École des sciences avancées de Luchon

School for advanced sciences of Luchon

Networks and data mining
Session II, June 27 - July 11, 2015

Integrating multi-omics

Luciano Milanesi

Consiglio Nazionale delle Ricerche

InterOmics
Flagship Project

Outline

- Introduction
- Omics challenges
- Data Integration
- Big Data
- Personalized system medicine
- International Initiatives
- Conclusions

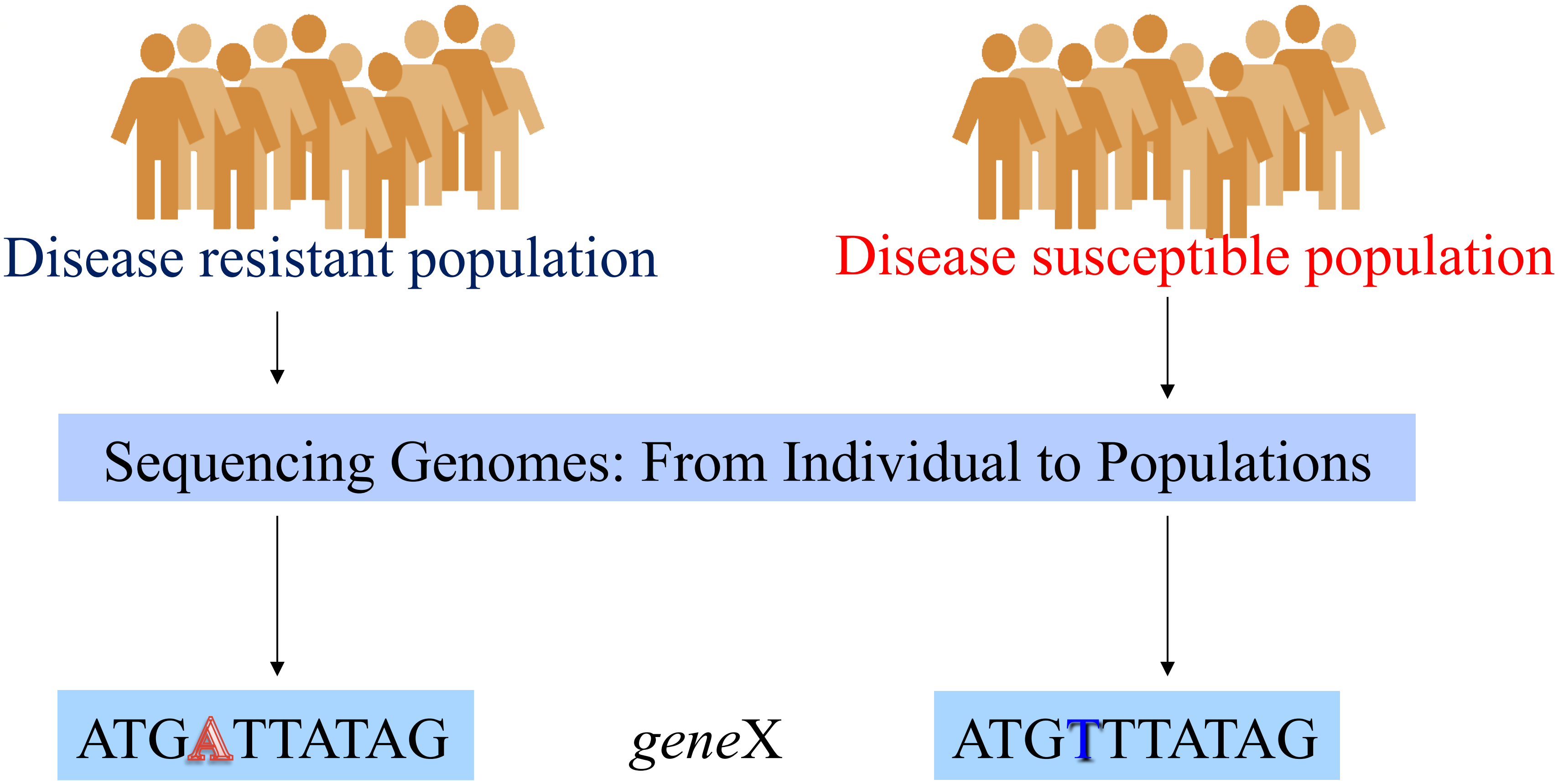
Big Data in “Omics Sciences”

The *"Omics Sciences"* consist of several areas of investigation :

- **Genomics,**
- **Proteomics,**
- **Interactomics,**
- **Bioinformatics,**
- **Neuroinformatics**
- **System Biology**
- **Metabolomics**
- Ecc.

These and the correlated disciplines constitute the paradigm around which all the research in the fields of biomedicine, biotechnology and ICT generally applicable to the biomedical sciences

Omics Applications



Disease resistant population

Disease susceptible population

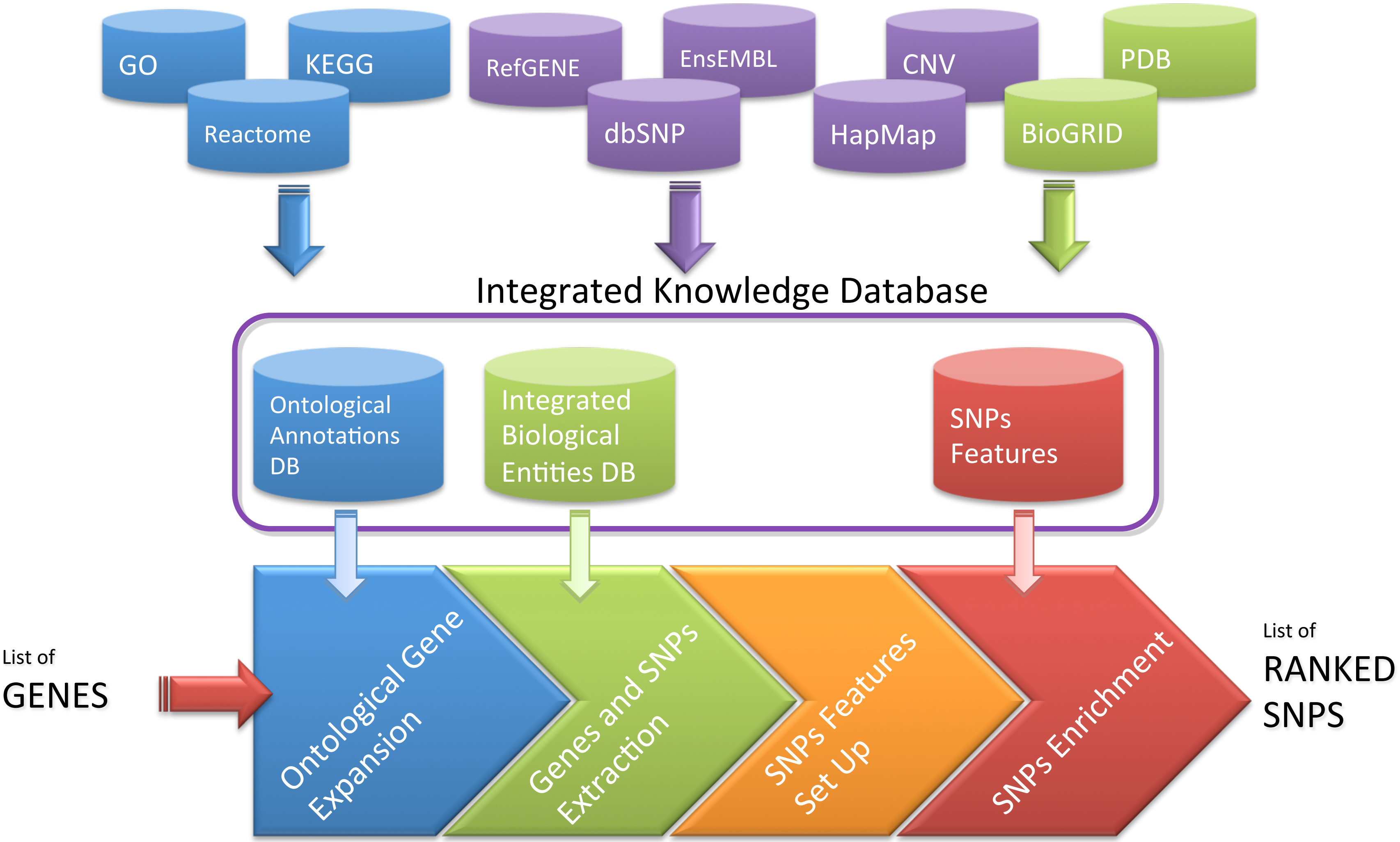
Sequencing Genomes: From Individual to Populations

ATGATTATAG

*gene*X

ATGTTTATAG

SNP and Biomarkers Analysis



GO
KEGG
Reactome
RefGENE
EnsEMBL
dbSNP
CNV
HapMap
PDB
BioGRID
Integrated Knowledge Database
Ontological Annotations DB
Integrated Biological Entities DB
SNPs Features
List of GENES
Ontological Gene Expansion
Genes and SNPs Extraction
SNPs Features Set Up
SNPs Enrichment
List of RANKED SNPS

Omics Technology

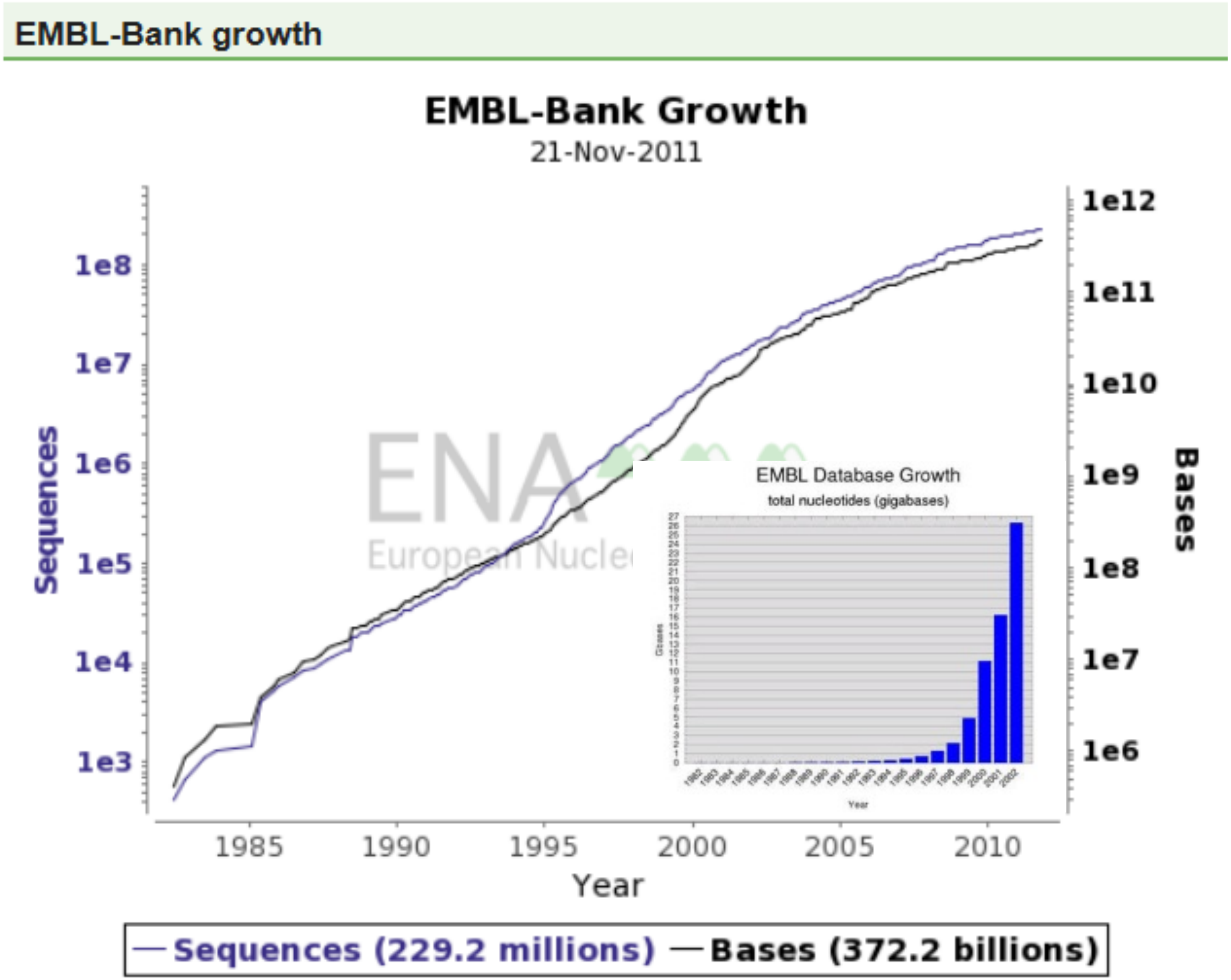
Targeted Resequencing

Viral Genomics

Transcriptome Sequencing miRNA

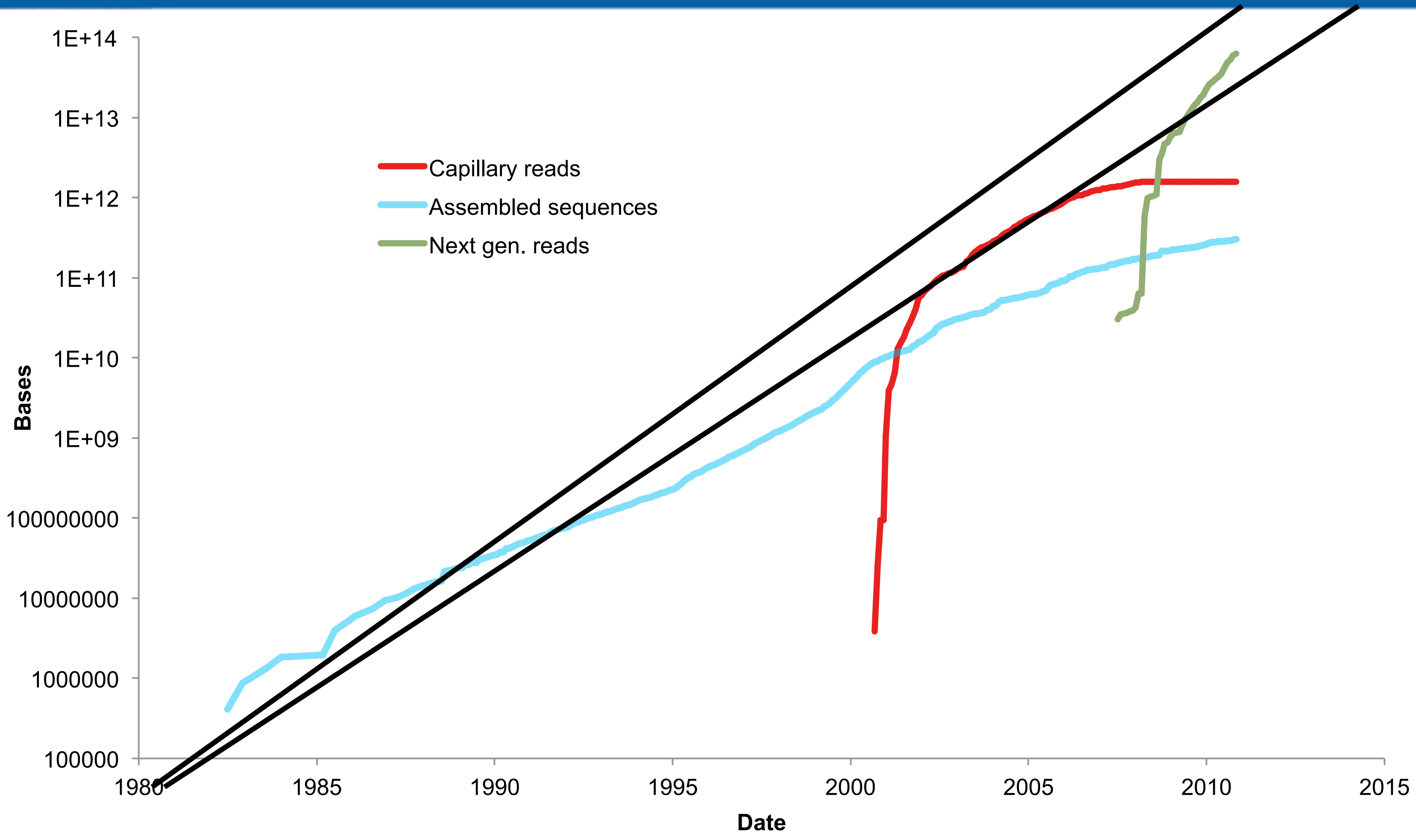
Microbial Genomics

Omics Data Explosion

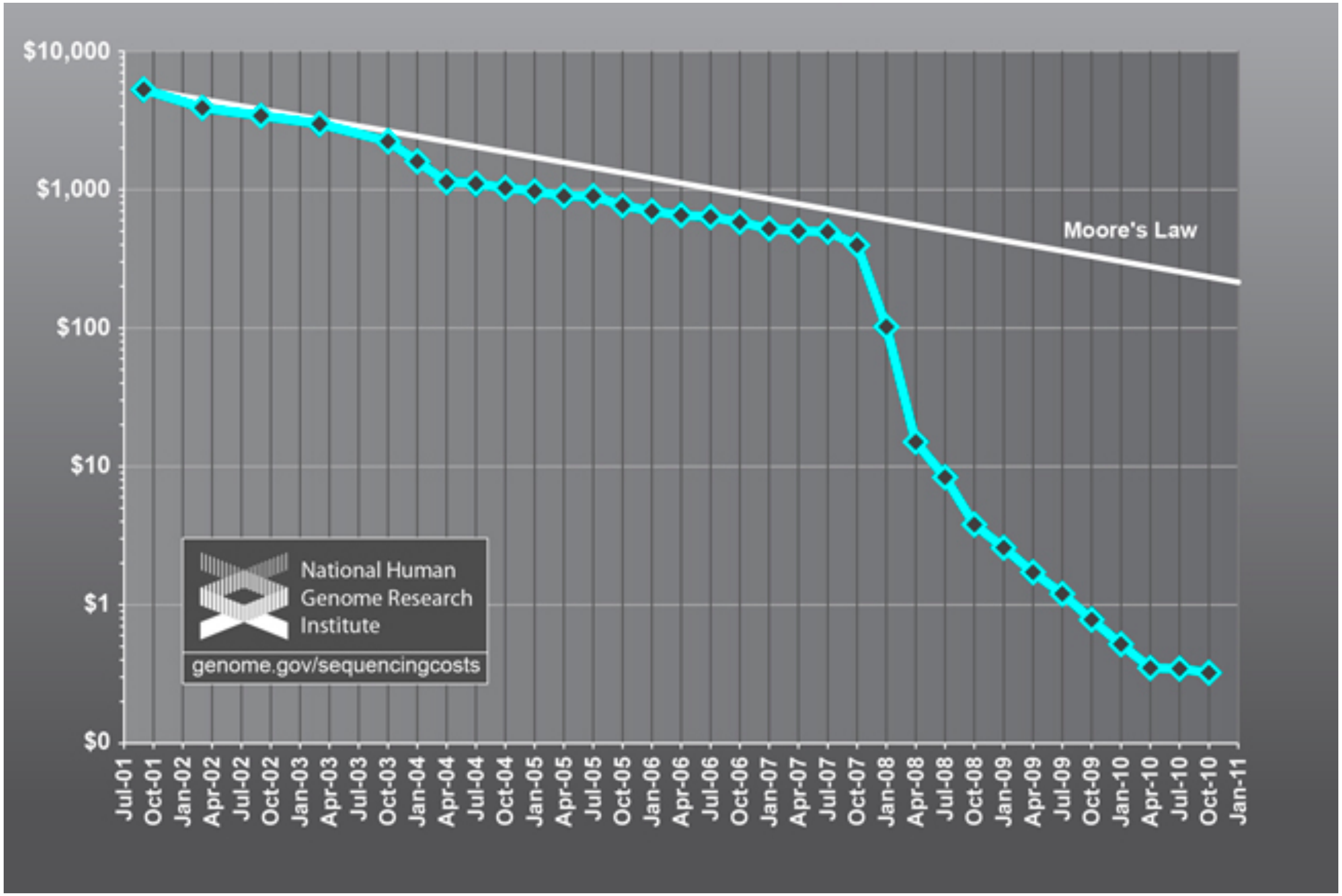


EMBL-Bank growth

Rate of sequence data generation



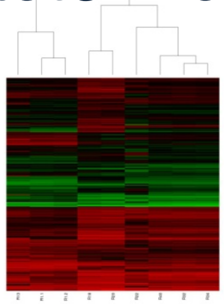
Cost of sequence data generation



InterOmics CNR

Omics Complexity Explosion

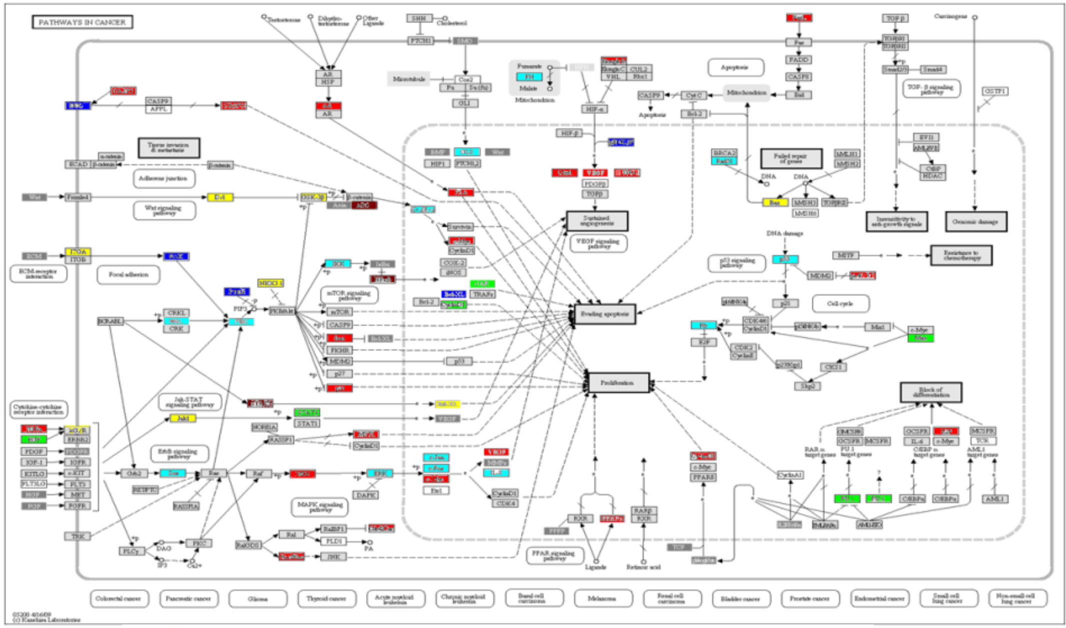
Interactomics and Pathways Discovery

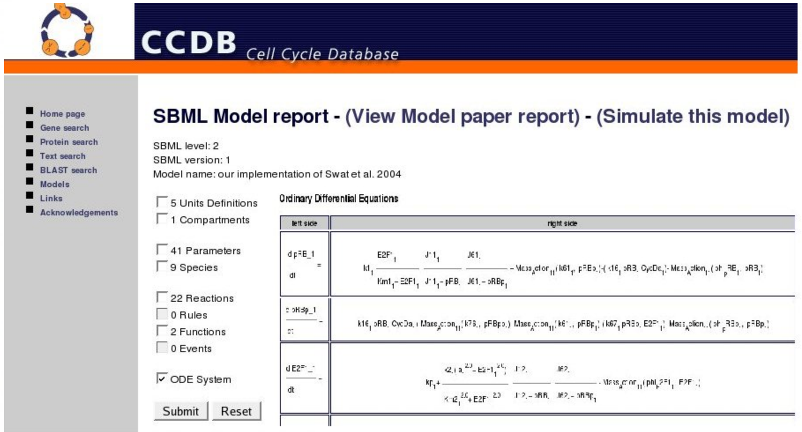


KEGG Kyoto Encyclopedia of Genes and Genomes

STRING

BioGRID





CCDB *Cell Cycle Database*

- Home page
- Gene search
- Protein search
- Text search
- BLAST search
- Models
- Links
- Acknowledgements

SBML Model report - (View Model paper report) - (Simulate this model)

SBML level: 2
SBML version: 1
Model name: our implementation of Swat et al. 2004

☐ 5 Units Definitions
☐ 1 Compartments

☐ 41 Parameters
☐ 9 Species

☐ 22 Reactions
☐ 0 Rules
☐ 2 Functions
☐ 0 Events

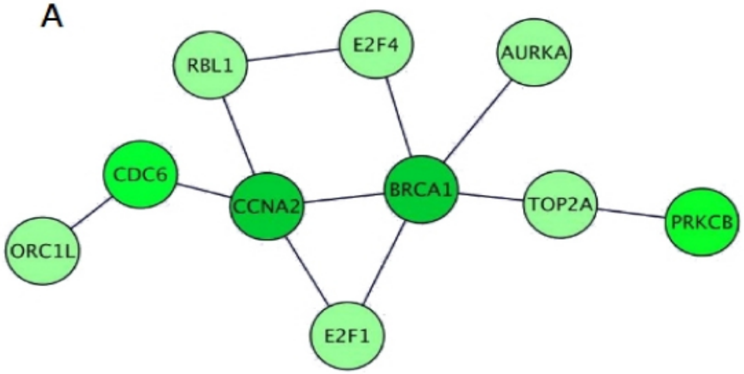
☑ ODE System

Submit Reset

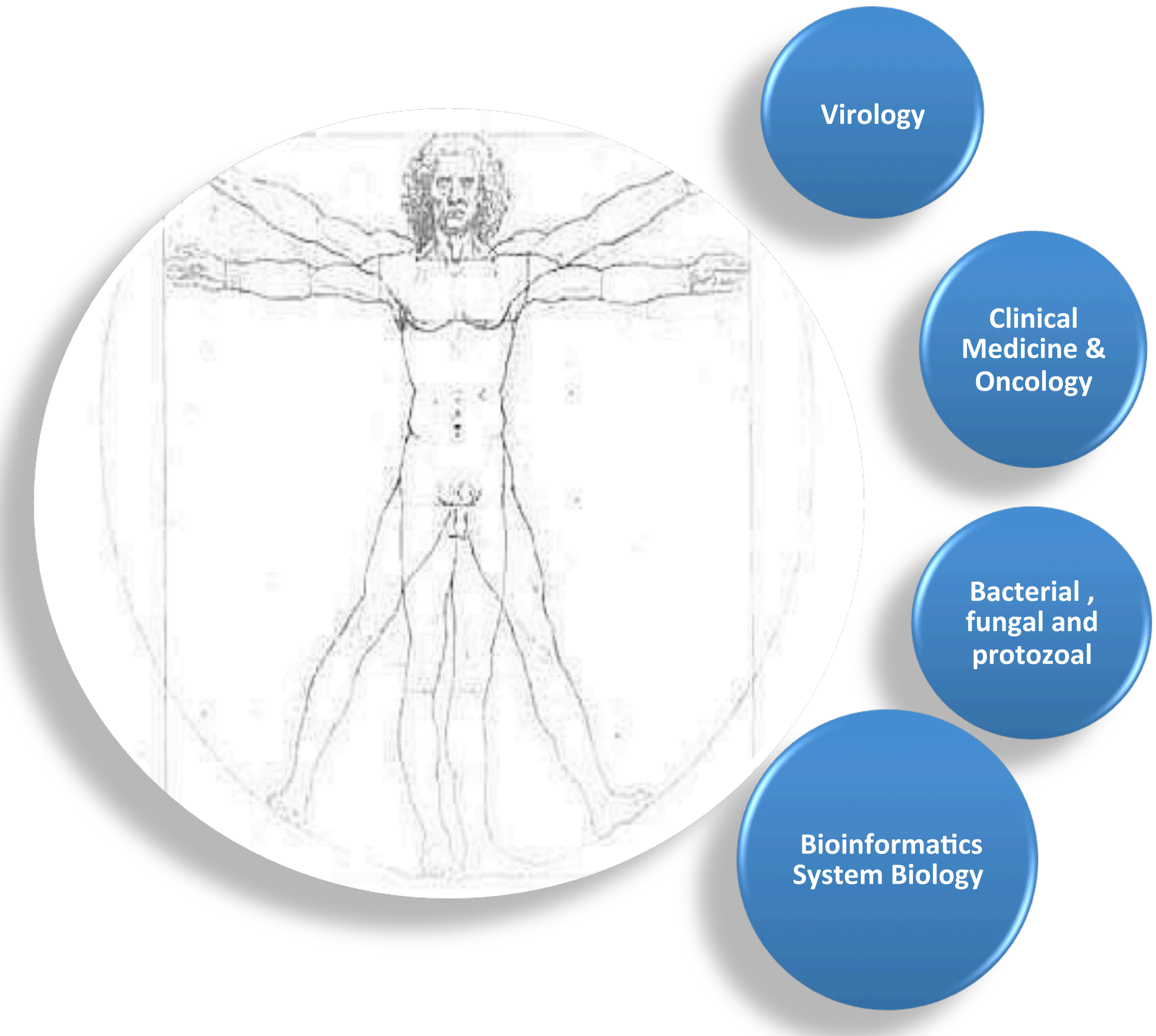
Ordinary Differential Equations

left side	right side
$\frac{d\,pRB\_1}{dt} =$	$k1_1 \frac{E2F1_1}{Km1_1 - E2F1_1} \frac{J11_1}{J11_1 - pRB_1} \frac{J61_1}{J61_1 - pRBp_1} - \text{Mass\_action}_{11}(k61_1, pRBp_1) - (k16_1\, pRB_1\, CycDa_1) - \text{Mass\_action}_1(phi\_pRB_1, pRB_1)$
$\frac{d\,pRBp\_1}{dt} =$	$k16_1\, pRB_1\, CycDa_1\, \text{Mass\_action}_{11}(k78_1, pRBpp_1)\, \text{Mass\_action}_{11}(k61_1, pRBp_1)\, (k67_1\, pRBp_1\, E2F1_1)\, \text{Mass\_action}_1(phi\_pRBp_1, pRBp_1)$
$\frac{d\,E2F1\_1}{dt} =$	$kp_1 + \frac{k2_1(a_1^{2.0} - E2F1_1^{2.0})}{Km2_1^{2.0} + E2F1_1^{2.0}} \frac{J12_1}{J12_1 - pRB_1} \frac{J62_1}{J62_1 - pRBp_1} - \text{Mass\_action}_{11}(phi\_E2F1_1, E2F1_1)$

A



Omics Applications



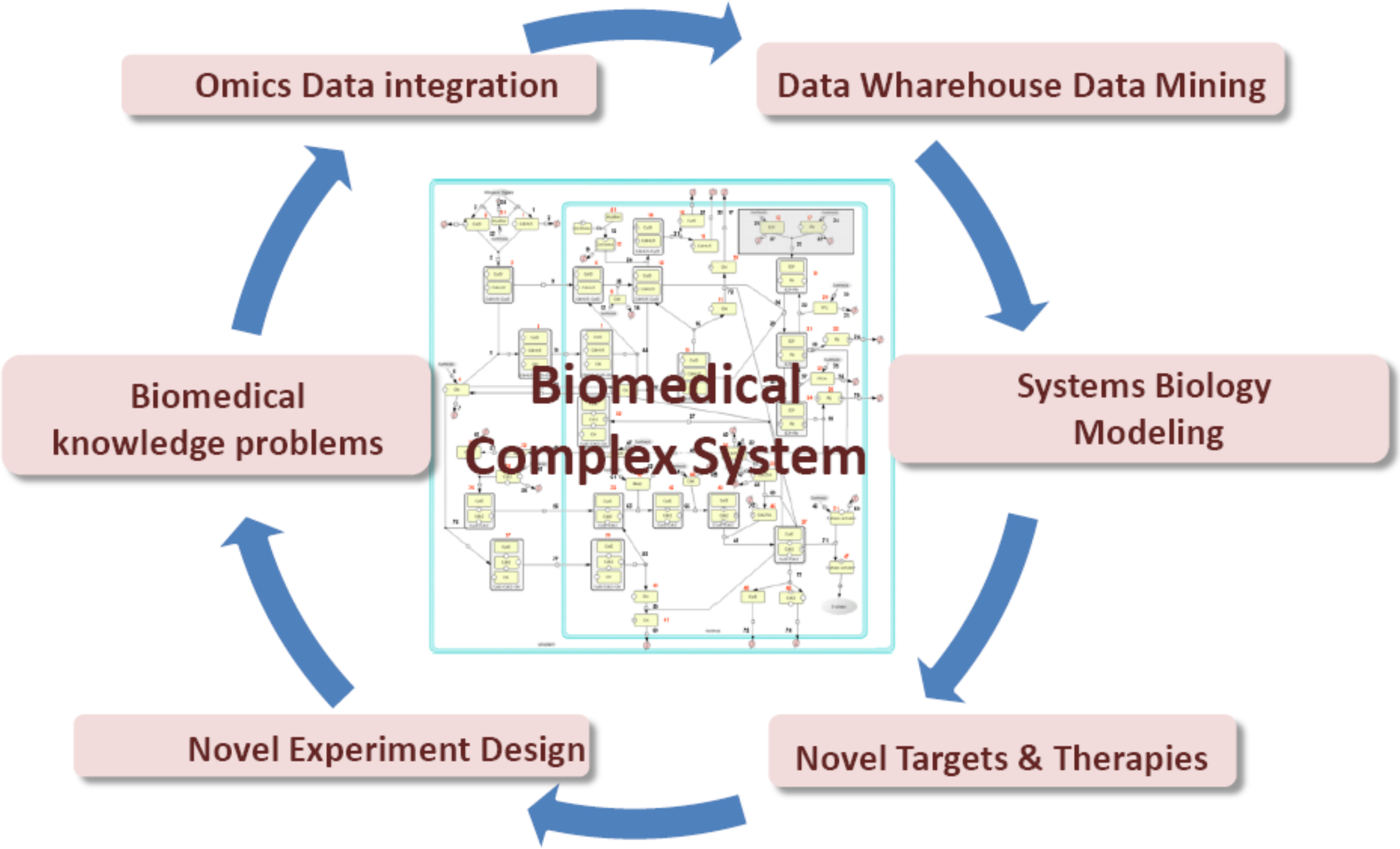
Virology

Clinical Medicine & Oncology

Bacterial , fungal and protozoal

Bioinformatics System Biology

Biomedical Complex System



Omics Data integration

Data Wharehouse Data Mining

Systems Biology Modeling

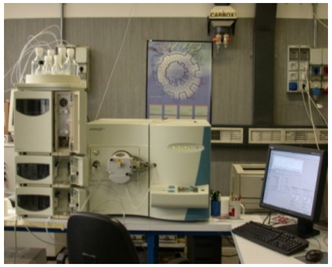
Novel Targets & Therapies

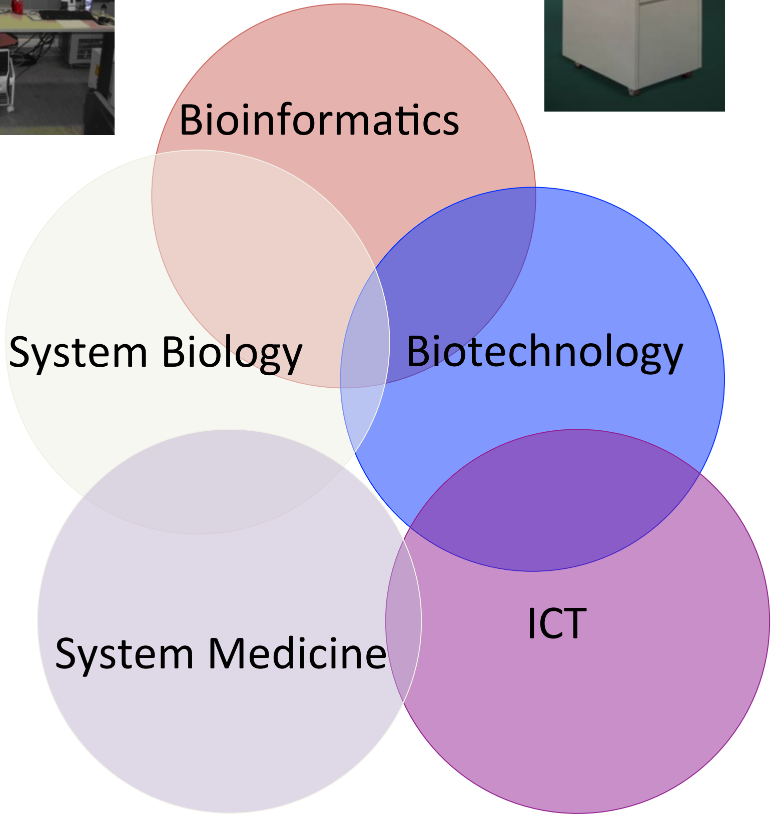
Novel Experiment Design

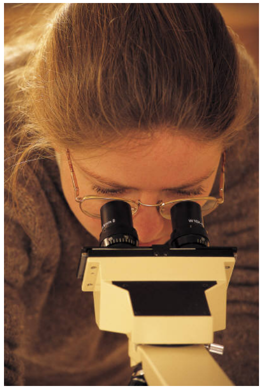
Biomedical knowledge problems

Biomedical Complex System

Omics Data Intergration





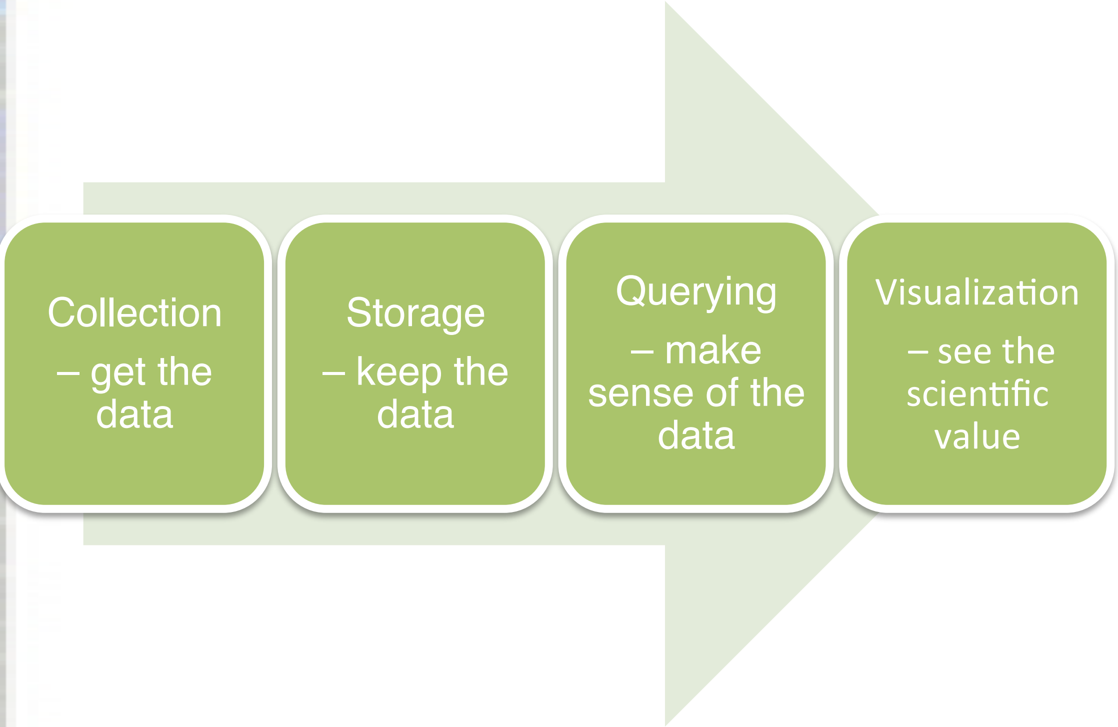


What is Big Data?

- Definition
  - **Big Data** refers to a collection of data sets so large and complex that it's impossible to process them with the usual databases and tools.
  - Because of its size and associated numbers, Big Data is hard to capture, store, search, share, analyze and visualize.
  - The three V's: **Volume, Velocity, Variety**
  - **High-Volume**: Amount of data
  - **High-Velocity**: Speed rate in collecting or acquiring or generating or processing of data
  - **High-Variety**: Different data type such as audio, video, image data, sequence data
- Processing
  - Parallel processing (eg. *Hadoop)*
  - Processing of data sets too large for transactional databases
  - Analyzing *interactions*, rather than *transactions*

L

Big Data



Collection – get the data

Storage – keep the data

Querying – make sense of the data

Visualization – see the scientific value

Who is collecting all of this data?

Medical Science

- Data bases from
- e-Health
- Patient Records
- Medical ImagingMRI & CT scans,, …
- Telemedicine
- Genomics
- Environmental data
- Food science
- Biosensors

Big Pharmaceutical Companies

Pfizer

BAYER

Roche

MERCK

AstraZeneca

gsk GlaxoSmithKline

Data integration

- **Cloud computing** in combination with **Big Data Tools** can be used to obtain the power and the scale of computation required to facilitate large-scale efforts required in **translational medicine data integration** and to perform analysis in more efficient and economical way.

InterOmics CNR

CNR-ITB Data Center

- Resources:
  - HPC (High Performance Computing) Cluster
  - HPSI (High Performace Storage Infrastructure) DDN –
  - WRVM (Web Remote Virtual Machine)
  - Databases: MySQL, ORACLE, SQL Server
  - Cluster Intel Servers: 44
  - Total RAM: 2.080 GB
  - Total Disk space: 1.164 TB
  - 192 CPU and 1.216 core
  - GPU Server : 16 GPU, 16 CPU and 96 core
  - Operating system: Ubuntu 13.04, Centos 6.5, Window Server, Mac OS
  - Portal technology: Java portal (LIFERAY)
  - GRID Node
  - Virtual Node
  - Cloud Computing
  - Hadoop

BBMRI-IT
Biobanking and BioMolecular Resources Research Infrastructure Italy

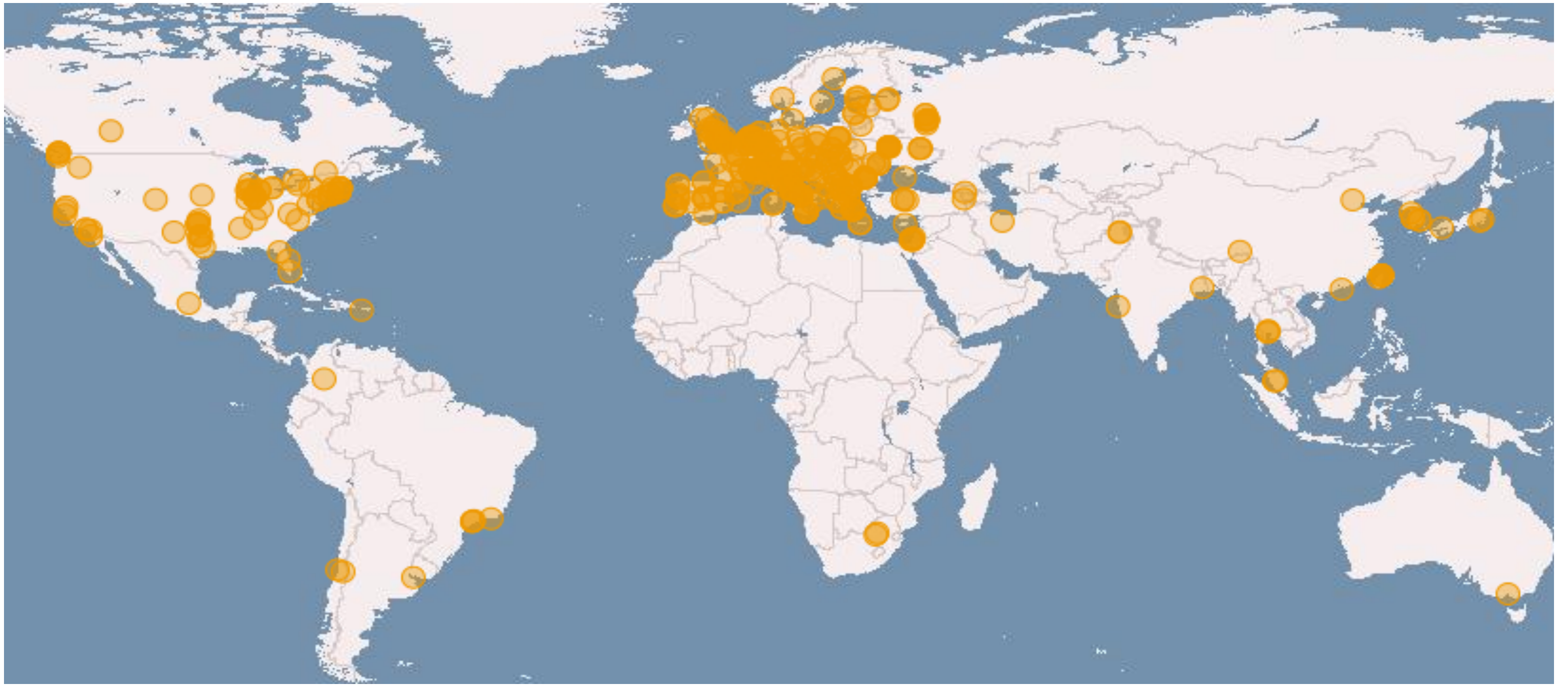
REPVBBLICA ITALIANA
Ministero dell'Istruzione dell'Università e Ricerca

InterOmics
Flagship Project

MIMOmics

HIRMA Hepatocarcinoma Innovative Research MArkers

European Grid and Cloud Infrastructure



- Distributed, federated storage and compute facilities
- Grid and Cloud compute platforms
- Virtual Research Environments
- > 200 user research projects

- 350 resource centres in 40 countries
- 400,000 logical CPU cores
- 190 PB disk, 180 PB tape
- > 99.6% reliability

WLCG Worldwide LHC Computing Grid

we-nmr

agINFRA A data infrastructure for agriculture

EISCAT SCIENTIFIC ASSOCIATION

VERCE

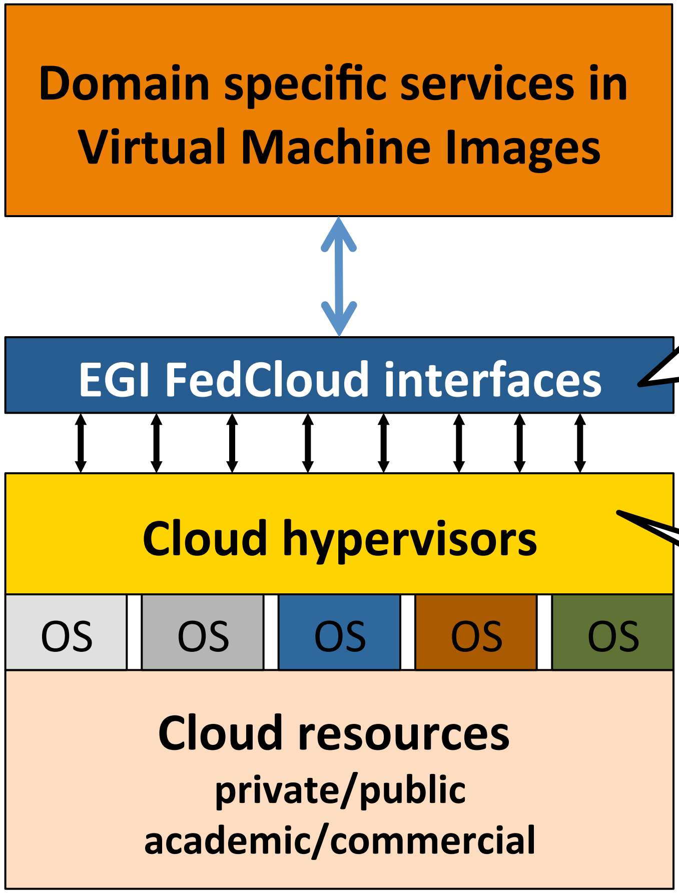
Life Science Grid Community

EUDAT

BioVeL

⋮

European Cloud Infrastructure



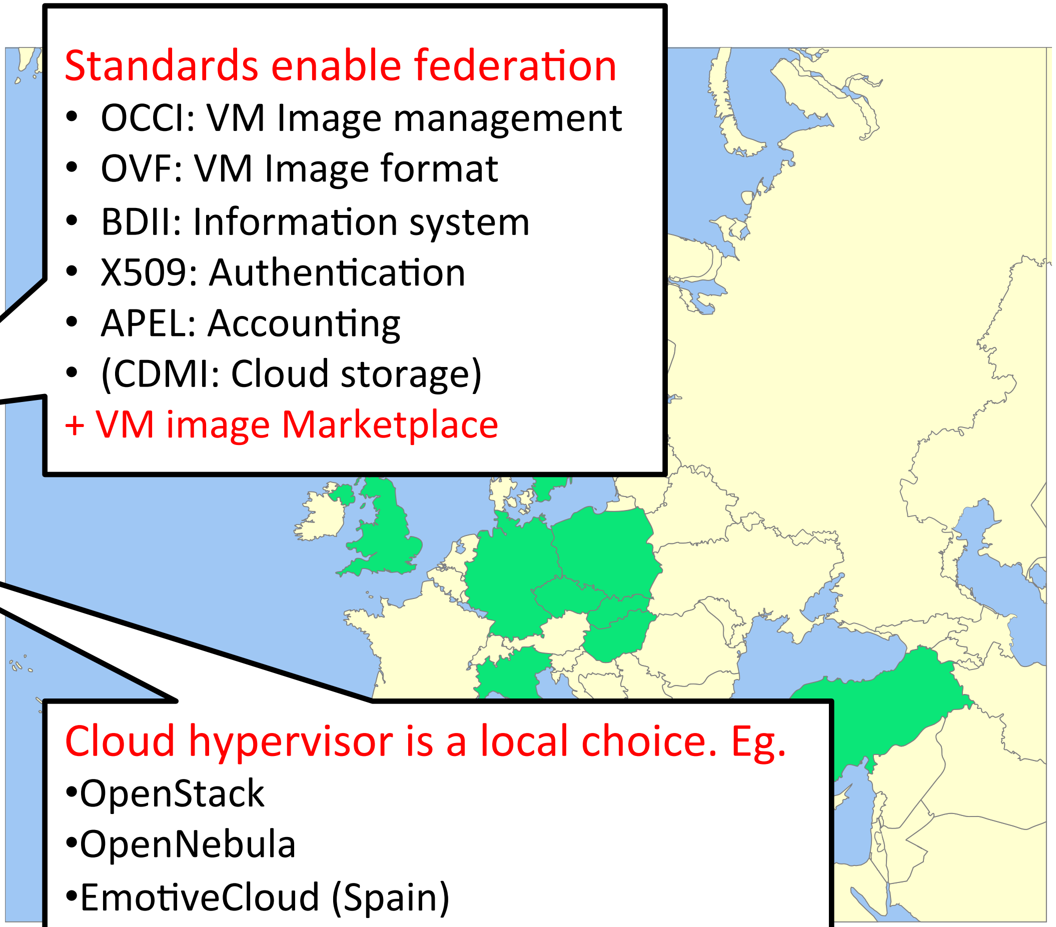
Domain specific services in Virtual Machine Images

EGI FedCloud interfaces

Cloud hypervisors

OS	OS	OS	OS	OS

Cloud resources
private/public
academic/commercial



Standards enable federation

- OCCI: VM Image management
- OVF: VM Image format
- BDII: Information system
- X509: Authentication
- APEL: Accounting
- (CDMI: Cloud storage)

+ VM image Marketplace

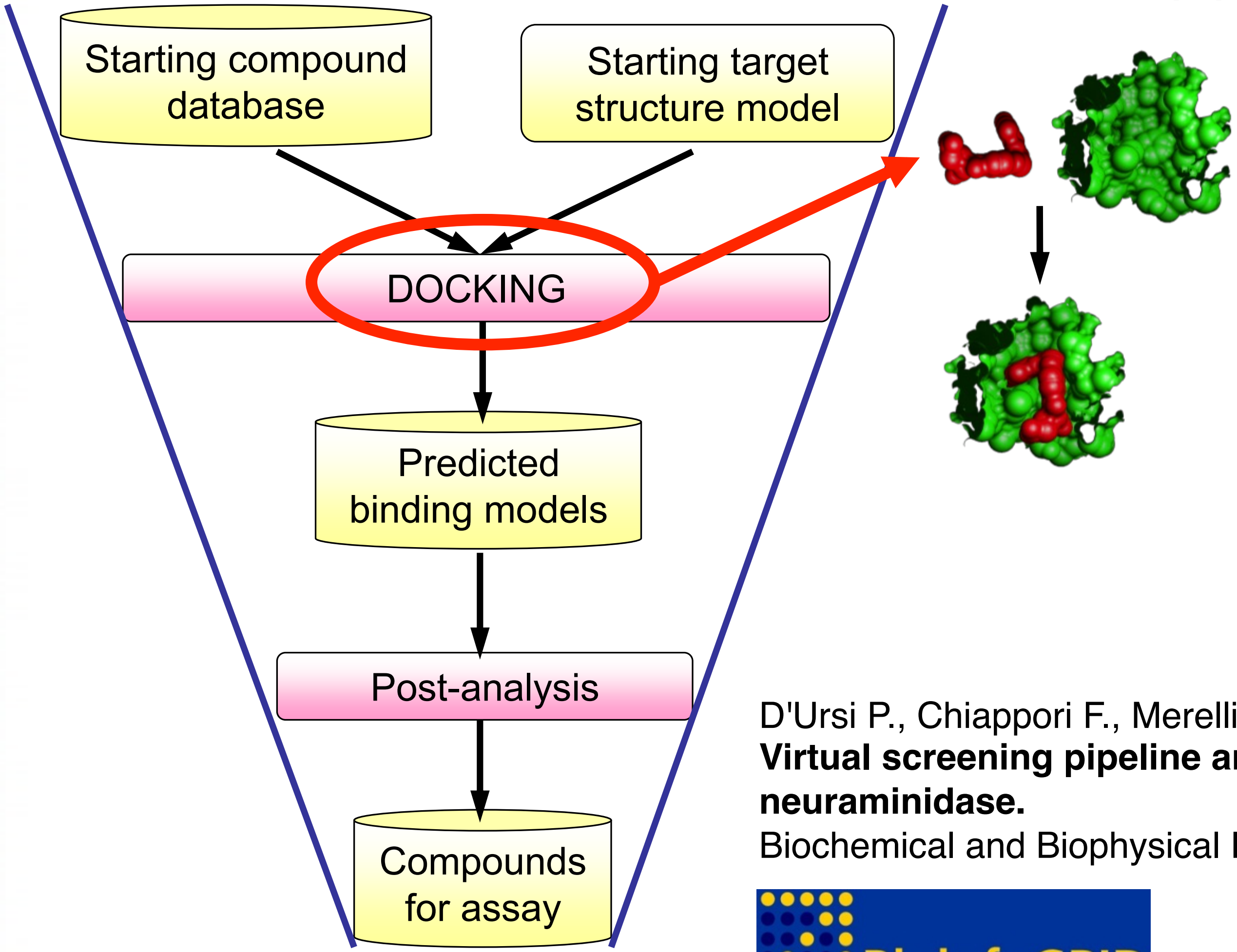
Cloud hypervisor is a local choice. Eg.

•OpenStack
•OpenNebula
•EmotiveCloud (Spain)
•Okeanos (OpenStack impl. in GR)
•WNoDeS (Italy)
•…

http://go.egi.eu/cloud

In Silico Drug Discovery

Millions of chemical compounds → 100 CPU years, 1 TB disk space


Starting compound database

Starting target structure model

DOCKING

Predicted binding models

Post-analysis

Compounds for assay

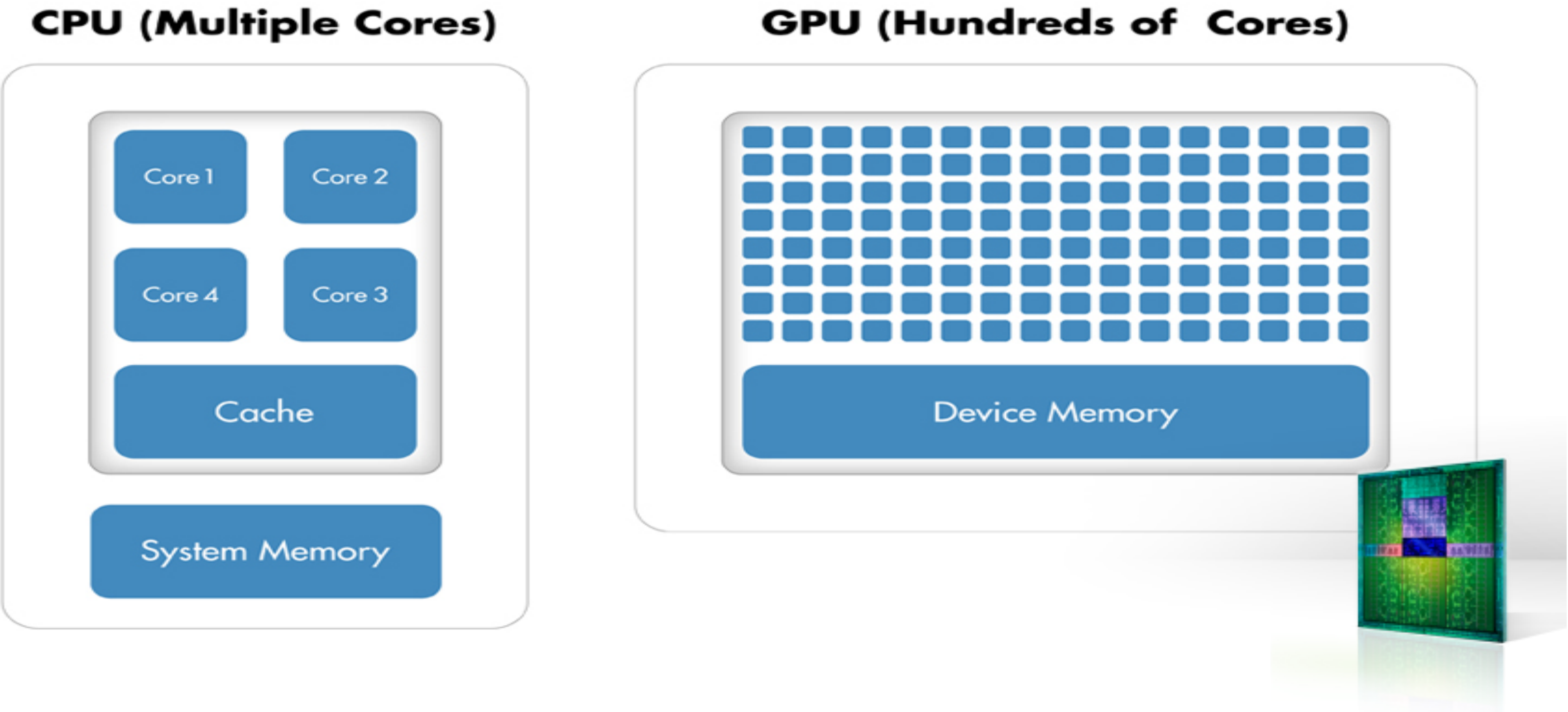
Docking: predict how small molecules bind to a receptor of known 3D structure

D'Ursi P., Chiappori F., Merelli I., Cozzi P., Rovida E., Milanesi L. **Virtual screening pipeline and ligand modelling for H5N1 neuraminidase.** Biochemical and Biophysical Research Communications. 2009

BioinfoGRID
Bioinformatics Grid Application for life science

GPU – Graphics Processing Unit

GPUs implement a SIMD (Single Instruction Multiple Data) many-core architecture, providing a very high level of parallelism on intense data-parallel computation problems.



GPU – Graphics Processing Unit

- GPU-based solution in bioinformatics for:
  - Sequence Database Searching
    - CUDASW++
  - Multiple Sequence Alignment
    - CUDA-BLASTP
  - Next-Generation Sequencing
    - DecGPU, CUDA-EC, Musket, SOAP3-dp, CUSHAW
  - Genome-Wide Association Studies
    - Mendel_GPU, GENIE, SWIFTLINK
  - Motif Finding
    - mCUDA-MEME

GPU – Graphics Processing Unit

- SNP genotyping analysis is very susceptible to SNPs chromosomal position errors;
- SNP mapping data are provided along the SNP arrays without information to assess in advance their accuracy;
- moreover, mapping data are related with a given build of a genome and need to be updated when a new build is available.

MIMOmics EU Project

- The aim of **MIMOmics** is to develop new statistical methods for the integrated analysis for metabolomics, proteomics, glycomics and genomic datasets in large studies.

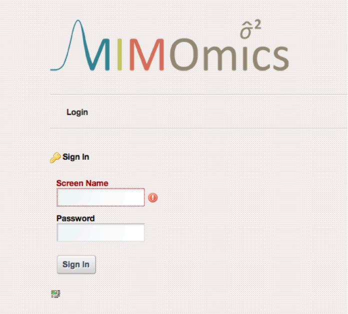
- Our partners are involvement involve in EU funded projects, i.e. **GEHA, IDEAL, Mark-Age, ENGAGE, EuroSpan, and BBMRI**

BBMRI
Biobanking and
Biomolecular
Resources Research
Infrastructure

- In these consortia the primary goal is to **identify molecular profiles that monitor and explain complex traits** with novel findings so far.

- **MIMOmics** web site http://www.mimomics.eu at CNR (Milan, Italy)

Omics Scientific Web Portal



MIMOmics

Login

Sign In

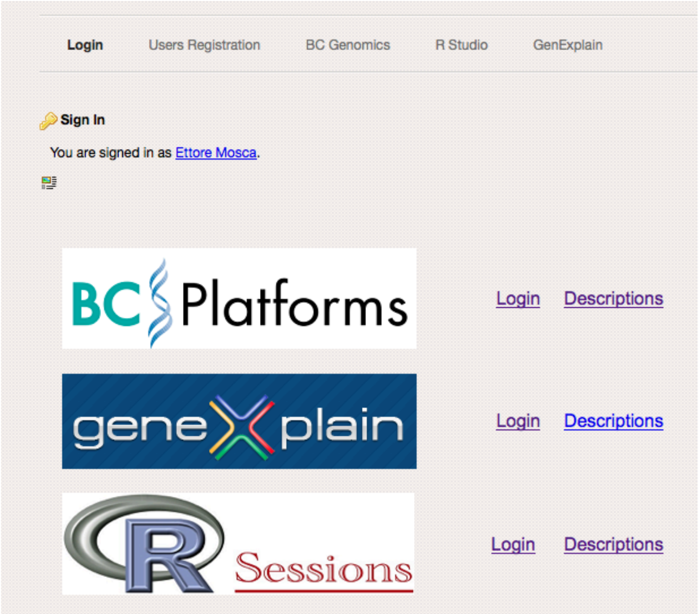
Screen Name

Password

Sign In

MIMOmics
authorized users

MIMOmics resources
(data sets and computational tools)



Login | Users Registration | BC Genomics | R Studio | GenExplain

Sign In

You are signed in as Ettore Mosca.

BC Platforms — Login Descriptions

geneXplain — Login Descriptions

R Sessions — Login Descriptions

Project Web Portal to:

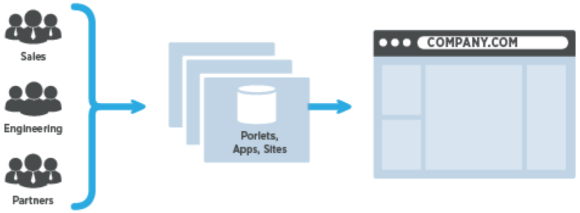
- create define the users credentials for all MIMOmics resources
- access MIMOmics resources
- develop, test and use tools on the data sets available
- create pipeline of analysis combining tools and data sets

Omics Scientific Web Portal

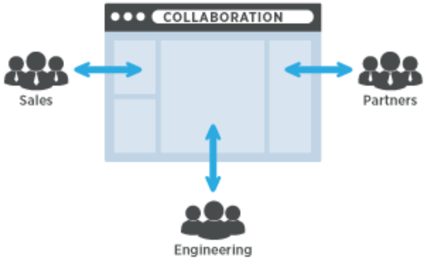
- The Omics Scientific Web Portal is based on **Liferay Portal tecnology**
- Liferay is a **robust** technology, fully supported in terms of **accessibility** and **scalability**
- Liferay provides a flexible template interface
- With Liferay the users can **manage contents** and documents in a distribuited and dinamic way over internet
- Liferay is compliabt with the **Java Portlet API 2.0**

WEB CMS
Case Study
Sales Doc
Marketing Doc

Documents Management

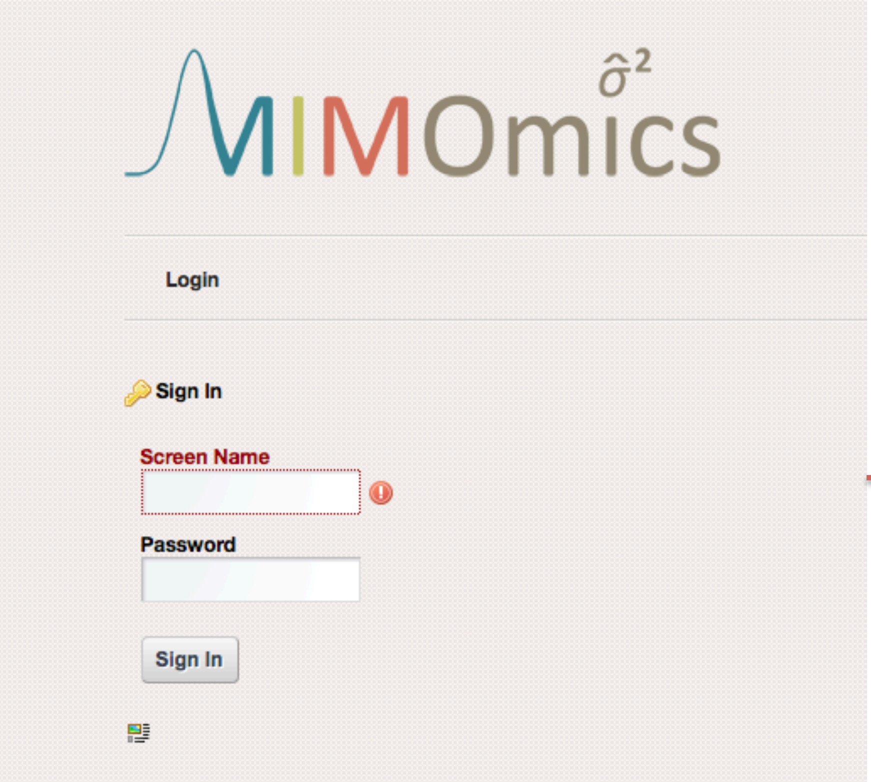


Web Editing



Collaboration, Services

User Registration



MIMOmics $\hat{\sigma}^2$

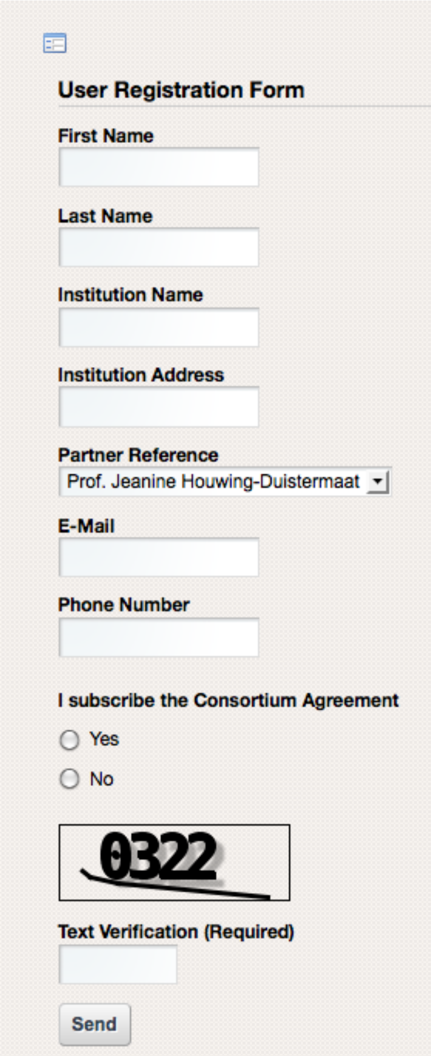
Login

Sign In

Screen Name

Password

Sign In



User Registration Form

First Name	
Last Name	
Institution Name	
Institution Address	
Partner Reference	Prof. Jeanine Houwing-Duistermaat
E-Mail	
Phone Number	

I subscribe the Consortium Agreement

○ Yes

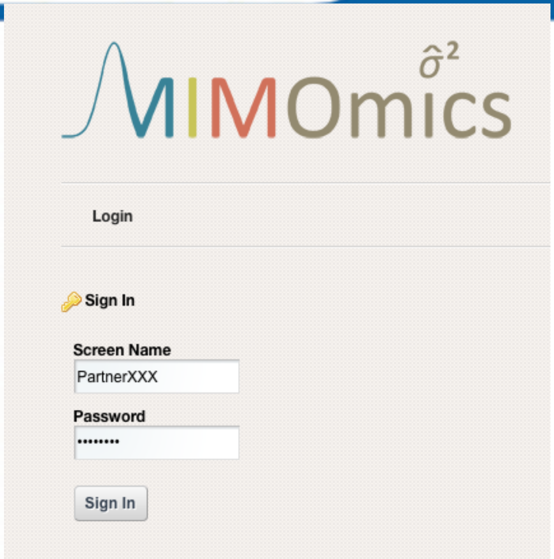
○ No

0322

Text Verification (Required)

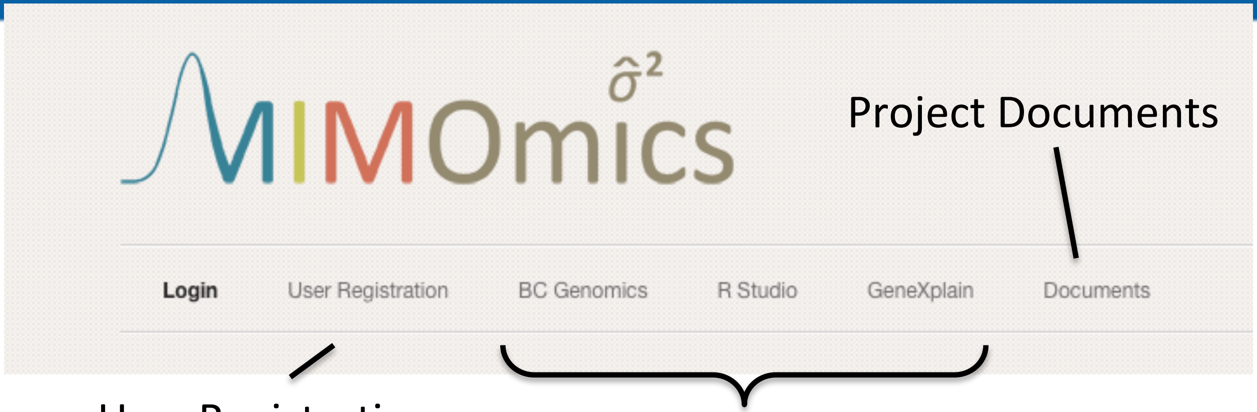
Send

Omics Scientific Web Portal



Screen Name: PartnerXXX

Password: ••••••••



Project Documents

User Registration

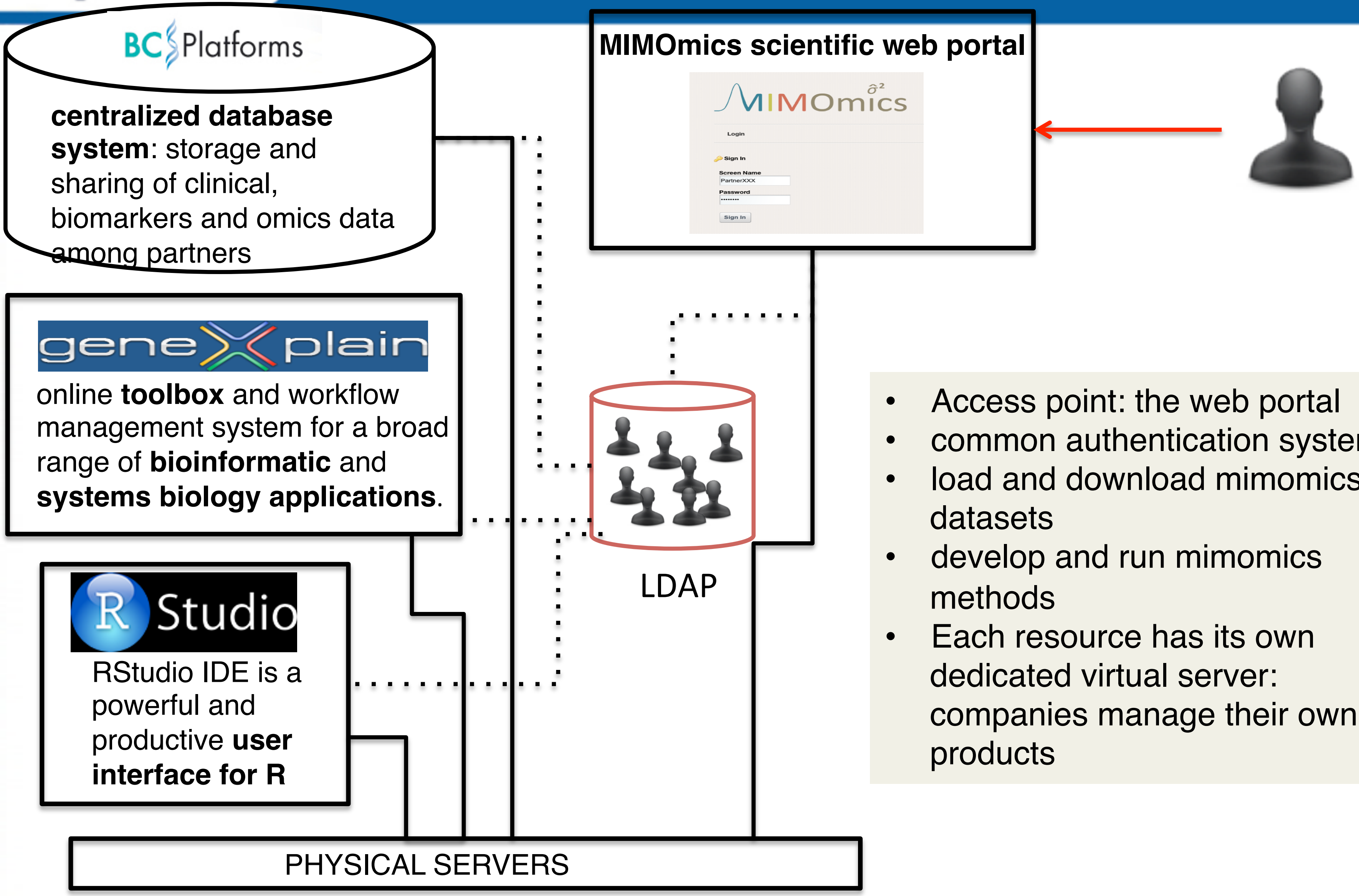
Link to MIMOmics resources

Omics scientific web portal:

- partner references can create new users with the same credentials for all MIMOmics resources
- access MIMOmics resources
- load and download MIMOmics datasets
- develop, test and use MIMOmics methods
- create pipeline of analysis combining tools and data sets

InterOmics CNR

Omics Scientific Web Portal



BC Platforms

centralized database system: storage and sharing of clinical, biomarkers and omics data among partners

genexplain

online **toolbox** and workflow management system for a broad range of **bioinformatic** and **systems biology applications**.

RStudio

RStudio IDE is a powerful and productive **user interface for R**

MIMOmics scientific web portal

MIMOmics

Login

Sign In

Screen Name

PartnerXXX

Password

Sign In

LDAP

PHYSICAL SERVERS

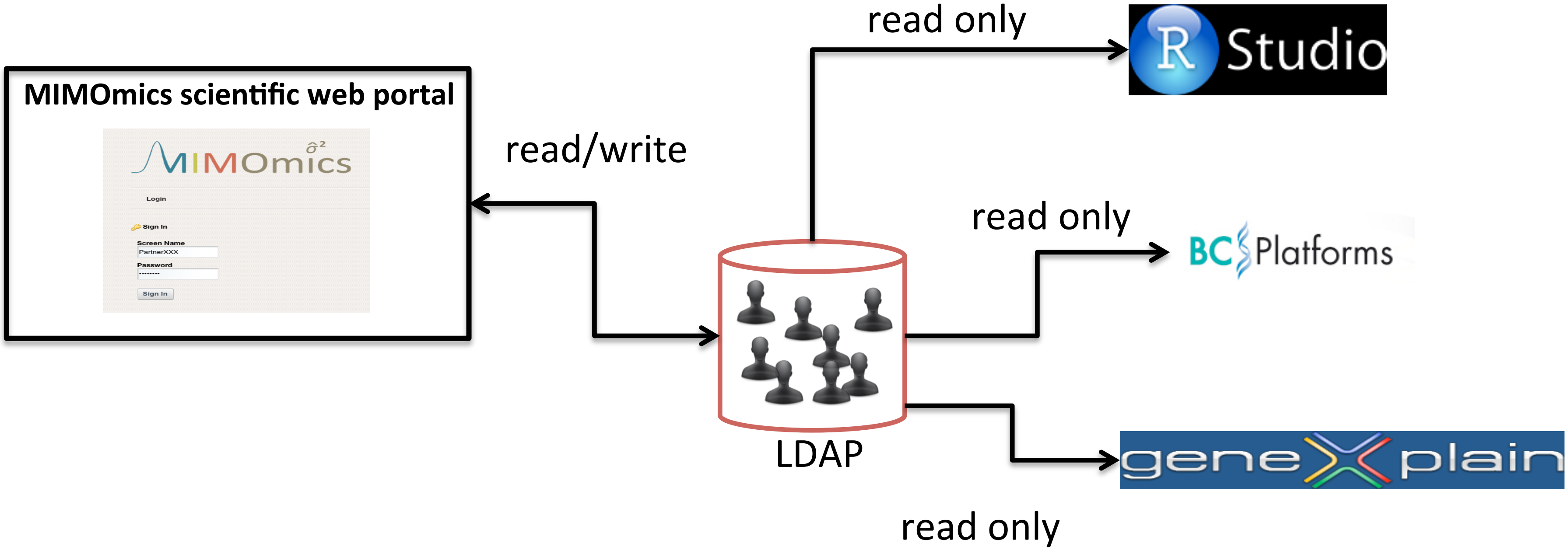
- Access point: the web portal
- common authentication system
- load and download mimomics datasets
- develop and run mimomics methods
- Each resource has its own dedicated virtual server: companies manage their own products

R packages in RStudio

- R packages available in RStudio server
  - core Bioconductor packages
  - R packages for multi-omics data analysis
    - **iCLuster**, a joint latent variable model for integrative clustering, (Shen et al., Bioinformatics, 2009)
    - **RISA**, converting experimental metadata from ISA-tab into Bioconductor data structures, (Gonzalez-Beltran et al., Bioconductor)
    - **OmicKriging**, Poly-Omic Prediction of Complex Traits, (Wheeler et al., 2013, arXiv:1303.1788)
    - *__ABEL,__ facilitate statistical analyses of polymorphic genomes data (Yurii Aulchenko)
    - **iNEMO,** integration of NEtworks with Multi-Omics (E. Mosca, L. Milanesi)

User management

Users are managed by the MIMOmics Scientific Web portal through the *Lightweight Directory Access Protocol* (LDAP).



MIMOmics scientific web portal

read/write

LDAP

read only

RStudio

read only

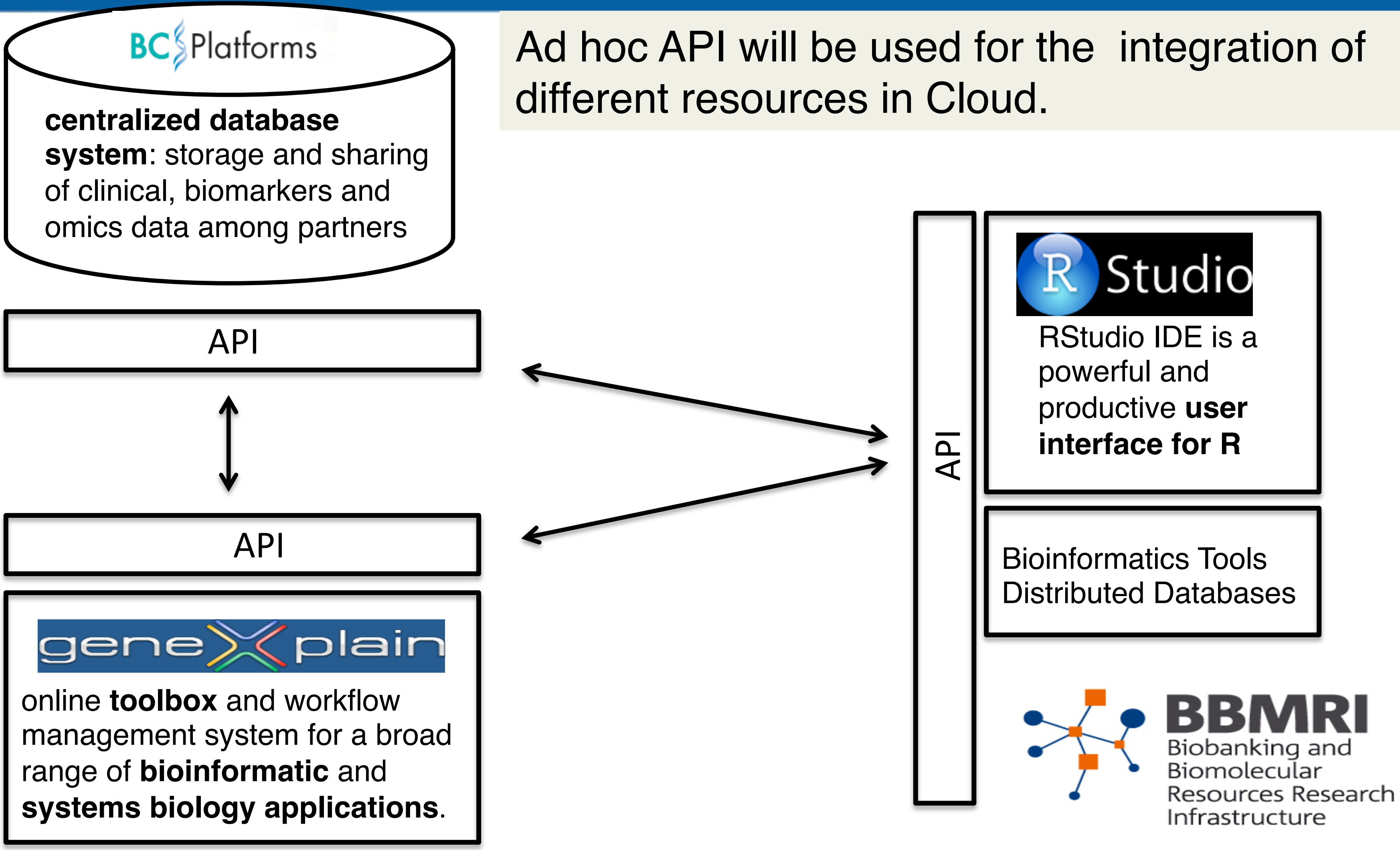
BC Platforms

read only

geneXplain

Omics Scientific Web Portal

Ad hoc API will be used for the integration of different resources in Cloud.



BC Platforms

centralized database system: storage and sharing of clinical, biomarkers and omics data among partners

API

API

geneXplain

online **toolbox** and workflow management system for a broad range of **bioinformatic** and **systems biology applications**.

API

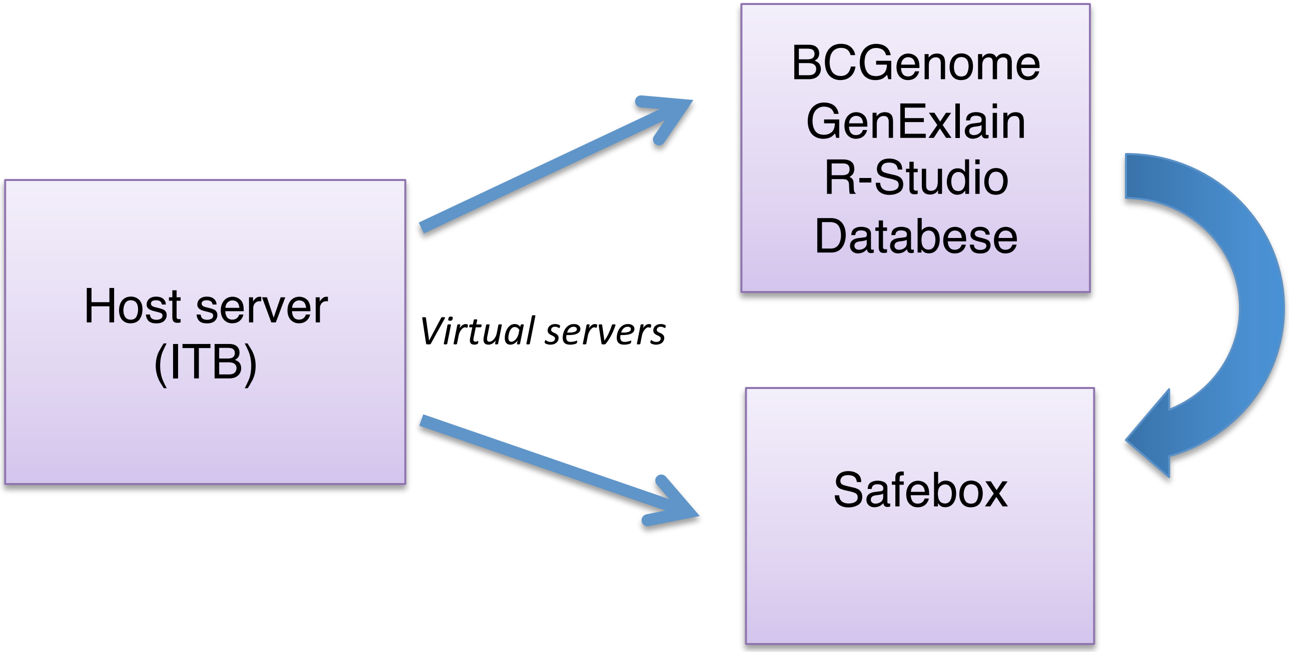
RStudio

RStudio IDE is a powerful and productive **user interface for R**

Bioinformatics Tools
Distributed Databases

BBMRI
Biobanking and Biomolecular Resources Research Infrastructure

Safebox set-up



Host server
(ITB)

Virtual servers

BCGenome
GenExlain
R-Studio
Databese

Safebox

Users can access read–only data Using Remote desktop protocol
User can execute the

Datasets, Studies, Biobanks

Several Omics Datasets:
Genomics, Glycomics, Proteomics, Metabolomics/Lipidomics

Several Studies:
Aging, Cancer, Isolated Populations studies, Multiple Sclerosis, Obesity and Metabolic sSyndrome

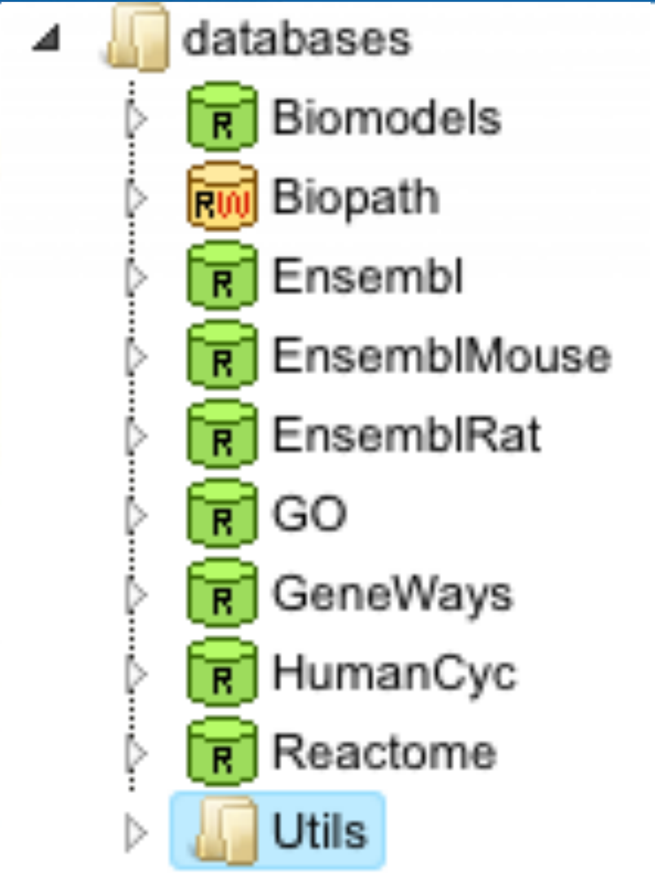
Biological Resource based on the BBMR standard Infrastructure:

BBMRI
Biobanking and
Biomolecular
Resources Research
Infrastructure

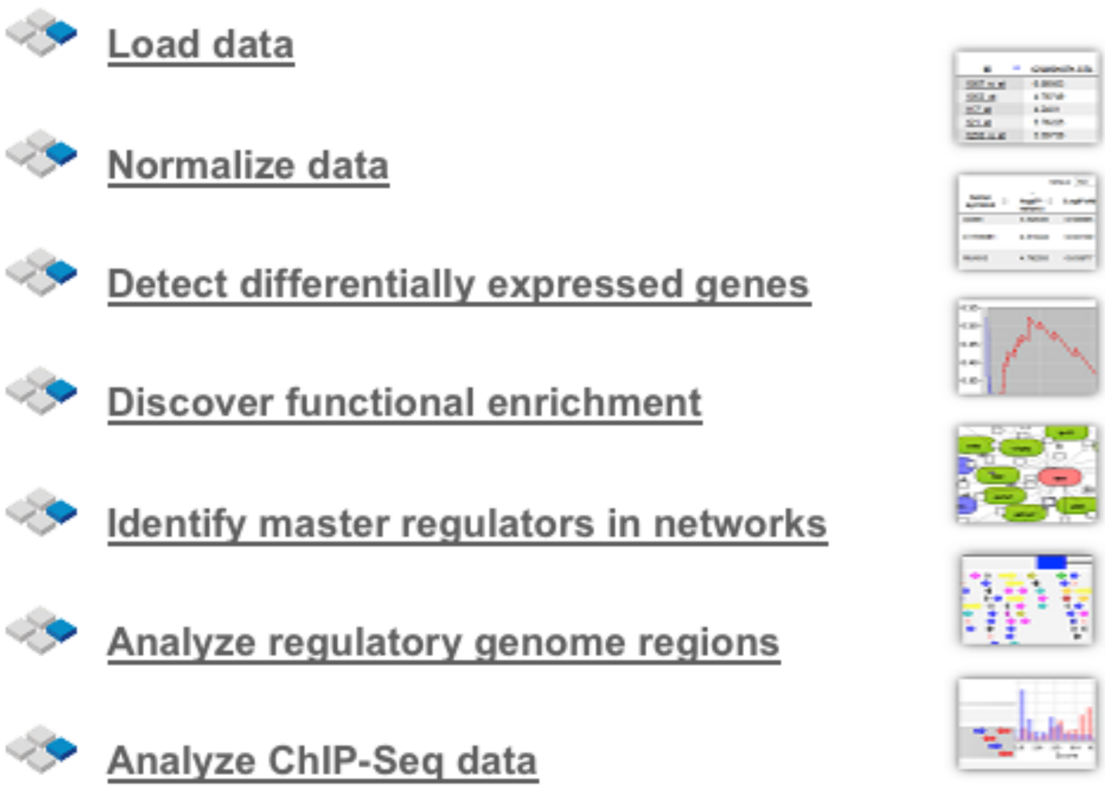
Tools

- **SAM Tools** provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.
- **The Genome Analysis Toolkit** or **GATK** is a software package developed at the Broad Institute to analyse next-generation resequencing data.
- **Granvil**: Gene- or Region-based ANalysis of Variants of Intermediate and Low frequency
- **Annovar**: Functional annotation of genetic variants from high-throughput sequencing data.
- **PLINK** is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.
- **IMPUTE** is a program for estimating ("imputing") unobserved genotypes in SNP association studies.

GeneXplain Data and Tools



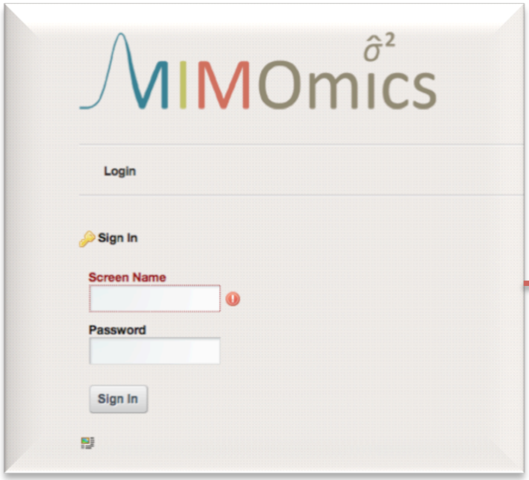
- databases
  - Biomodels
  - Biopath
  - Ensembl
  - EnsemblMouse
  - EnsemblRat
  - GO
  - GeneWays
  - HumanCyc
  - Reactome
  - Utils

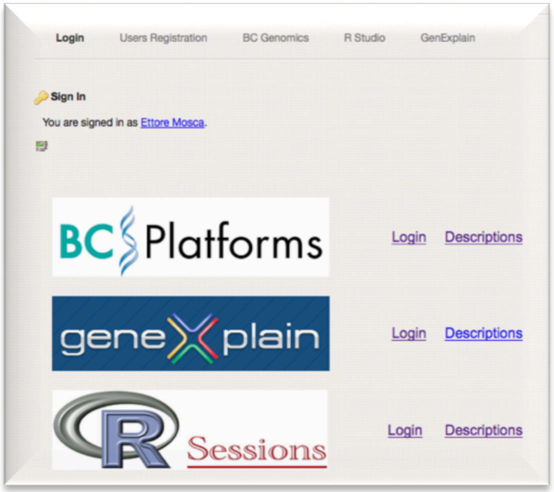


- Load data
- Normalize data
- Detect differentially expressed genes
- Discover functional enrichment
- Identify master regulators in networks
- Analyze regulatory genome regions
- Analyze ChIP-Seq data

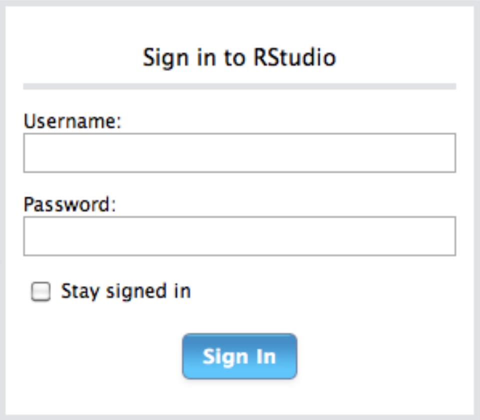
Some of data and analysis tools based on GeneXplain

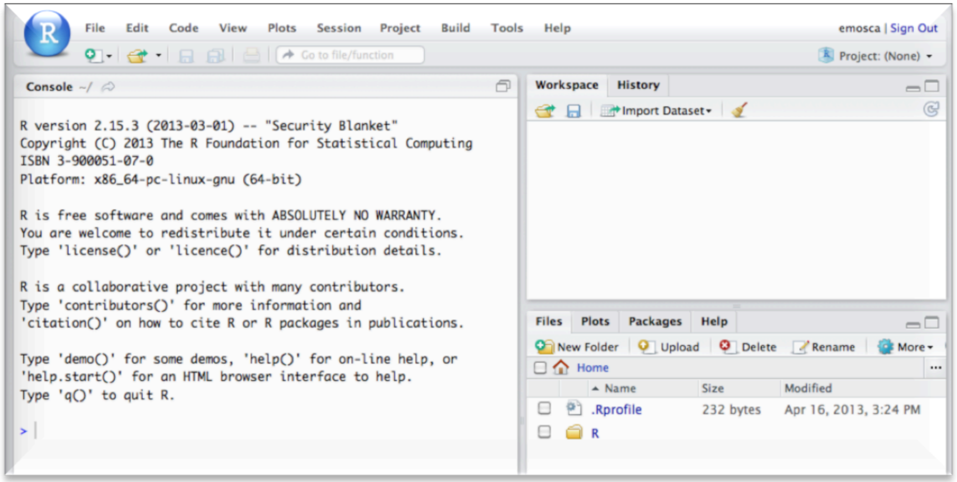
RStudio Server





You are signed in as Ettore Mosca.





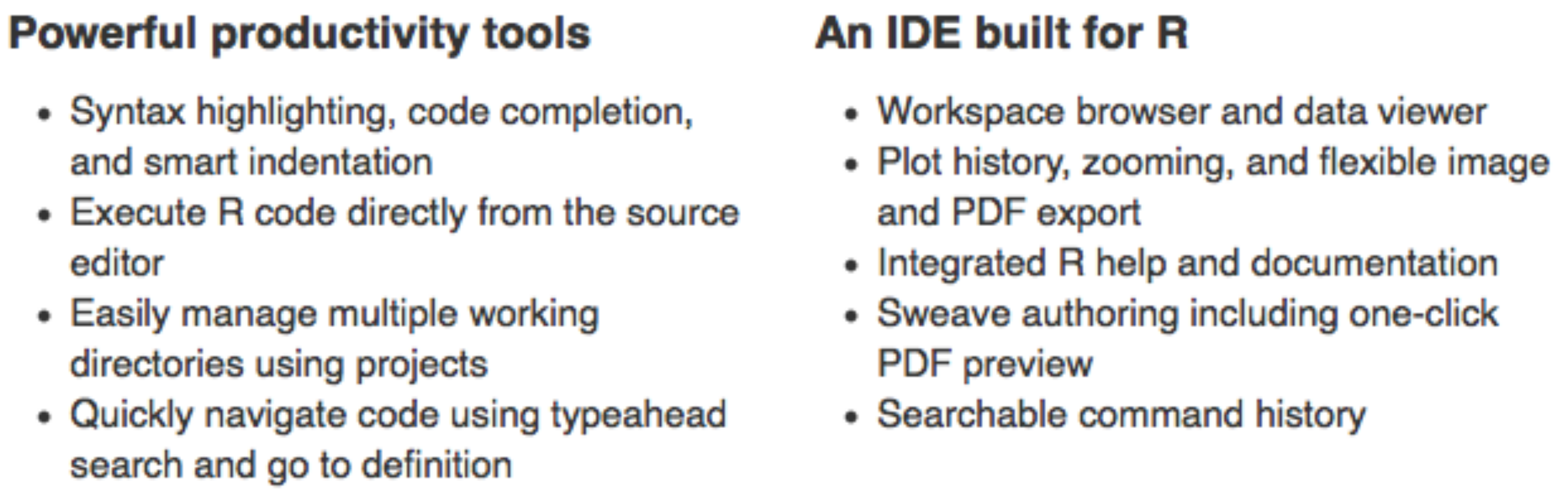
```
R version 2.15.3 (2013-03-01) -- "Security Blanket"
Copyright (C) 2013 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Name	Size	Modified
.Rprofile	232 bytes	Apr 16, 2013, 3:24 PM
R		

RStudio Virtual Server

- An instance of RStudio server has been installed and available for the MIMOmics users

- RStudio Integrated Development Environment is a powerful and productive user interface for R (http://www.rstudio.com/)



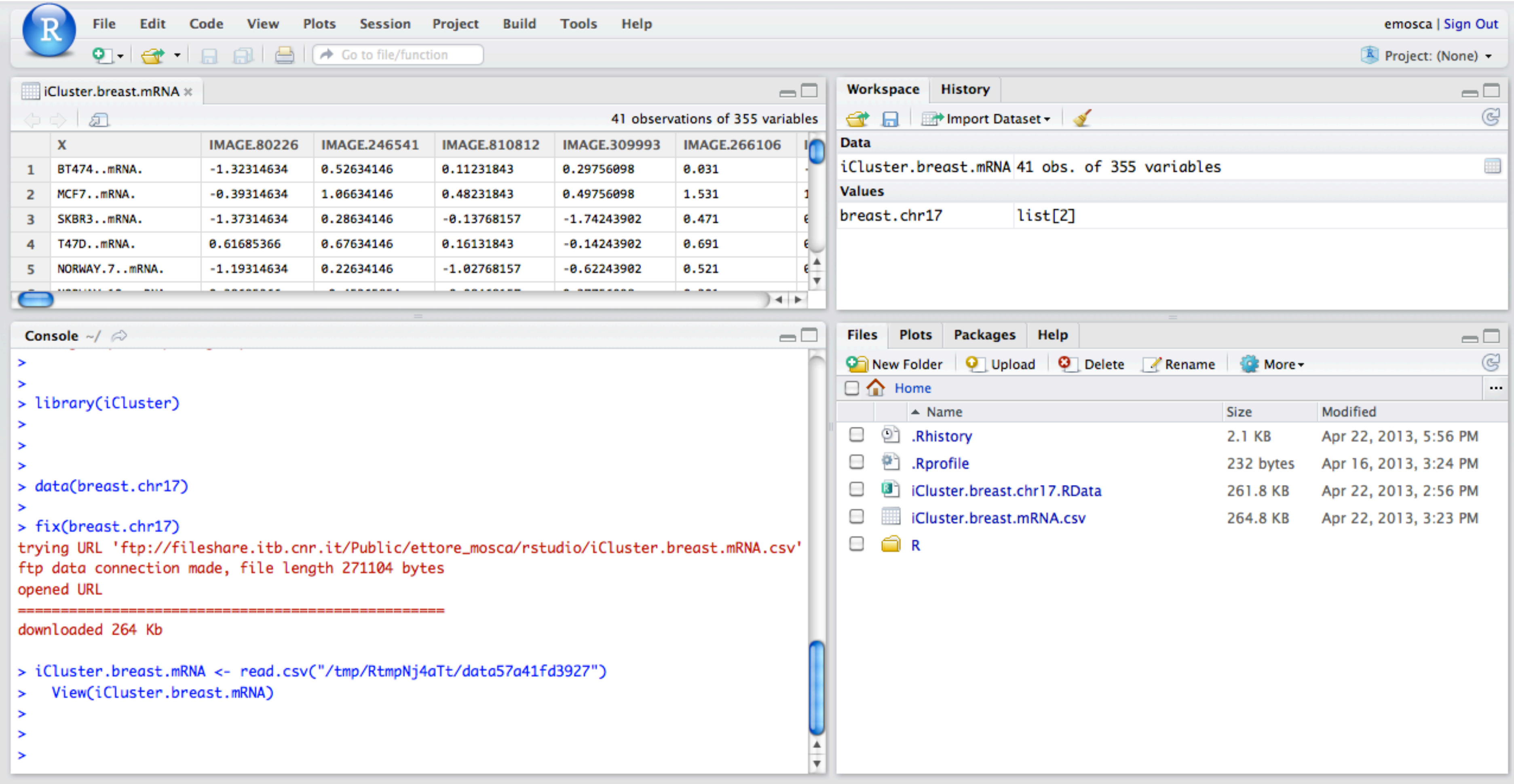
Powerful productivity tools

- Syntax highlighting, code completion, and smart indentation
- Execute R code directly from the source editor
- Easily manage multiple working directories using projects
- Quickly navigate code using typeahead search and go to definition

An IDE built for R

- Workspace browser and data viewer
- Plot history, zooming, and flexible image and PDF export
- Integrated R help and documentation
- Sweave authoring including one-click PDF preview
- Searchable command history

RStudio



41 observations of 355 variables

	X	IMAGE.80226	IMAGE.246541	IMAGE.810812	IMAGE.309993	IMAGE.266106
1	BT474..mRNA.	-1.32314634	0.52634146	0.11231843	0.29756098	0.031
2	MCF7..mRNA.	-0.39314634	1.06634146	0.48231843	0.49756098	1.531
3	SKBR3..mRNA.	-1.37314634	0.28634146	-0.13768157	-1.74243902	0.471
4	T47D..mRNA.	0.61685366	0.67634146	0.16131843	-0.14243902	0.691
5	NORWAY.7..mRNA.	-1.19314634	0.22634146	-1.02768157	-0.62243902	0.521

Data

iCluster.breast.mRNA	41 obs. of 355 variables

Values

breast.chr17	list[2]

```
>
>
> library(iCluster)
>
>
>
> data(breast.chr17)
>
> fix(breast.chr17)
trying URL 'ftp://fileshare.itb.cnr.it/Public/ettore_mosca/rstudio/iCluster.breast.mRNA.csv'
ftp data connection made, file length 271104 bytes
opened URL
==================================================
downloaded 264 Kb

> iCluster.breast.mRNA <- read.csv("/tmp/RtmpNj4aTt/data57a41fd3927")
>   View(iCluster.breast.mRNA)
>
>
>
```

	Name	Size	Modified
□	.Rhistory	2.1 KB	Apr 22, 2013, 5:56 PM
□	.Rprofile	232 bytes	Apr 16, 2013, 3:24 PM
□	iCluster.breast.chr17.RData	261.8 KB	Apr 22, 2013, 2:56 PM
□	iCluster.breast.mRNA.csv	264.8 KB	Apr 22, 2013, 3:23 PM
□	R		

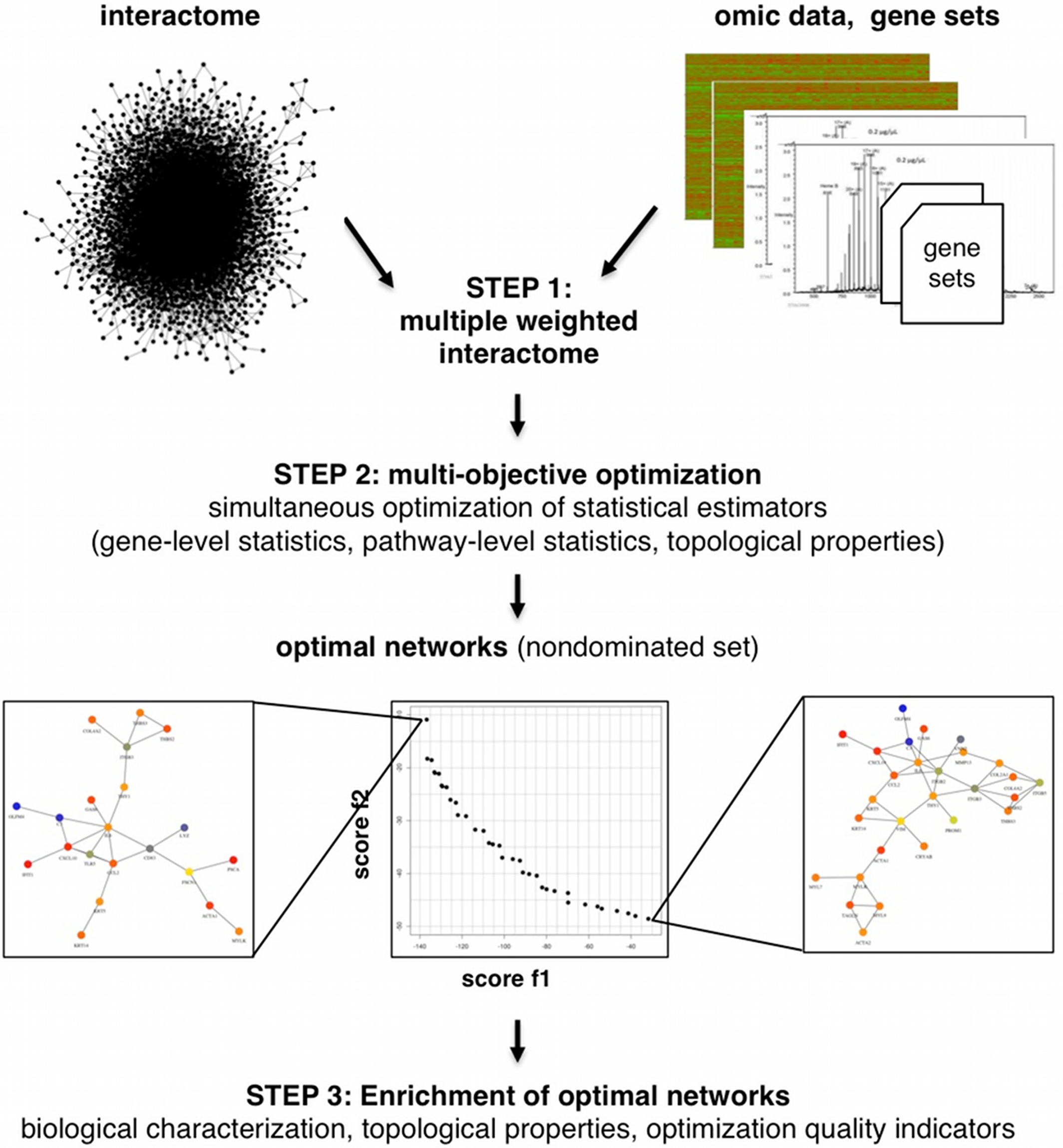
R based packages

Examples of R packages for multi-omic data analysis:

- from the literature
  - **iCLuster**, a joint latent variable model for integrative clustering, (Shen et al., Bioinformatics, 2009)
  - **RISA**, converting experimental metadata from ISA-tab into Bioconductor data structures, (Gonzalez-Beltran et al., Bioconductor)
  - **OmicKriging**, Poly-Omic Prediction of Complex Traits, (Wheeler et al., 2013, arXiv:1303.1788)
  - **piano**, Platform for integrative analysis of omics data (Varemo, et al., 2013, NAR)

- from MIMOmics parters
  - *__ABEL (GenABLE, OmicABLE, ProbABLE, …)__ facilitate statistical analyses of polymorphic genomes data (Yurii Aulchenko)
  - **network-based integration of omics** (Mosca E, Milanesi L, *et al*. submitted)
  - Ecc.

New Tools: Network based integration of Omic



interactome

omic data, gene sets

gene sets

STEP 1:
multiple weighted
interactome

STEP 2: multi-objective optimization
simultaneous optimization of statistical estimators
(gene-level statistics, pathway-level statistics, topological properties)

optimal networks (nondominated set)

STEP 3: Enrichment of optimal networks
biological characterization, topological properties, optimization quality indicators

Integrating omic data:

- Analyze the biological components and their interactions,
- Define a multiple-weighted **network**
- Find the **optimal modules** on the basis of the simultaneous optimization of **several statistical estimators**

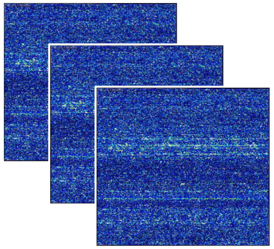
Mosca E, Milanesi L, *et al.*

InterOmics CNR

Application: HCV and Hepatocellular Carcinoma

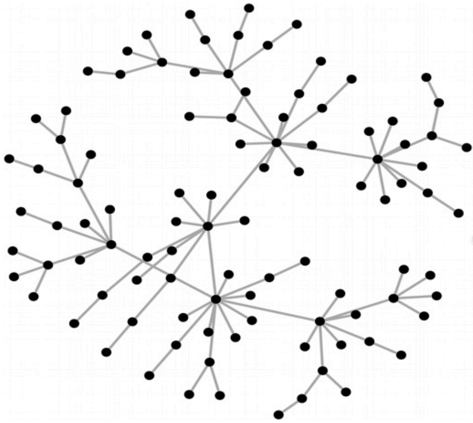
HCV and Host protein-protein interactions

Expression data of stepwise hepatocarcinogenic process



GSE6764 (Geo Database)
Affymetrix HG-U133A
75 tissue samples

Normal, Cirrhosis,
Dysplasia, Hepatocellular carcinoma



HCV – Host interactome with multiple transcriptomic data

HCV - Host PPI

Kwofie SK et al. Infect Genet Evol 2011
DeChassey B et al. Mol Syst Biol. 2008

Tot: 542 HCV- Host interactions

Host PPI

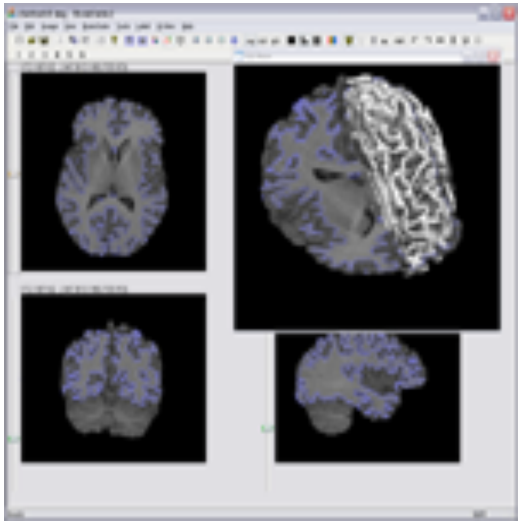
Franceschini A et al. Nucl Acid Res 2013
STRING v9.1

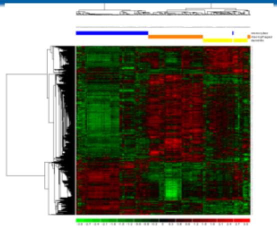
Tot: ~224000 human interactions

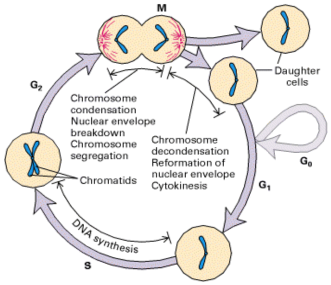
OBJECTIVE

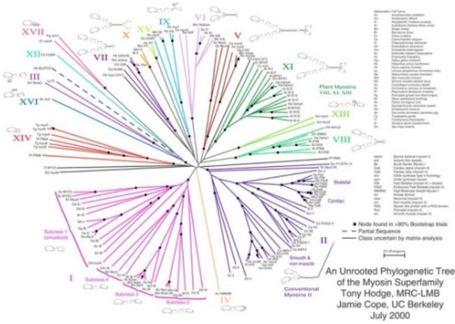
Identification of subnetworks enriched in differentially expressed genes and HCV-host protein-protein interactions

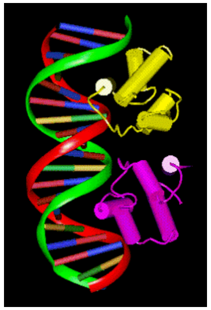
Precision Medicine

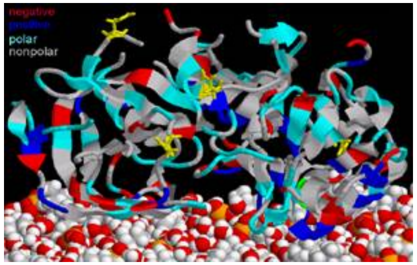







XVII
XII
III
XVI
XIV
VII
X
XV
IX
VI
V
XI
Plant Myosins VIII, XI, XIII
XIII
VIII
II
I
IV
Node found in >90% Bootstrap trials
Partial Sequence
Class uncertain by matrix analysis
An Unrooted Phylogenetic Tree of the Myosin Superfamily
Tony Hodge, MRC-LMB
Jamie Cope, UC Berkeley
July 2000





Big Data : Personalised medicine

- **Personalised medicine** will require sequencing of the genomes of large numbers of patients and volunteers
- It will be necessary to compare at least some of these genomes with the reference data collections
- Most hospitals and clinical research institutes will not wish to maintain up-to-date copies of the reference data collections
- It will be therefore be necessary to send these genomes to the institutes that hold the reference data collections
- It seems likely that this will be achieved using **secure VMs and secure clouds** holding the reference data collections
- EMBL-EBI is engaging with stakeholders to evaluate opportunities in this area.

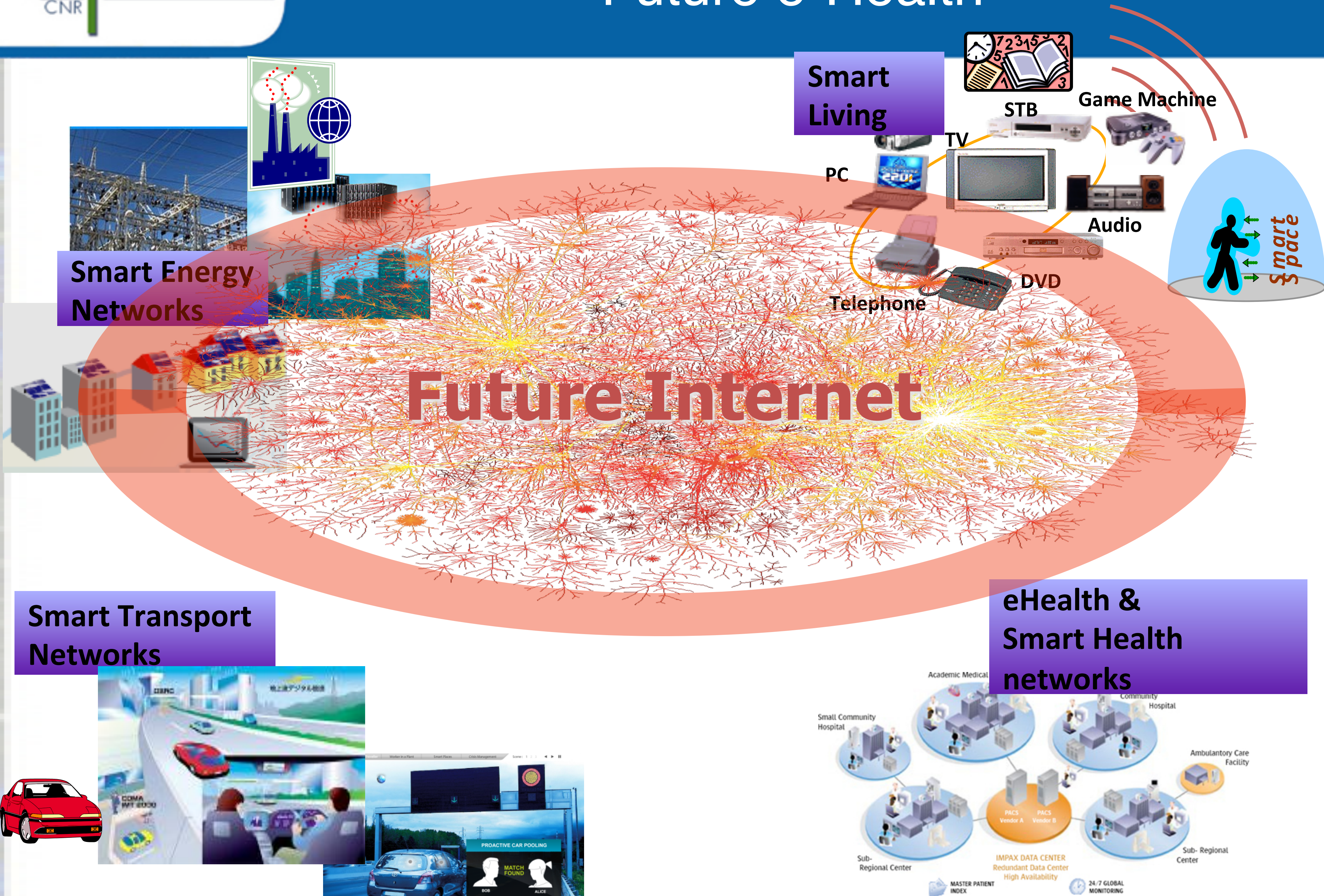
Future e-Health

Smart Living

Smart Energy Networks

Smart Transport Networks

eHealth & Smart Health networks



Conclusions

- The use of **Big Data** and the **Omics technologies** will improve the research for the future **personalized system medicine** since the disease phenotypes arise from complex interactions among genetic factors and environment.

- The use of public's bioinformatics resources data center in connection with specialized **BioBanks** will be progressively used for **large-scale population biomarker discovery** and validation by integrating clinical and genetic databases and providing an integrated access to this huge amount of information.

- A range of new applications in biomedical data mining based on **Cloud Computing** are in fast development.

HandsOn:Biobanks 2015

Local organizing committee:

M. Lavitrano, E. Bravo, MG Daidone, R. Lawlor, L. Milanesi, B. Parodi, D. Pistillo, G. Stanta.

Acknowledgments

BioinfoGRID
Bioinformatics Grid Application for life science

REPVBBLICA ITALIANA
Ministero dell'Istruzione dell'Università e Ricerca

InterOmics
Flagship Project

Consiglio Nazionale delle Ricerche

LITBIO

ItalBioNet

egee
Enabling Grids for E-sciencE

elixir

MIMOmics

BBMRI-IT
Biobanking and BioMolecular Resources Research Infrastructure Italy

HIRMA
Hepatocarcinoma
Innovative
Research
MArkers