Data Analysis and Visualization with R

Carlos Jaime Barrios Hernández, PhD

Gilberto Javier Diaz Toro, MSc

MAESTRIA EN MICROBIOLOGIA 2017

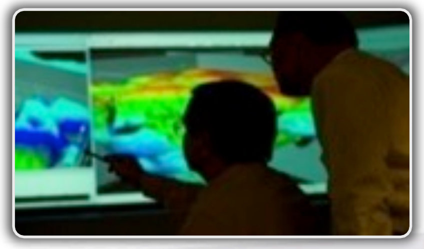
HPC Data Reduction

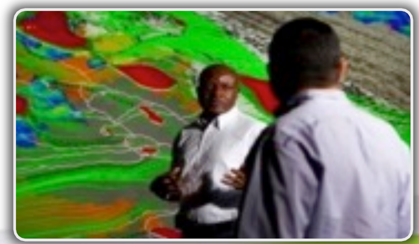
Big Data + Large Size Data = Great Problem

"… In the last 5 years, ONLY the astronomy research has produce more of 200PBytes of data/day… we have a lot of data to process… more than all the history of the humanity…"

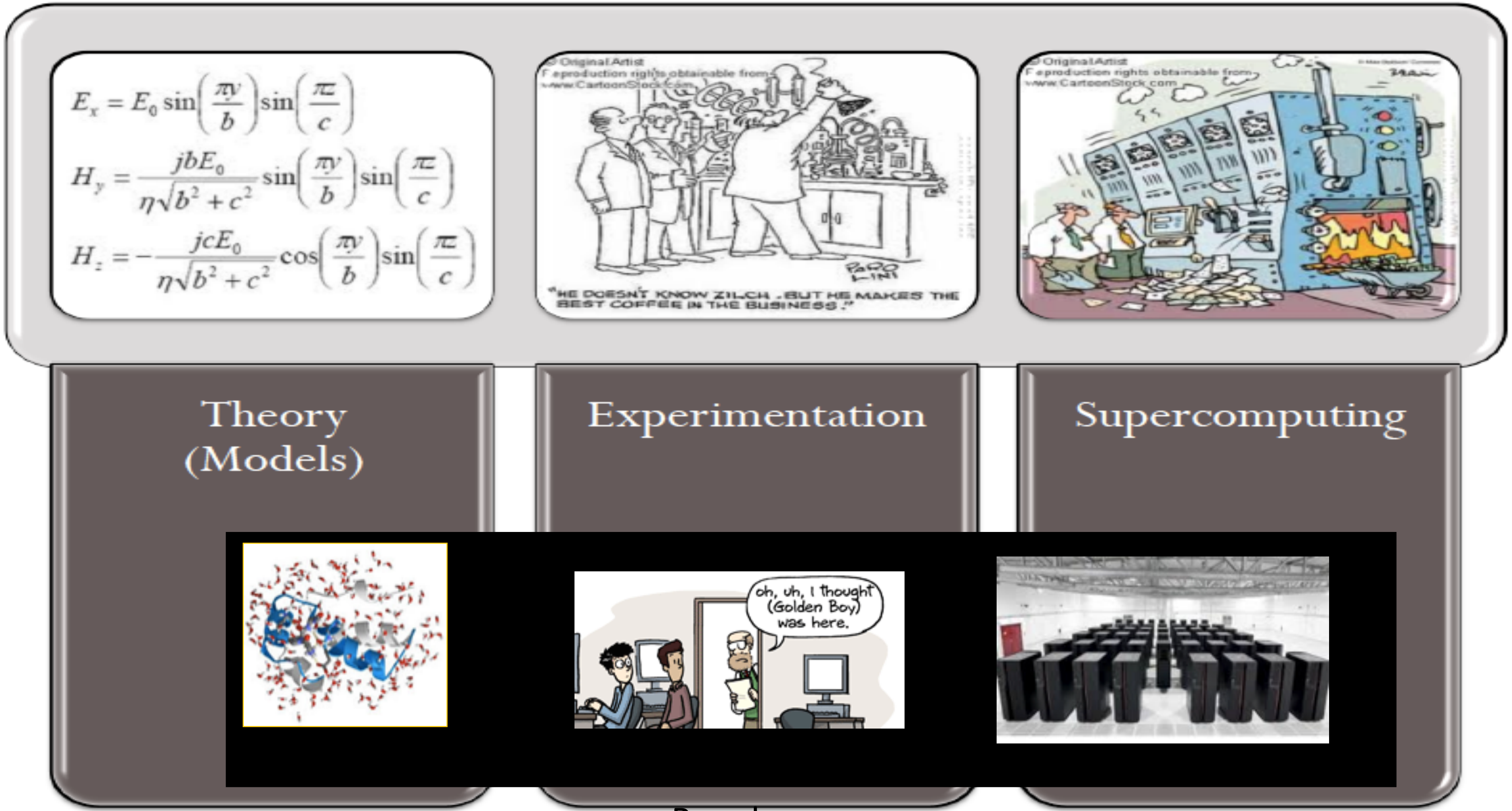
"… aja y esto es una oportunidad".

- Large Data Sets
- Data Reduction
- Collaboration
- High Resolution
- Real Time (almost)





Three Pillars and Three Actors



$$E_x = E_0 \sin\left(\frac{\pi y}{b}\right)\sin\left(\frac{\pi z}{c}\right)$$

$$H_y = \frac{jbE_0}{\eta\sqrt{b^2+c^2}}\sin\left(\frac{\pi y}{b}\right)\sin\left(\frac{\pi z}{c}\right)$$

$$H_z = -\frac{jcE_0}{\eta\sqrt{b^2+c^2}}\cos\left(\frac{\pi y}{b}\right)\sin\left(\frac{\pi z}{c}\right)$$

© Original Artist
Reproduction rights obtainable from www.CartoonStock.com
"HE DOESN'T KNOW ZILCH, BUT HE MAKES THE BEST COFFEE IN THE BUSINESS."
© Original Artist
Reproduction rights obtainable from www.CartoonStock.com
oh, uh, I thought (Golden Boy) was here.

Theory (Models)

Experimentation

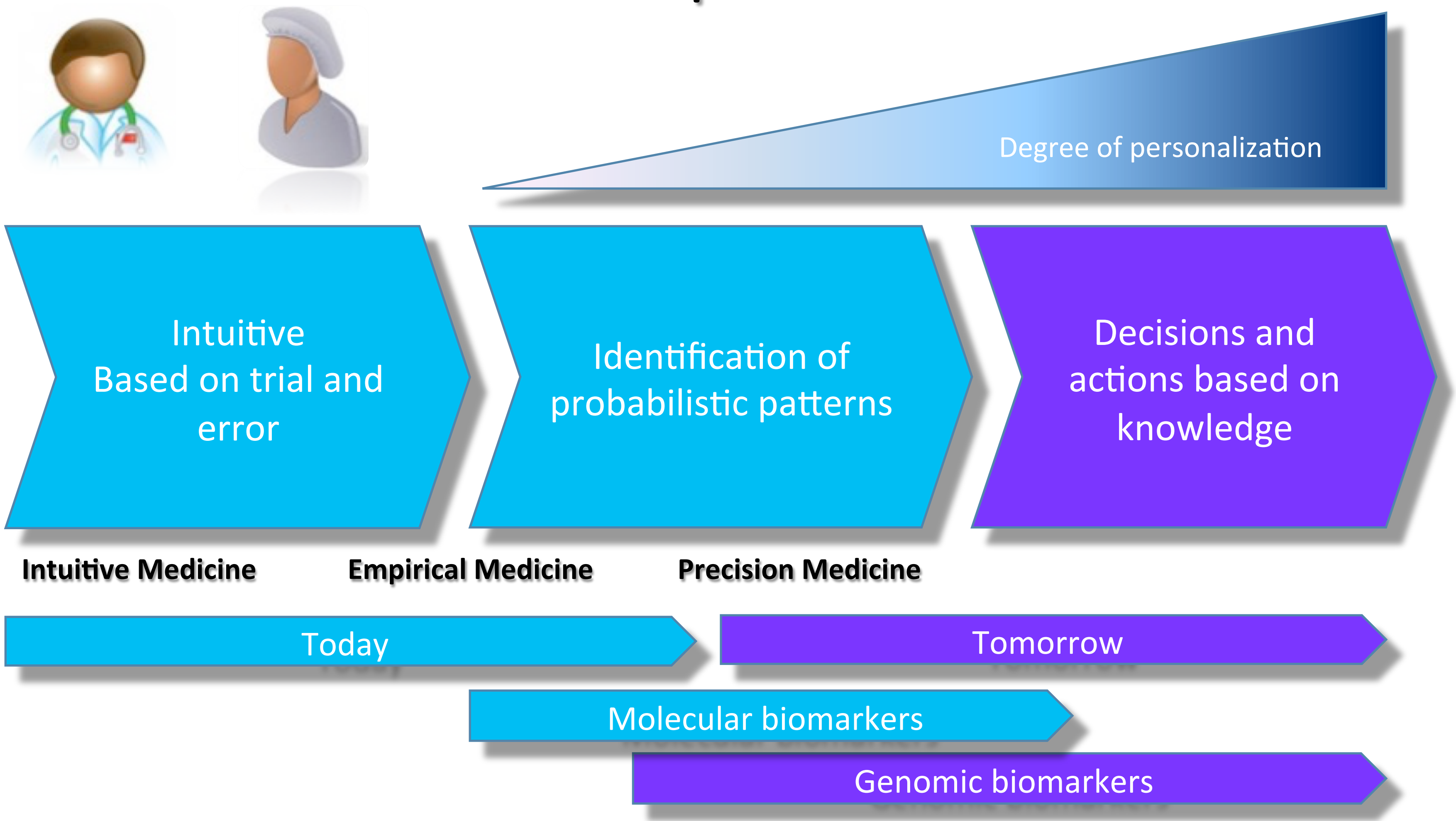
Supercomputing

Problems

People

Infrastructure / Tools

The transition to precision medicine



Degree of personalization

Intuitive
Based on trial and error

Identification of probabilistic patterns

Decisions and actions based on knowledge

Intuitive Medicine **Empirical Medicine** **Precision Medicine**

Today

Tomorrow

Molecular biomarkers

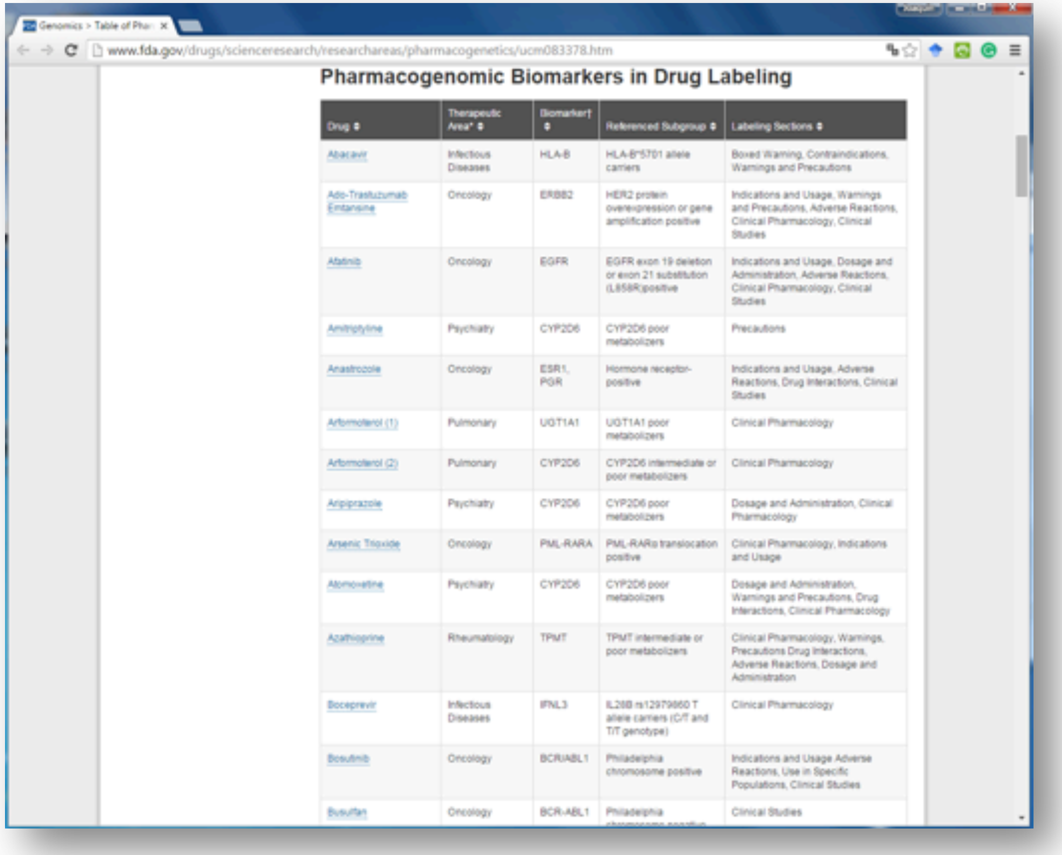
Genomic biomarkers

Precision medicine is based on a better knowledge of phenotype-genotype relationships. That is the knowledge od **disease** and **drug action mechanisms**

With the introduction of **molecular biomarkers** we are living now the **transition** from **intuitive** to **empirical** medicine

From Dopazo

And how do we identify patterns?
Using single-gene biomarkers


Genomics > Table of Pha...
www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm

Pharmacogenomic Biomarkers in Drug Labeling

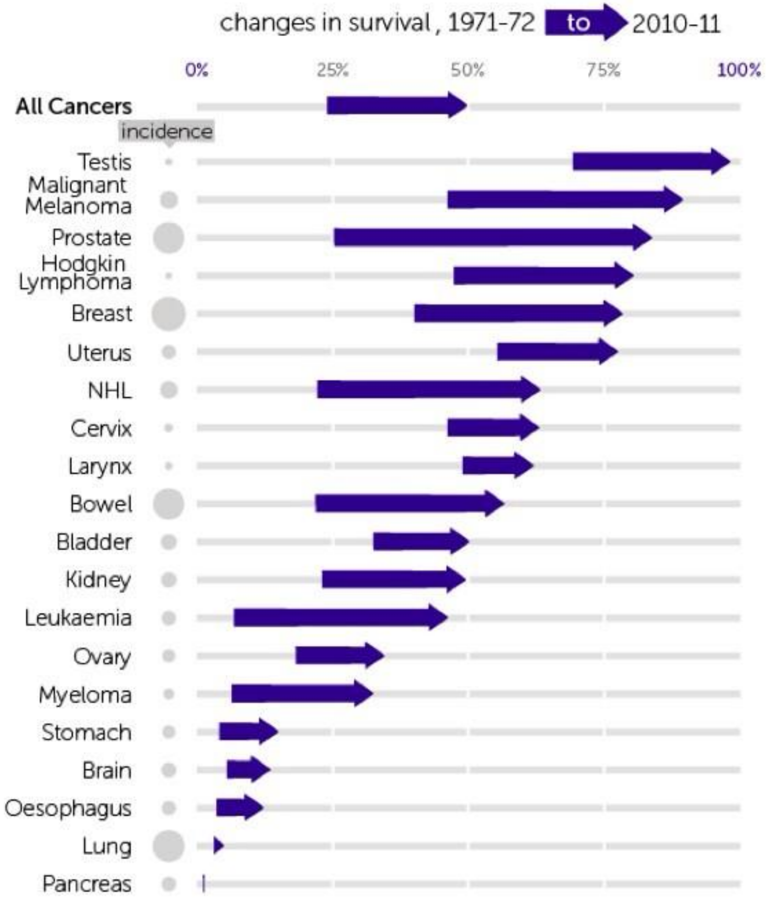
Drug ⇕	Therapeutic Area* ⇕	Biomarker† ⇕	Referenced Subgroup ⇕	Labeling Sections ⇕
Abacavir	Infectious Diseases	HLA-B	HLA-B*5701 allele carriers	Boxed Warning, Contraindications, Warnings and Precautions
Ado-Trastuzumab Emtansine	Oncology	ERBB2	HER2 protein overexpression or gene amplification positive	Indications and Usage, Warnings and Precautions, Adverse Reactions, Clinical Pharmacology, Clinical Studies
Afatinib	Oncology	EGFR	EGFR exon 19 deletion or exon 21 substitution (L858R)positive	Indications and Usage, Dosage and Administration, Adverse Reactions, Clinical Pharmacology, Clinical Studies
Amitriptyline	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Precautions
Anastrozole	Oncology	ESR1, PGR	Hormone receptor-positive	Indications and Usage, Adverse Reactions, Drug Interactions, Clinical Studies
Arformoterol (1)	Pulmonary	UGT1A1	UGT1A1 poor metabolizers	Clinical Pharmacology
Arformoterol (2)	Pulmonary	CYP2D6	CYP2D6 intermediate or poor metabolizers	Clinical Pharmacology
Aripiprazole	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Dosage and Administration, Clinical Pharmacology
Arsenic Trioxide	Oncology	PML-RARA	PML-RARα translocation positive	Clinical Pharmacology, Indications and Usage
Atomoxetine	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Dosage and Administration, Warnings and Precautions, Drug Interactions, Clinical Pharmacology
Azathioprine	Rheumatology	TPMT	TPMT intermediate or poor metabolizers	Clinical Pharmacology, Warnings, Precautions Drug Interactions, Adverse Reactions, Dosage and Administration
Boceprevir	Infectious Diseases	IFNL3	IL28B rs12979860 T allele carriers (C/T and T/T genotype)	Clinical Pharmacology
Bosutinib	Oncology	BCR/ABL1	Philadelphia chromosome positive	Indications and Usage Adverse Reactions, Use in Specific Populations, Clinical Studies
Busulfan	Oncology	BCR-ABL1	Philadelphia chromosome positive	Clinical Studies

http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm

Most “personalized” therapies are based on this type of biomarkers

From Dopazo

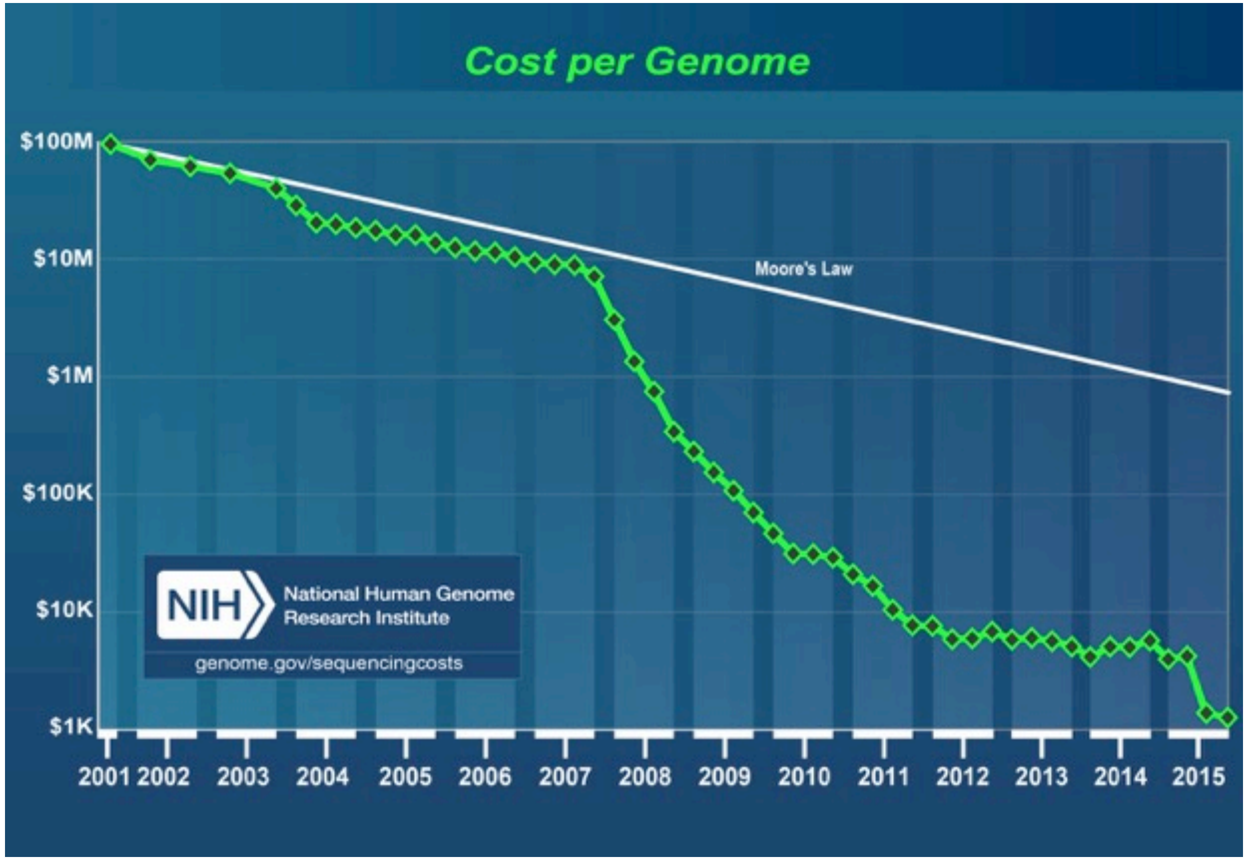
In spite of its simplicity, Empirical medicine based on biomarkers really works: Increasingly personalized treatments increase patient survival



From Dopazo

New sequencing technologies change the rules of the game.

http://www.genome.gov/sequencingcosts/



With exome sequencing costs ranging 600-800€ and clinical panels below 300€ the use of **NGS for routine diagnosis** matches the price of many other clinical tests.

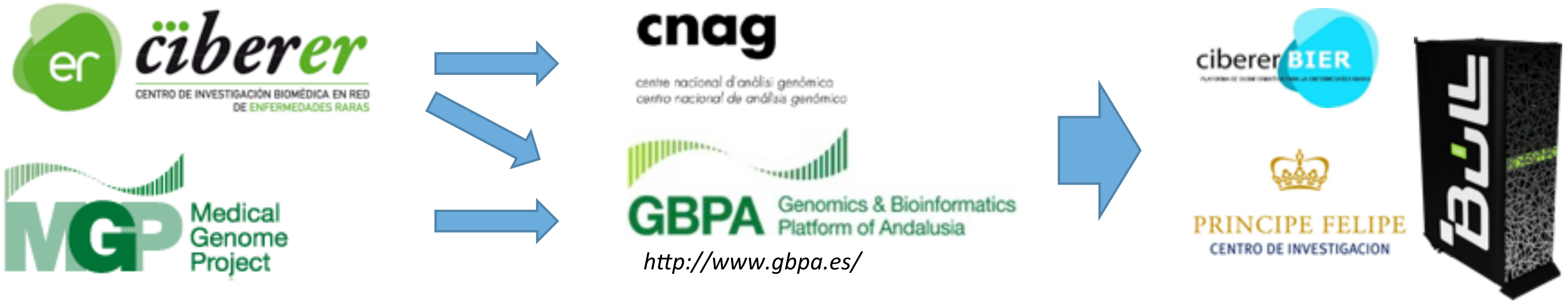
While **cost** is in continuous **reduction, data volume and complexity increases**

Medicine becomes more and more computational

From Dopazo

The Spanish "1000 genomes"

Initiative to sequence rare disease patients


ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED DE ENFERMEDADES RARAS
MGP Medical Genome Project
cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico
GBPA Genomics & Bioinformatics Platform of Andalusia
cibererBIER
PRINCIPE FELIPE
CENTRO DE INVESTIGACION
BULL

http://www.gbpa.es/

Sample providers **Sequencing platforms** **Data analysis**

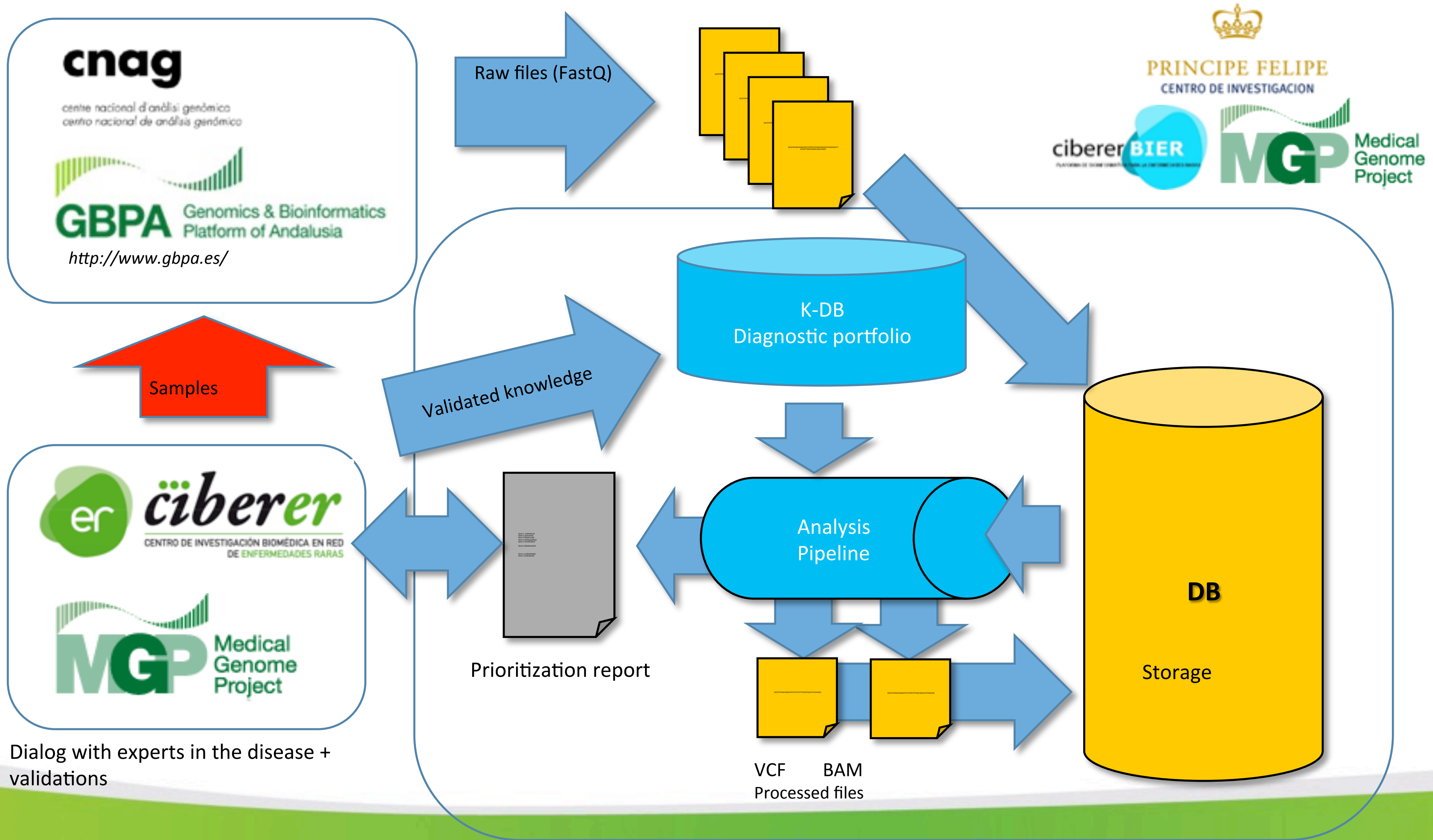
Diseases with
- Unknown genes
- No mutations in known genes

Search for:
- New genes
- Known genes with unknown modifier genes
- Susceptibility genes

A total of 1044 patients (including 300 controls) of more than 30 diseases were sequenced

From Dopazo

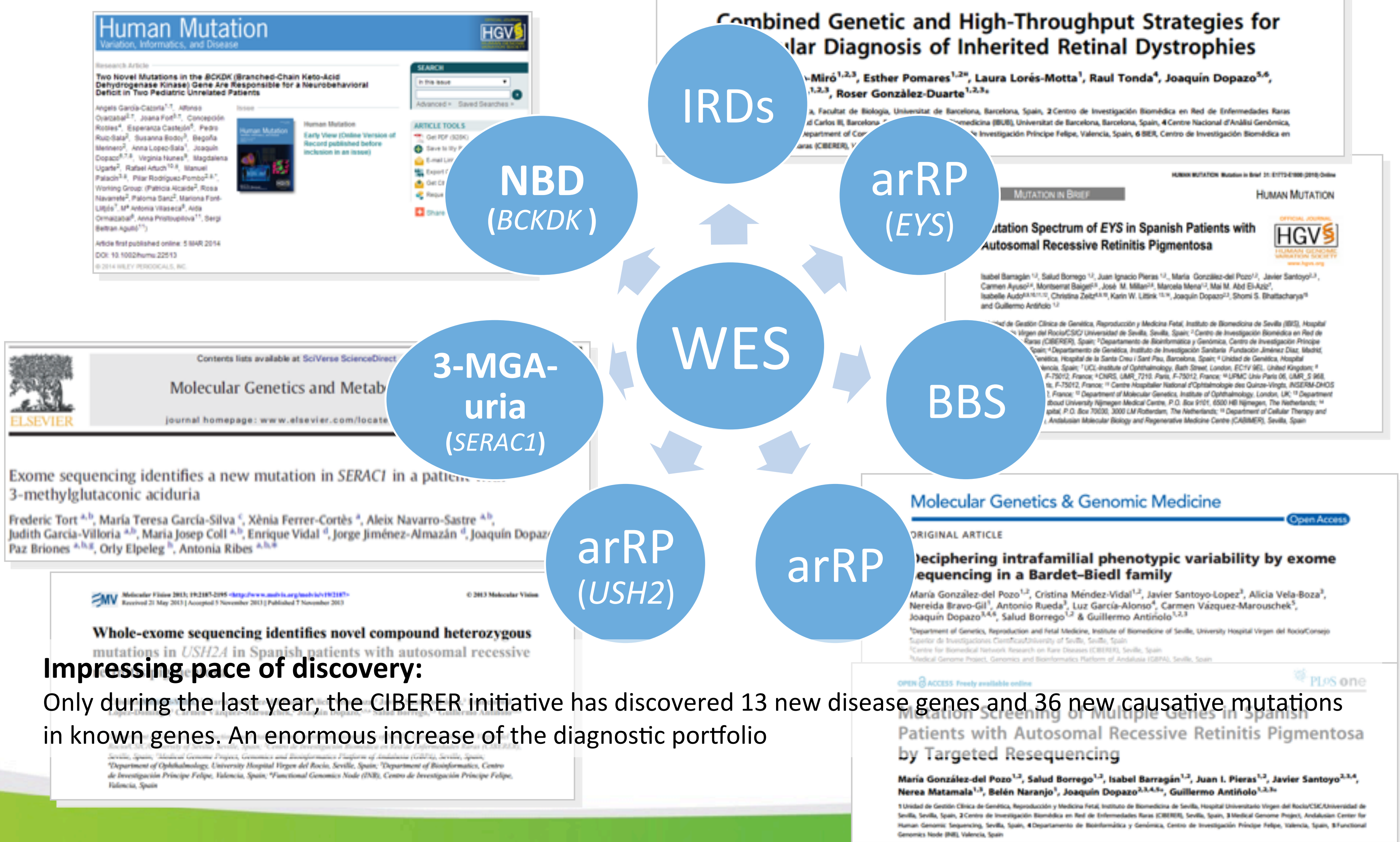
Data management, analysis and storage = knowledge increase


cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico
GBPA Genomics & Bioinformatics Platform of Andalusia
http://www.gbpa.es/
Samples
Raw files (FastQ)
PRINCIPE FELIPE
CENTRO DE INVESTIGACION
ciberer BIER
MGP Medical Genome Project
K-DB
Diagnostic portfolio
Validated knowledge
ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED DE ENFERMEDADES RARAS
MGP Medical Genome Project
Analysis Pipeline
Prioritization report
DB
Storage
VCF
BAM
Processed files

Dialog with experts in the disease + validations

From Dopazo

Discovery of new disease genes by whole exome sequencing



Human Mutation
Variation, Informatics, and Disease

Research Article

Two Novel Mutations in the *BCKDK* (Branched-Chain Keto-Acid Dehydrogenase Kinase) Gene Are Responsible for a Neurobehavioral Deficit in Two Pediatric Unrelated Patients

Angels García-Cazorla[1,7], Alfonso Oyarzabal[2,7], Joana Fort[3,7], Concepción Robles[4], Esperanza Castejón[5], Pedro Ruiz-Sala[2], Susanna Bodoy[3], Begoña Merinero[2], Anna Lopez-Sala[1], Joaquín Dopazo[6,7,8], Virginia Nunes[9], Magdalena Ugarte[2], Rafael Artuch[10,8], Manuel Palacín[3,8], Pilar Rodríguez-Pombo[2,8,*], Working Group: (Patricia Alcaide[2], Rosa Navarrete[2], Paloma Sanz[2], Mariona Font-Llitjós[1], Mª Antonia Vilaseca[8], Aida Ormazabal[8], Anna Pristoupilova[11], Sergi Beltran Agulló[11])

Article first published online: 5 MAR 2014

DOI: 10.1002/humu.22513

© 2014 WILEY PERIODICALS, INC.

Issue

Human Mutation

Early View (Online Version of Record published before inclusion in an issue)

SEARCH

In this issue

Advanced › Saved Searches ›

ARTICLE TOOLS

Get PDF (928K)

Save to My P

E-mail Lin

Export C

Get Ci

Reque

Share

OPEN ACCESS Freely available online PLOS ONE

Combined Genetic and High-Throughput Strategies for …lar Diagnosis of Inherited Retinal Dystrophies

…-Miró[1,2,3], Esther Pomares[1,2¤], Laura Lorés-Motta[1], Raul Tonda[4], Joaquín Dopazo[5,6], …[1,2,3], Roser Gonzàlez-Duarte[1,2,3*]

…a, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain, 2 Centro de Investigación Biomédica en Red de Enfermedades Raras …d Carlos III, Barcelona, … …medicina (IBUB), Universitat de Barcelona, Barcelona, Spain, 4 Centre Nacional d'Anàlisi Genòmica, …epartment of Com… …e Investigación Príncipe Felipe, Valencia, Spain, 6 BIER, Centro de Investigación Biomédica en …aras (CIBERER), V…

HUMAN MUTATION Mutation in Brief 31: E1772-E1800 (2010) Online

MUTATION IN BRIEF

HUMAN MUTATION

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

…utation Spectrum of *EYS* in Spanish Patients with Autosomal Recessive Retinitis Pigmentosa

Isabel Barragán[1,2], Salud Borrego[1,2], Juan Ignacio Pieras[1,2], María González-del Pozo[1,2], Javier Santoyo[2,3], Carmen Ayuso[2,4], Montserrat Baiget[2,5], José M. Millán[2,6], Marcela Mena[1,2], Mai M. Abd El-Aziz[7], Isabelle Audo[8,9,10,11,12], Christina Zeitz[8,9,10], Karin W. Littink[13,14], Joaquín Dopazo[2,3], Shomi S. Bhattacharya[15] and Guillermo Antiñolo[1,2]

…idad de Gestión Clínica de Genética, Reproducción y Medicina Fetal, Instituto de Biomedicina de Sevilla (IBIS), Hospital … Virgen del Rocío/CSIC/ Universidad de Sevilla, Sevilla, Spain; [2] Centro de Investigación Biomédica en Red de … Raras (CIBERER), Spain; [3] Departamento de Bioinformática y Genómica, Centro de Investigación Príncipe … Spain; [4] Departamento de Genética, Instituto de Investigación Sanitaria Fundación Jiménez Díaz, Madrid, … Genética, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain; [6] Unidad de Genética, Hospital … Valencia, Spain; [7] UCL-Institute of Ophthalmology, Bath Street, London, EC1V 9EL, United Kingdom; [8] … F-75012, France; [9] CNRS, UMR_7210, Paris, F-75012, France; [10] UPMC Univ Paris 06, UMR_S 968, … Paris, F-75012, France; [11] Centre Hospitalier National d'Ophtalmologie des Quinze-Vingts, INSERM-DHOS … France; [12] Department of Molecular Genetics, Institute of Ophthalmology, London, UK; [13] Department … Radboud University Nijmegen Medical Centre, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands; [14] … Hospital, P.O. Box 70030, 3000 LM Rotterdam, The Netherlands; [15] Department of Cellular Therapy and … Andalusian Molecular Biology and Regenerative Medicine Centre (CABIMER), Sevilla, Spain

Contents lists available at SciVerse ScienceDirect

Molecular Genetics and Metab…

journal homepage: www.elsevier.com/locate…

ELSEVIER

Exome sequencing identifies a new mutation in *SERAC1* in a patie… 3-methylglutaconic aciduria

Frederic Tort[a,b], María Teresa García-Silva[c], Xènia Ferrer-Cortès[a], Aleix Navarro-Sastre[a,b], Judith García-Villoria[a,b], Maria Josep Coll[a,b], Enrique Vidal[d], Jorge Jiménez-Almazán[d], Joaquín Dopaz… Paz Briones[a,b,g], Orly Elpeleg[h], Antonia Ribes[a,b,*]

MV Molecular Vision 2013; 19:2187-2195 <http://www.molvis.org/molvis/v19/2187>
Received 21 May 2013 | Accepted 5 November 2013 | Published 7 November 2013 © 2013 Molecular Vision

Whole-exome sequencing identifies novel compound heterozygous mutations in *USH2A* in Spanish patients with autosomal recessive …

…Rocío/CSIC/University of Seville, Seville, Spain; [2] Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Seville, Spain; [3] Medical Genome Project, Genomics and Bioinformatics Platform of Andalusia (GBPA), Seville, Spain; [4] Department of Ophthalmology, University Hospital Virgen del Rocío, Seville, Spain; [5] Department of Bioinformatics, Centro de Investigación Príncipe Felipe, Valencia, Spain; [6] Functional Genomics Node (INB), Centro de Investigación Príncipe Felipe, Valencia, Spain

Molecular Genetics & Genomic Medicine

Open Access

ORIGINAL ARTICLE

…eciphering intrafamilial phenotypic variability by exome …equencing in a Bardet–Biedl family

María González-del Pozo[1,2], Cristina Méndez-Vidal[1,2], Javier Santoyo-Lopez[3], Alicia Vela-Boza[3], Nereida Bravo-Gil[1], Antonio Rueda[3], Luz García-Alonso[4], Carmen Vázquez-Marouschek[5], Joaquín Dopazo[3,4,6], Salud Borrego[1,2] & Guillermo Antiñolo[1,2,3]

[1] Department of Genetics, Reproduction and Fetal Medicine, Institute of Biomedicine of Seville, University Hospital Virgen del Rocío/Consejo Superior de Investigaciones Científicas/University of Seville, Seville, Spain
[2] Centre for Biomedical Network Research on Rare Diseases (CIBERER), Seville, Spain
[3] Medical Genome Project, Genomics and Bioinformatics Platform of Andalusia (GBPA), Seville, Spain

OPEN ACCESS Freely available online PLOS one

Mutation Screening of Multiple Genes in Spanish Patients with Autosomal Recessive Retinitis Pigmentosa by Targeted Resequencing

María González-del Pozo[1,2], Salud Borrego[1,2], Isabel Barragán[1,2], Juan I. Pieras[1,2], Javier Santoyo[2,3,4], Nerea Matamala[1,3], Belén Naranjo[1], Joaquín Dopazo[2,3,4,5*], Guillermo Antiñolo[1,2,3*]

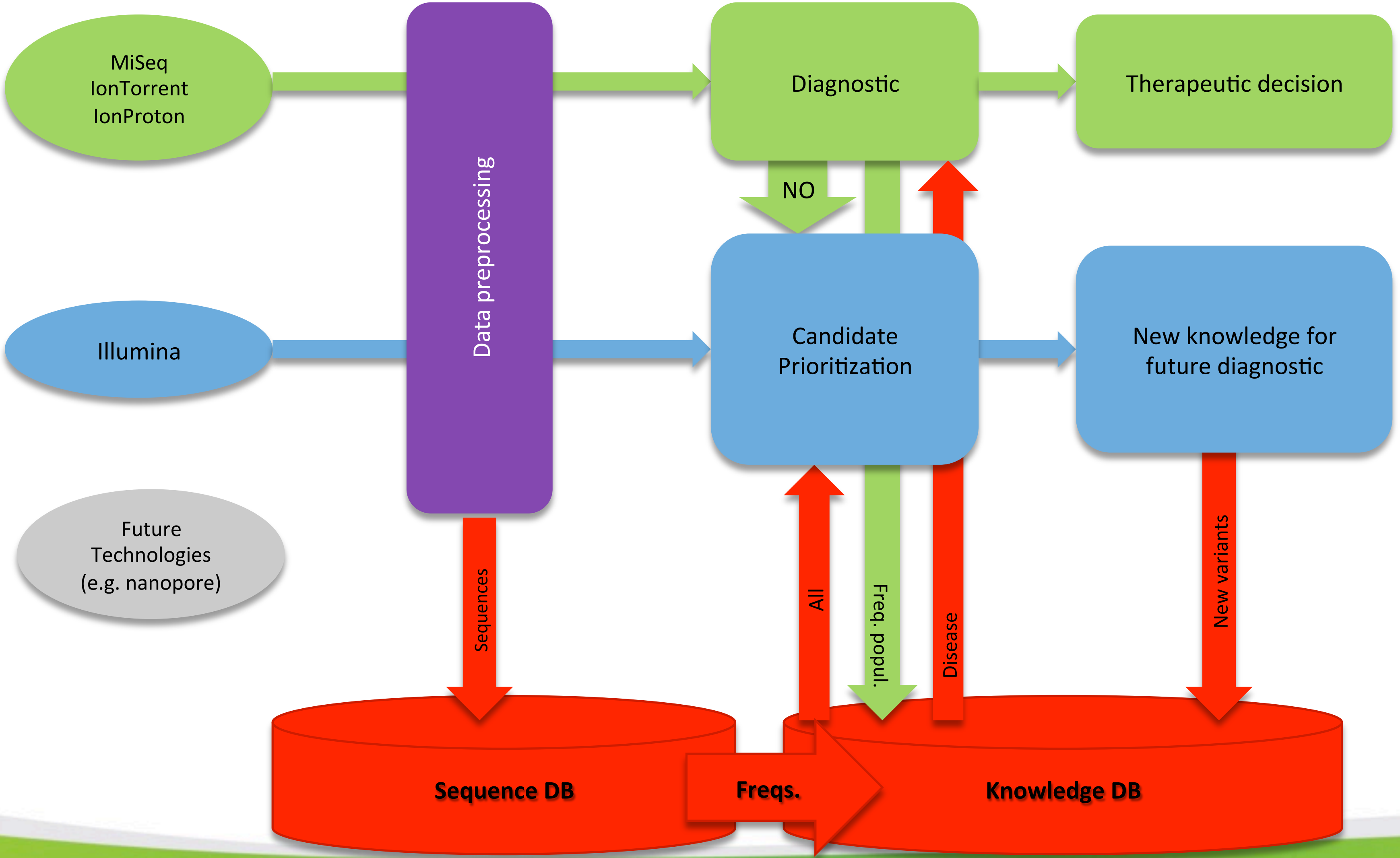
1 Unidad de Gestión Clínica de Genética, Reproducción y Medicina Fetal, Instituto de Biomedicina de Sevilla, Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Sevilla, Spain, 2 Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Sevilla, Spain, 3 Medical Genome Project, Andalusian Center for Human Genomic Sequencing, Sevilla, Spain, 4 Departamento de Bioinformática y Genómica, Centro de Investigación Príncipe Felipe, Valencia, Spain, 5 Functional Genomics Node (INB), Valencia, Spain

Impressing pace of discovery:

Only during the last year, the CIBERER initiative has discovered 13 new disease genes and 36 new causative mutations in known genes. An enormous increase of the diagnostic portfolio

From Dopazo

Scaling-up the developments to support the rare diseases initiative to the development of a comprehensive system for diagnosis and gene discovery



From Dopazo

Implementation of a system for genomic data management in the supercomputing center IT4I (Czech Republic)

This pilot project has been set up in the IT4I supercomputing center, where the genomic data analysis of the country will be centralized.

Obviously not yet challenged with 10M genomes

IT4Innovations
národní
superpočítačové
centrum

Bull

OMICS MASTER 24.7.2014

PI 332 Budova Podnikatelského inkubátoru v areálu VŠB – Technické univerzity Ostrava 9:00-16:00

Obsahem kurzu je praktické seznámení se sadou nástrojů OMICS MASTER, které budou nasazeny v IT4Innovations národním superpočítačovém centru. Cílem instalace je vytvoření standardního prostředí, nástrojů a postupů (pipeline) pro analýzu dat - NGS (Next Generation Sequencing). Zaměřením připravované pipeline je především lidský genomický výzkum, zejména pak sestavení genomu, identifikace genů a anotace jejich variant pro diagnostické účely.

Seznámení s pipeline proběhne prostřednictvím případové studie. První část bude zaměřena na analýzu reálného vzorku primárních dat. Druhá část se soustředí na práci s diagnostickými moduly. Kurz přináší zručnosti potřebné pro efektivní zpracování a analýzu NGS dat na infrastruktuře IT4Innovations, od primárních dat až po anotaci variant včetně identifikace jejich biologické signifikance a diagnostické hodnoty.

Zaváděná pipeline je postavena na otevřeném softwaru a zabezpečí efektivní provedení výpočetně intenzivních úloh spojených se zpracováním primárních dat ze NGS přístrojů. Mezi výstupy analýzy je přímo VCF formát a nástroje pro jeho další analýzu.

Pipeline je postavená na těchto nástrojích:

- FASTQC
- HPG-aligner
- Samtools
- Picard
- GATK
- hpg-variant
- variant (annotation)
- OpenCGA indexer
- BiERapp (for gene discovery)
- TEAM (for diagnosis)

Na seminář se můžete přihlásit do pátku 18. července 2014 na emailové adrese zuzana.kolarkova@vsb.cz

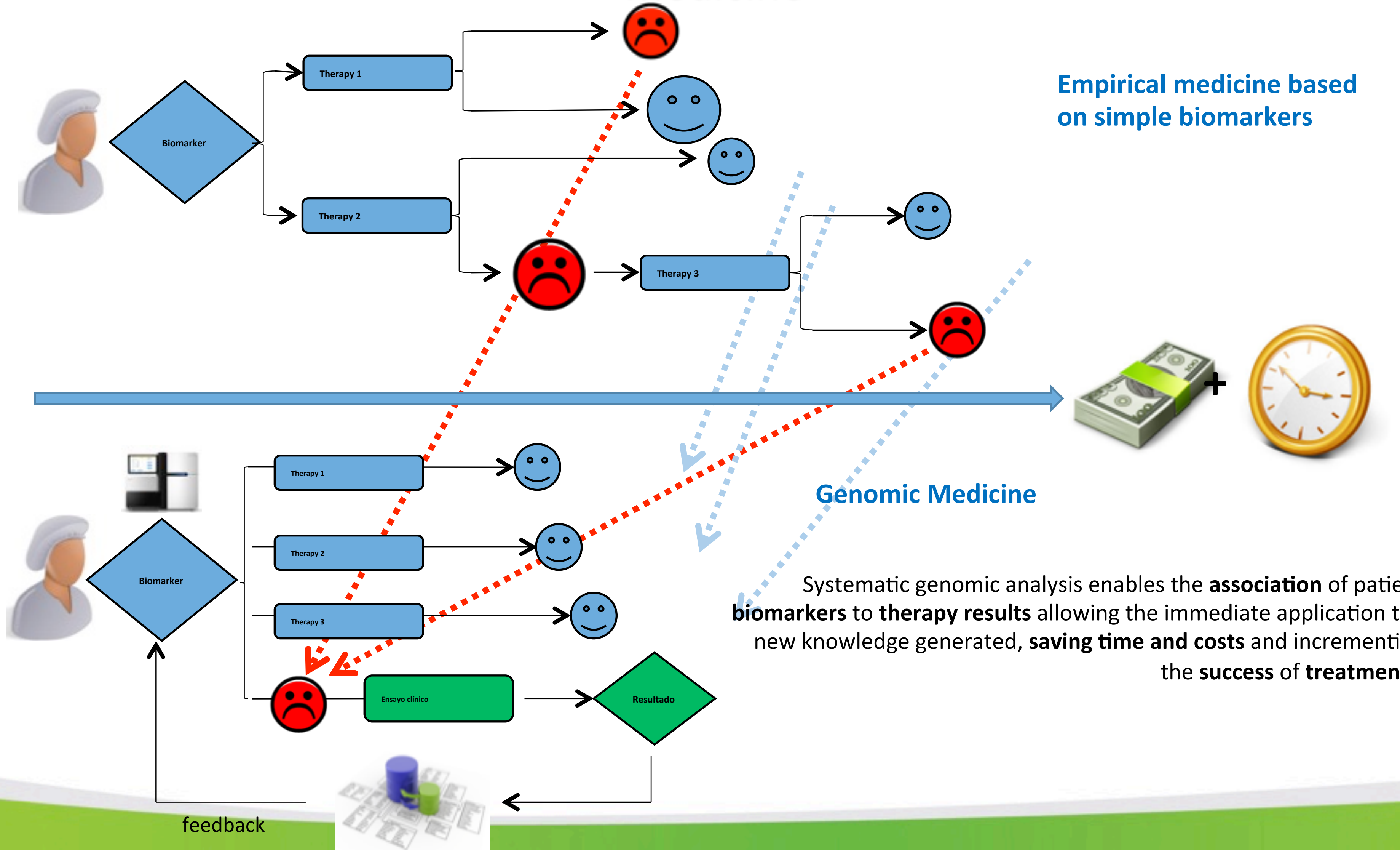
www.it4i.cz

OP Výzkum a vývoj pro inovace

www.bull.cz

From Dopazo

Genomic medicine bridges the gap between empirical and precision medicine



Empirical medicine based on simple biomarkers

Genomic Medicine

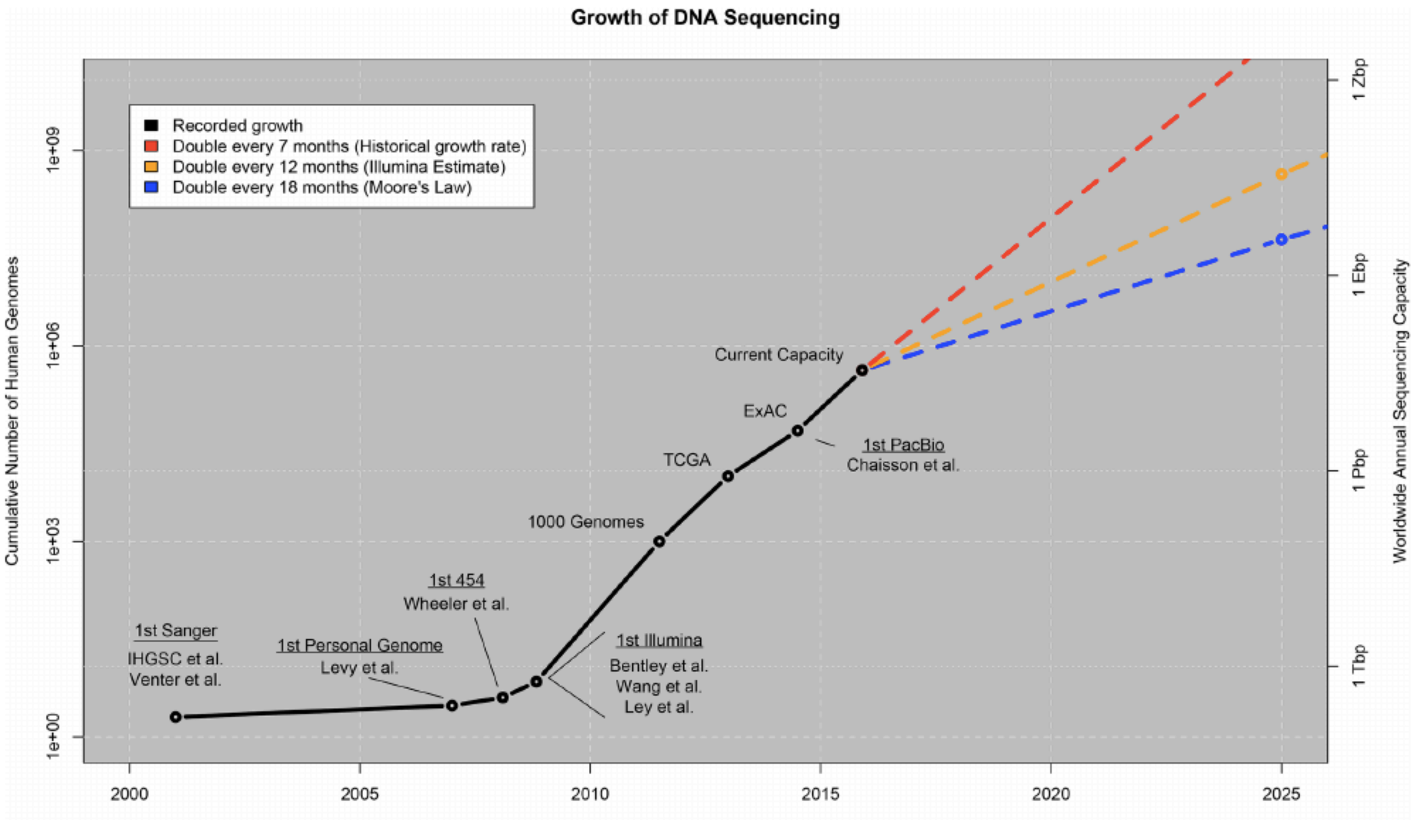
Systematic genomic analysis enables the **association** of patient **biomarkers** to **therapy results** allowing the immediate application the new knowledge generated, **saving time and costs** and incrementing the **success** of **treatments**.

From Dopazo

What are the real dimensions of the challenge?

Big Data: Astronomical or Genomical?
Stephens ZD. et al., PLOS Biology, July 2015

ABSTRACT - Genomics is a Big Data science and is going to get much bigger, very soon, but it is not known whether the needs of genomics will exceed other Big Data domains. ***Projecting to the year 2025****, we compared genomics with three other major generators of Big Data:* ***astronomy, YouTube, and Twitter****. Our estimates show that* ***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***



From Dopazo

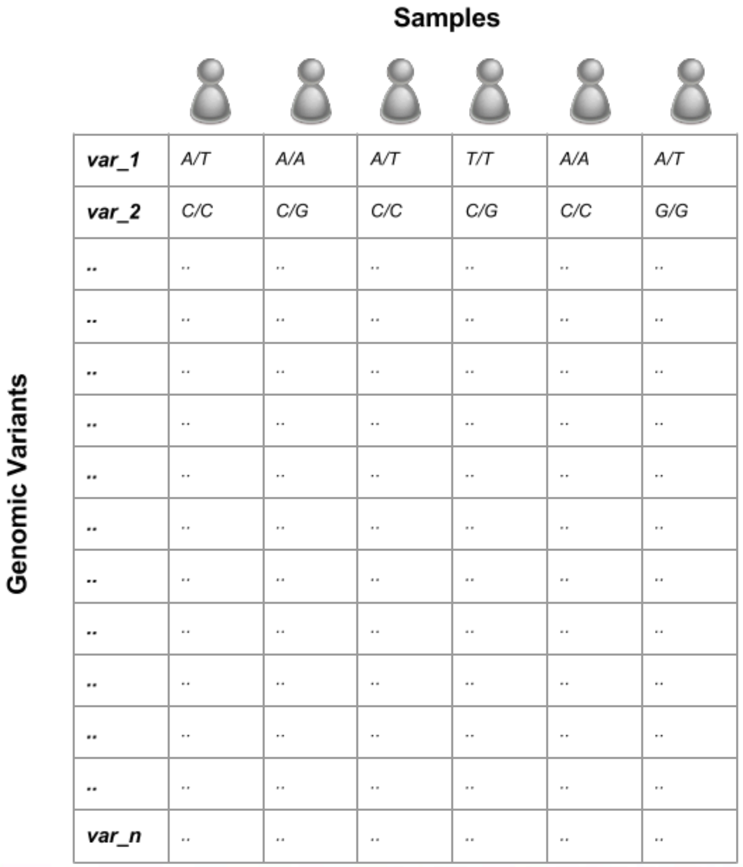
NGS genomic variation data, big and complex

Logical view of genomic variation data, real data comes in **different VCF files**.

Each cell represents one specific genotype for one mutation in one sample

Meta data: **Sample annotation**
- Phenotype
- Family pedigree, Population
- Clinical variables
- ...

Hundreds of millions of mutations, some meta data needed: **Variant annotation**
- Clinical info
- Consequence type
- Conservation scores
- Population frequencies
- ...



Samples

Genomic Variants

var_1	A/T	A/A	A/T	T/T	A/A	A/T
var_2	C/C	C/G	C/C	C/G	C/C	G/G
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
var_n	..	..	..	..	..	..

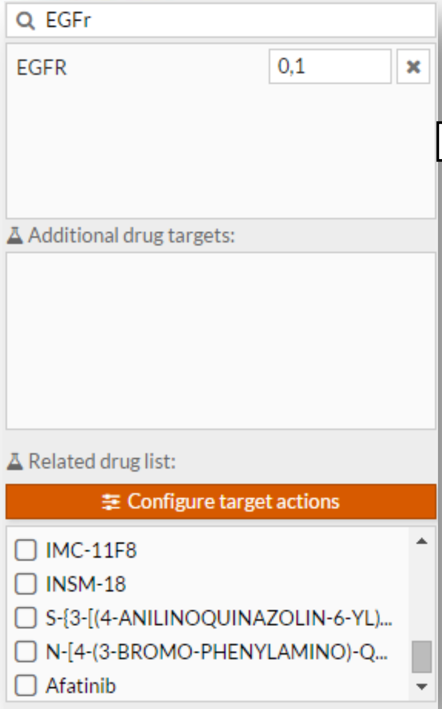
Heterogeneous data analysis and algorithms, different technologies and solutions required:
- Search and filter using data and meta data
- Data mining, correlation
- Statistic tests
- Machine learning
- Interactive analysis
- Network-based analysis
- Visualization
- Encryption
- ...

Genomics England project:
200M variants x 100K samples.
About **20 trillion** points with a lot of meta data. About **500-1000TB** to be indexed.

Applications:
- Personalized medicine
- ...

From Dopazo

Simulating drug inhibition



EGFr

EGFR 0,1 ✖

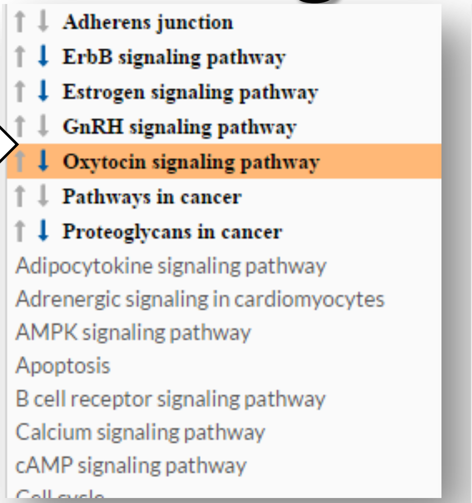
Additional drug targets:

Related drug list:

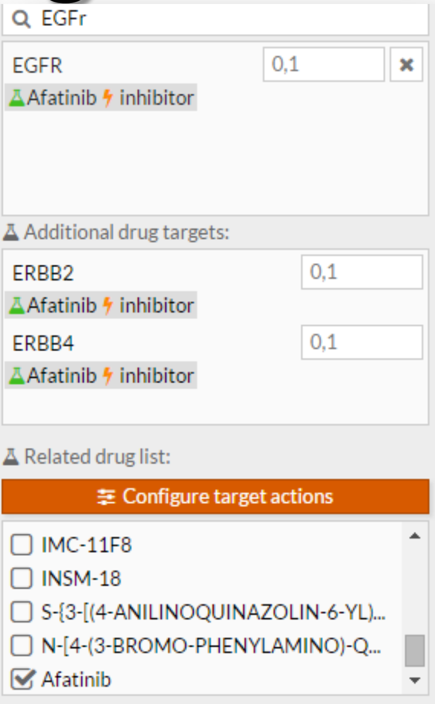
Configure target actions

□ IMC-11F8
□ INSM-18
□ S-{3-[(4-ANILINOQUINAZOLIN-6-YL)...
□ N-[4-(3-BROMO-PHENYLAMINO)-Q...
□ Afatinib

“Ideal” KO of EGFR affects 7 pathways



↑ ↓ Adherens junction
↑ ↓ ErbB signaling pathway
↑ ↓ Estrogen signaling pathway
↑ ↓ GnRH signaling pathway
↑ ↓ Oxytocin signaling pathway
↑ ↓ Pathways in cancer
↑ ↓ Proteoglycans in cancer
Adipocytokine signaling pathway
Adrenergic signaling in cardiomyocytes
AMPK signaling pathway
Apoptosis
B cell receptor signaling pathway
Calcium signaling pathway
cAMP signaling pathway
Cell cycle



EGFr

EGFR 0,1 ✖
Afatinib inhibitor

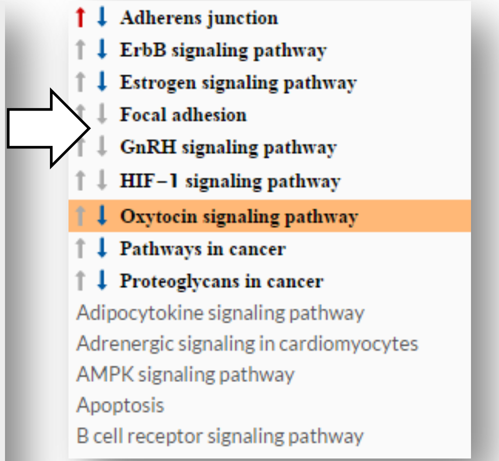
Additional drug targets:

ERBB2 0,1
Afatinib inhibitor
ERBB4 0,1
Afatinib inhibitor

Related drug list:

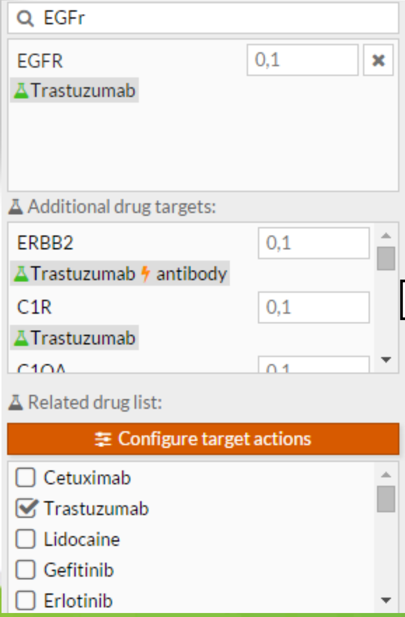
Configure target actions

□ IMC-11F8
□ INSM-18
□ S-{3-[(4-ANILINOQUINAZOLIN-6-YL)...
□ N-[4-(3-BROMO-PHENYLAMINO)-Q...
☑ Afatinib



↑ ↓ Adherens junction
↑ ↓ ErbB signaling pathway
↑ ↓ Estrogen signaling pathway
↑ ↓ Focal adhesion
↑ ↓ GnRH signaling pathway
↑ ↓ HIF−1 signaling pathway
↑ ↓ Oxytocin signaling pathway
↑ ↓ Pathways in cancer
↑ ↓ Proteoglycans in cancer
Adipocytokine signaling pathway
Adrenergic signaling in cardiomyocytes
AMPK signaling pathway
Apoptosis
B cell receptor signaling pathway

Real inhibition with Afanatib affects 11 pathways



EGFr

EGFR 0,1 ✖
Trastuzumab

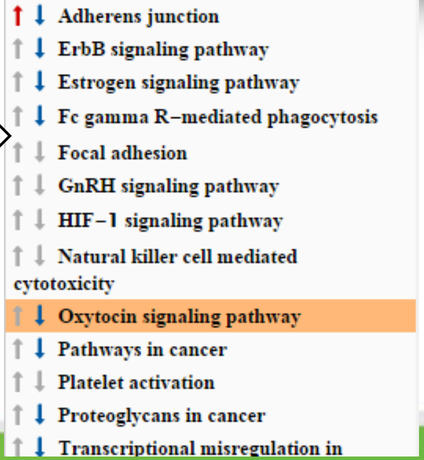
Additional drug targets:

ERBB2 0,1
Trastuzumab antibody
C1R 0,1
Trastuzumab
C1QA 0,1

Related drug list:

Configure target actions

□ Cetuximab
☑ Trastuzumab
□ Lidocaine
□ Gefitinib
□ Erlotinib



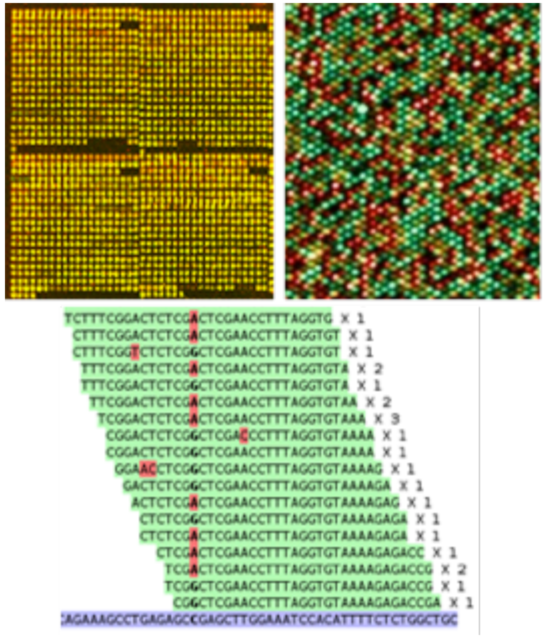
↑ ↓ Adherens junction
↑ ↓ ErbB signaling pathway
↑ ↓ Estrogen signaling pathway
↑ ↓ Fc gamma R−mediated phagocytosis
↑ ↓ Focal adhesion
↑ ↓ GnRH signaling pathway
↑ ↓ HIF−1 signaling pathway
↑ ↓ Natural killer cell mediated cytotoxicity
↑ ↓ Oxytocin signaling pathway
↑ ↓ Pathways in cancer
↑ ↓ Platelet activation
↑ ↓ Proteoglycans in cancer
↑ ↓ Transcriptional misregulation in

Inhibition with broader spectrum Trastuzumab affects 13 pathways

From Dopazo

Omics views of genomes

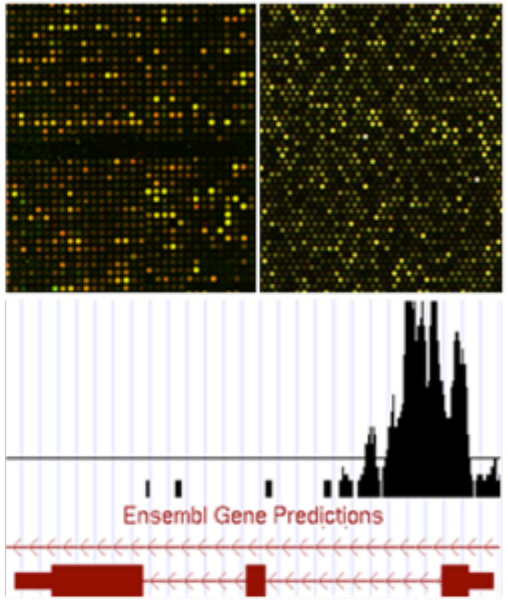
Genetic variation



```
TCTTTCGGACTCTCGACTCGAACCTTTAGGTG X 1
 CTTTCGGACTCTCGACTCGAACCTTTAGGTGT X 1
 CTTTCGGTCTCTCGGCTCGAACCTTTAGGTGT X 1
  TTTCGGACTCTCGACTCGAACCTTTAGGTGTA X 2
  TTTCGGACTCTCGGCTCGAACCTTTAGGTGTA X 1
   TTCGGACTCTCGACTCGAACCTTTAGGTGTAA X 2
    TCGGACTCTCGACTCGAACCTTTAGGTGTAAA X 3
     CGGACTCTCGGCTCGACCCTTTAGGTGTAAAA X 1
     CGGACTCTCGGCTCGAACCTTTAGGTGTAAAA X 1
      GGAACCTCGGCTCGAACCTTTAGGTGTAAAAG X 1
       GACTCTCGGCTCGAACCTTTAGGTGTAAAAGA X 1
        ACTCTCGACTCGAACCTTTAGGTGTAAAAGAG X 1
         CTCTCGGCTCGAACCTTTAGGTGTAAAAGAGA X 1
         CTCTCGACTCGAACCTTTAGGTGTAAAAGAGA X 1
           CTCGACTCGAACCTTTAGGTGTAAAAGAGACC X 1
            TCGACTCGAACCTTTAGGTGTAAAAGAGACCG X 2
            TCGGCTCGAACCTTTAGGTGTAAAAGAGACCG X 1
             CGGCTCGAACCTTTAGGTGTAAAAGAGACCGA X 1
AGAAAGCCTGAGAGCCGAGCTTGGAAATCCACATTTTCTCTGGCTGC
```

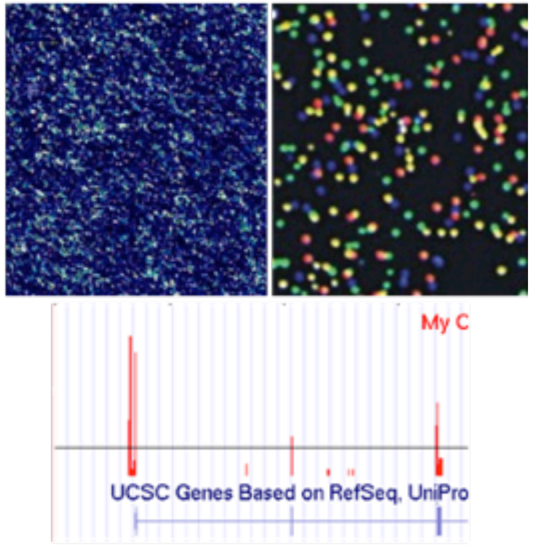
SNPs, loss-of-heterozygosity
Copy number variants

Epigenetic variation



DNA methylation
Chromatin

Expression variation



RNA expression
Gene structure

From LUMC

HPC Data Analysis, Simulation.. And Visualization

- Scientists need visualize simulations and data
- HPC platforms are necessary
  - Terascale/Petascale/Exascale
    - Energy Efficiency
    - Data Movement
    - Programmability
  - Workflows
  - Hybrid Computing
- Relevance in Simulations
  - Smart
  - Ultrascale

Bioinformatics Tools – Open source

Programming languages

- R, Perl, Bash scripting (Linux), MySQL, Apache, PHP, Python, Java, …

Software, e.g. Bioconductor, BioPerl, Ensembl Perl API, Bowtie, BWA, Velvet, Varscan, Rmap, …

- Alignment, analysis of next-generation sequencing and microarray data

Web browsers, e.g. UCSC, Ensembl

- visualize data in relation to genome features

Gene Ontology, e.g. DAVID

- functional annotation and enrichment

From Dopazo

And… why R?

R is powerfull, worldwide used and open source.

We can exploit easy HPC architectures (almost).

This course introduces some relatively new additions to the R programming language: advanced reduction and visualization. R packagess provide a powerful toolkit to make the process of manipulating and visualising data easy and intuitive, in this case for microbiology.

From Dopazo

Schedule and Topics

1. Introduction Course (Today)
2. Some Topics about Algorithms and Platform uses
   1. Personal Installation
   2. HPC UIS Platform Use
3. Introduction to R – Data Structures
4. Writing Analysis workflows with R
5. Summarizing and Combining Data
6. Plotting and Visualization

Goals of the Course

Participants will gain practical experience and skills to be able to:

- Introduce to scripting computational languages, in this case R
- Meet the challenges of data handling and reduction;
- Introduce to the use of R syntax, functions and packages;
- Understand best practices for scientific computational work.
- Introduce to use visualization tools
- Introduce to use HPC platforms in collaborative environments

Information on line and Bibliography

- www.sc3.uis.edu.co
  - http://wiki.sc3.uis.edu.co/
    - http://wiki.sc3.uis.edu.co/index.php/An%C3%A1lisis_y_Visualizaci%C3%B3n_de_Datos_con_R
- https://www.r-project.org/
- https://journal.r-project.org/

- https://bioinformatics.ca/workshops/2017/introduction-r-2017
- http://bioinformatics-core-shared-training.github.io/
- https://training.csx.cam.ac.uk/bioinformatics/course/bioinfo-intR
- http://bioinfotraining.bio.cam.ac.uk/postgraduate/specialized/bioinfo-intR
- … and other to find in the first link.

Something about US

Carlos Jaime Barrios Hernandez, PhD. (@carlosjaimebh cbarrios@uis.edu.co)

PhD in Computer Science (Nice, France) , MSc in Applied Mathematics and Computer Science (Grenoble, France) and Systems Engineering (Bucaramanga, Colombia)

Director of SC3UIS, Assistant Professor UIS.

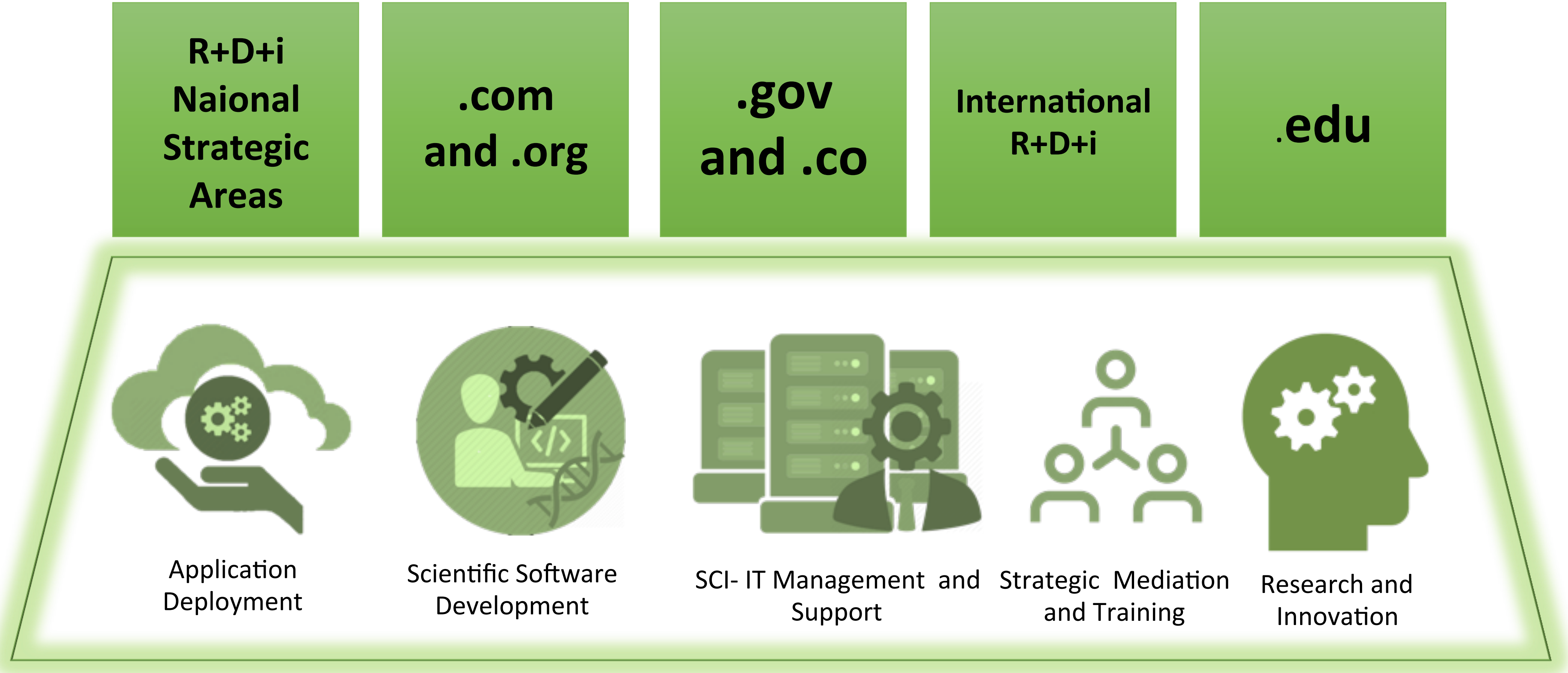
Gilberto Javier Diaz Toro (@gilbertodiazt gilberto.diaz@uis.edu.co)

MSc in Computer Science (Mérida, Venezuela) Systems and Computing Engineering (Mérida, Venezuela).

CTO of SC3UIS, International Instructor and Professor on HPC and Scientific Computing.

Super Computación y Cálculo Científico UIS

¿What is SC3UIS?



R+D+i Naional Strategic Areas

.com and .org

.gov and .co

International R+D+i

.edu

Application Deployment

Scientific Software Development

SCI- IT Management and Support

Strategic Mediation and Training

Research and Innovation

GUANE-1 and Yajé

- **GUANE-1 Reload**
  - High Density and Green HPE-HPC Platform
  - 128 NVIDIA M2070 TESLA GPUs
  - 32 Intel Xeon E5645 2.4 GHz Processors
  - 3 High Bandwth Networks
  - 1.6 TB RAM
  - General Purpose Platform
- **YAJE**
  - HPE ML150 G9 Development Platform
  - Intel Xeon E52609 1.9 GHz Processor
  - 64GB RAM
  - NVIDIA GRID K2
    - 1536 GPU Cores
    - 2 x 8 GB Memory

Parnertship

Universidad Industrial de Santander UIS CONSTRUIMOS FUTURO

Super Computación y Cálculo Científico UIS

Alcaldía de Bucaramanga

uni.lu UNIVERSITÉ DU LUXEMBOURG

SnT securityandtrust.lu

Fonds National de la Recherche Luxembourg

SCALAC

CNCA Colaboratorio Nacional de Computación Avanzada

CATAÏ
Collaboration France-Colombie

RedCLARA

Bull atos technologies

NVIDIA.

UNIRED

SISSA

ICTP The Abdus Salam International Centre for Theoretical Physics

Sistema FIEB SENAI Federação das Indústrias do Estado da Bahia

CIMATEC

Hewlett Packard Enterprise

Universidad de los Andes

un

RISC

IVIC INSTITUTO VENEZOLANO DE INVESTIGACIONES CIENTIFICAS

UNIVERSIDAD DE COSTA RICA

CCT Val

UFRGS UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Laboratório Nacional de Computação Científica

SINAPAD

BSC

LIG

BIOS CENTRO DE BIOINFORMÁTICA Y BIOLOGÍA COMPUTACIONAL

UNIVERSIDAD EAFIT Abierta al mundo

ecopetrol ENERGÍA PARA EL FUTURO

IBM

i3S sophia antipolis

Schlumberger

CRAY

UNIVERSIDAD DE ANTIOQUIA

intel

renata

inria INVENTEURS DU MONDE NUMÉRIQUE

Grid'5000

Aladdin

TACC

CNS CENTRO NACIONAL DE SUPERCÓMPUTO

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

UTP Universidad Tecnológica de Pereira

CETICS

Seagate

« Quien no computa, no compite »

Mateo Valero, PhD.
Director BSC-CNS, Spain

Soon Colombia Advanced Computer Center

Follow us: @sc3uis

Or visit us: www.sc3.uis.edu.co

Thanks!