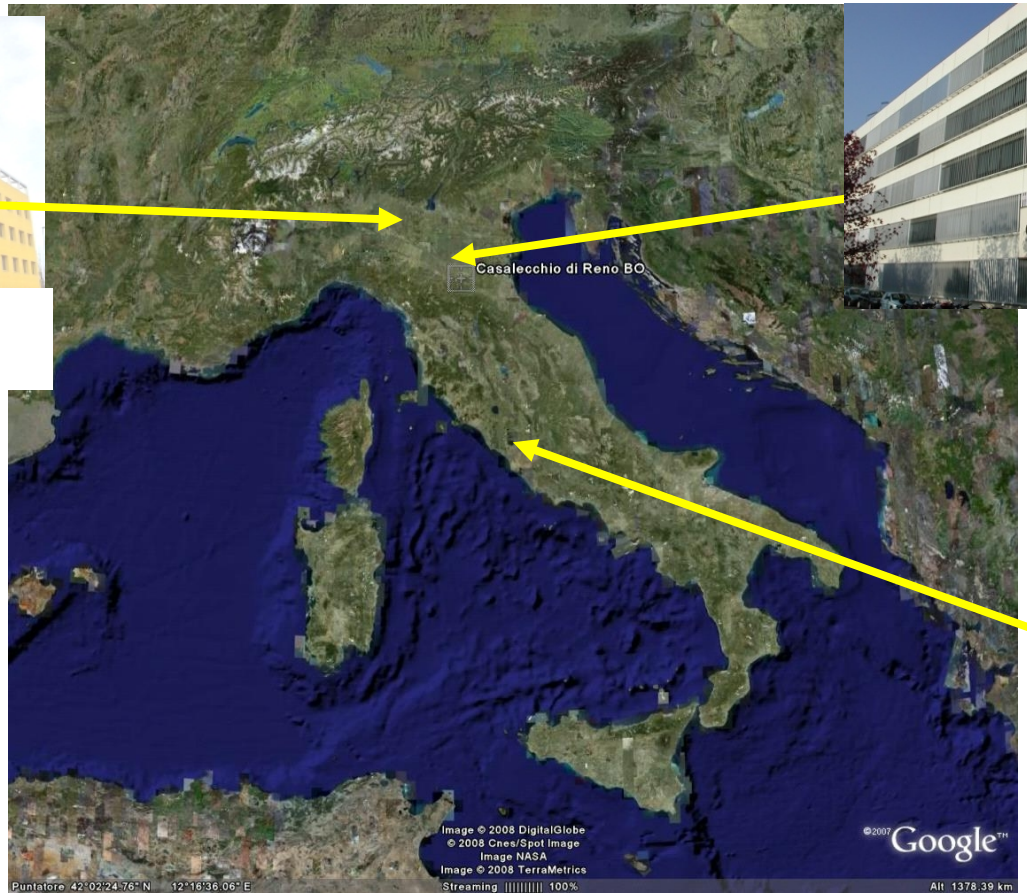
Gene Expression profiling with HTS: RNA-Seq data analysis

Giovanni Chillemi
CINECA – Supercomputing Application & Innovation

21 Ottobre 2015

CINECA today



CINECA is a non profit Consortium, made up of 69 Italian universities, and three Institutions (CNR, OGS and MIUR).

CINECA today it is the largest Italian computing centre, one of the most important worldwide. The High Performance Systems department (SCAI: SuperComputing Applications and Innovation) offers support to scientific and technological research through supercomputing and its applications.

FERMI

High-end system, only for extremely scalable applications

Name: Fermi
Architecture: BlueGene/Q (10 racks)
Processor type: IBM PowerA2 @1.6 GHz
Computing Nodes: 10.240
Each node: 16 cores and 16GB of RAM
Computing Cores: 163.840
RAM: 1GByte / core (163 TByte total)
Internal Network: 5D Torus
Disk Space: 2PByte of scratch space
Peak Performance: 2PFlop/s
Power Consumption: 820 kWatts

N. 7 in Top 500 rank (June 2012)

National and PRACE Tier-0 calls

GALILEO

X86 based system for production of medium scalability applications

Name: Galileo
Model: IBM NeXtScale
Architecture: IBM NeXtScale

Processor type: Intel Xeon Haswell@ 2.4 GHz
Computing Nodes: 516
Each node: 16 cores, 128 GB of RAM
Computing Cores: 8.256
RAM: 66 TByte
Internal Network: Infiniband 4xQDR switches (40 Gb/s)
Accelerators: 768 Intel Phi 7120p (2 per node on 384 nodes
+ 80 Nvidia K80

Peak Performance: 1.2 PFlops

National and PRACE Tier-1 calls

PICO

Storage and processing of large volumes of data. Data Analytics.
BIOINFORMATICS

Name: Pico
Model: IBM NeXtScale
Architecture: Linux Infiniband cluster
Processor type: Intel Xeon E5 2670 v2 @2,5Ghz
Computing Nodes: 66+
Each node: 20 cores, 128 GB of RAM + 2 accelerators
Computing Cores: 1.320+
RAM: 6,4 GB/core

+2 Visualization nodes
+2 Big Mem nodes
+4 BigInsight nodes

Integrated with a multi tier storage system:

40 TByte of SSDs
5 PByte of Disks
16 PByte of Tapes

Computing Infrastructure

Tier0: Fermi
Tier1: Galileo
BigData: Pico

Tier0: new
(HPC Top10) ~16/18PFlops
BigData: Galileo/Pico

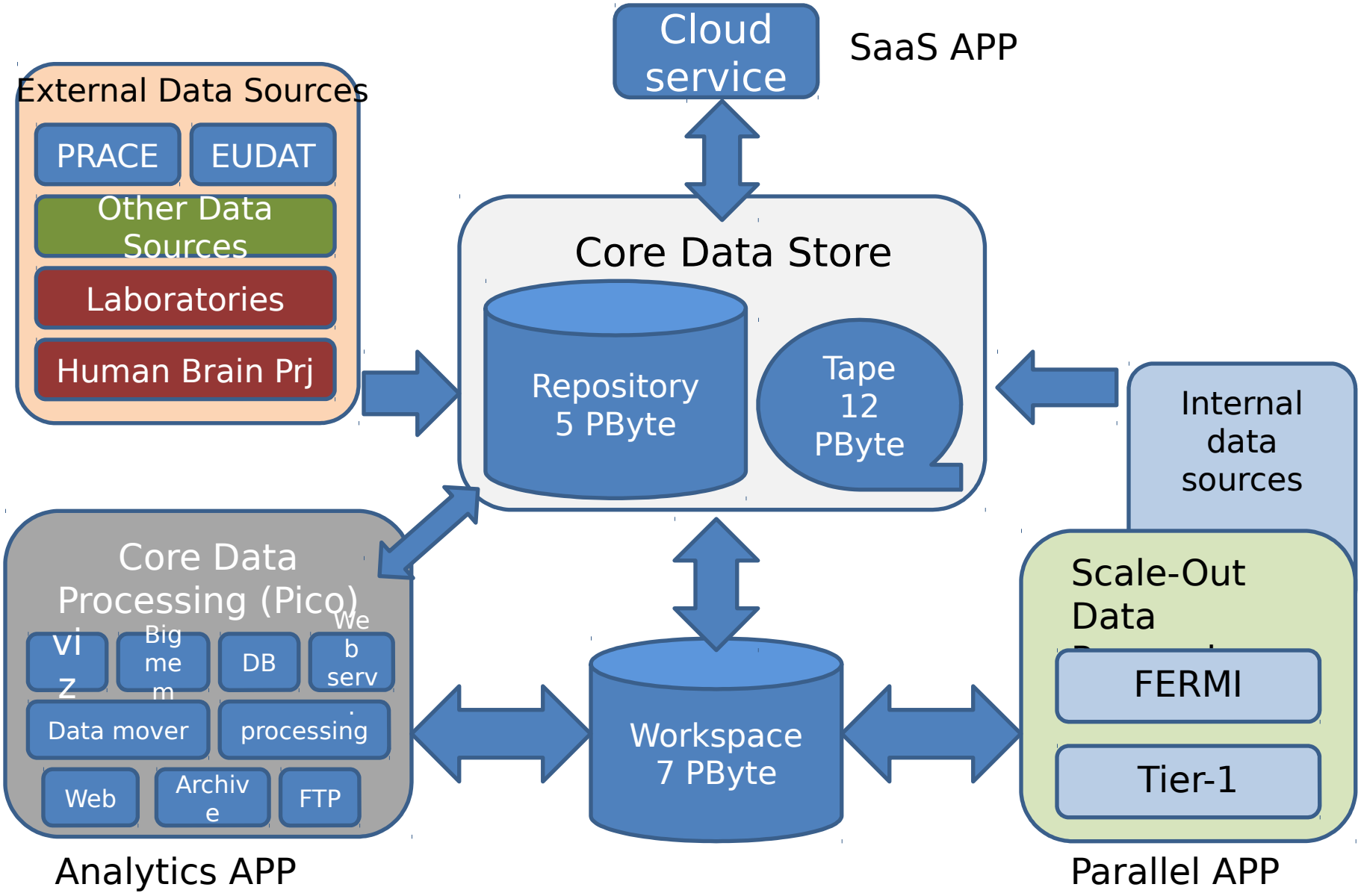
Tier0 BigData:
50PFlops
50PByte

today

1Q 2016

2018

(data centric) Infrastructure



How to get HPC resources

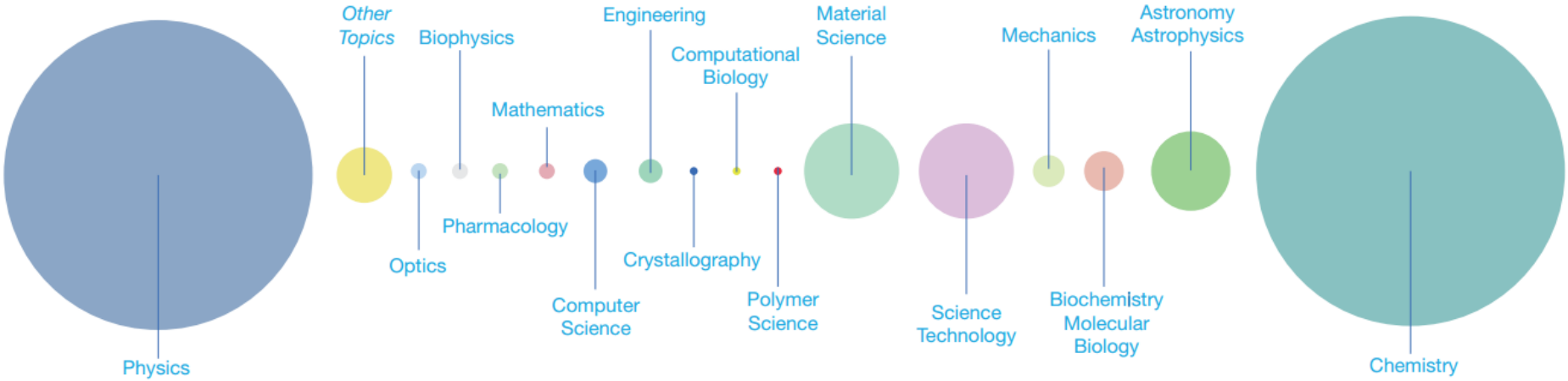
Peer reviewed projects:
you can submit a project that will be reviewed. If you win you will get the needed resources for free

National:
ISCRA http://www.hpc.cineca.it/services/iscra

European:
PRACE http://www.prace-ri.eu/Call-Announcements
EUDAT: http://eudat.eu/eudat-call-data-pilots
(storage resources)

Special agreement are signed with research institutions in a co-funding scheme

Distribuzione per disciplina

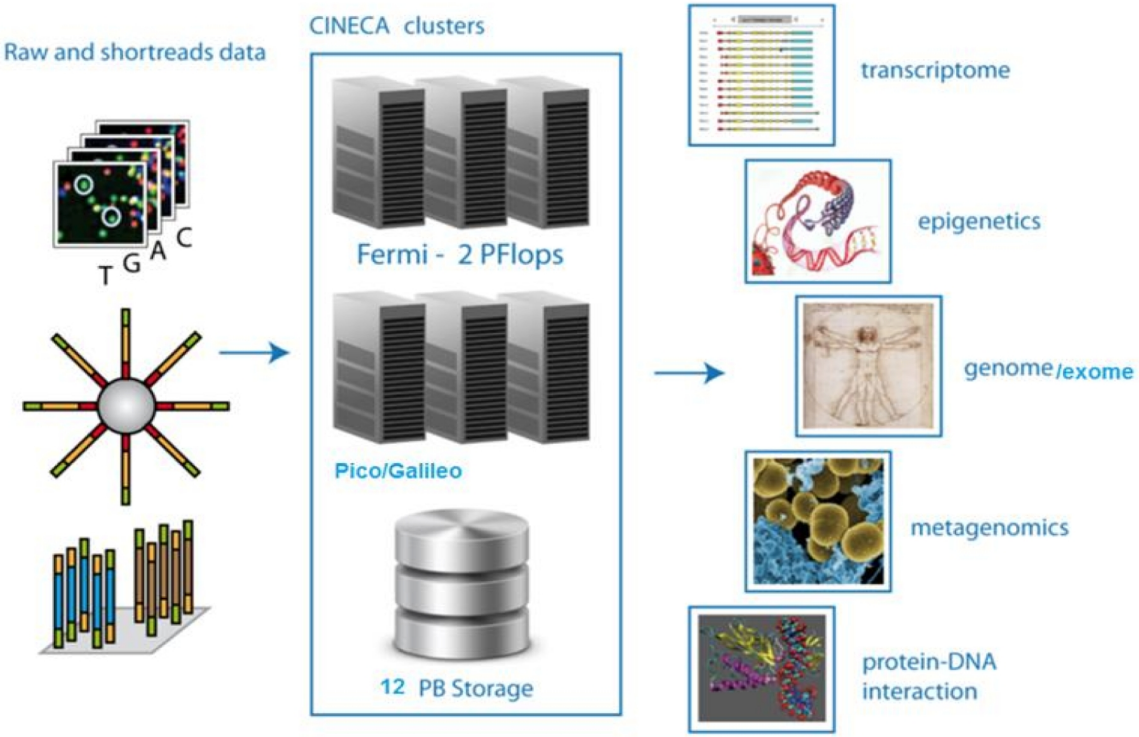


È auspicabile un aumento delle risorse consumate dalle discipline computazionali emergenti, come la bioinformatica

Bioinformatics environment on HPC infrastructures

All HPC infrastructures can be used for bioinformatics research and analysis.

CINECA has built a user-friendly bioinformatics environment supporting data analysis for several organisms on Pico architecture (http://www.hpc.cineca.it/hardware/pico):



Command line access (standard linux module environment)

gchillem@node001:~

```
gchillem@node001.pico:[~]$ module load profile/bio
gchillem@node001.pico:[~]$ module av
------------ /cineca/prod/modulefiles/profiles ------------
profile/advanced      profile/bio
profile/base(default) profile/gmql
------------ /cineca/prod/modulefiles/bio/biodata ------------
D_melanogaster/dm3   Mus_musculus/mm9     Z_mays/ZmB73
Homo_Sapiens/hg18    R_norvegicus/rn4     Z_mays/maize3
Homo_Sapiens/hg19    S_cerevisiae/sacCer3
Mus_musculus/mm10    Z_mays/Mo17_v1
------------ /cineca/prod/modulefiles/bio/environment ------------
autoload/0.1
------------ /cineca/prod/modulefiles/bio/libraries ------------
plink-seq/0.10--gnu--4.8.3
------------ /cineca/prod/modulefiles/bio/compilers ------------
gnu/4.8.3    jre/1.7.0_72 perl/5.x
------------ /cineca/prod/modulefiles/bio/tools ------------
plink/1.07
------------ /cineca/prod/modulefiles/bio/applications ------------
abra/0.86           cluto/2.1.1        mauve/2.3.1         samtools/0.1.19
annovar/2014Nov12   concoct/0.4.0      mosaik/2.2.3        samtools/1.1
annovar/2014Sep15   cufflinks/2.2.1    mrjob/0.4.2         shrimp/2.2.3
annovar/2015Mar22   diamond/0.7.9      mrjob/0.4.3-dev     sicer/1.1
bedtools/2.21.0     fastqc/0.11.2      mummer/3.23         snpeff/4.1b
bedtools/2.24       fastqc/0.11.3      mutect/1.1.4        spades/3.5.0
bioconductor/2.14   gatk/3.3.0         ngsqctoolkit/2.3.3  splicetrap/0.95
bioconductor/3.0    gatk/3.4.46        ngsrich/0.7.8       sra/2.4.2-4
biopython/1.65      homer/4.7          pagit/1.1           star/2.4.0d
bismark/0.14.2      htseq/0.6.1        peakranger/1.17     tophat/2.0.11
blast+/2.2.30       idl/8.1            picard/1.119        tophat/2.0.12
bowtie/1.0.1        igvtools/2.3.40    pintron/1.3.0       trimmomatic/0.33
bowtie2/2.2.3       lofreq/2.1.1       qiime/1.9.0         ucsc/1.0
bwa/0.7.10          macs/1.4.0         r/3.1.2             varscan2/2.3.7
chimerascan/0.4.5a  macs/2.0.9         ray/2.3.1           velvet/1.2.10
cisa/1.3            macs/2.1.0         reditools/1.0.3
gchillem@node001.pico:[~]$
```

web pipeline

HPC for Next Generation Sec

https://bioinformatics.cineca.it

HPC for Next Generation Sequencing

High-Performance Bioinformatics Services for Next Generation Sequencing data analysis in Public Health and Research

Next-generation DNA sequencing (NGS) has incredibly accelerated biological and biomedical research, by allowing the comprehensive analysis of genomes, transcriptomes and interactomes. Managing the huge amount of data from new sequencing platforms requires non trivial skills, strong computational power and storage capacity which are generally not available in most research labs. Our consortium has been recognized as big data center and HPC analysis for the Italian epigenomic flag project Epigen.
The CINECA centralized bioinformatics core facility provides shared resources for the computational and IT requirements.

Whole-Exome Sequencing

Whole Exome Sequencing (WES) analysis is now available for several research purposes. A frequently updated pipeline, WEP, is used to call variants, both SNPs and indels. Variants are then filtered with many public databases including dbSNP, the 1000 Genomes project, HapMap exomes and more. Variant prioritization is obtained by comparing disease and healthy controls and performing their functional annotation (e.g. the functional relevance of a protein variant is assessed by SITF software). Moreover, for family-based samples, the advanced analysis of haplotype phasing and complex heterozygous or homologous mutations detection is available as well.

UDT-seq

A new sequencing platform, the MiSeq Illumina sequencer, allows to identify known causative mutations by producing a Ultra-Deep coverage on a selected list of Targeted genomic regions Sequencing (UDT-Seq). UDT-Seq is becoming particularly suitable for clinical diagnostic applications since it implies full coverage of sequenced regions and guarantees that no other mutation was lost by the analysis. ODESSA (Online Deep Exome Sequencing Software Analysis) is a new automated high-performance bioinformatics web platform, developed for targeting genes at high coverage through deep sequencing with the maximum usability, and focused on rational diagnosis of targeted therapies. It identifies Single Nucleotide Variants (SNVs) and Deletion/Insertion Variants (DIVs) classified by different useful scores (e.g. depth coverage).

CINECA
Consorzio Interuniversitario

web pipeline

HPC for Next Generation Sec

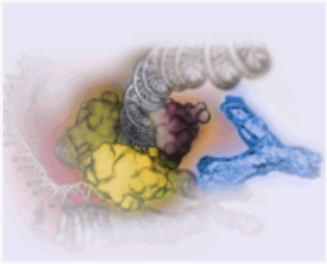
https://bioinformatics.cineca.it

RNA-Seq

RNA-Seq (Transcriptome) analysis is now avaliable for transcriptome structural analysis and quantification. The transcriptome analysis allows the identification of known or novel expressed transcript variants, and their quantification.
RNA-Seq, unlike microarrays, does not require prior knowledge of the genome and therefore offers several advantages. Our facility, RAP, can study the transcriptome profiling of each sample, performs differential gene expression analysis, cassette exons, chimeric transcripts and polyA sites detection.

RNA editing (from RNA-seq data)

RNA editing is a post-transcriptional mechanism challenging the central dogma of molecular biology. Nowadays, the term RNA editing is also used to indicate post-transcriptional changes due to specific base substitutions. Such alterations may affect coding as well as non-coding RNAs located in different cellular compartments and occur in a variety of organisms. ExpEdit is a web application for assessing RNA editing in human at known or user-specified sites supported by transcript data obtained by RNA-Seq experiments. Mapping data or directly sequence reads can be provided as input to carry out a comparative analysis against a large collection of known editing sites collected in DARNED database as well as other user-provided potentially edited positions. Results are shown as dynamic tables containing University of California, Santa Cruz (UCSC) links for a quick examination of the genomic context.



ChIP-Seq

ChIP-Seq is widely used to analyze DNA-protein interactions. It combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify binding sites of DNA-associated proteins, and can be used to precisely map global binding sites for any protein of interest. Our bioinformatic service, CAST, provides Genome-wide distribution of ChIP sequencing reads, peak identification and differential analysis across different samples.

Whole Exome sequencing Pipeline web tool

Whole-Exome sequencing Pipeline web tool

The **WEP resource** performs a complete **whole-exome sequencing pipeline** and provides easy access through interface to intermediate and final results.

The pipeline is composed of several steps:

1. Verification of input integrity, quality checks, read trimming and primer contamination removal;
2. Gapped alignment;
3. BAM conversion, sorting and indexing;
4. Duplicates removal, as they result as PCR amplification bias;
5. A local realignment around known IN-DELs position, carried on to delete the other artifacts;
6. Quality score recalibration to refine some oddness caused by sequencing and mapping on quality scores;
7. Variants (SNV and DIP) calling from the filtered mapping data obtained from the previous steps;
8. Association of as many annotation as possible to the variant list (i.e. annotation stored in database like dbSNP, 1000 Genomes Project, etc.);
9. Data post processing: raw outputs are parsed and stored into custom databases to allow cross-linking and intersections, statistics and much more.

Through our tool a user can perform the whole analysis without knowing the underlying hardware and software architecture, dealing with both paired and single end data. The interface provides an easy and intuitive access for data submission and user-friendly web pages for annotated variant visualization.

Non-IT mastered users can access through WEP to the most updated and tested whole exome sequencing algorithms, ad-hoc tuned to maximize the quality of variants called while minimizing artifacts and false positives.

BMC Bioinformatics. 2013 Apr 22;14 Suppl 7:S11. doi: 10.1186/1471-2105-14-S7-S11. Epub 2013 Apr 22.

WEP: a high-performance analysis pipeline for whole-exome data.

D'Antonio M, D'Onorio De Meo P, Paoletti D, Elmi B, Pallocca M, Sanna N, Picardi E, Pesole G, Castrignanò T.

PMID 23815231

Login

E-mail

Password

Login

Click here to register a new account

Forgot your password? Click here to reset

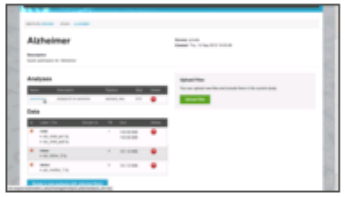
News

No news available for this service at the moment

VALID RSS

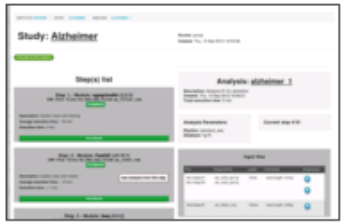
How to: analysis monitoring and results

Do you need help submitting an analysis? Go back to our help page



1. Access your analysis

You can access your analysis right after the submission, or from
Project list > your project > your analysis



2. Analysis page

From the analysis page, you can monitor the state of your analysis step by step.
You can also display additional information regarding every single step of your analysis, e.g. the command line and the running time.

3. Step Monitoring

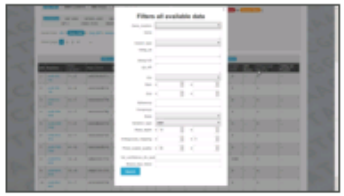
For each completed step, click on the "view results" green bar, to access the single step reports.

4. Results Summary

At the end of your analysis, you'll be able to read the results for each file. The results will be linked in your analysis page, or you can access our sample results from the *Results example* page.

5. Variants list

From the Results page (*link* button or *variants* page) is possibile to interact with the variants, sort the annotation, export the results as CSV (comma-separates-values) file and much more.



6. Filters

Each variant is hyperlinked to public databases for the visualization of read alignments and variant calling information at the variant position.
You can browse the list of variants, and filter them by position, type (SNPs, Indels), zygosity, presence in dbSNP, etc.

7. Intersections

Finally, WEP provides also an "intersections" section allowing the user to search for variants shared between samples.
The variants are classified as homozygous or heterozygous by GATK algorithm within each analyzed individual