CINECA
consorzio interuniversitario

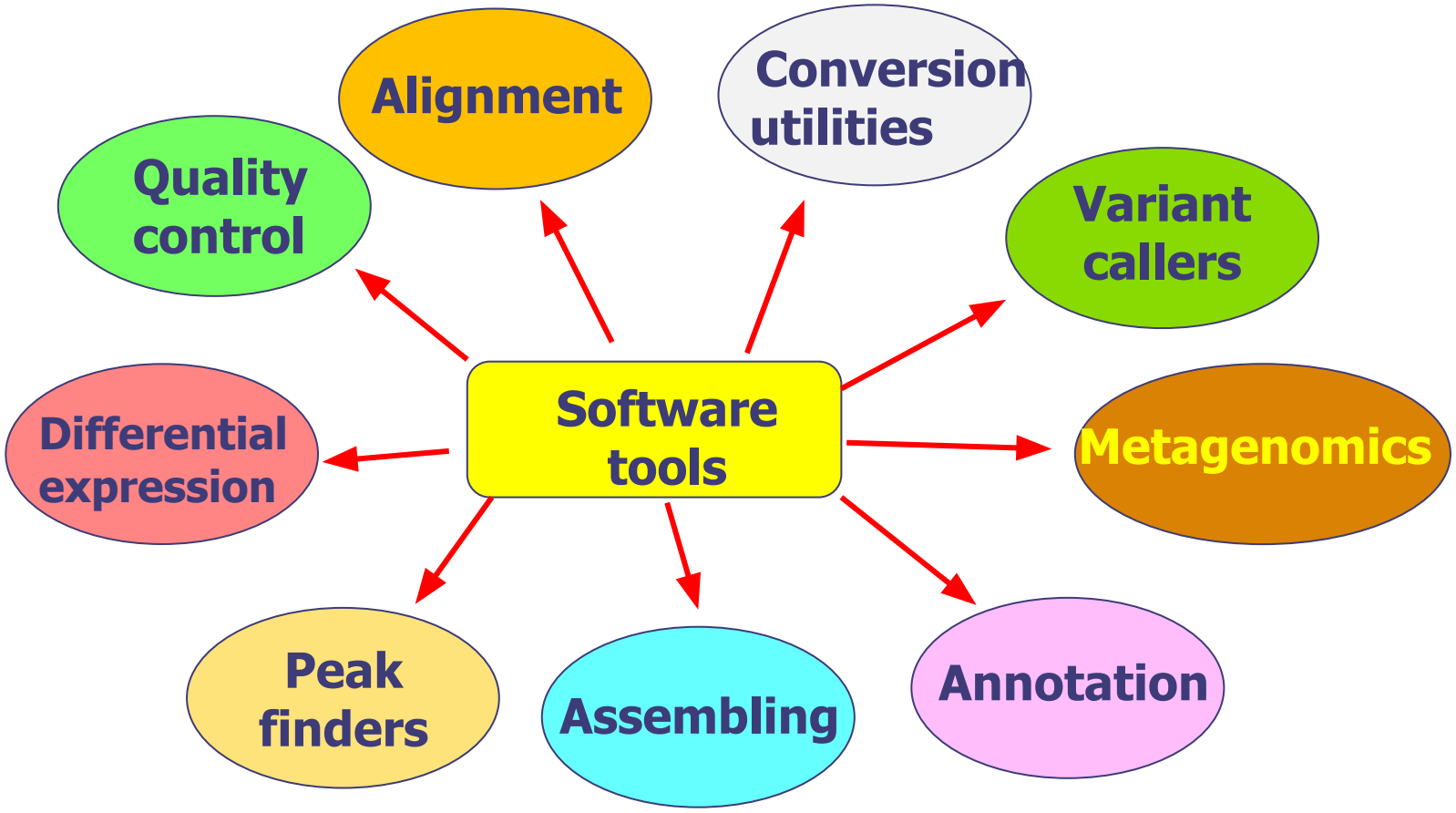
High-Performance Bioinformatics Environment @ CINECA

HPC Department

Brescia, 21st March 2017

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.

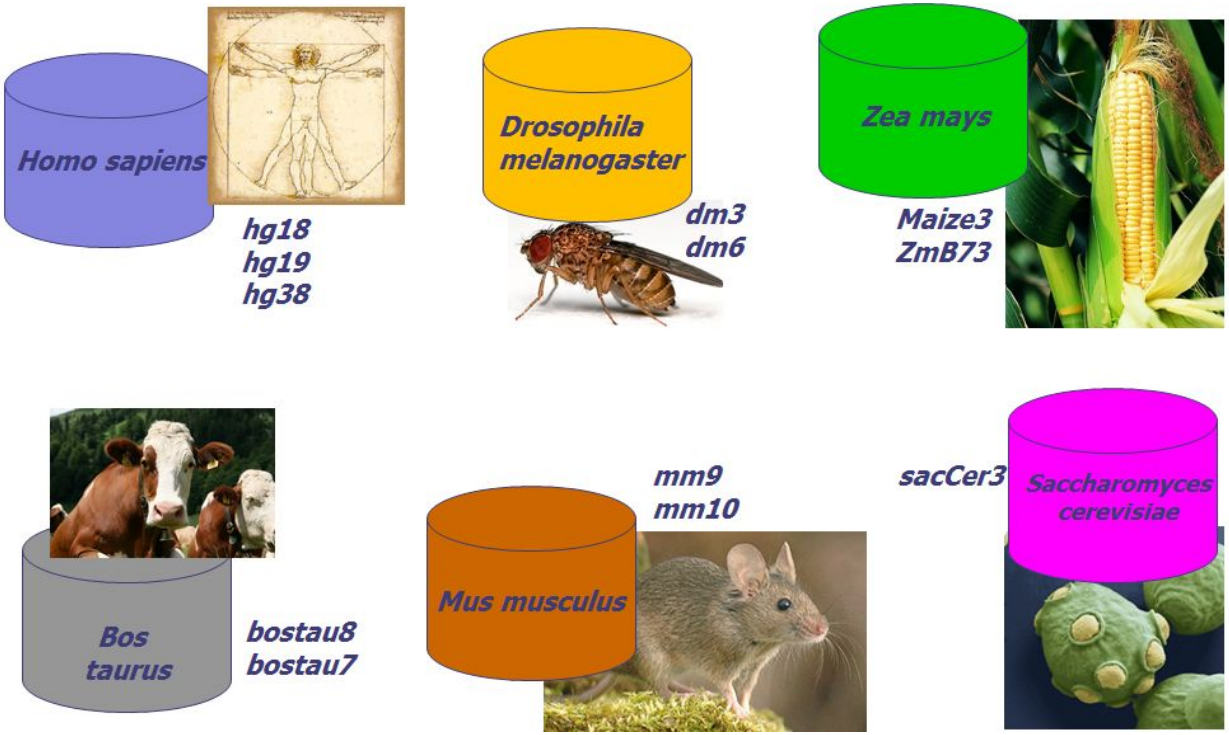


Genomes and Databases

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



Big Data – Bioinformatics
Cineca Calls

CINECA is offering tens of Millions of HPC core hours and hundreds of TBs for the following national and international bioinformatics research projects:

- **ELIXIR** (Italy-Europe)
- **LISA** (Lombardia)
- **PRACE** (Europe)
- **ISCRA** (Italy)

Example of Elixir development project

IVaCS:The Integrated Variant Calling System

https://bioinformatics.cineca.it/ivacs

IVaCS is an automated, highly accurate workflow for genotyping, variant calling and annotation for Next Generation Sequencing data.

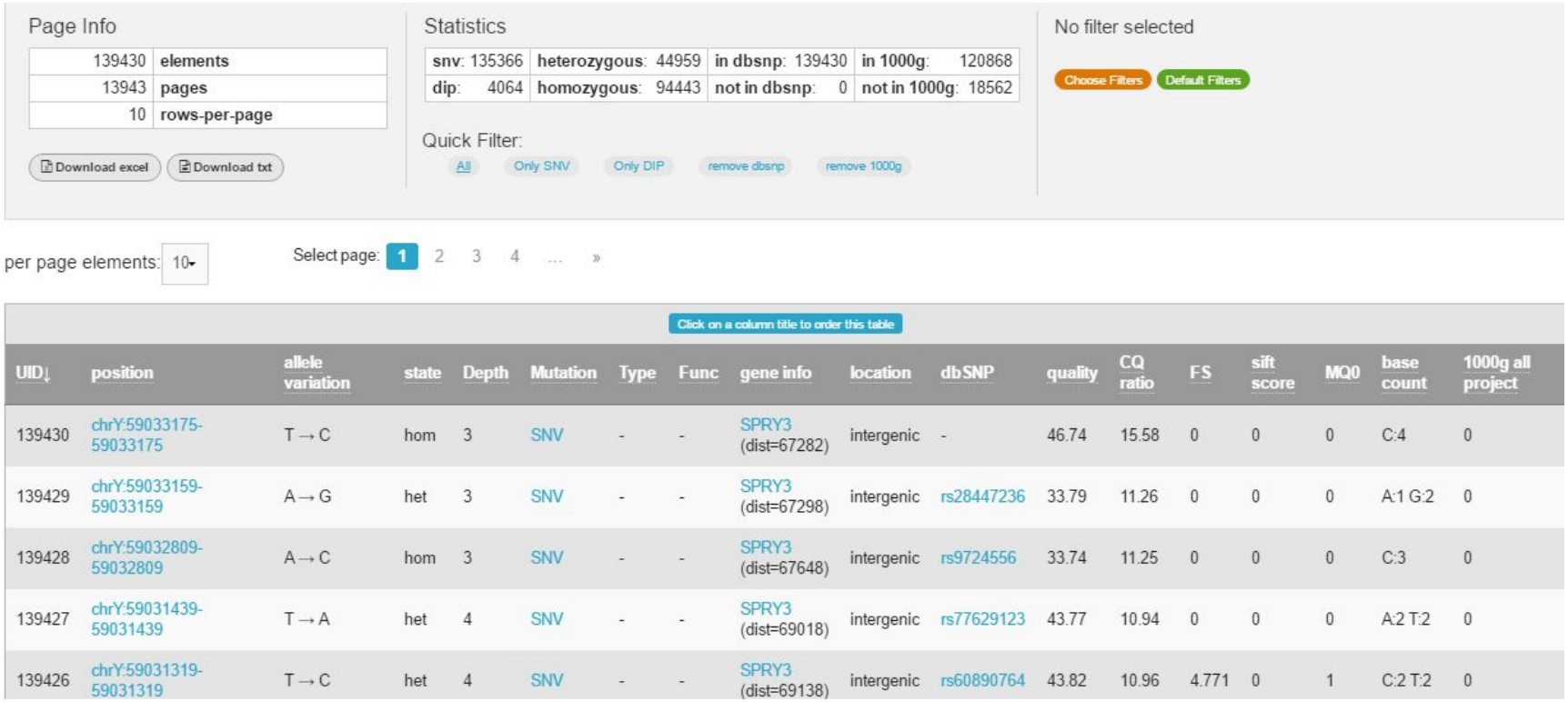
By integrating state of the art tools and algorithms in a swift and effective manner and providing a extensive collection of reference genomes, IVaCS offer the possibilty to perform elaborate and comprehensive genotyping of NGS data even to users with little/no bioinformatic background.

IVaCS is currently hosted at CINECA and benefits all the advantages of HPC computing.

In its current form IVaCS offers pre-configured workflows for the analyses of WGS, WES and targeted resequencing data.

Example of Elixir development project: IVaCS:The Integrated Variant Calling System

https://bioinformatics.cineca.it/ivacs



Page Info

139430	elements
13943	pages
10	rows-per-page

Download excel | Download txt

Statistics

snv: 135366	heterozygous: 44959	in dbsnp: 139430	in 1000g: 120868
dip: 4064	homozygous: 94443	not in dbsnp: 0	not in 1000g: 18562

Quick Filter: All | Only SNV | Only DIP | remove dbsnp | remove 1000g

No filter selected

Choose Filters | Default Filters

per page elements: 10

Select page: 1 2 3 4 ... »

Click on a column title to order this table

UID↓	position	allele variation	state	Depth	Mutation	Type	Func	gene info	location	dbSNP	quality	CQ ratio	FS	sift score	MQ0	base count	1000g all project
139430	chrY:59033175-59033175	T→C	hom	3	SNV	-	-	SPRY3 (dist=67282)	intergenic	-	46.74	15.58	0	0	0	C:4	0
139429	chrY:59033159-59033159	A→G	het	3	SNV	-	-	SPRY3 (dist=67298)	intergenic	rs28447236	33.79	11.26	0	0	0	A:1 G:2	0
139428	chrY:59032809-59032809	A→C	hom	3	SNV	-	-	SPRY3 (dist=67648)	intergenic	rs9724556	33.74	11.25	0	0	0	C:3	0
139427	chrY:59031439-59031439	T→A	het	4	SNV	-	-	SPRY3 (dist=69018)	intergenic	rs77629123	43.77	10.94	0	0	0	A:2 T:2	0
139426	chrY:59031319-59031319	T→C	het	4	SNV	-	-	SPRY3 (dist=69138)	intergenic	rs60890764	43.82	10.96	4.771	0	1	C:2 T:2	0

Matteo Chiara, Giovanni Chillemi, Mattia D'Antonio, Paolo D'Onorio De Meo, Tiziano Flati, Silvia Gioiosa, Ernesto Picardi, Graziano Pesole, Tiziana Castrignanò

https://peerj.com/preprints/2213/

Bioinformatics web software for Next Generation Sequencing (NGS)

Cineca provides the following NGS pipelines implemented with a user-friendly web interface:

1. Deep targeted exome sequencing;
2. RNA sequencing (transcriptome analysis);
3. Whole exome sequencing;
4. Identification of DNA protein interactions by ChiP-seq;
5. RNA editing prediction by RNA-seq

Bioinformatics web software for Next Generation Sequencing (NGS)

Most of the web NGS pipelines have been published on:

Deep targeted exome sequencing;

M. D'Antonio, P. D'Onorio De Meo, T. Castrignanò, G. Erbacci, M. Pallocca and G.Pesole. "ODESSA: a High Performance Analysis Pipeline for Ultra Deep Targeted Exome Sequencing Data". Published in "International Conference on High Performance Computing & Simulation (HPCS), 2014" *IEEE. Special Issue HPCS* 2014. doi: 10.1109/HPCSim.2014.6903743

RNA sequencing (transcriptome analysis);

D'Antonio M., D'Onorio De Meo P., Pallocca M., Picardi E., D'Erchia AM., Calogero R., Castrignanò T., Pesole G. RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application. *BMC Genomics.* 2015;16:S3. doi: 10.1186/1471-2164-16-S6-S3. Epub 2015 Jun 1.

Whole exome sequencing;

D'Antonio M., D'Onorio De Meo P., Daniele Paoletti, Elmi B., Pallocca M., Sanna N., Picardi E., Pesole G. and Castrignanò T. "WEP: an high-performance analysis pipeline for whole-exome data". January 2013 *BMC Bioinformatics*.

RNA editing prediction by RNA-seq;

Picardi E., D'Antonio M., Carrabino D., Castrignanò T. and Pesole G. "ExpEdit: a web server to explore human RNA editing in RNA-Seq experiments". *Bioinformatics* 2011 May 1;27(9):1311-2.

Example of support to ELIXIR project

Elixir project 8. : Gene fusions and retinoid sensitivity
(Istituto Farmacologico Negri, P.I. Maddalena Fratelli)

We have provided support to develop a parallel high performance tool to investigate fusion genes with the top-5 fusion algorithms published on

Angie Duy Vo, Fujun Qin & Hui Li. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. http://www.nature.com/articles/srep21597 *Scientific Reports* **6**, Article number: 21597 (2016), doi:10.1038/srep21597

This HPC application will be used to perform large scale analysis on:

- a compilation of massively parallel sequencing data from 935 human cancer cell lines from «The Cancer Cell Line Encyclopedia» (CCLE)
- a compilation of RNA-seq samples from the «The Cancer Genome Atlas" repository (TCGA)

Example of support to LISA project

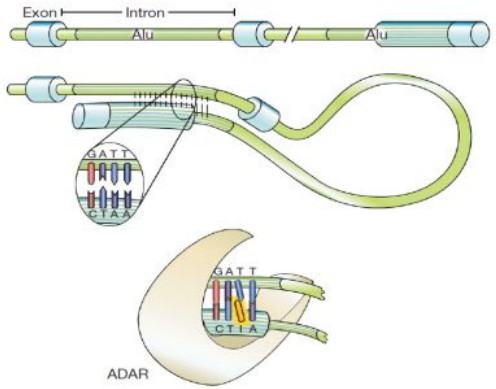
Support to the project:
"Detecting of genetic variation of peach genome data".
(Facoltà di Agraria, Universita' Statale di Milano. P.I. Filippo Geuna)

- We have provided support to develop a parallel high performance tool to analysed genetic variations in 150 peach cultivars at low coverage mapped against the reference genome at high coverage.

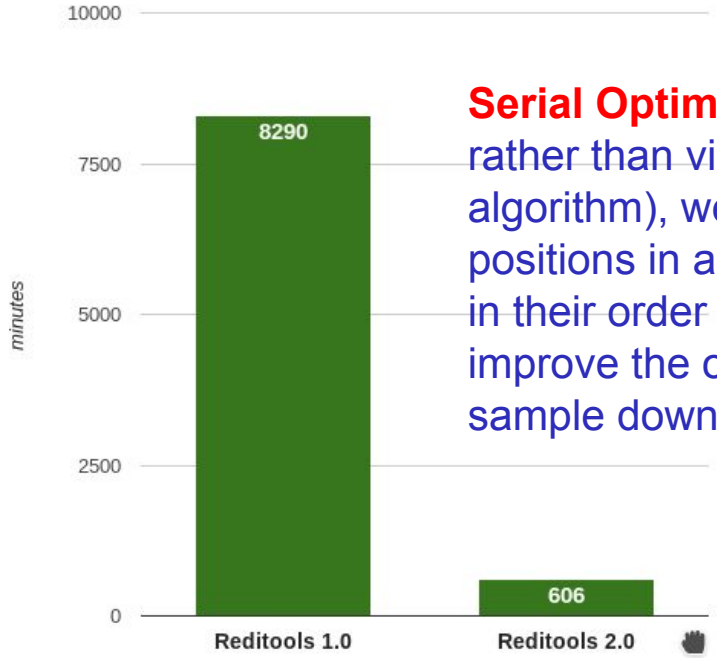
- We have also developed high parallel Python tools to manage the result, a vcf file of 150 GB.

Example of support to preparatory-PRACE project

Support to the project: "Developing HPC-Reditools, a parallel algorithm to detect Rna-editing events".
(Dipartimento di Bioscienze, Biotecnologie e Biofarmaceutica, Università di Bari. P.I. Ernesto Picardi)



Aim of this project is to improve the speed of REDItools, a tool the analysis of genomic data, with particular regard to RNA-Seq data.

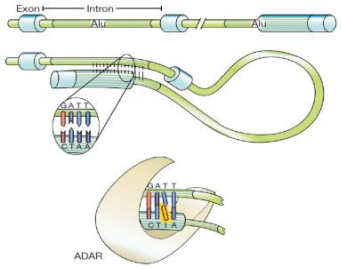


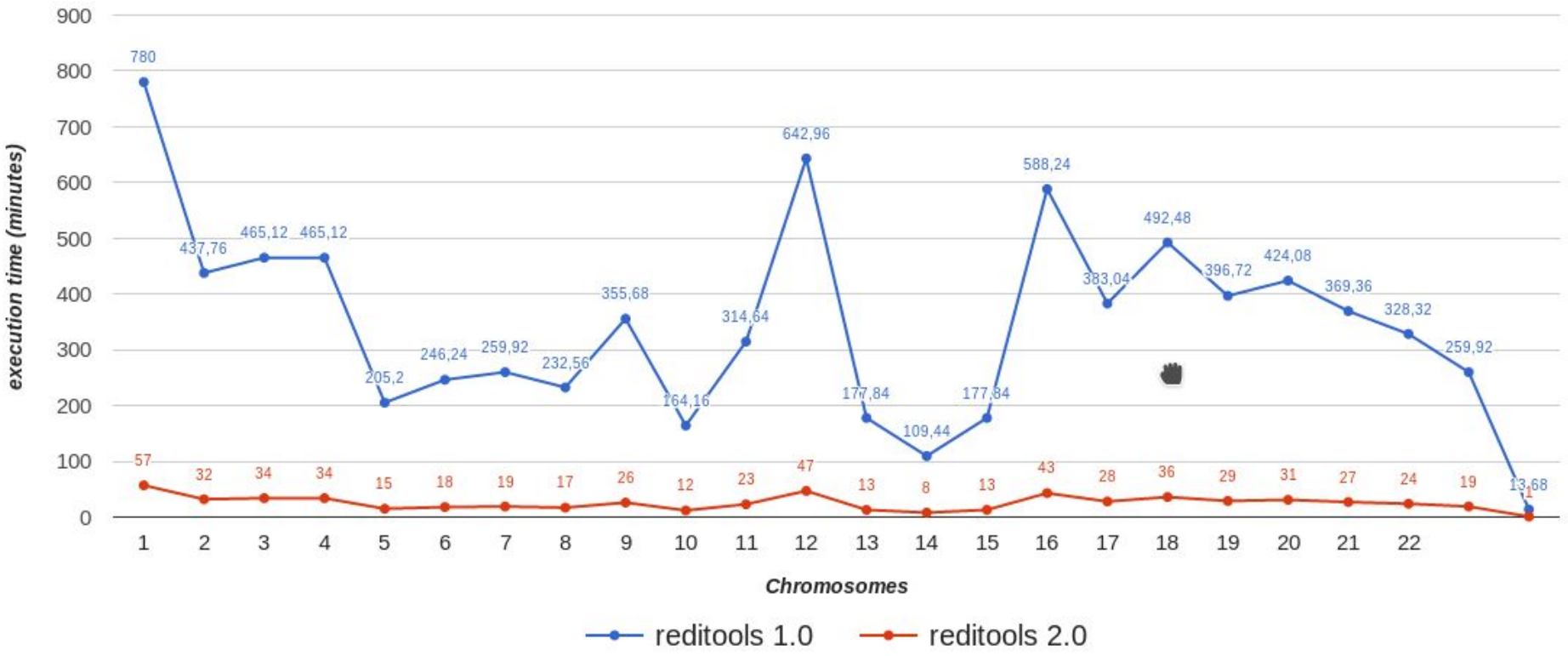
Serial Optimization:
rather than visiting genomic positions in serial (as in the original serial algorithm), we developed preliminary code which iterates through the positions in an optimized manner, by smartly loading supporting reads in their order of appearance in the BAM file. We thus managed to improve the overall running time, cutting it down from 30-40 hours per sample down to less than half an hour.

Example of support to preparatory-PRACE project

Results for parallel implementation

as further code will be developed during the PRACE project, we foresee to implement a further parallelization over chromosomes or chromosomal range. This coupled with the previous optmizations will make possible to analyze a single sample in approximately less than 5 minutes.





Involvent of Cineca in several community projects

THE TELETHON-CINECA GENOMICS REPOSITORY

Non-profit organisation recognised by the Ministry of the university and scientific and technological research founded to foster research that leads to cures for rare genetic diseases.

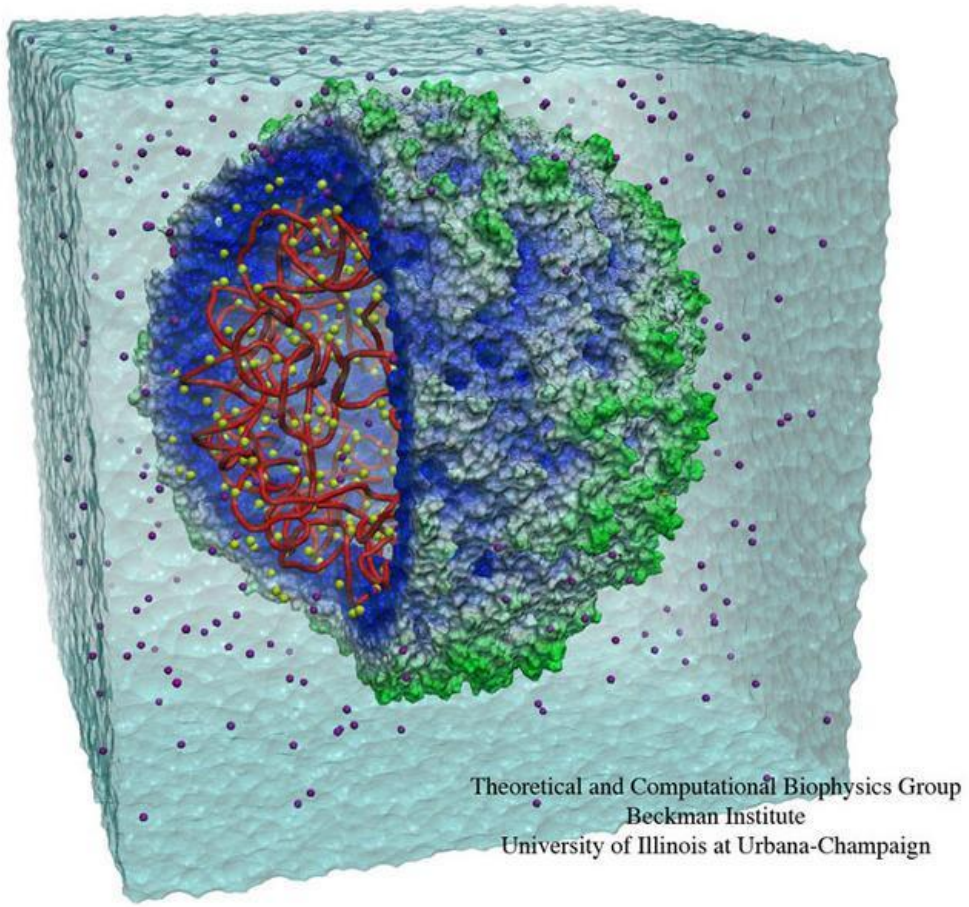
DIVA PROJECT

Italian database of BRCA1 and BRCA2 variants realised with the collaboration of six Italian oncological centres (IRCCS).

NIG – Network for Italian Genomes

The NIG fosters genetics and genomics research in broad sense with special reference to Italian population.

Satellite Tobacco Mosaic virus



STMV (virus) benchmark (1,066,628 atoms, periodic, PME)

hpc-bioinformatics@cineca.it

- Giovanni Chillemi

g.chillemi@cineca.it

- Tiziana Castrignanò

t.castrignano@cineca.it

- Ilario Tagliaferri

i.tagliaferri@cineca.it