Introduction

GDA
International Course on
Genomic Data Analysis

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Who we are

The Computational Genomics Department at the Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain, and…

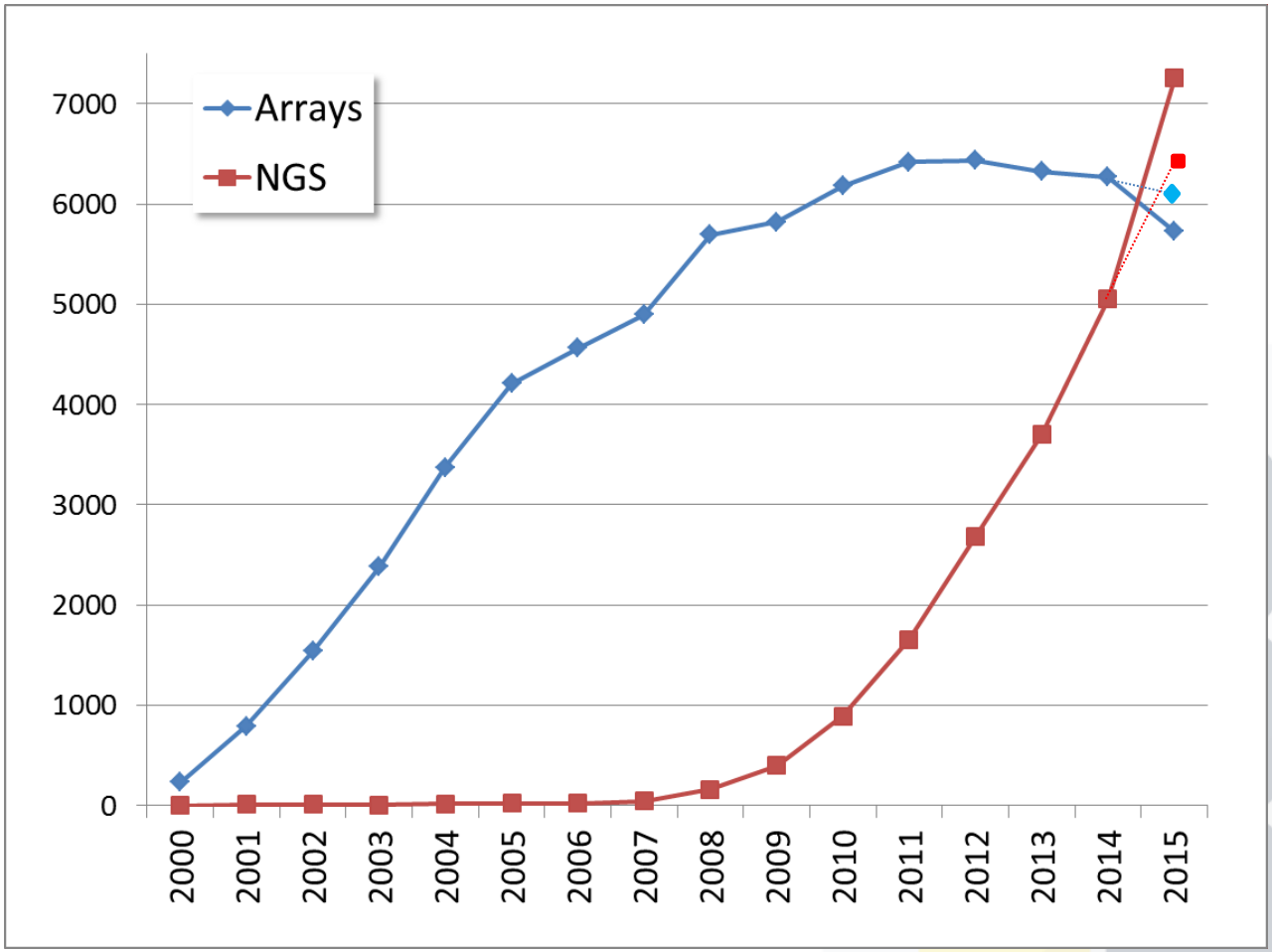
…the INB, National Institute of Bioinformatics (Functional Genomics Node) and the BiER (CIBERER Network of Centers for Rare Diseases)

http://bioinfo.cipf.es
http://www.babelomics.org

@xdopazo @bioinfocipf

PRINCIPE FELIPE CENTRO DE INVESTIGACION
Proyectos PROMETEO Generalitat Valenciana
MGP Medical Genome Project
HPC4G
INB
ciberer CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED DE ENFERMEDADES RARAS
MINISTERIO DE CIENCIA E INNOVACIÓN
F.E.D.E.R.
Fundació La Marató de TV3
ciberer BIER
BULL
mlpm Machine Learning for Personalized Medicine
PEOPLE
MARIE CURIE ACTIONS

Papers published in microarrays and NGS technologies

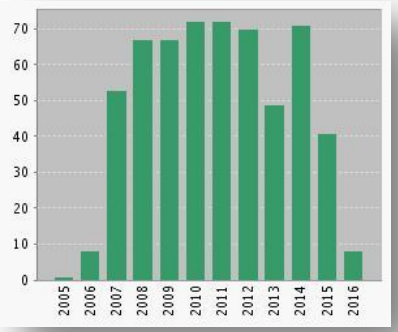


Source Pubmed. Query: "high-throughput sequencing"[Title/Abstract] OR "next generation sequencing"[Title/Abstract] OR "rna seq"[Title/Abstract]) AND year[Publication Date]

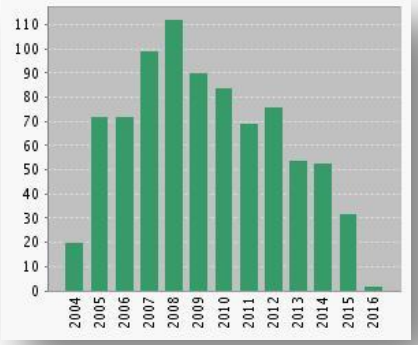
Tools used in the course

579 papers cite Babelomics (plus 835FatiGO cites)

(source ISI Web of Knowledge, Februry2016)



Babelomics



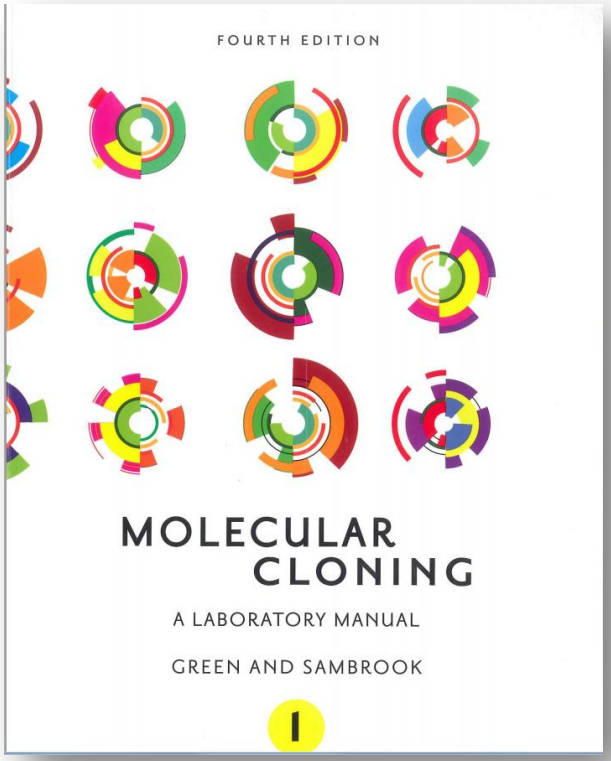
FatiGO

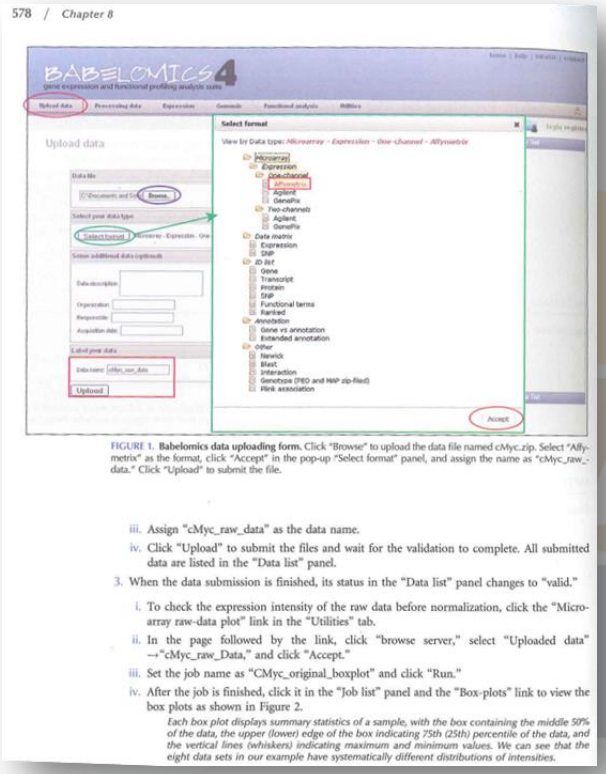
More than 150,000 experiments analysed during the last year.

More than 1000 experiments per day.

Babelomics in the textbooks

The Babelomics suite of programs becomes a classic. Now is cited as a method in the last edition of **Molecular Cloning**. The protocol 4 of chapter 8, Expression Profiling by Microarray and RNA-seq, contains a description on how to use Babelomics to analyze expression data.

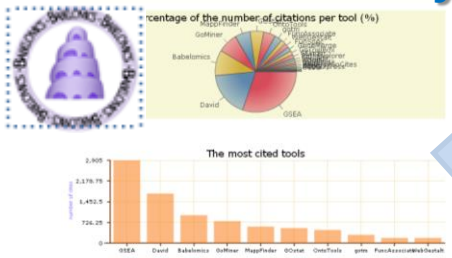




High impact developments

Software development

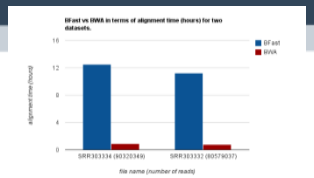
Functional analysis



Babelomics is the third most cited tool for functional analysis. Includes more than 30 tools for advanced, systems-biology based data analysis

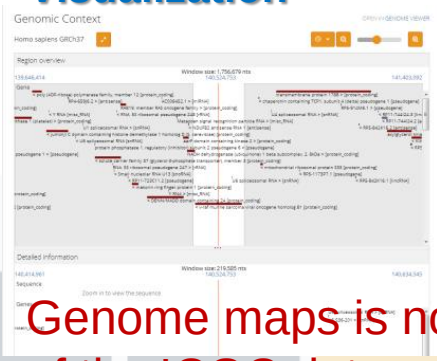
See interactive map of for the last 24h use http://bioinfo.cipf.es/toolsusage

Mapping



HPC on CPU, SSE4, GPUs on NGS data processing
Speedups up to 40X

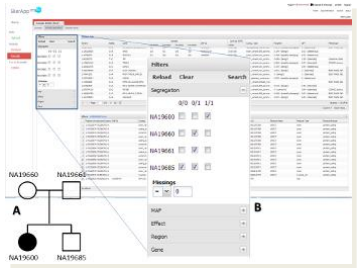
Visualization



Ultrafast genome viewer with google technology

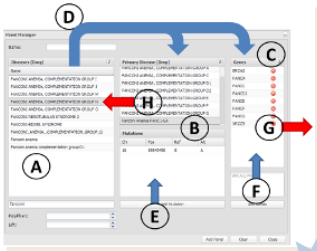
Genome maps is now part of the ICGC data portal

Variant prioritization

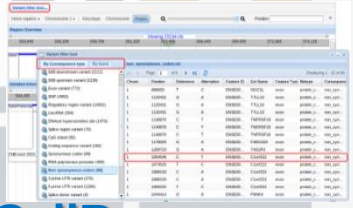


Diagnostic

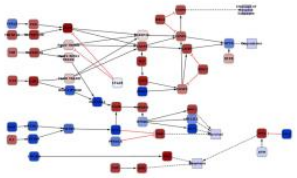
NGS panels



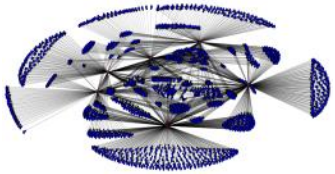
Variant annotation



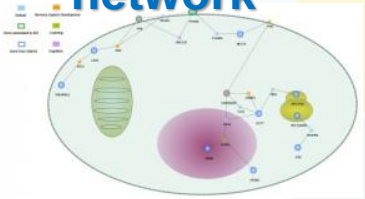
Signaling network



Regulatory network



Interaction network



CellBase

Knowledge database

More than 150.000 experiments were analyzed in our tools during the last year

Background

The road of excess leads to the palace of wisdom

(*William Blake, 28 November 1757 – 12 August 1827, poet, painter, and printmaker*)

The introduction and popularisation of high-throughput techniques has drastically changed the way in which biological problems **can** be addressed and hypotheses **can** be tested.

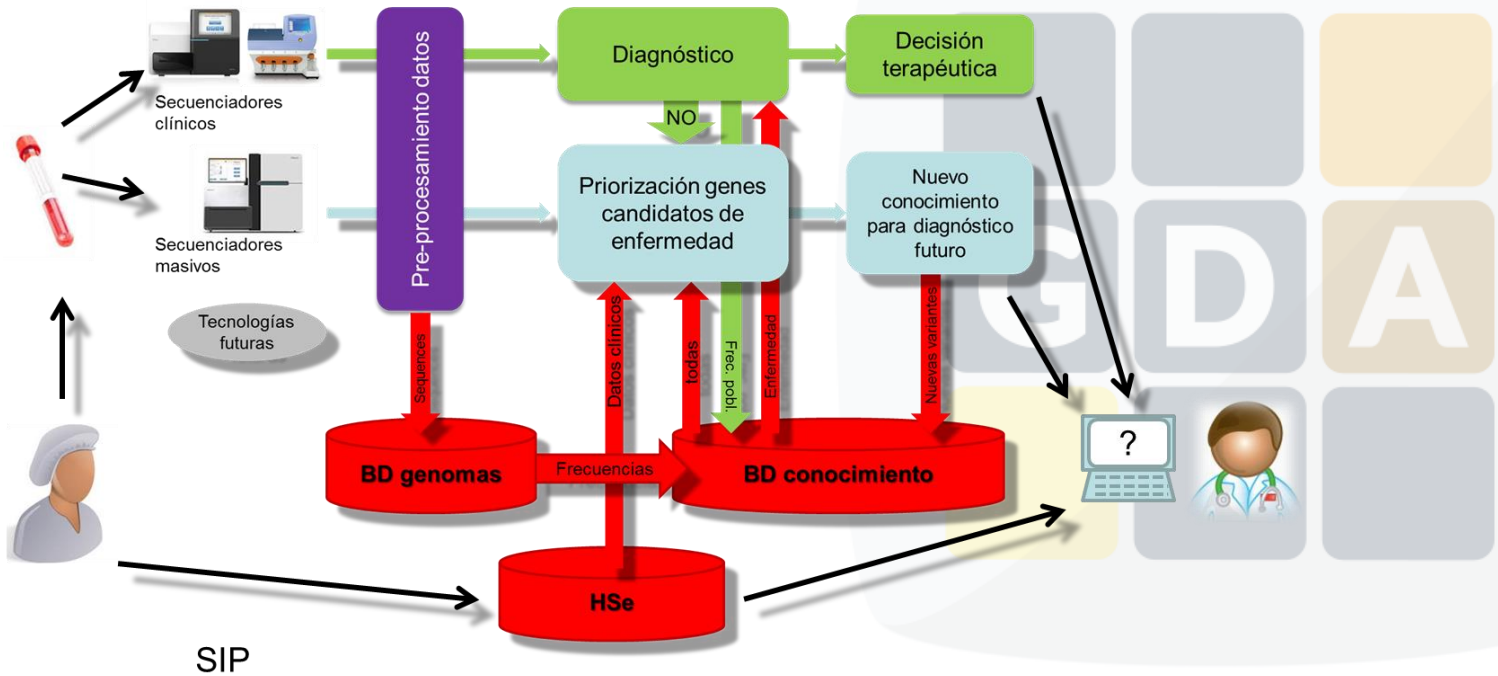
But not necessarily the way in which we really address or test them…

Here you will learn how to do so using state-of-the-art methods and software.

NGS genomic data analysis

Two main applications:

- Diagnostic
- Disease gene finding



Disease gene finding

Typically, an exome renders between 40 and 60K variants (and a genome about 1 million). Only one or a few among all of them are expected to be the causative factors of the disease.

***Casablanca**: Round up all suspicious characters and search them for stolen documents.*

40-60K suspects?
We need…

Agatha Christie

The prioritization process is like a police investigation in which suspected are discarded by their alibies

Thus, through sequential heuristic filtering steps, unlikely candidates are discarded and a final, reduced list with one or a few candidate genes is (hopefully) produced

Pipeline of data analysis

Primary processing	Secondary analysis (Heuristic filtering)	Knowledge-based prioritization
Initial QC FASTQ file	Variant annotation	Proximity to other known disease genes
Mapping BAM file	Filtering by effect	Functional proximity
Variant calling VCF File	Filtering by MAF	Network proximity
	Filtering by family segregation	Burden tests
		Other prioritization methods

Primary analysis

Gene prioritization

Pipeline of data analysis

Primary processing	Secondary analysis (Heuristic filtering)	Knowledge-based prioritization
Initial QC FASTQ file	Variant annotation	Proximity to other known disease genes
Mapping BAM file	Filtering by effect	Functional proximity
Variant calling VCF File	Filtering by MAF	Network proximity
	Filtering by family segregation	Burden tests
		Other prioritization methods

Primary analysis

Gene prioritization

HPG suite

High-Performance Genomics

More info at:

http://bioinfo.cipf.es/docs/compbio/projects/hpg/doku.php

Fastq file, up to hundreds of GB per run

QC and preprocessing

QC stats, filtering and preprocessing options

HPG Aligner, short read aligner

Double mapping strategy: Burrows-Wheeler Transform (*GPU Nvidia CUDA*) + Smith-Waterman (*CPU OpenMP+SSE/AVX*)

SAM/BAM file

QC and preprocessing

QC stats, filtering and preprocessing options

Other analysis (HPC4Genomics consortium)

- RNA-seq
- DNA assembly
- Methyl-seq
- Copy Number
- Structural variation
- Transcript isoform
- Etc.

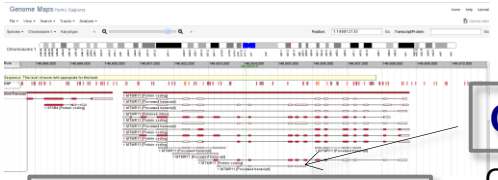
Variant calling analysis

GATK and SAM mPileup HPC Implementation.
Statistics genomic tests

VCF file

QC and preprocessing

QC stats, filtering and preprocessing options



Variant VCF viewer

HTML5+SVG Web based viewer

HPG Variant, Variant analysis

Consequence type, GWAS, regulatory variants and system biology information

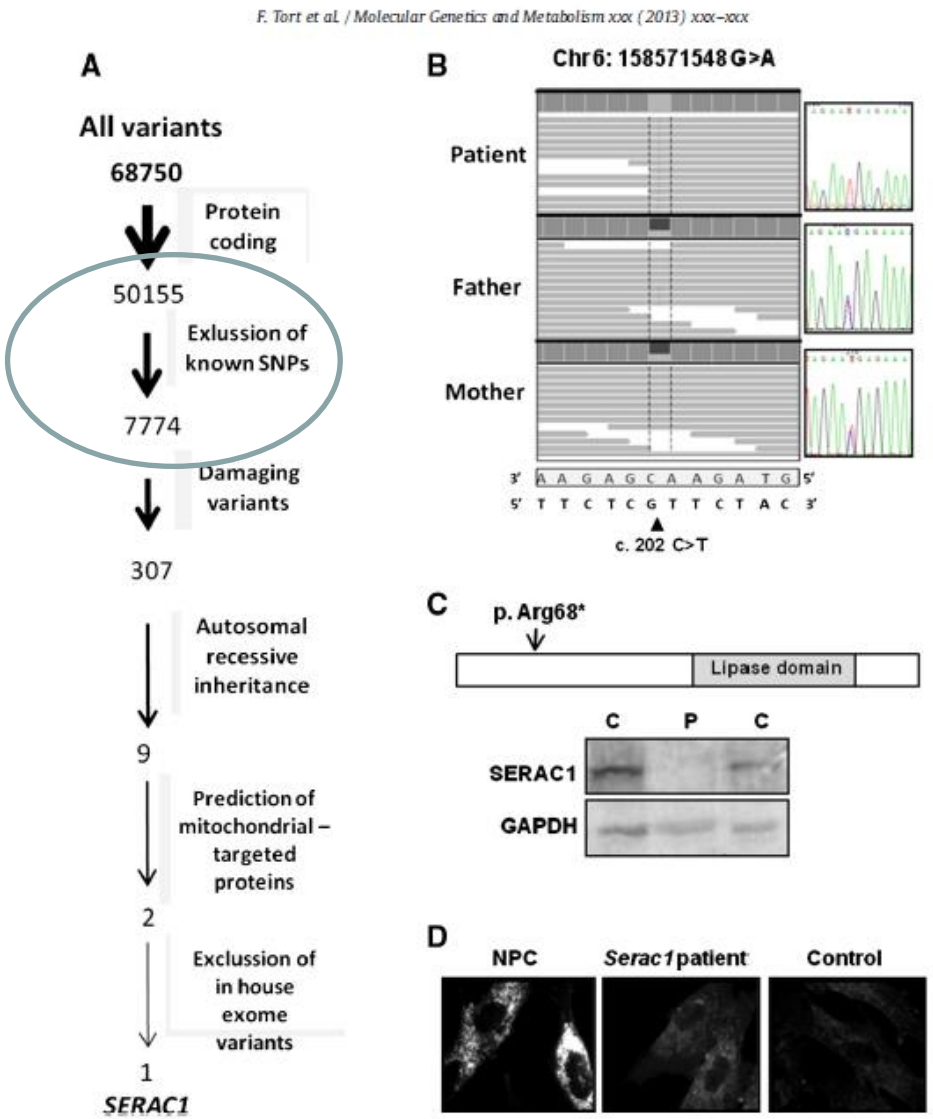
Pipeline of data analysis

Primary processing	Secondary analysis (Heuristic filtering)	Knowledge-based prioritization
Initial QC FASTQ file	Variant annotation	Proximity to other known disease genes
Mapping BAM file	Filtering by effect	Functional proximity
Variant calling VCF File	Filtering by MAF	Network proximity
	Filtering by family segregation	Burden tests
		Other prioritization methods

Primary analysis

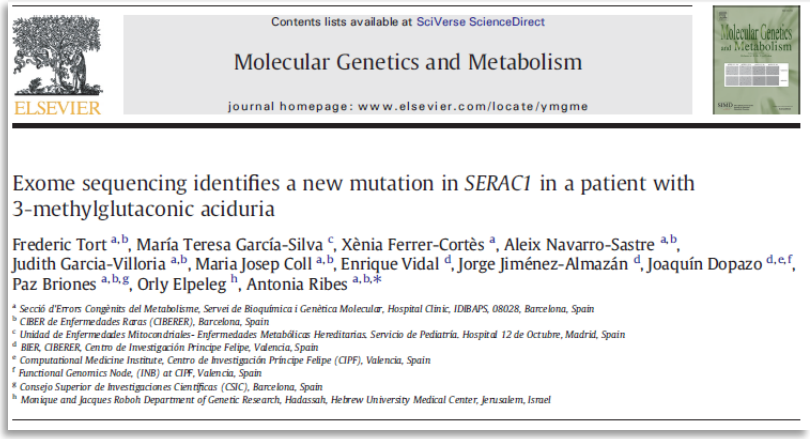
Gene prioritization

Heuristic filtering



3-Methylglutaconic aciduria (3-MGA-uria) is a heterogeneous group of syndromes characterized by an increased excretion of 3-methylglutaconic and 3-methylglutaric acids.

WES with a consecutive filter approach is enough to detect the new mutation in this case.



Contents lists available at SciVerse ScienceDirect

Molecular Genetics and Metabolism

journal homepage: www.elsevier.com/locate/ymgme

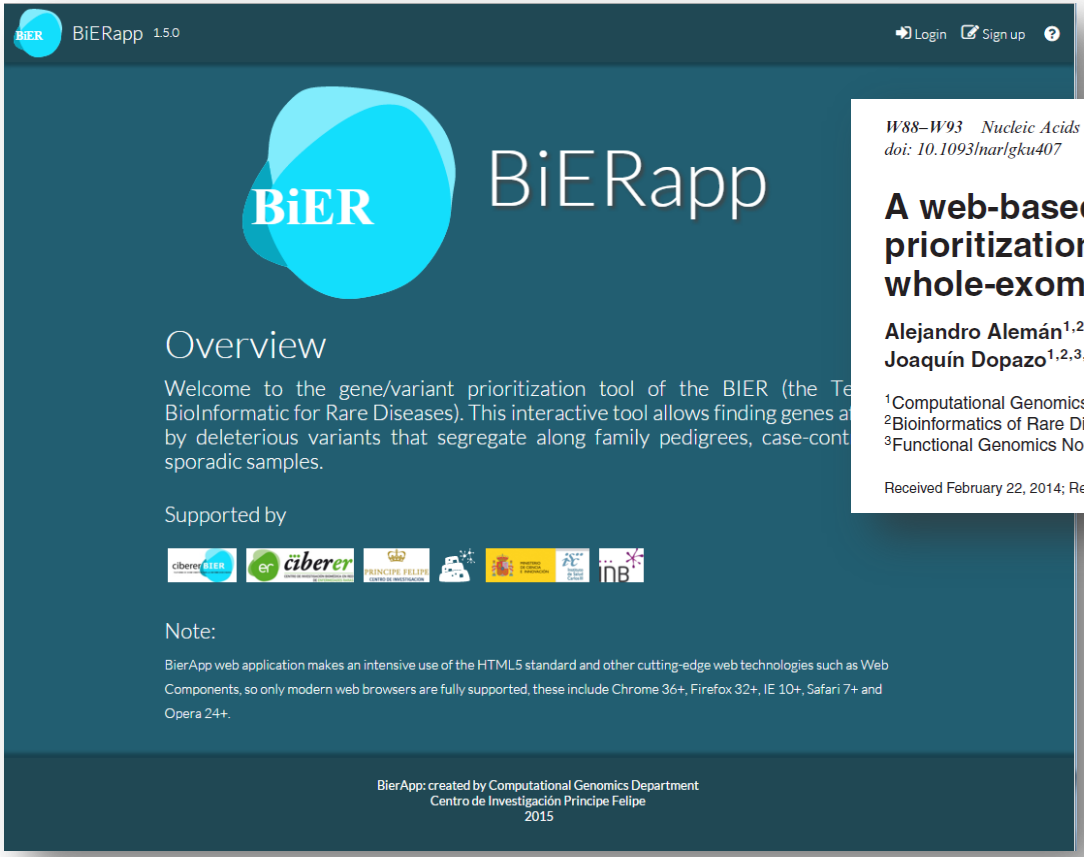
ELSEVIER

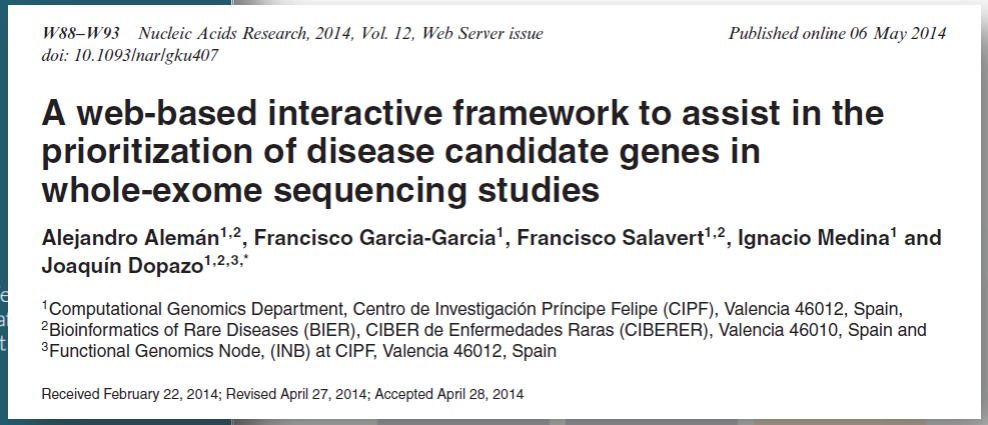
Exome sequencing identifies a new mutation in *SERAC1* in a patient with 3-methylglutaconic aciduria

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

[a] Secció d'Errors Congènits del Metabolisme, Servei de Bioquímica i Genètica Molecular, Hospital Clínic, IDIBAPS, 08028, Barcelona, Spain
[b] CIBER de Enfermedades Raras (CIBERER), Barcelona, Spain
[c] Unidad de Enfermedades Mitocondriales- Enfermedades Metabólicas Hereditarias. Servicio de Pediatría, Hospital 12 de Octubre, Madrid, Spain
[d] BIER, CIBERER, Centro de Investigación Príncipe Felipe, Valencia, Spain
[e] Computational Medicine Institute, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain
[f] Functional Genomics Node, (INB) at CIPF, Valencia, Spain
[g] Consejo Superior de Investigaciones Científicas (CSIC), Barcelona, Spain
[h] Monique and Jacques Roboh Department of Genetic Research, Hadassah, Hebrew University Medical Center, Jerusalem, Israel

The BiERapp





W88–W93 Nucleic Acids Research, 2014, Vol. 12, Web Server issue *Published online 06 May 2014*
doi: 10.1093/nar/gku407

A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies

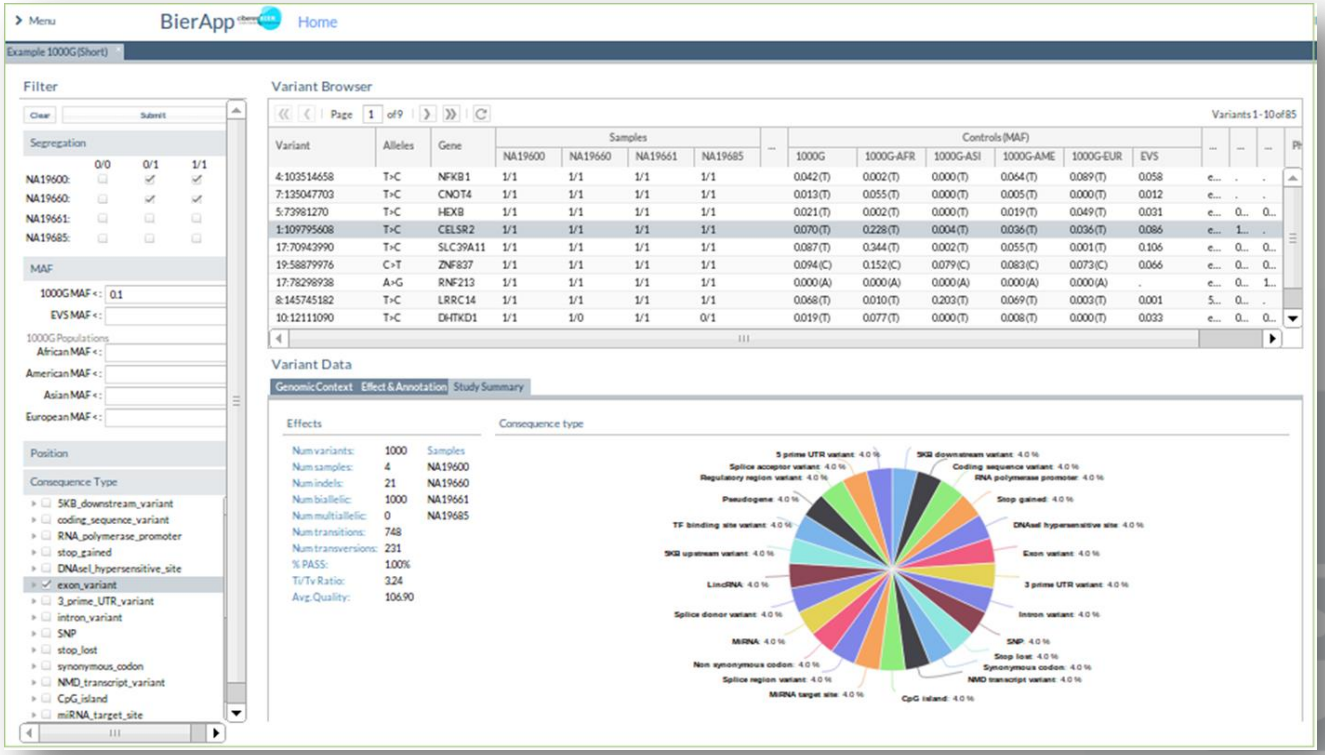
Alejandro Alemán[1,2], Francisco Garcia-Garcia[1], Francisco Salavert[1,2], Ignacio Medina[1] and Joaquín Dopazo[1,2,3,*]

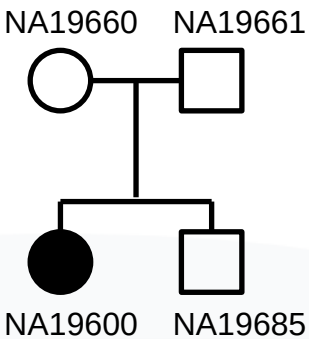
[1]Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia 46012, Spain, [2]Bioinformatics of Rare Diseases (BIER), CIBER de Enfermedades Raras (CIBERER), Valencia 46010, Spain and [3]Functional Genomics Node, (INB) at CIPF, Valencia 46012, Spain

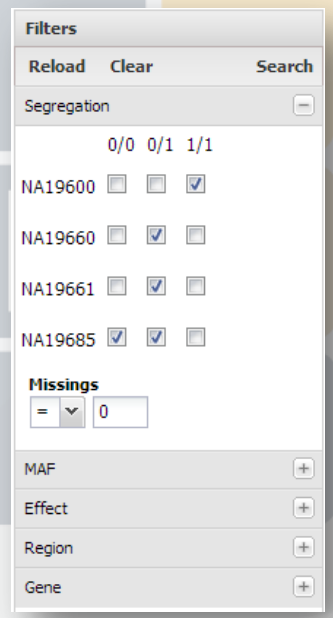
Received February 22, 2014; Revised April 27, 2014; Accepted April 28, 2014

An interactive web tool that implements different heuristic filters for disease variant/gene prioritization

The BiERapp

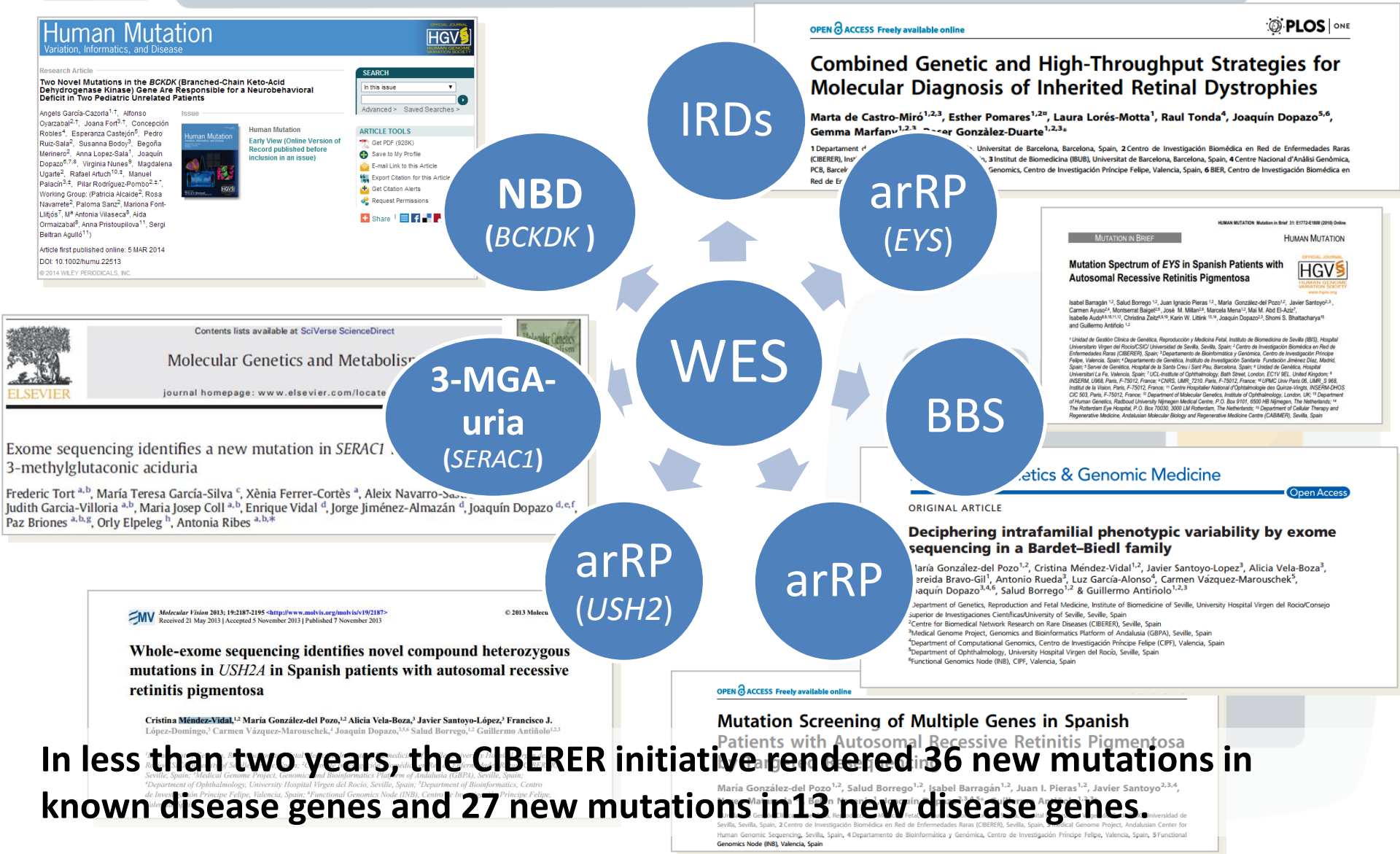






Filters include family pedigree segregation, population frequencies, pathogenic indexes, etc.

Successful use of BiERapp



In less than two years, the CIBERER initiative rendered 36 new mutations in known disease genes and 27 new mutations in13 new disease genes.

Pipeline of data analysis

Primary processing

Initial QC
FASTQ file

Mapping
BAM file

Variant calling
VCF File

Secondary analysis (Heuristic filtering)

Variant annotation

Filtering by effect

Filtering by MAF

Filtering by family segregation

Knowledge-based prioritization

Proximity to other known disease genes

Functional proximity

Network proximity

Burden tests

Other prioritization methods

Primary analysis

Gene prioritization

Knowledge-based prioritization

Network analysis

Research

Open Access

Four new loci associations discovered by pathway-based and network analyses of the genome-wide variability profile of Hirschsprung's disease

Raquel Ma Fernández[1,2], Marta Bleda[2,3], Rocío Núñez-Torres[1,2], Ignacio Medina[3,4], Berta Luzón-Toro[1,2], Luz García-Alonso[3], Ana Torroglosa[1,2], Martina Marbà[3,4], Ma Valle Enguix-Riego[1,2], David Montaner[3], Guillermo Antiñolo[1,2], Joaquín Dopazo[2,3,4]* and Salud Borrego[1,2]*

* Corresponding authors: Joaquín Dopazo jdopazo@cipf.es - Salud Borrego salud.borrego.sspa@juntadeandalucia.es

▸ Author Affiliations

For all author emails, please log on.

Orphanet Journal of Rare Diseases 2012, **7**:103 doi:10.1186/1750-1172-7-103

Published: 28 December 2012

Nucleic Acids Research Advance Access published May 19, 2009

Nucleic Acids Research, 2009, 1–6
doi:10.1093/nar/gkp402

SNOW, a web-based tool for the statistical analysis of protein–protein interaction networks

Pablo Minguez[1], Stefan Götz[1,2], David Montaner[1], Fatima Al-Shahrour[1] and Joaquin Dopazo[1,2,3,*]

[1]Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), [2]CIBER de Enfermedades Raras (CIBERER) and [3]Functional Genomics Node (INB) at CIPF, Valencia, Spain

Received January 21, 2009; Revised April 22, 2009; Accepted May 2, 2009

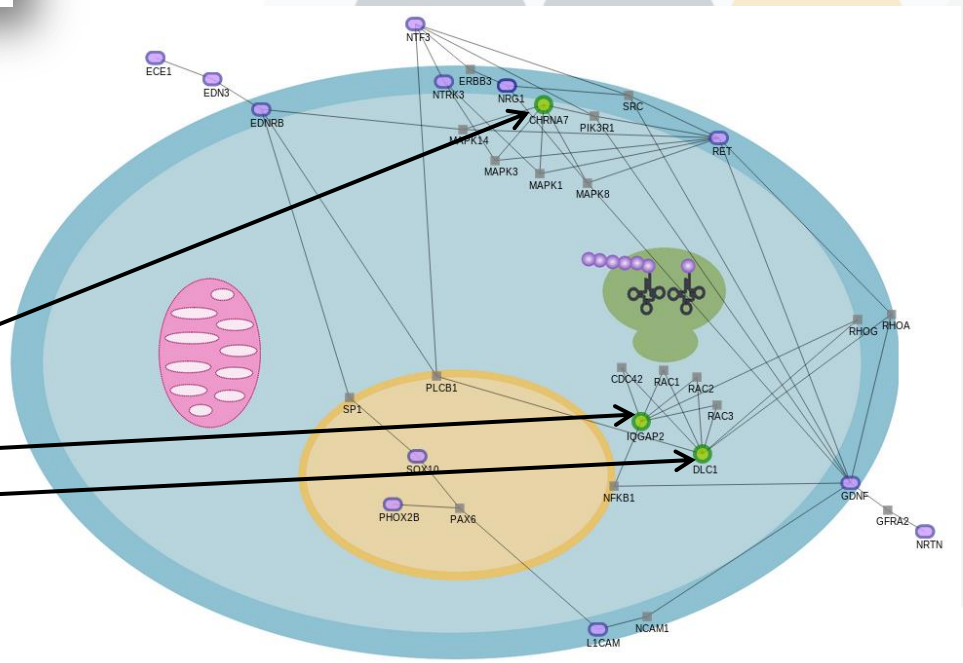
Published online 27 July 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 20 e158*
doi:10.1093/nar/gks699

Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments

Luz García-Alonso[1], Roberto Alonso[1], Enrique Vidal[1], Alicia Amadoz[1], Alejandro de María[1], Pablo Minguez[2], Ignacio Medina[1,3] and Joaquín Dopazo[1,3,4,*]

[1]Department of Bioinformatics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain, [2]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany, [3]Functional Genomics Node (INB) at CIPF, Valencia and [4]CIBER de Enfermedades Raras (CIBERER), Valencia, Spain

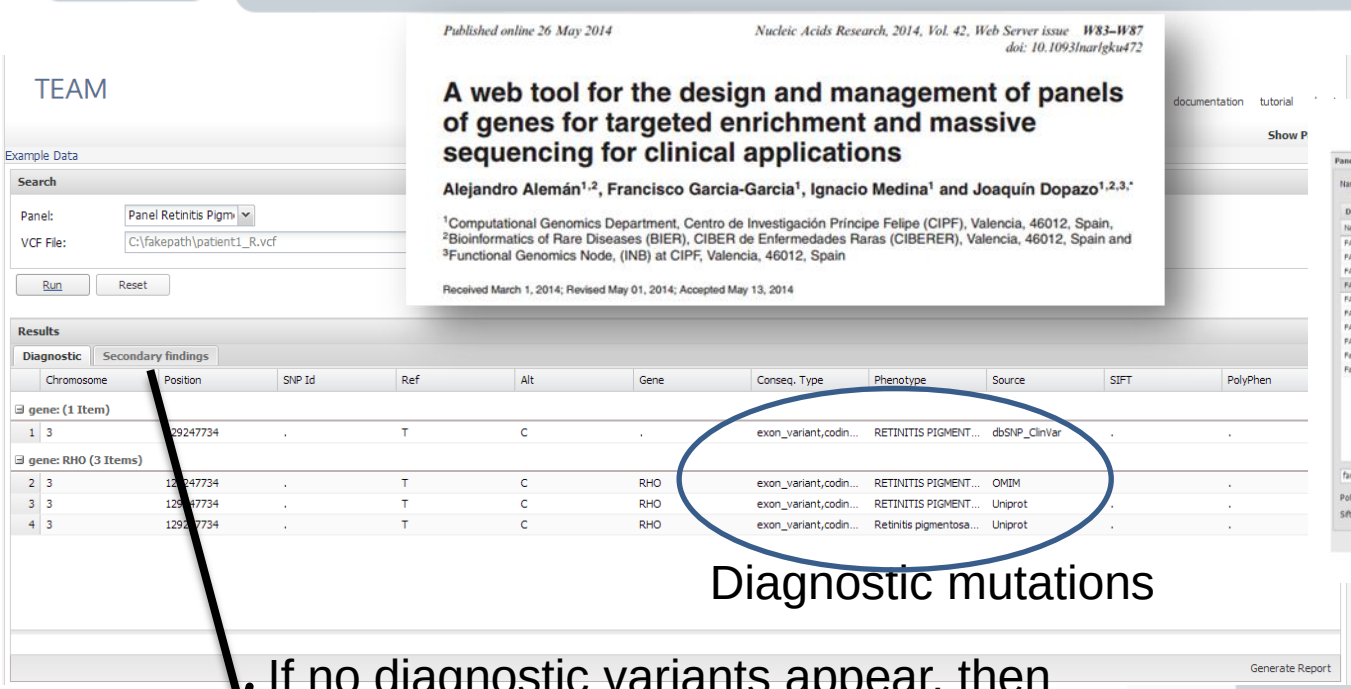
Received March 14, 2012; Revised June 1, 2012; Accepted June 26, 2012



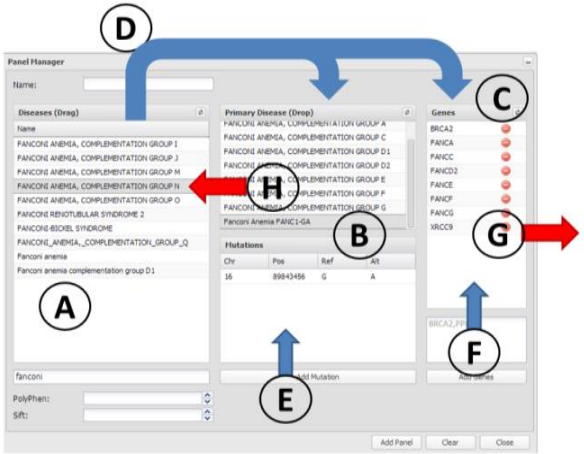
CHRNA7 (rs2175886 p = 0.000607)
IQGAP2 (rs950643 p = 0.0003585)
DLC1 (rs1454947 p = 0.007526)

SNPs validated in independent cohorts

NGS for diagnostic

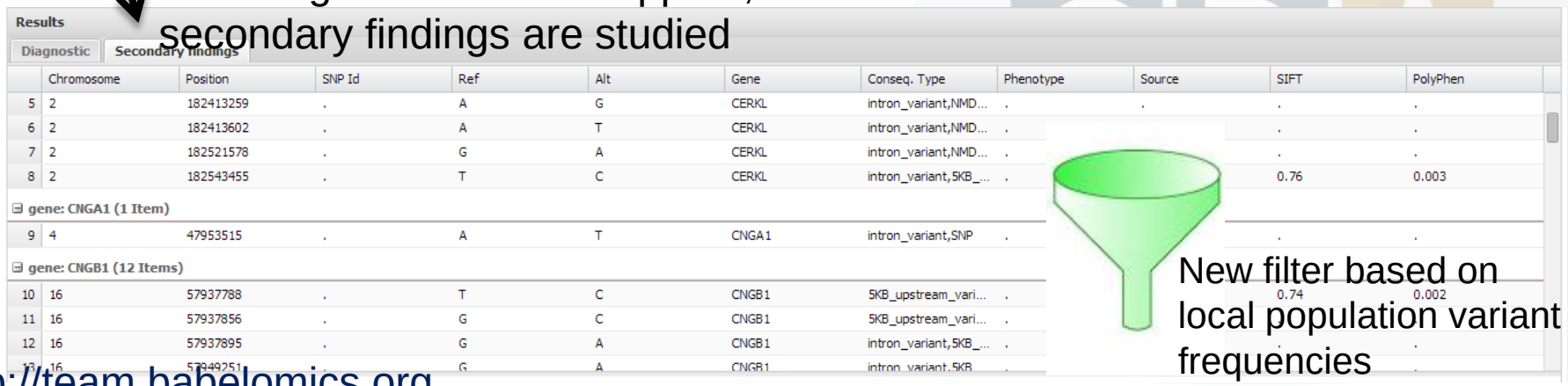


Diagnostic mutations



Tool for defining panels

If no diagnostic variants appear, then secondary findings are studied



New filter based on local population variant frequencies

http://team.babelomics.org

Uses of BiERapp and TEAM

Implementation of tools in the IT4I Supercomputing Center (Czech Republic)

The pipelines of primary and secondary analysis developed by the Computational Genomics Department of the CIPF in close collaboration with the Bull Chair of computational genomics has proven its efficiency in the analysis of more than 1000 exomes in a joint collaborative project of the CIBERER and the MGP

A first pilot implementation has been done in the IT4I supercomputing center, which aims to centralize the analysis of genomics data in the country.

IT4Innovations
národní
superpočítačové
centrum

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)
- OpenOGA indexer
- BiERapp (for gene discovery)
- TEAM (for diagnosis)

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

www.it4i.cz www.bull.cz

Course schedule

GDA2016: 29 Feb - 4 Mar

	Monday 29	Tuesday 1	Wednesday 2	Thursday 3	Friday 4
9:00-9:30	Registration				
9:30-10:00	Course presentation	Biological and Clinical Databases. *CSVS*	Panel of genes: design and analysis for clinical applications. *TEAM*	Differential Expression Analysis. *Babelomics 5*	Pathways Analysis
10:00-10:30	Introduction to NGS Technologies for Genomic Analysis				
10:30-11:00					
11:00-11:30	Coffee break				
11:30-12:00	Introduction to linux	Prioritization of variants and genes: *BiERapp*	Panel of genes: design and analysis for clinical applications. *TEAM*	Clustering and Supervised Classification Analysis. *Babelomics 5*	Pathways Analysis
12:00-12:30	Primary Analysis: Quality contol for raw data				
12:30-13:00			Closing		
13:00-13:30	Lunch				
13:30-14:00					
14:00-14:30					
14:30-15:00	Primary Analysis: Mapping NGS Reads and visualization for Genomics Studies	Prioritization of variants and genes: *BiERapp* and *Network tools*	Course presentation	Functional Profiling from *Babelomics 5*: FatiGO and Gene Set Analysis	Enrichment Analysis for microRNAS
15:00-15:30			Introduction to NGS Technologies for Expression Analysis		
15:30-16:00					
16:00:16:30	Coffee break				
16:30-17-00	Primary Analysis: Variant Calling SNPs and INDELs. Variant Annotation	Bring your own data!	Primary Analysis for RNA-Seq data: quality control, mapping and quantification	Functional Profiling from *Babelomics 5*: Network Analysis	Bring your own data!
17:00-17:30					
17:30-18:00					Closing
21:00		SOCIAL DINNER		SOCIAL DINNER	

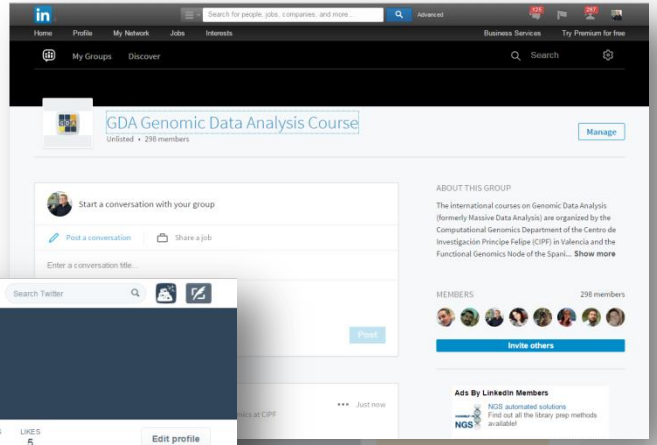
GENOMICS: 29 February, 1 and 2 March

TRANSCRIPTOMICS: 2, 3 and 4 March

Social

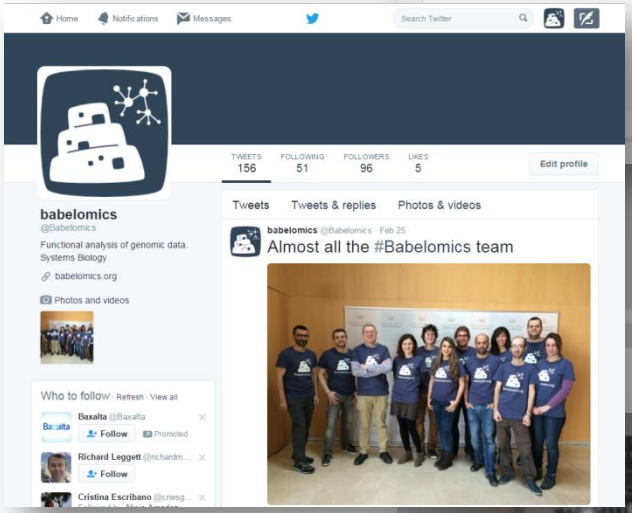
GDA group in Linked-in

https://www.linkedin.com/groups/1934338



Babelomics group in twitter

@xdopazo
@bioinfocipf
@babelomics



http://bioinfo.cipf.es
http://www.babelomics.org

And the social dinner (and *mascletás*)… Networking is very important for your career. Keep in touch with fellows and instructors