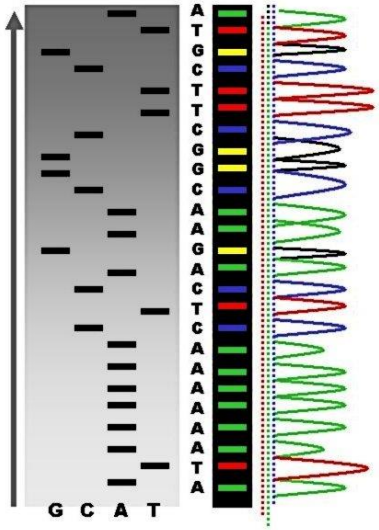
CINECA
Consorzio Interuniversitario

Ruolo del Supercalcolo nell'analisi dei dati genetici

Tiziana Castrignanò
Milano, April **21th , 2015**

www. cineca.it



Sanger sequencing is a method of **DNA sequencing** based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication.

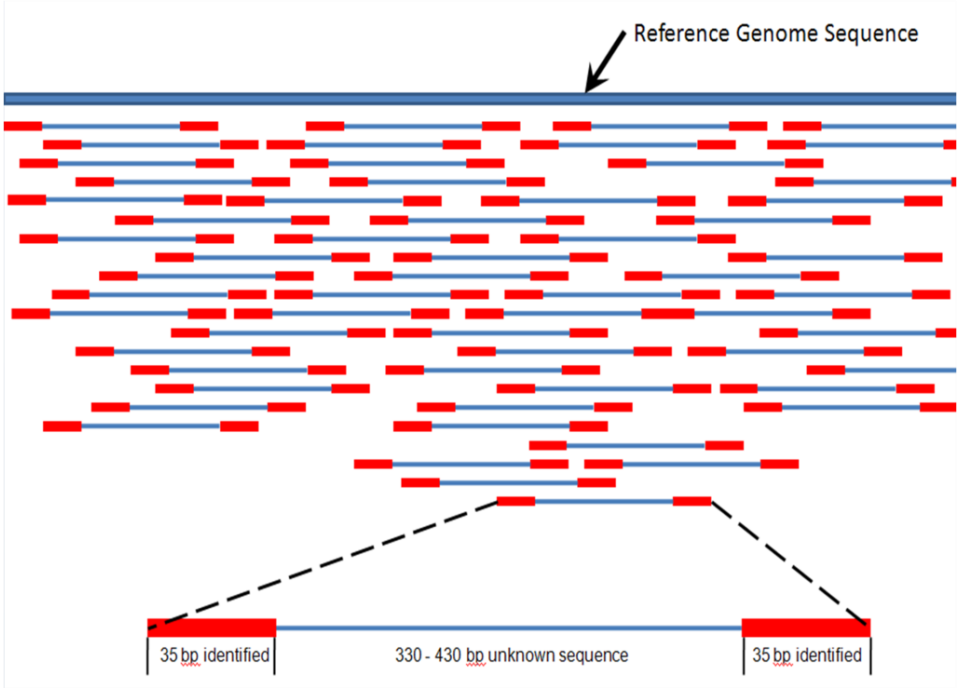
Developed in 1977, it was the most widely used sequencing method for approximately 25 years thanks to its relative ease and reliability.

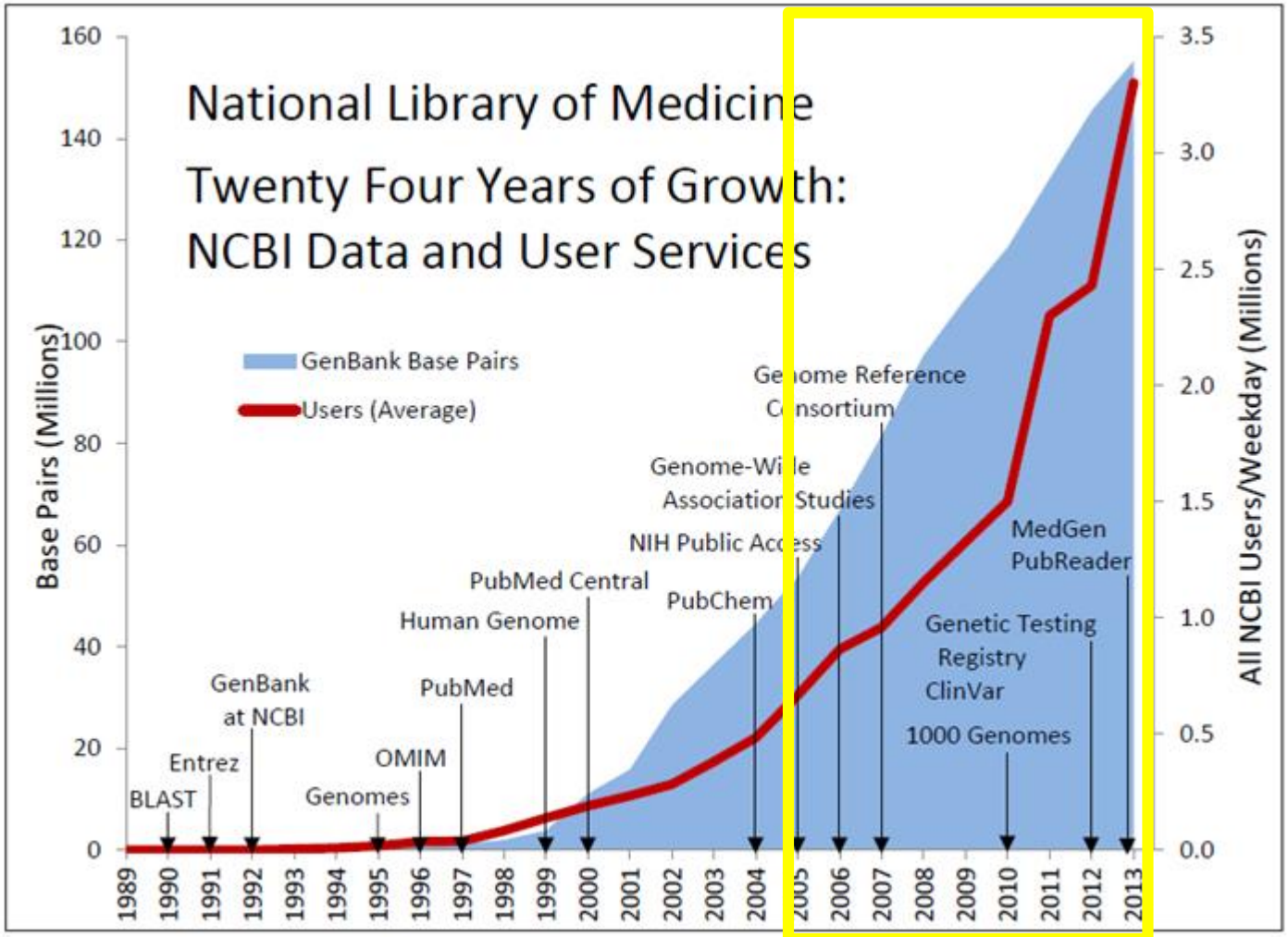
More recently, Sanger sequencing has been supplanted by "Next Generation Sequencing" (NGS) methods, especially for large-scale, automated genome analyses.

However, the Sanger method remains in wide use, for smaller-scale projects, validation of NGS results and for obtaining especially long contiguous DNA sequence reads (>500 nucleotides).

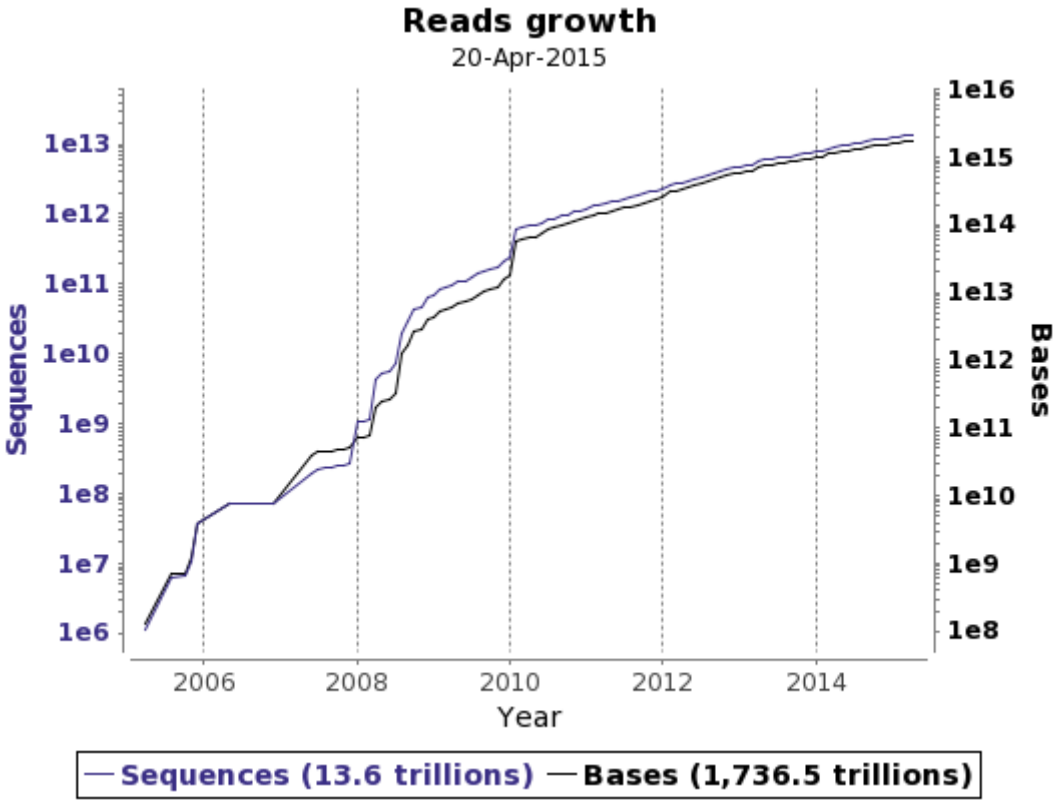
Demand has never been greater for revolutionary technologies that deliver fast, inexpensive and accurate genome information. This challenge has catalysed the development of next-generation sequencing (NGS) technologies.

The high demand for low-cost sequencing has driven the development of high-throughput sequencing
(or next-generation sequencing) technologies that parallelize the sequencing process, producing thousands or millions of sequences concurrently.

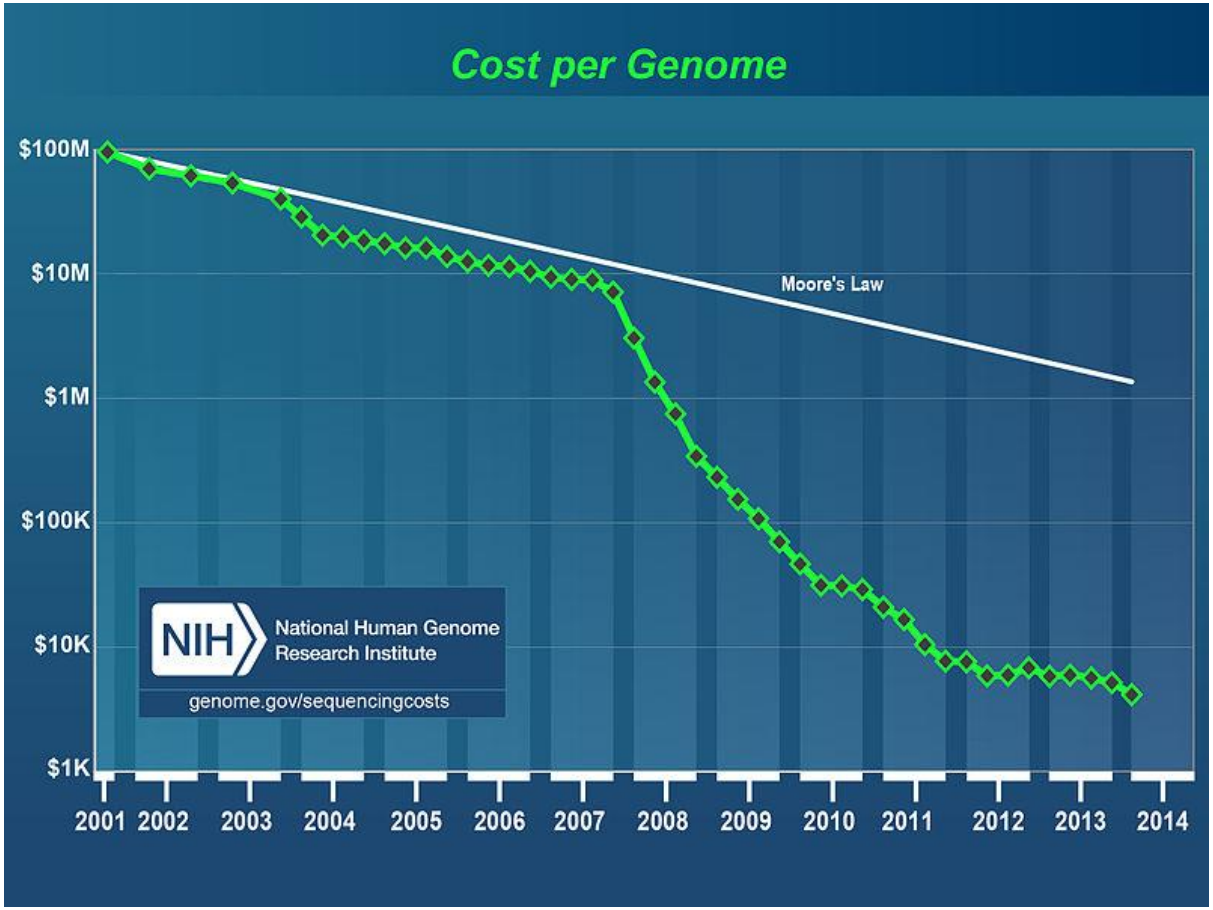



National Library of Medicine
Twenty Four Years of Growth:
NCBI Data and User Services
GenBank Base Pairs
Users (Average)
Base Pairs (Millions)
All NCBI Users/Weekday (Millions)
BLAST
Entrez
GenBank at NCBI
Genomes
OMIM
PubMed
Human Genome
PubMed Central
PubChem
NIH Public Access
Genome-Wide Association Studies
Genome Reference Consortium
MedGen
PubReader
Genetic Testing Registry
ClinVar
1000 Genomes
0
20
40
60
80
100
120
140
160
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013

http://www.ebi.ac.uk/ena/about/statistics



Statistics regarding data growth in European Nucleotide Archive



Next-generation DNA sequencing (NGS) has incredibly accelerated the comprehensive analysis of genomes, transcriptomes and interactomes.

NGS applications are resource-hungry

HiSeq 2000 Output:

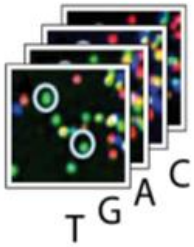
- 300 Gb (fastq)
- 375 Million/lane PE reads

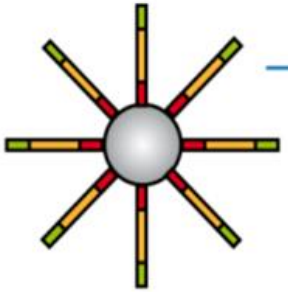
Increased size due to replicates and PE

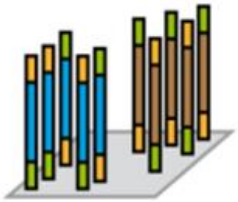
This huge quantity of data requires computational clusters

- Support of Computational Centers (e.g. CINECA)

Raw and shortreads data





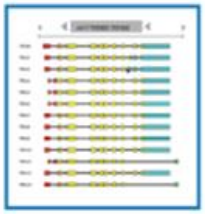


CINECA clusters

Fermi - 2 PFlops

Pico/Galileo

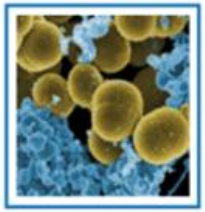
12 PB Storage



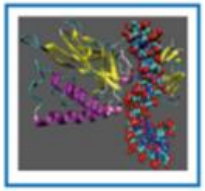
transcriptome

epigenetics

genome /exome



metagenomics



protein-DNA
interaction

What CINECA offers for bioinformatics?

1. **Computing resources**

   Bioinformatics software available through command line

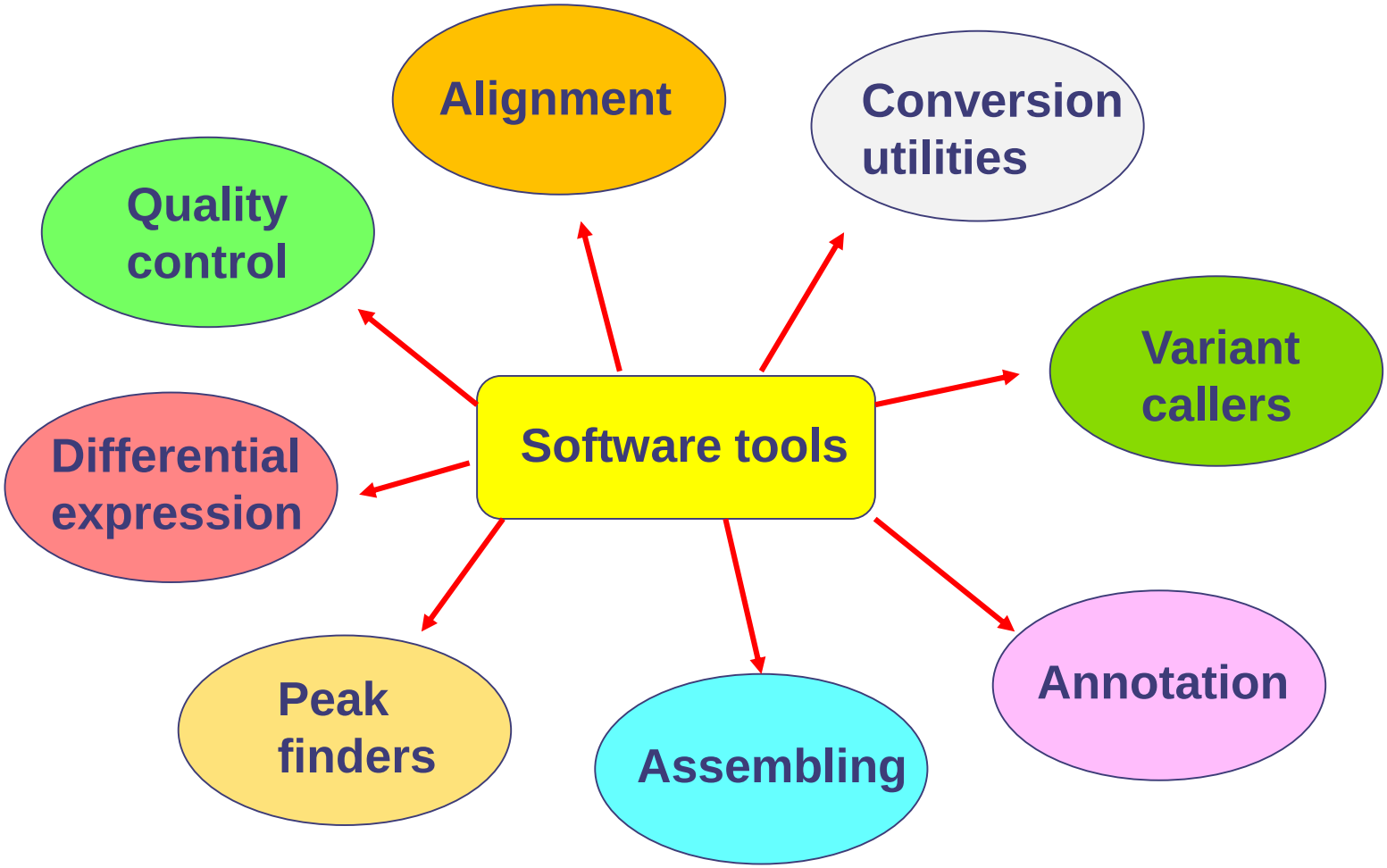
2. **Advanced services**

   Automated web workflows for Next Generation Sequencing

3. **Bioinformatics Expertise**

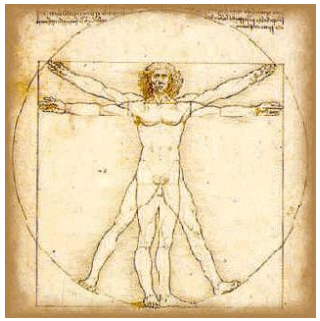
   To customize solutions or implement new systems and tools

Bioinformatics software and tools



DATA: Available genomes

Homo sapiens



Drosophila melanogaster

Zea mays

Rattus norvegicus

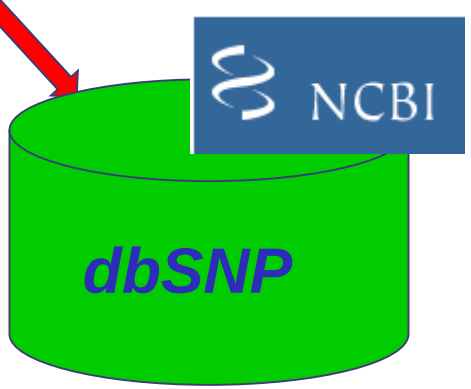
Mus musculus

Saccharomyces cerevisiae

Annotation databases

ANNOVAR is an efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes

ExAC Data Set:
exome sequencing data from a wide variety of large-scale sequencing projects



a free public archive for short genetic variation within and across different species

What CINECA offers for bioinformatics?

1. Computing resources

   Bioinformatics software available through command line

2. **Advanced services**

   Automated web workflows for Next Generation Sequencing

3. Bioinformatics Expertise

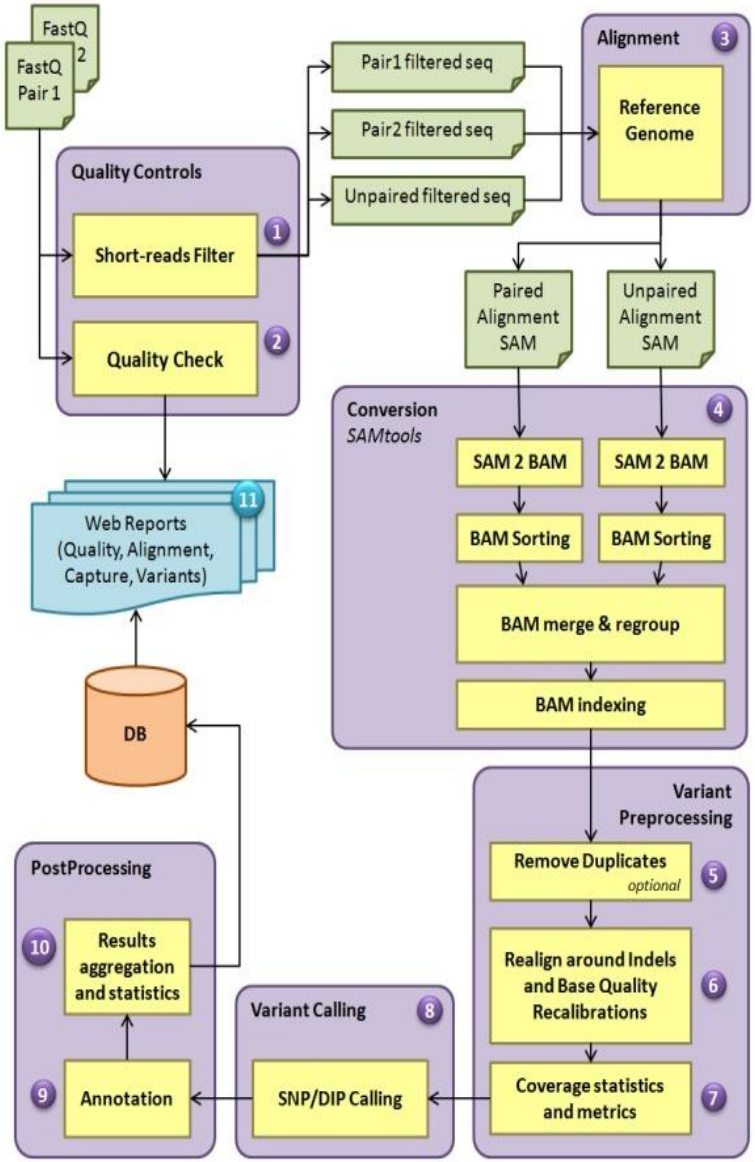
   To customize solutions or implement new systems and tools

Bioinformatics NGS Pipelines

Automated workflows (pipelines) for Next Generation Sequencing are available through a web interface and are able to perform analyses for several NGS applications:

- Deep targeted exome sequencing;
- RNA sequencing (trascriptome analysis);
- Whole exome sequencing;
- Identification of DNA protein interactions by ChIP-seq;

Ultra Deep Exome Sequencing Pipeline



Online Deep Exome Sequencing Software Analysis (ODESSA)

Handles genes targeted at high coverage

Specifically focused for clinical diagnostics

Identifies (SNPs) and (DIPs) classified by different scores (e.g. depth, SIFT, MAV, MEQ).

Results are supported with genomic information, functional annotations, cross-linking databases and quality and relevance scores, graphics, tables and browsing, filtering and download.

Optimized for MiSeq Illumina platform

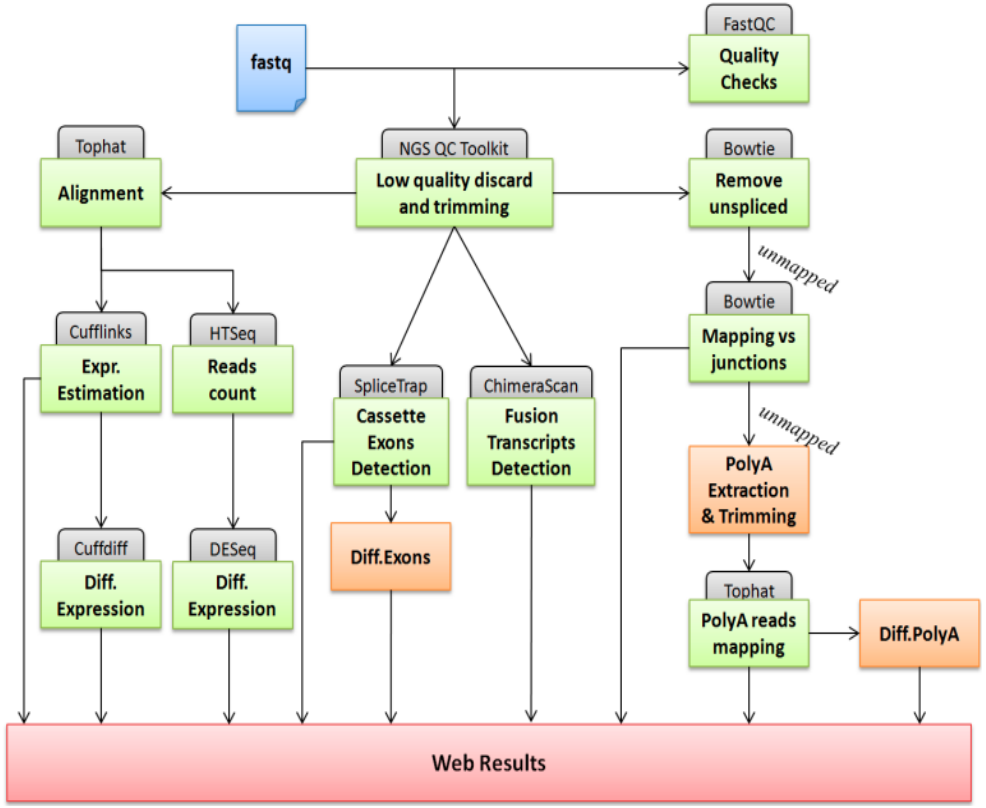
Variant results

position	allele variation	state	Depth	Mutation	Type	Func	gene info	location	dbSNP
chr16:23360199-23360199	T → C	het	66	SNV	synonymous SNV	-	SCNN1B	exonic	rs238547
chr16:27373915-27373915	G → T	het	147	SNV	synonymous SNV	-	IL4R	exonic	rs2234898
chr16:85706047-85706047	A → C	het	62	SNV	synonymous SNV	-	GSE1	exonic	rs9940601
chr16:15818141-15818141	A → C	het	115	SNV	synonymous SNV	-	MYH11	exonic	rs2075511
chr16:89836323-89836323	C → T	het	140	SNV	nonsynonymous SNV	-	FANCA	exonic	rs7195066
chr16:20554248-20554248	G → A	het	166	SNV	synonymous SNV	-	ACSM2B	exonic	rs140717461
chr16:20489919-20489919	G → A	het	47	SNV	nonsynonymous SNV	-	ACSM2A	exonic	rs147314845
chr16:15811023-15811023	C → T	het	120	SNV	synonymous SNV	-	MYH11	exonic	rs1050163

RNA Sequencing Pipeline

The RNA-Seq Analysis Pipeline (RAP)

Performs a complete and customizable RNA-seq pipeline, allowing users to examine NGS data under many points of view:



- Gene and transcript expression
- Differential expression
- Splicing junctions
- Cassette exons
- Poly(A) sites
- Fusion transcripts
- RNA editing

RNA Sequencing Pipeline

Gene and transcript expression summary

Click on the colored-box numbers to open the expression overview

File	Label		Expressed FPKM>0	Expressed FPKM>10	Expressed FPKM>20	Expressed FPKM>100	#HIDATA Loci
1	Embryonic1	transcripts	22852	7374	4265	640	
		genes	16963	7180	4355	680	0
2	Embryonic2	transcripts	23096	7436			
		genes	17160	7196			
3	Embryonic3	transcripts	23104	7332			
		genes	17160	7126			
4	Embryonic4	transcripts	23182	7408			
		genes	17223	7203			
5	Adult1	transcripts	23989	7198			
		genes	17866	6987			
6	Adult2	transcripts	23874	7262			
		genes	17782	7045			

Click on a column title to order this table

UID	Gene	Transcript	Genomic Position	Strand	TLen	#Exons	FPKM↓	Coverage
1268	MIR4461	NR_039666	chr5:134291628-134291701	+	74	1	237307.93	9918.79
637	MIR548AC	NR_039621	chr17:28547066-28547096	-	31	1	64029.67	2676.26
987	MIR3687	NR_037458	chr21:1678868-1678928	-	61	1	42134.91	1761.12
1206	MIR1267	NR_031671	chr4:177196342-177331125	+	57	3	39547.53	1652.97
672	MIR548O2	NR_039605	chr17:60821546-60847231	-	52	3	34715.01	1450.99
941	MIR663A	NR_030386	chr20:26136822-26136914	-	93	1	16631.98	695.17
1282	MIR548O2	NR_030385	chr5:159002885-159095000	+	81	4	14808.62	618.96
1214	MIR4454	NR_039659	chr5:7322416-7322467	-	52	1	12569.28	525.36
1603	MIR548D1	NR_030382	chr9:123415763-123798763	-	59	4	11998.16	501.49
1207	MIR548AB	NR_039611	chr4:183713766-183720064	-	56	2	11737.12	490.58

What CINECA offers for bioinformatics?

1. Computing resources

   Bioinformatics software available through command line

2. Advanced services

   Automated web workflows for Next Generation Sequencing

3. **Bioinformatics Expertise**

   To customize solutions or implement new systems and tools

Bioinformatics specialistic support

Bioinformatics specialistic support to develop and optimize

- configuration parameters
- command-line programs
- complex bash scripts

on thousands of computing cores

Data integration

A centralized system for data storage,
metadata assignment,
data aggregation,
information sharing
and data analytics

with an eye on security and availability

Metadata - sequences

Example of metadata from sequencing sample

- age
- phenotype
- state of health
- technical / biological replication
- disease information (OMIM, MeSH)

Example of metadata from sequencing platform

- name of platform
- read length
- library preparation protocol (stranded, unstranded)
- sequencing target (manifest?)

Metadata - sequences

Id	Label / File	Sample	Run	PE	read length	File Size	Download
☑	**adipose_female_73y_1** ERR030880_1.fastq ERR030880_2.fastq	adipose f73c		yes	50bp 50bp	17.08 GiB 17.08 GiB	
☑	**adipose_female_73y_2** ERR030888.fastq	**adipose f7**		no	75bp	20.4 GiB	
☑	**adrenal_male_60y_1** ERR030881_1.fastq ERR030881_2.fastq	**adrenal m6**		yes	50bp 50bp	16.46 GiB 16.46 GiB	
☑	**adrenal_male_60y_2** ERR030889.fastq	**adrenal m6**		no	75bp	20.38 GiB	
☑	**brain_female_77y_1** ERR030882_1.fastq ERR030882_2.fastq	**brain f77c**		yes	50bp 50bp	16.25 GiB 16.25 GiB	

adipose f73c
Scientific Name: Homo Sapiens
Common Name: Human
TaxonID: 9606
Tissue: Adipose
Cell: -
Phenotype: female 73y caucasian
Strain: -

Data Aggregation

Data aggregation is the real strenght of a repository

More the data, more the derived information, more the information significance

Data can generate more data

- statistics, frequencies
- identification of recurrent patterns

I found a SNP, it is common in related patients?

I found a SNP, it is uniquely related to a disease?

Quality and Security

The **Information Security Management System** of Cineca is compliant with the international standard **ISO 27001** (since 2005)

The **Quality Management System** of Cineca is compliant with the international standard **ISO 9001** (since November 2001) extended to **bioinformatics services** (since September 2013)

Links

For further information

- Official web site http://www.hpc.cineca.it
- Bio & Genomics http://www.hpc.cineca.it/content/hpc-bioinformatics
- Bioinformatics user support hpc-bioinformatics@cineca.it