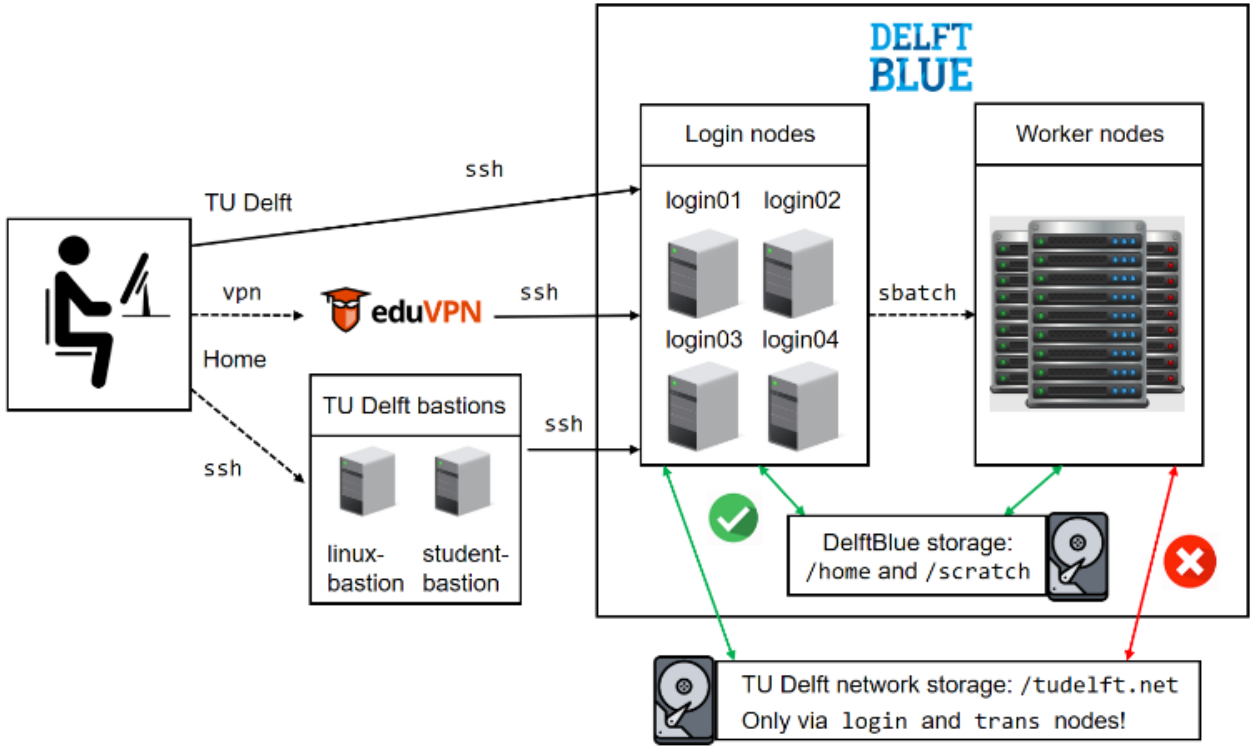
Educational Clouds

SURF SARA — SURF National HPC (High Performance Computing)

TU Delft — Delft HPC – DelftBlue
Available since 2022
https://www.tudelft.nl/dhpc

imb — Own server at IMB
64 Cores
384 GB RAM

High Performance Computing example



Summary and performance:

CPU total	**Compute nodes**	**338**
	CPU's	**676**
	Compute cores	**17,842**
	Rpeak (theoretical, in DP PFlop/s)	**1.45**

GPU total	**GPU nodes**	**20**
	GPU's	**80**
	Tensor cores	**42,880**
	CUDA cores	**481,280**
	Rpeak (theoretical, in DP PFlop/s)	**0,61**

Operating System: Red Hat Enterprise Linux 8

SLURM workload manager

- Job scheduling system

```
#!/bin/bash

#SBATCH --job-name="Py_pi"
#SBATCH --time=00:10:00
#SBATCH --ntasks=8
#SBATCH --cpus-per-task=1
#SBATCH --partition=compute
#SBATCH --mem-per-cpu=1GB
#SBATCH --account=research-<faculty>-<department>

module load 2023r1
module load openmpi
module load python
module load py-numpy
module load py-mpi4py

srun python calculate_pi.py > pi.log
```

#SBATCH --job-name="Py_pi“:	Name of the job
#SBATCH --time=00:10:00:	set run duration.
#SBATCH --ntasks=8:	set number of cores
#SBATCH --cpus-per-task=1:	set number of threads
#SBATCH --partition=compute:	set the partition
#SBATCH --mem-per-cpu=1GB:	set the amount of RAM/core
#SBATCH --account=innovation:	set your account

Load central installed software via module load

srun will start 8 instances of python that can communicate via MPI. Each instance willbe allowed to use one CPU core with a single thread.

Amazon EC2

- In this course we will use Amazon Elastic Compute Cloud (EC2).
- Scalable infrastructure; from one simple computer to a whole data centre.

- Used AWS instances: t3.large, 2 vCPU, 8 GB RAM / user

- Login example: $ ssh user@3.70.236.93