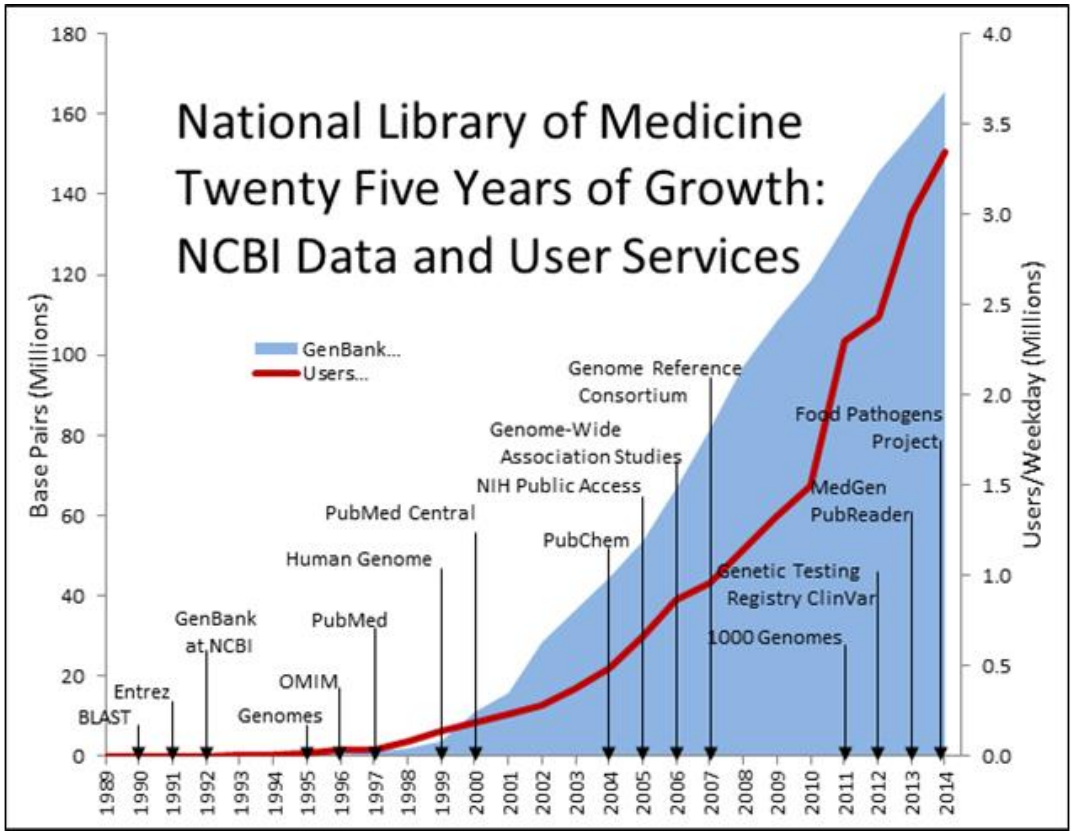

High Performance
Bioinformatics and
Biomedical research

Bioinformatics is entering the Big Data hera.

With the advent of ever-evolving, high throughput -omics technologies, we are observing an explosive growth in the volume of biological and biomedical data.

The monumental completion of human genome sequencing ignited the generation of big biomedical data.

BIG DATA ----> BIG COMPUTE !!!



Big data bioinformatics examples are the growth of DNA sequence information in NCBI's GenBank database and the growth of prote sequences in the UniProt database.

Emerging next-generation sequencing technologies have broken many experimental barriers to genome scale sequencing, facilitatir the extraction of huge quantities of sequences, which will further promote the future growth of biological databases.

Two example of huge massive sequencing repositories

1. The Cancer Genomics Hub – CGHub (https://cghub.ucsc.edu/) is a secure repository for storing, cataloguing and accessing cancer genome sequences, alignments and mutation information from The Cancer Genome Atlas (TCGA) consortium and related projects. The TCGA Data Portal (https://tcga-data.nci.nih.gov) provides access to biological relevant information about the molecular changes in cancer genome datasets.

This allows researchers to attack the molecular complexity of cancer by combining analysis of hundreds of clinical tumours. The aim is to interpret molecular profiles at the DNA, RNA (ribonucleic acid), protein and epigenetic levels as markers of several tumour types and their subtypes.

2. The Genotype-Tissue Expression – GTEx database (https://commonfund.nih.gov/GTEx) has been provided by NIH. It allows to study the relationship between genetic variation and gene expression in human tissues. Today the release V6 of the database contains 8555 samples.
This resource enables large-scale analyses on genetic variation and regulation of gene expression in multiple reference human tissues in order to identify normal variations that do not contribute to disease and to study how gene variations affect pharmacodynamics and individualised responses to therapy.

Four domains of Big Data in 2025

In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

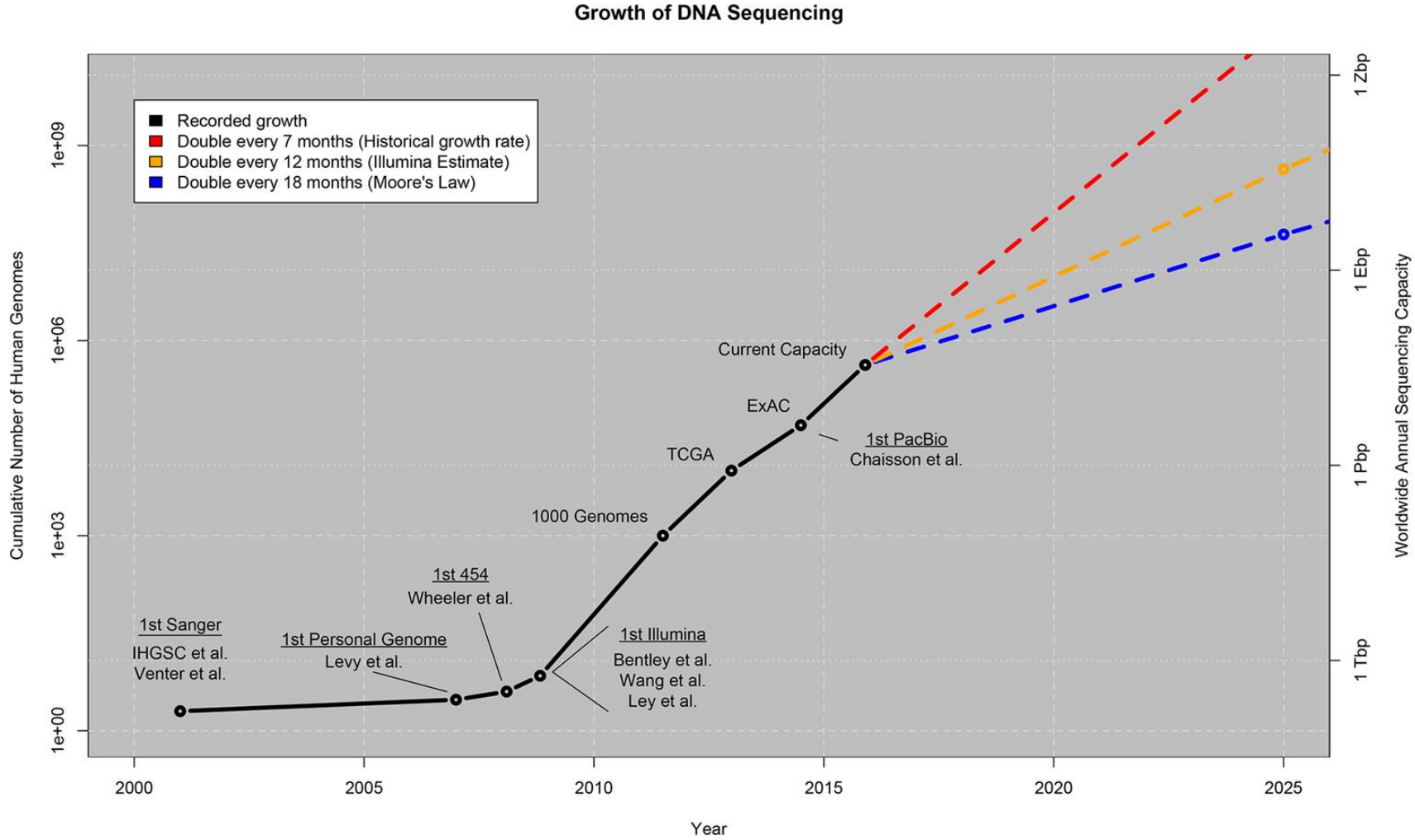
"Genomics is a "four-headed beast"--it is either on par with or the most demanding of the domains analyzed here in terms of data acquisition, storage, distribution, and analysis."

Big Data: Astronomical or Genomical? Zachary D. et al.

PLoS Biol. 2015 Jul 7;13(7):e1002195. doi: 10.1371/journal.pbio.1002195. eCollection 2015.

Decimal	
Value	Prefix
1000	k kilo
1000^2	M mega
1000^3	G giga
1000^4	T tera
1000^5	P peta
1000^6	E exa
1000^7	Z zetta
1000^8	Y yotta

Prefixes for multiples of bytes.
Image from Wikipedia.


Growth of DNA Sequencing
Recorded growth
Double every 7 months (Historical growth rate)
Double every 12 months (Illumina Estimate)
Double every 18 months (Moore's Law)
Cumulative Number of Human Genomes
1e+00
1e+03
1e+06
1e+09
Worldwide Annual Sequencing Capacity
1 Tbp
1 Pbp
1 Ebp
1 Zbp
1st Sanger
IHGSC et al.
Venter et al.
1st Personal Genome
Levy et al.
1st 454
Wheeler et al.
1st Illumina
Bentley et al.
Wang et al.
Ley et al.
1000 Genomes
TCGA
ExAC
Current Capacity
1st PacBio
Chaisson et al.
2000
2005
2010
2015
2020
2025
Year

Sanger sequencing or 1st Generation Sequencing

Sanger Sequencing (1977 ⇒ present)

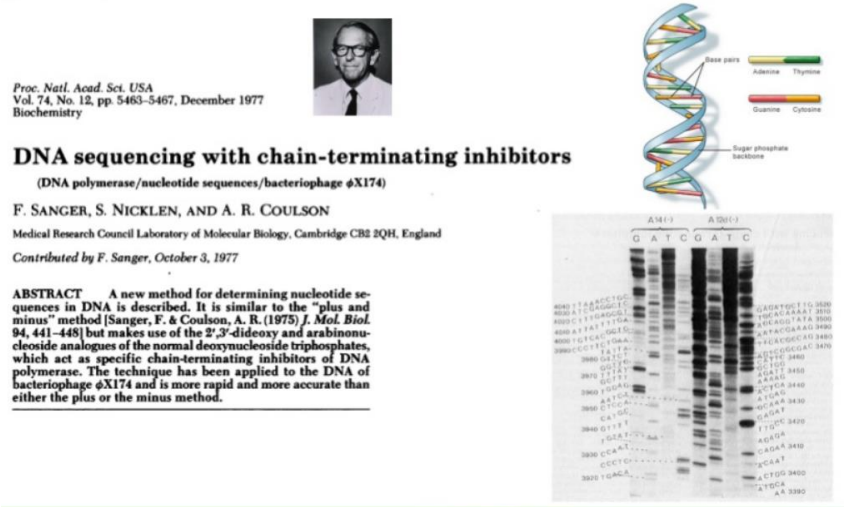
First method for DNA sequencing.

Pros:

Precise: sequenced mapped base to base;

Cons:

1. Time consuming;
2. Short sequences produced (500-600 bp);
3. Not suitable for big sequencing projects; (i.e. 2.4 millions bp/year: 1000 years needed to sequence the human genome!)



Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage φX174)

F. SANGER, S. NICKLEN, AND A. R. COULSON

Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977

ABSTRACT A new method for determining nucleotide sequences in DNA is described. It is similar to the "plus and minus" method [Sanger, F. & Coulson, A. R. (1975) *J. Mol. Biol.* 94, 441–448] but makes use of the 2′,3′-dideoxy and arabinonucleoside analogues of the normal deoxynucleoside triphosphates, which act as specific chain-terminating inhibitors of DNA polymerase. The technique has been applied to the DNA of bacteriophage φX174 and is more rapid and more accurate than either the plus or the minus method.

NGS 2nd generation (2005 ⇒ present)

Ion Torrent

Ion PGM
- 3 types of chips
- 200 or 400 bp reads
- Up to 5.5 million reads / Ion 318 chi
- 4 – 7 h run time

Ion Proton
- Up to 200 bp reads
- Up to 60-80 million reads
- 2 – 4 h run time

Illumina

MiSeq
- Up to 2 x 300 bp reads
- 15 million single reads
- 2.7 days run time

HiSeq2500
- Up to 2 x 125 bp reads
- 2 billion single reads
- 6 days run time

Roche 454

GS Junior
- ~ 400 bp reads
- Up to 100.000 reads / run
- 10h run time

GS FLX Titanium XL+
- ~ 700 bp reads
- Up to 1 million reads/run
- 23h run time

Pros:

High coverage: each part of the sample is read multiple times;
Relatively short run time;

Cons:

Error prone in repeated sequences (accuracy problems);
Short sequence must be assembled in longer pieces; Relatively expensive.

3rd Generation Sequencing (2011 ⇒ present)

Pacific Biosystem

PacBio RSII

- Up to 30 kb reads
- 50.000 reads
- Up to 3h run time

Often used in combination with Illumina sequencing

Oxford Nanopore (2014)

- **MinION** portable sequencer
  - Currently a beta version is being tested in 500 selected labs
  - Accuracy: 60-85%
  - First reports on read lengths: up to 79000 bases!
- Under development
  - **GridION**
  - **PromethION**

Pros:

Longer sequences produced: less errors in *in silico* analyses such as assembly;
Can detect nucleotide modifications (e.g. methylation)

Cons:

High Error rate but expected to improve;

Sangers vs. NGS

	Sanger	NGS
Sequencing samples	Clones, PCR	DNA Libraries
Sample Tracking	Many samples in 96, 384 well plates	Few
Preparation steps	Few, Sequencing reactions clean up	Many, Complex procedures
Data Collection	Samples in plates 96, 384	Samples on slides 1 – 16+
Data	One read/ sample	Thousands and Millions of reads/Samples.

Technologies Comparison

Technology	Read Length	Accuracy	Time per Run	Cost /Mb
Sanger	400-900 bp	99.9%	20 m - h	~ 2400 $
Roche GS	700 bp	99.9%	23 h	~ 10 $
Illumina MiSeq	50-500 bp	99.9%	4-56 h	0.05 - 0.15 $
Nanopore MinION	5-10 kbp	70-90%	1m - 48 h	~ 1 $
PacBio SMRT	10-15 kbp	85-90%	30 m - 4h	~ 0.5 $


High Performance
Bioinformatics
High Performance
Computing

Big data versus the big C

(NATURE, VOL 509, 29 MAY 2014)

The torrents of data flowing out of cancer research and treatment are yielding fresh insight into the disease.

“Informatics researchers are developing algorithms to split data into smaller packets for parallel processing on separate processors,and to compress files without omitting any relevant information. And they are relying on advances in computer science to speed up processing and communications in general”.

Main bioinformatics algorithms demanding high performance computational resources

- ### alignment algorithms

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index (based on the Burrows-Wheeler Transform or BWT) to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 gigabytes of RAM. Bowtie 2 supports gapped, local, and paired-end alignment modes. Multiple processors can be used simultaneously to achieve greater alignment speed.

- ### variant callers

GATK, pronounced "Gee Ay Tee Kay" (*not* "Gat-Kay"), stands for **G**enome**A**nalysis**T**ool**k**it.

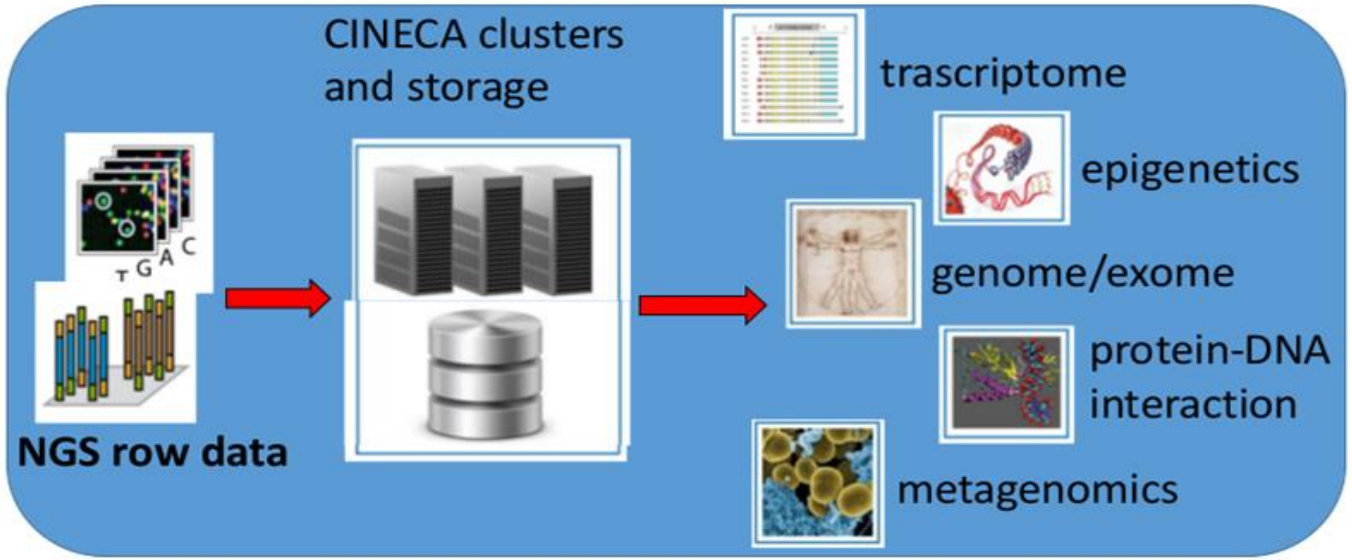
It is a collection of command-line tools for analyzing high-throughput sequencing (HTS) data in formats such as SAM/BAM/CRAM and VCF, with a focus on variant discovery.

- ### machine learnig for personalized medicine

Computing Infrastructure

https://www.cineca.it/it/content/hpc

CINECA academic users can access the High Performance Computing bioinformatics resources both through the web interface and a standard Unix module environment.



Bioinformatics software for Next Generation Sequencing (NGS)

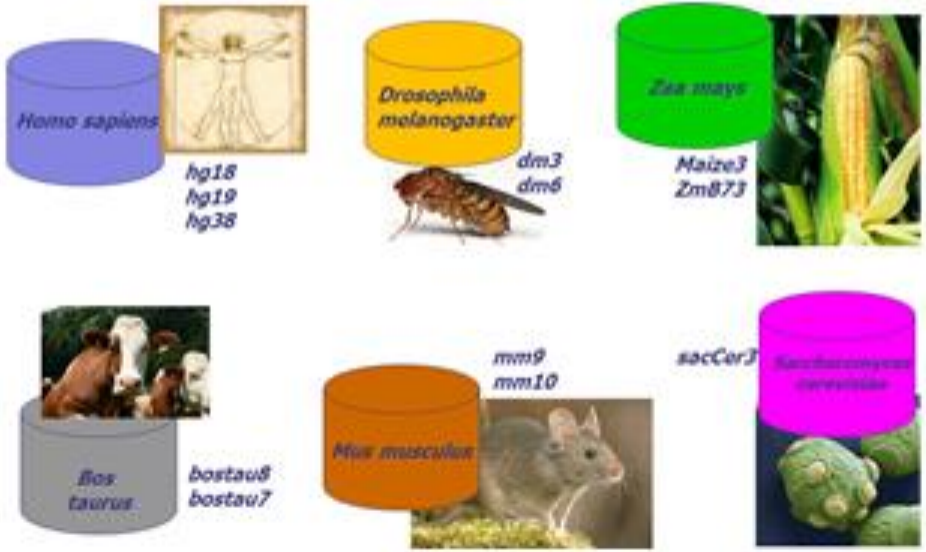
Our bioinformatics platforms contains most of the current and emerging applications for NGS, including alignment, variant callers, assembling, etc.

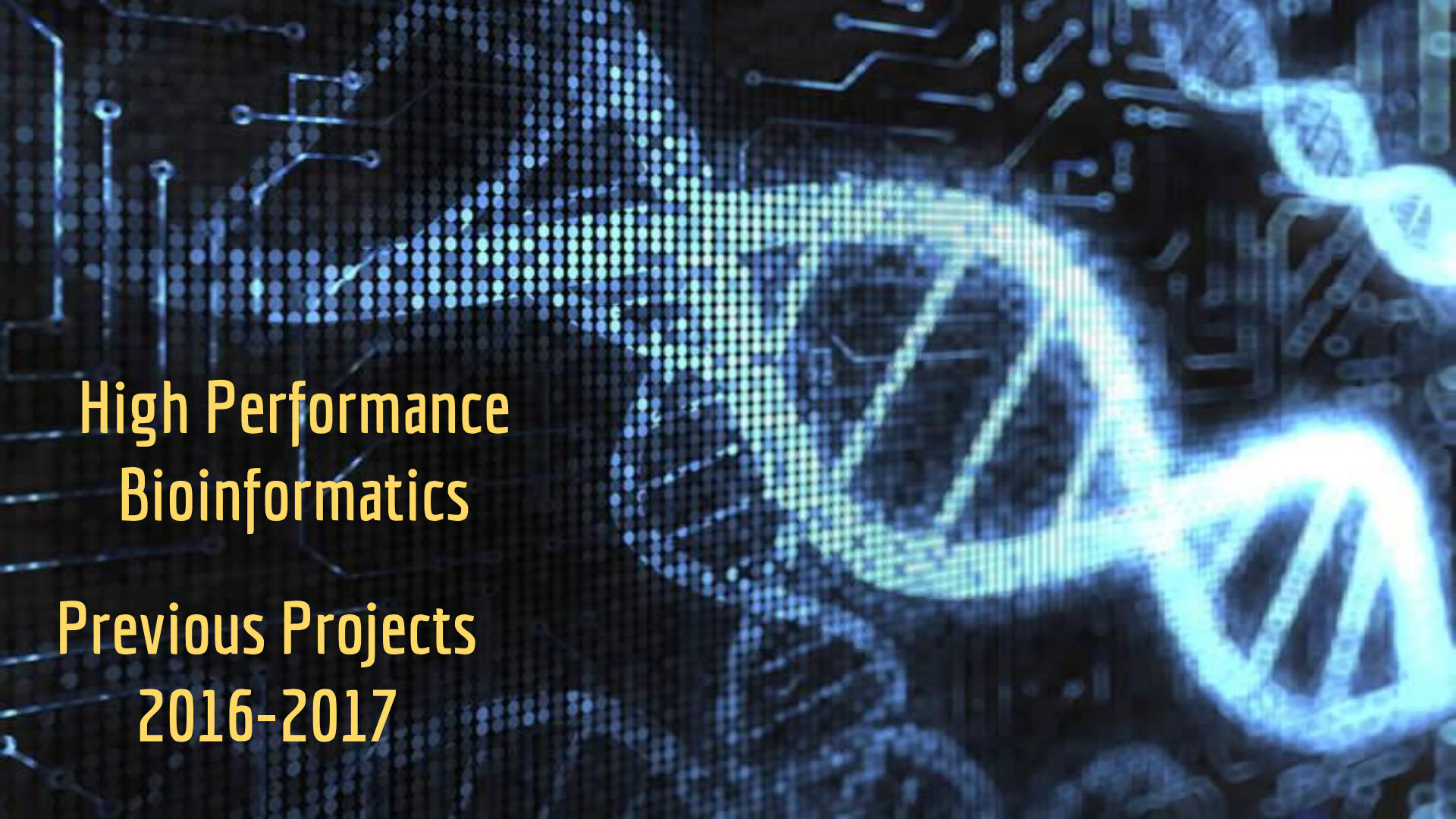
Genomes and Databases

The currently available bio-data primary resources are all the illumina genomes releases provided by NCBI, ENSEMBL and UCSC and available at http://support.illumina.com/sequencing/sequencing_software/igenome.html.

Furthermore, we provide dbSNPs, 1000Genomes databases and *'nr'*, *'nt'* ncbi databases.

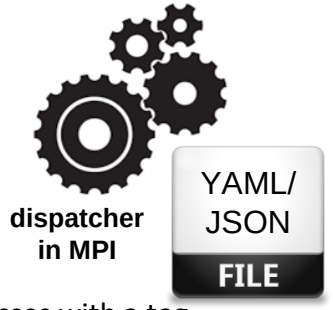
Some model organism available on clusters



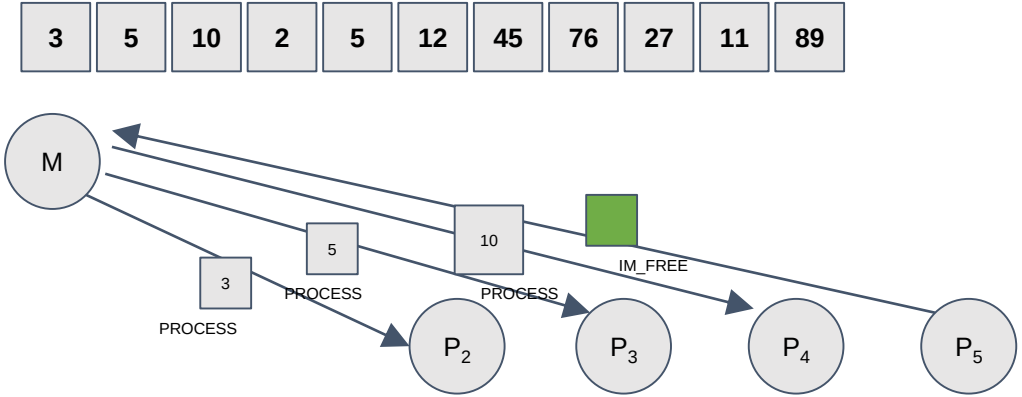

High Performance
Bioinformatics
Previous Projects
2016-2017

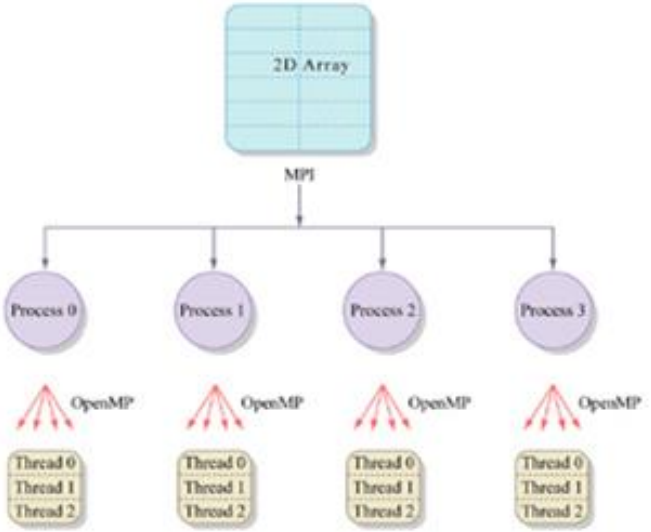
Previous projects (1/4)

NGS-Raptor (dispatcher) - Fabrizio Gattuso



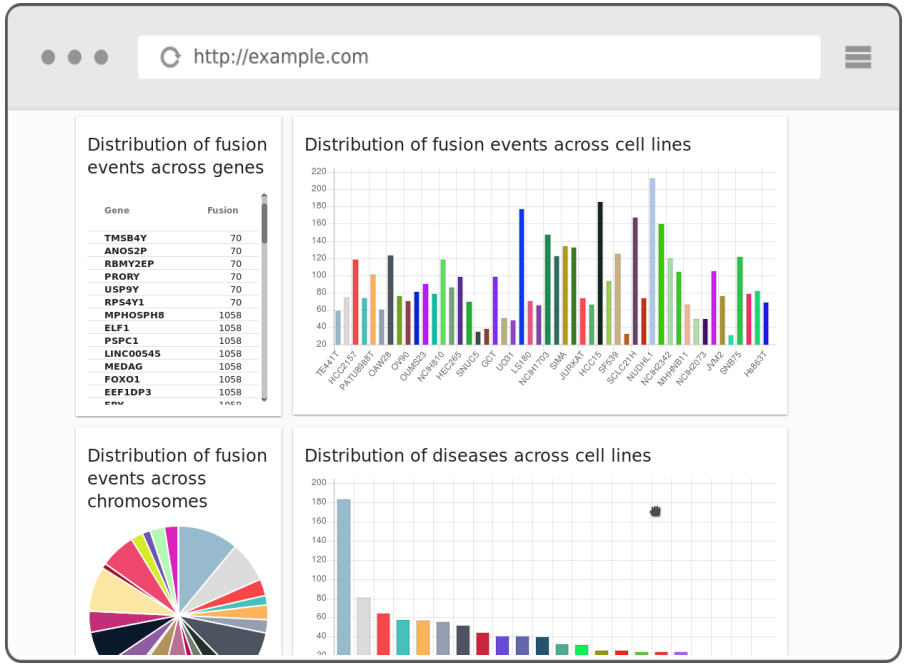
- 1 master process, n-1 slaves
- The master process computes the set of objects to work on and sends a single object to the available slave processes with a tag ‘PROCESS’
- On completion, each slave signals the master process with a special tag ‘IM_FREE’
- When all the input has been processed, the master sends a ‘STOP_WORKING’ tag to all the slaves

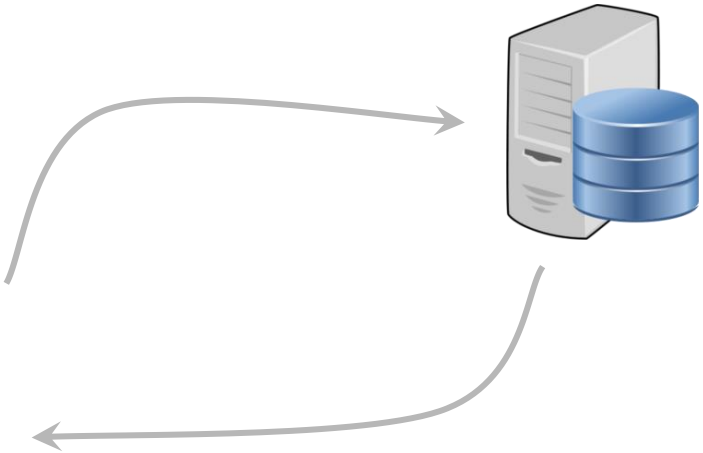


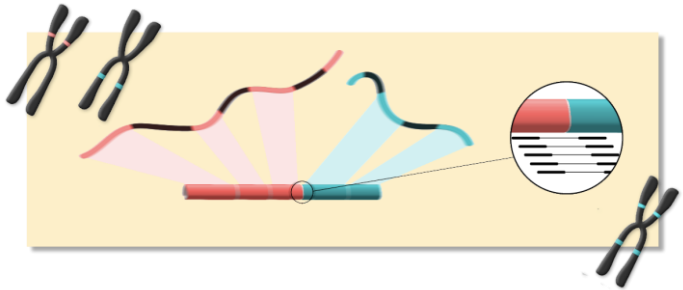


Previous projects (2/4)

Analysis, parsing and visualization of gene fusion events data (Andrea Micco)

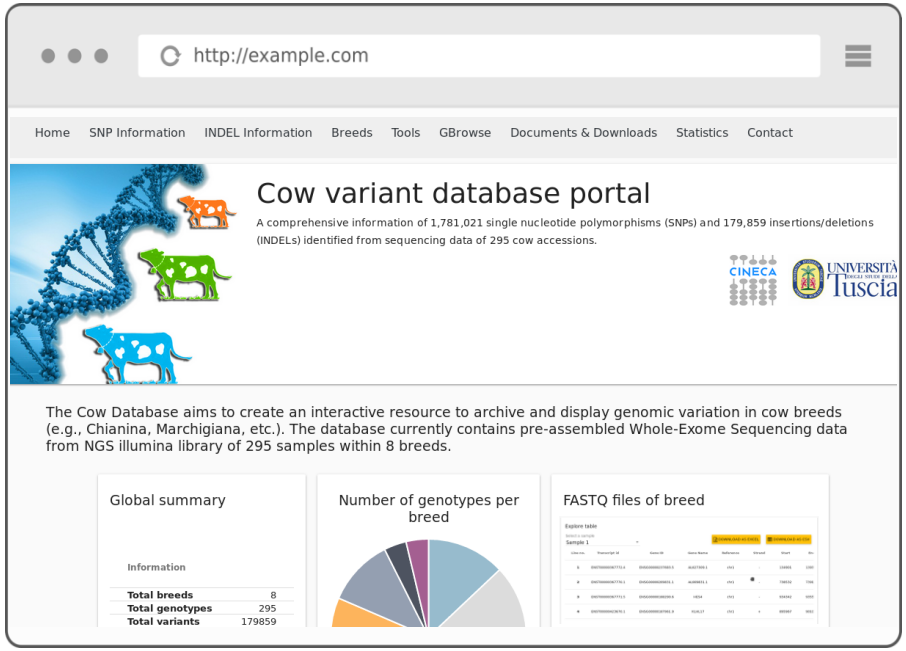


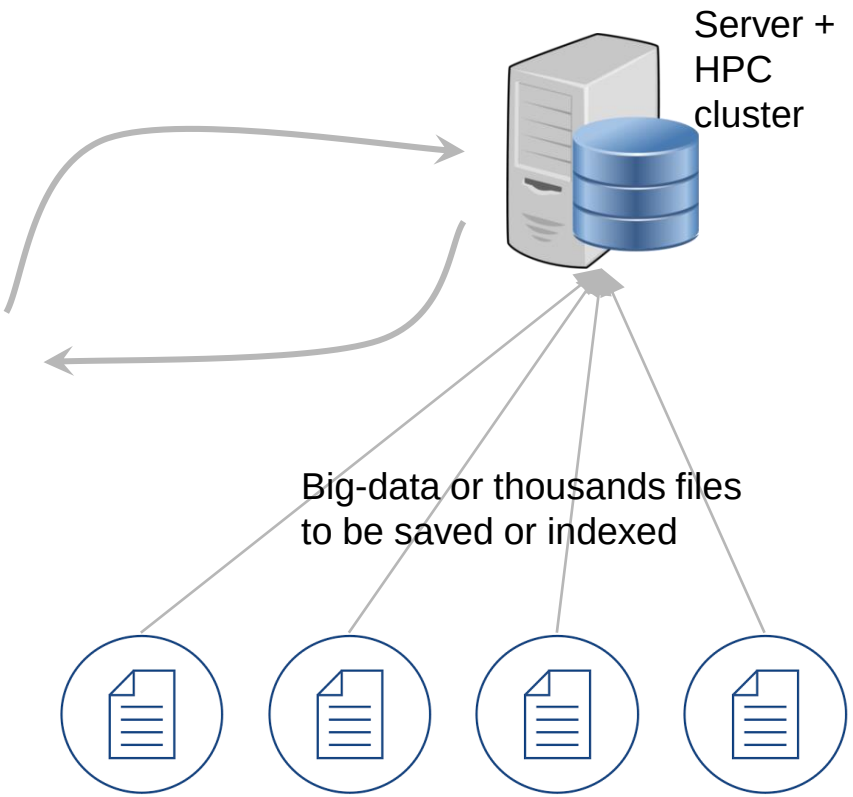




Previous projects (3/4)

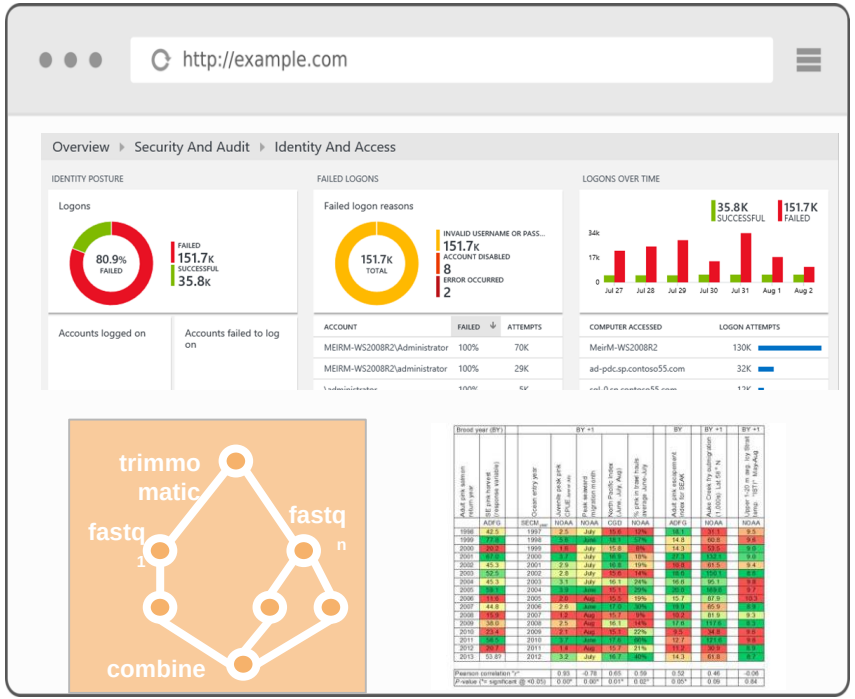
Querying and displaying multiple sources of genomic variant information (Stefano Di Biase)

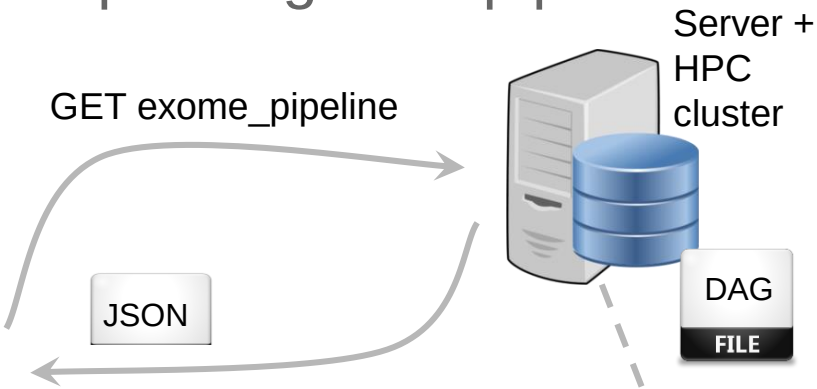




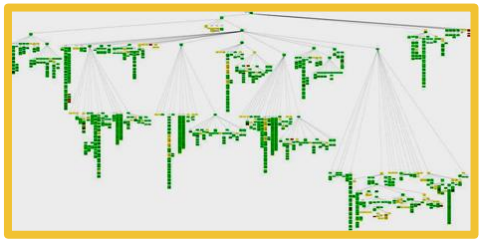
Previous projects (4/4)

Analysis and visualization of complex big-data pipelines (Luca Fochetti)





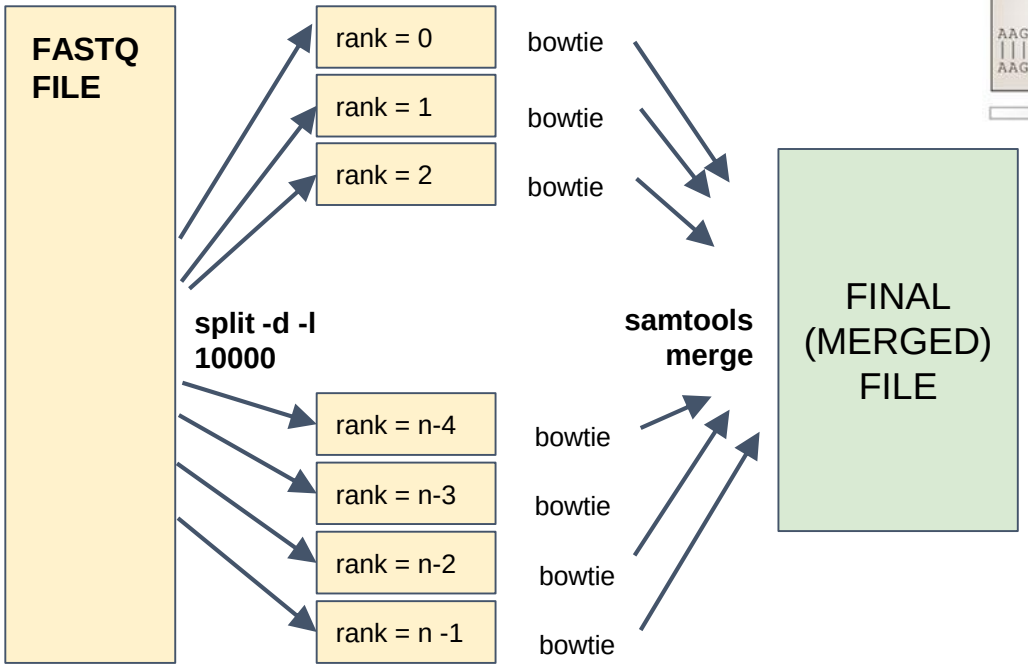
Experiments with **thousands** of files to be analyzed and **thousands** of jobs to be monitored

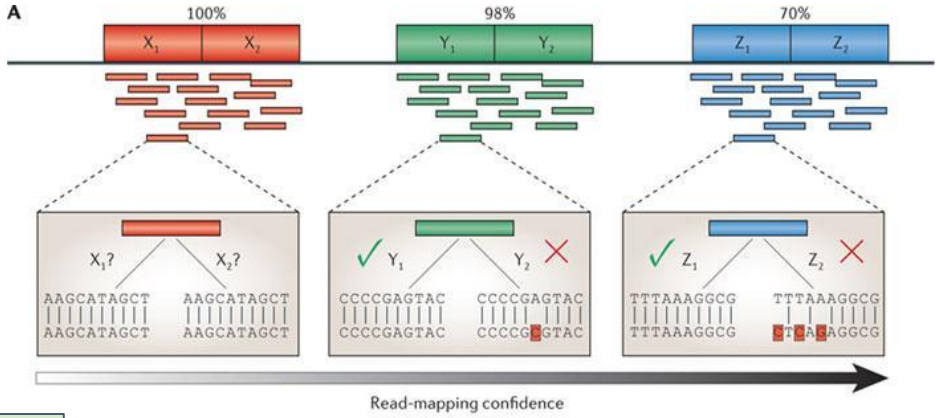


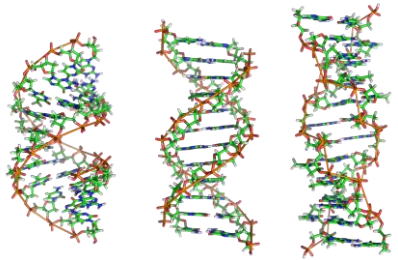

High Performance
Bioinformatics
Next Projects
2017-2018

Project #1

Parallel large-scale alignment







(Also with several aligners...)

Project #2

Parallel big-data VCF/GTF/BAM(/...) comparison

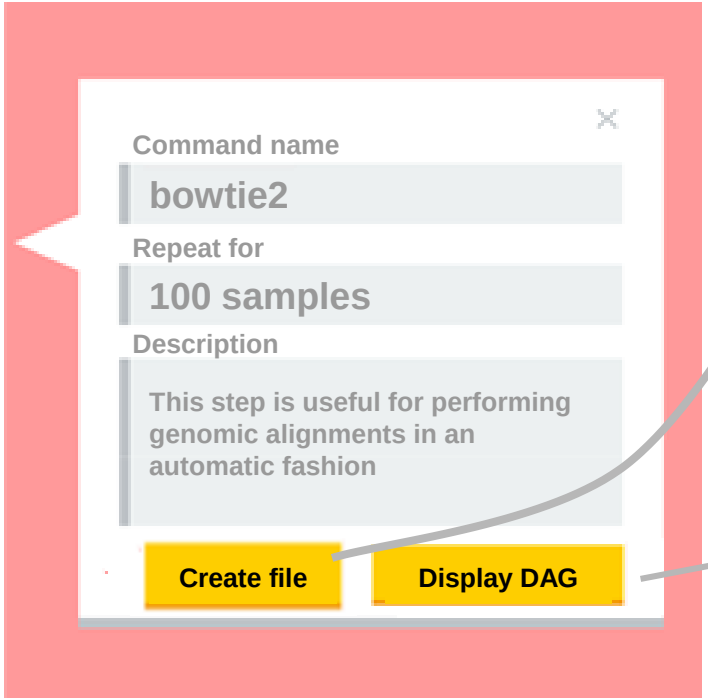
MPI code to handle big-data in parallel

For instance:

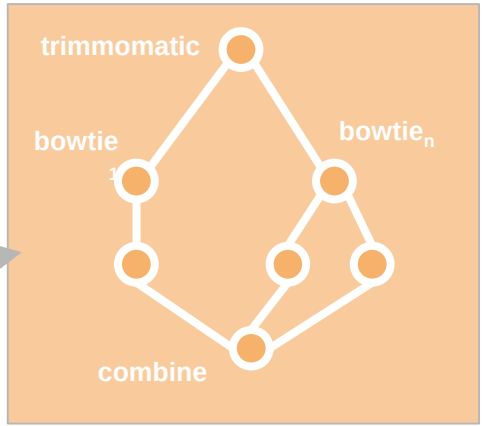
i) calculate the **variants in common** among thousands of samples in a study, or
ii) calculate the **average coverage in bam** files in a collection...
iil) etc.

Project #3

Simplifying big-data analysis setup (through the Web) for dispatching complex bioinformatics pipelines



Either ask the system to create the configuration file for the dispatcher...



...or visually display the DAG associated with the input data.

dispatcher in MPI