# A NGS data analysis course

# Introduction


## Joaquín Dopazo

Computational Genomics Department,
Centro de Investigación Príncipe Felipe (CIPF),
Functional Genomics Node, (INB),
Bioinformatics Group (CIBERER) and
Medical Genome Project,
Spain.

http://bioinfo.cipf.es
http://www.medicalgenomeproject.com
http://www.babelomics.org
http://www.hpc4g.org
@xdopazo

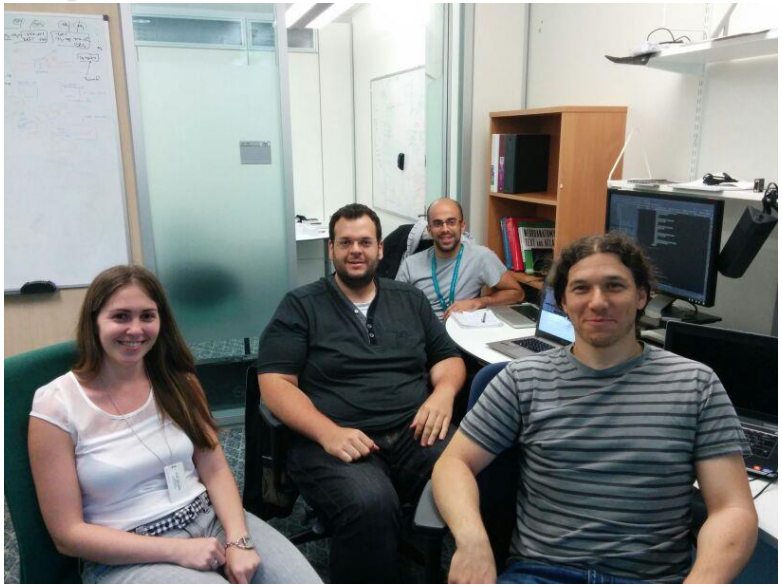

*University of Cambridge, 23-25 February 2015*

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

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

@xdopazo
@bioinfocipf

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



… Computational Biology Lab, HPCS University of Cambridge, UK…

…and EMBL-EBI

Machine Learning for Personalized Medicine

Fundació La Marató de TV3

Proyectos PROMETEO Generalitat Valenciana

MGP Medical Genome Project

HPC4G

INB

er ciberer CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED DE ENFERMEDADES RARAS

MINISTERIO DE CIENCIA E INNOVACIÓN

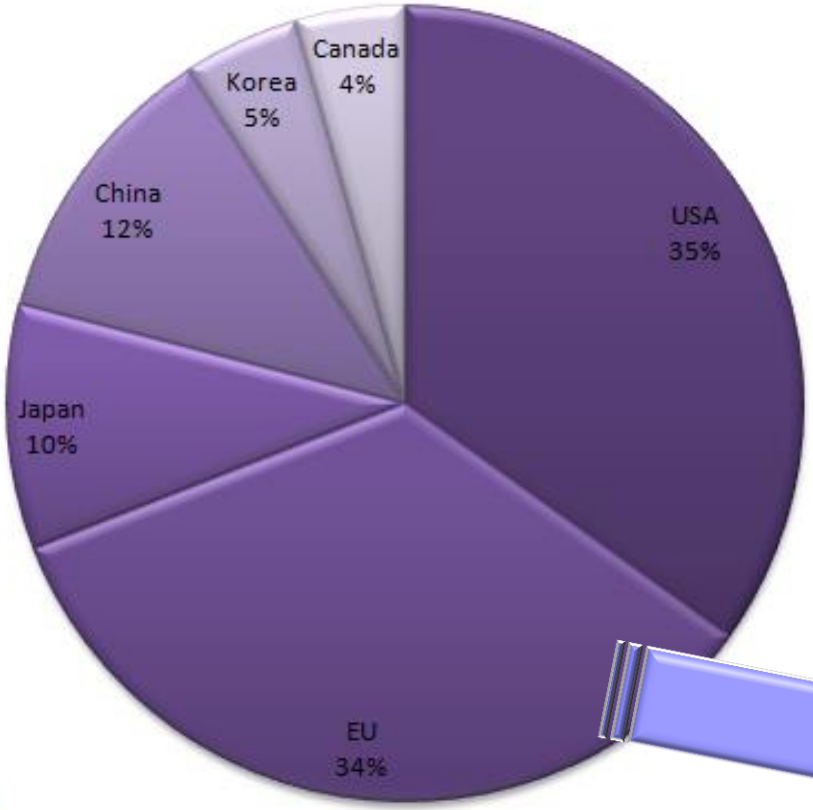
F.E.D.E.R.

ciberer BIER

BULL STATegra

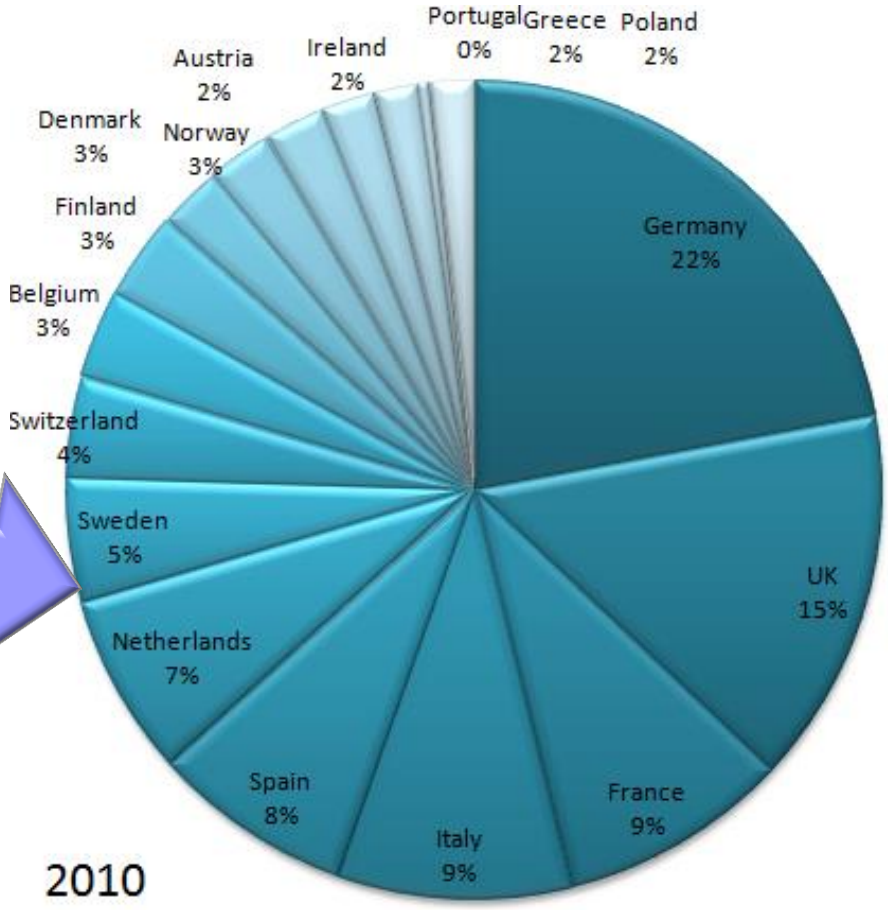
iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY AFFILIATED REGIONAL INSTITUTE

# Some bibliographic data: Microarray publications



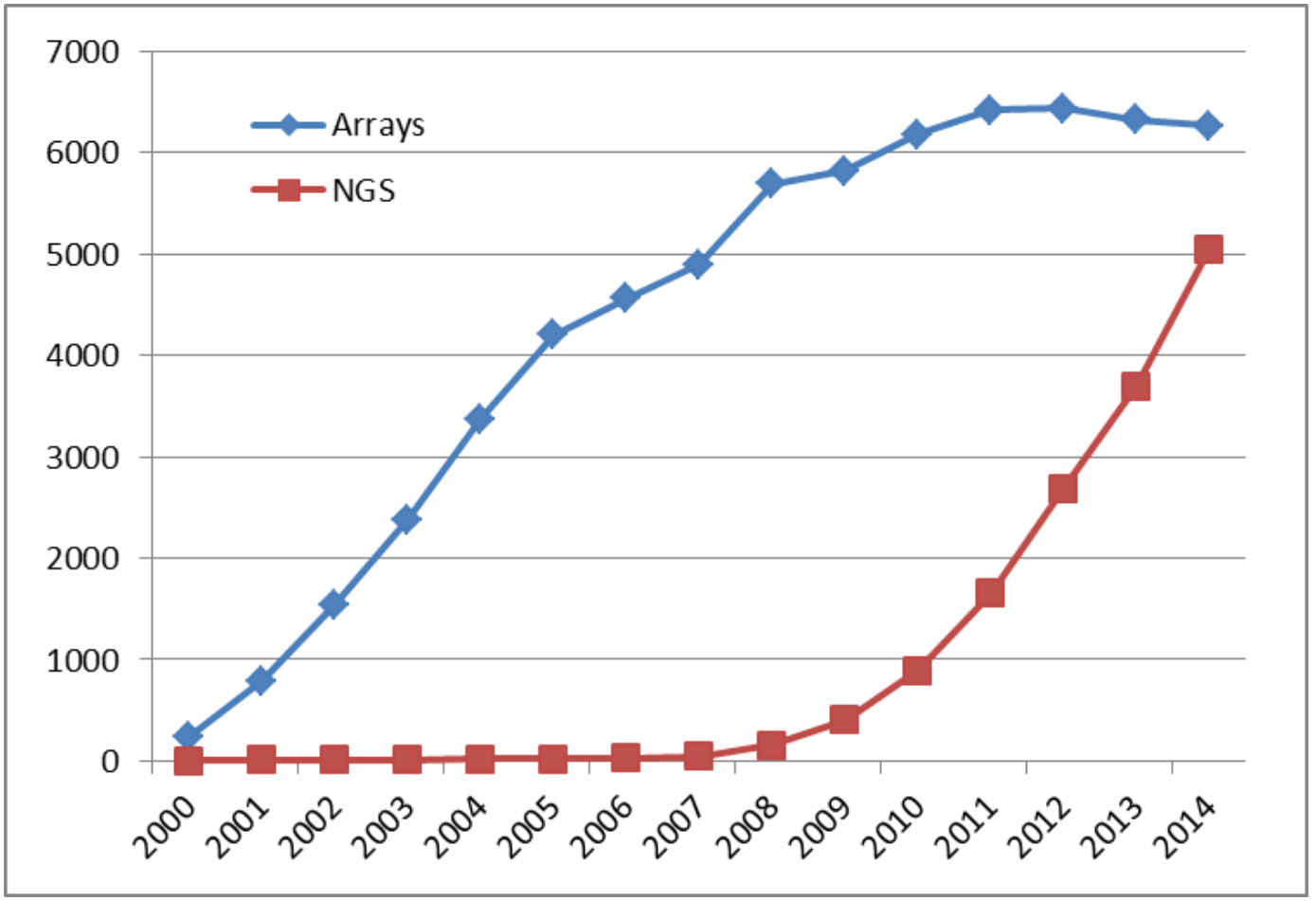

2010 Worlwide

2010 Europe



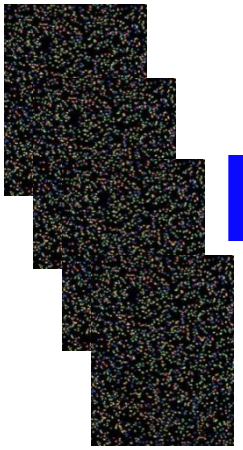

**Source Pubmed. Query: 2009[Entrez Date] AND country[Affiliation]AND microarray[Title/Abstract]**

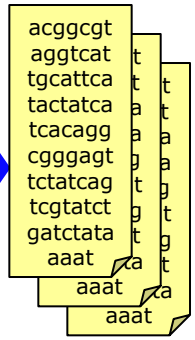
# Evolution of the papers published in microarray and NGS technologies



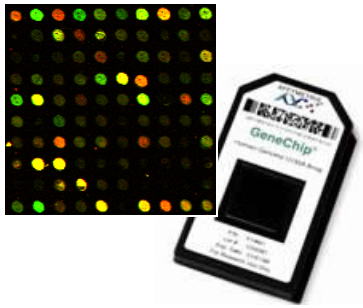

**Source Pubmed. Query:** "high-throughput sequencing"[Title/Abstract] OR "next generation sequencing"[Title/Abstract] OR "rna seq"[Title/Abstract]) AND year[Publication Date]

# Genomic data, the double challenge: Data processing and interpretation

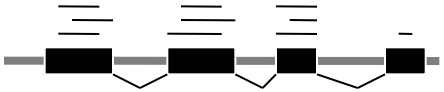




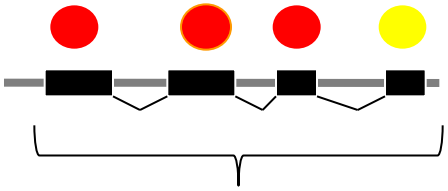

Fasta files and QC files




Raw image generation

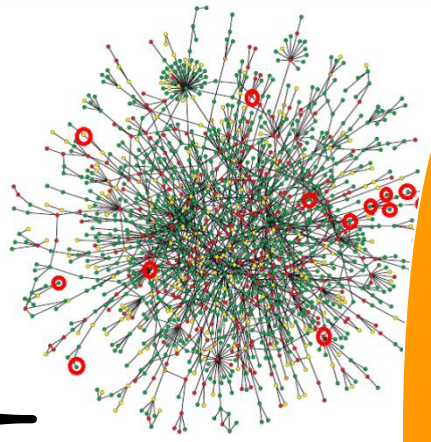


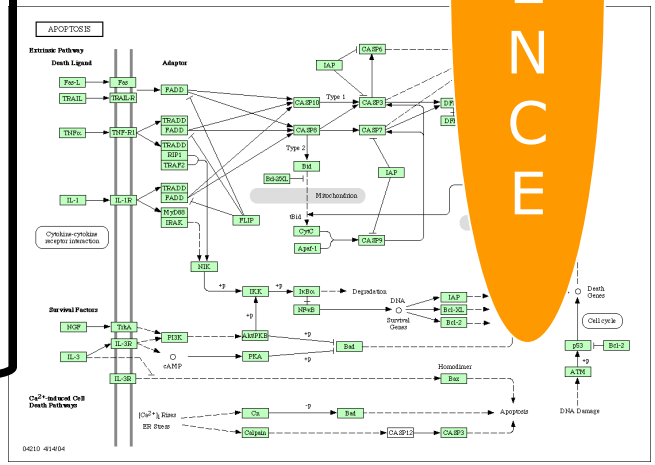
Read mapping and quantification



Intensity summary

Transcriptomics



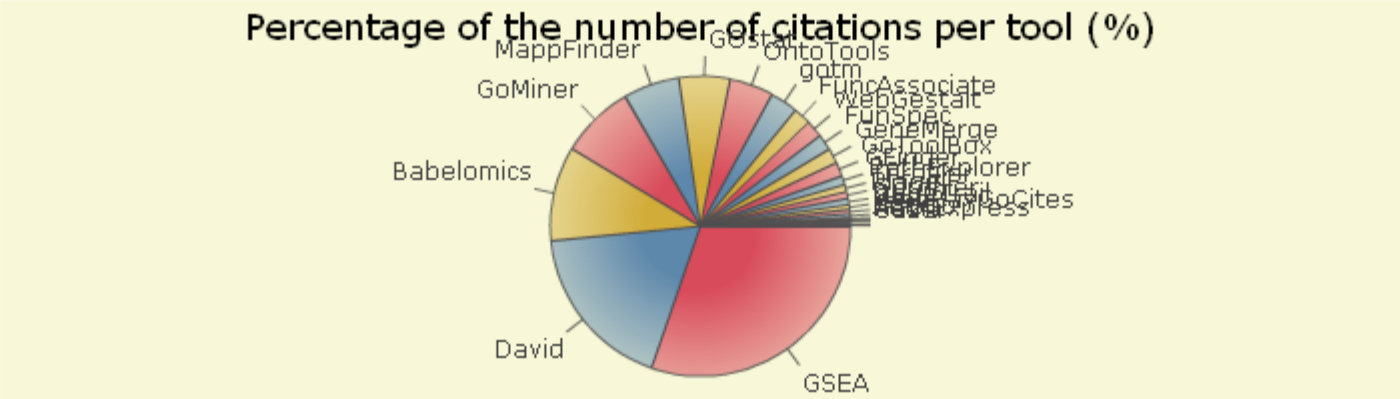


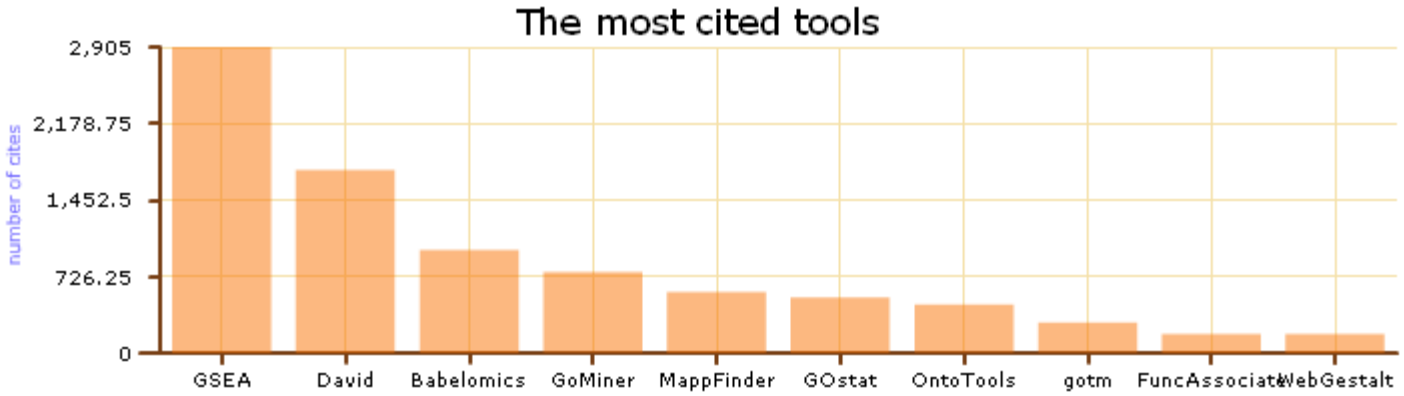

SCIENCE

Technology driven

Hypothesis driven

# Tools for genomic data analysis and functional profiling



# Some numbers

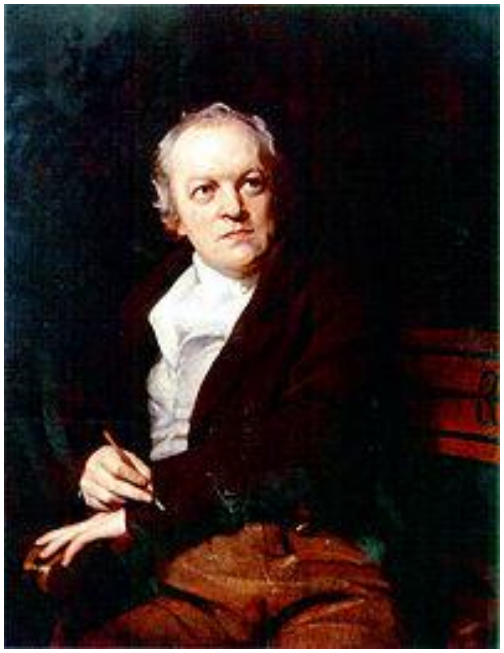
529 papers cite Babelomics (plus 798 FatiGO cites)

*(source ISI Web of Knowledge, Februry2015)*

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

More than 1000 experiments per day.

# Background



**The road of excess leads to the palace of wisdom**

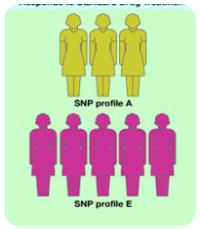
(*William Blake, 28 November 1757 – 12 August 1827, poet, painter, and printmaker*)

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…

# Where do we come from?
# The pre-genomics paradigm

Genes in the DNA...




...produces the final phenotype

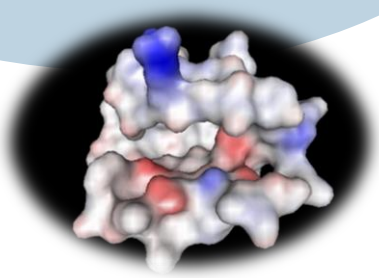
...code for proteins...

```
>protein kunase
acctgttgatggcgacagggactgtatgctg
atctatgctgatgcatgcatgctgactactga
tgtgggggctattgacttgatgtctatc....
```

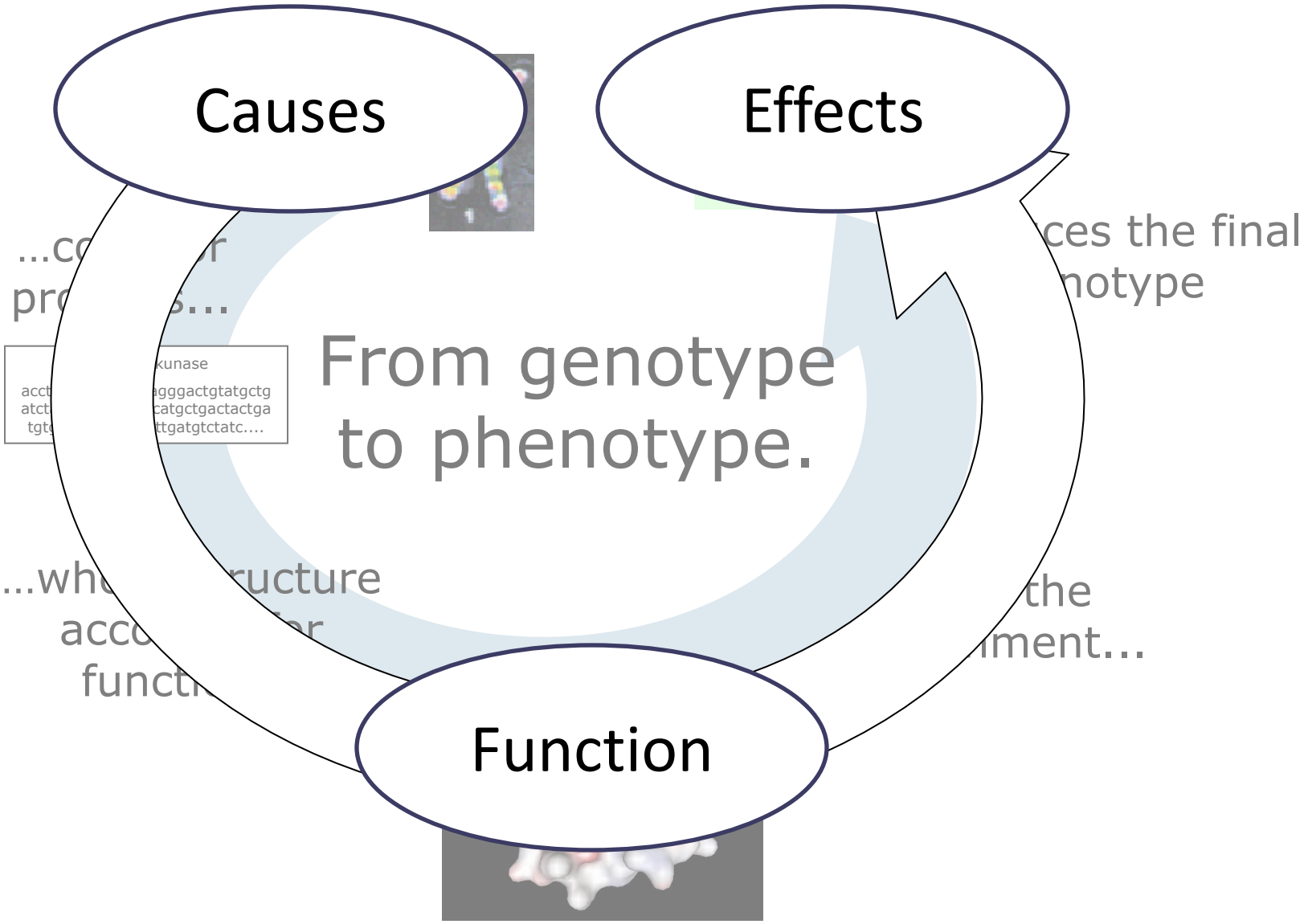
**From genotype to phenotype.**

...whose structure accounts for function...

...plus the environment...



# Reduccionistic approach to link causes (genome) to effects (phenotype) through actions (function)

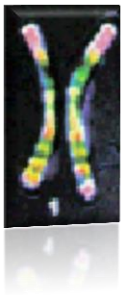


Next Generation Sequencing
$10^9$bp per round

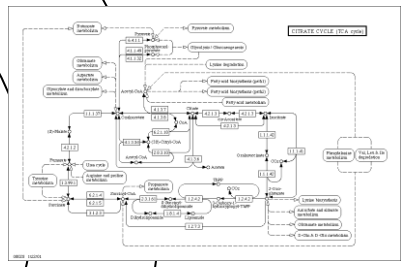
Genes in the DNA...


12 million SNPs in exonic regions

...with its complex variability...



...whose final effect configures the phenotype...

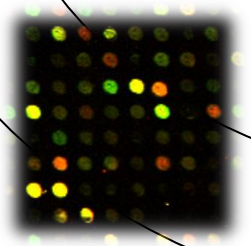


# From genotype to phenotype.

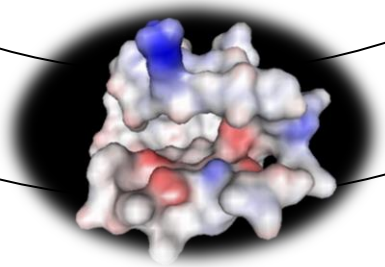
**(in the post-genomics scenario)**

...conforming complex interaction networks...

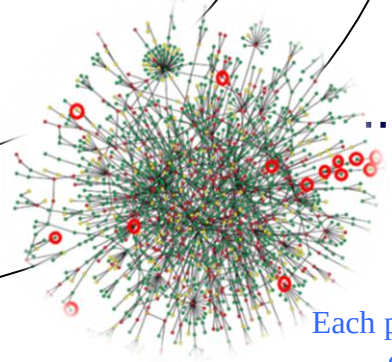
...when they are expressed in the proper moment and place...



...code for proteins...



That undergo post-translational modifications, somatic recombination...
100K-500K proteins

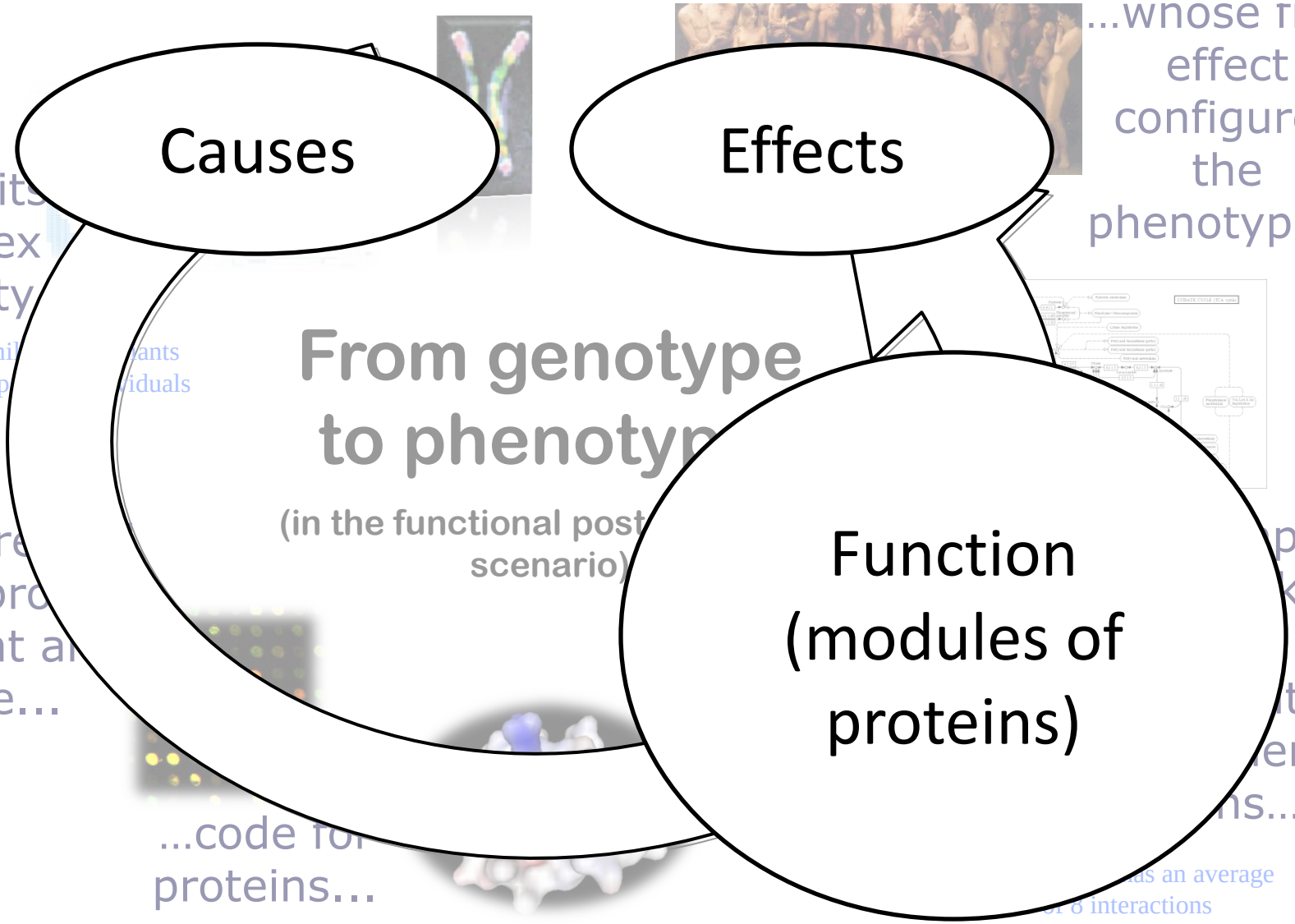


...in cooperation with other proteins...

Each protein has an average of 8 interactions

...that account for function if...

# Holistic approach. Causes and effects remain essentially the same. The concept of function has changed



# High-throughput data for functional genomics

Next Generation Sequencing
$10^9$bp per round

Genes in the DNA...

...whose final effect configures the phenotype...

Genome wide

...with its complete variability...

Half a million of variants between pairs of individuals

Genotyping

From genotype to phenotype.

(in the functional post-genomics scenario)

...when they are expressed in the proper moment and place...

Metabolomics

...conforming complex interaction networks...

Transcriptomics

...code for proteins...

Proteomics

...in cooperation with other proteins...

Each protein has an average of 6 interactions

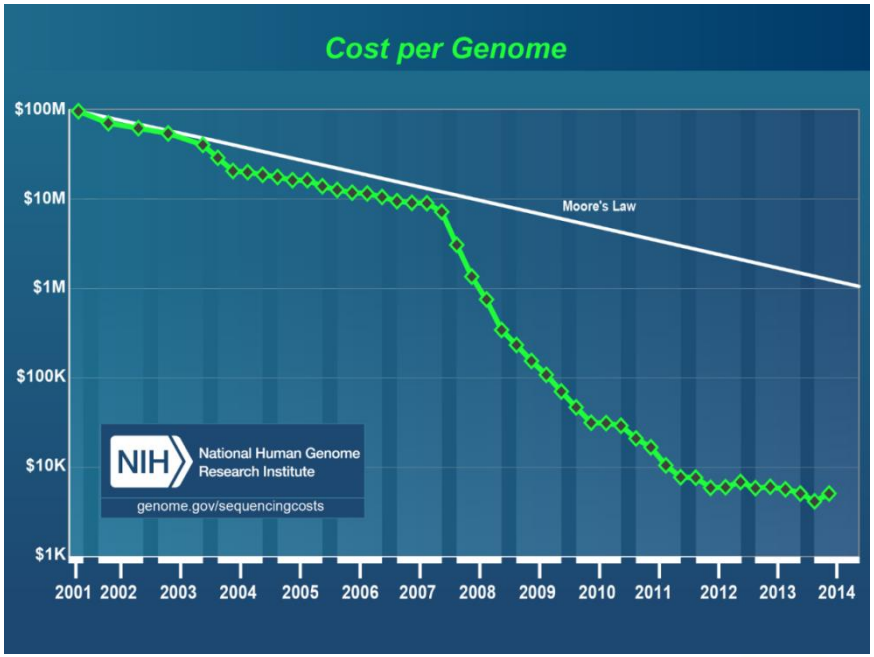
That undergo post-translational modifications, somatic recombination...
100K-500K proteins

...that account for function if...

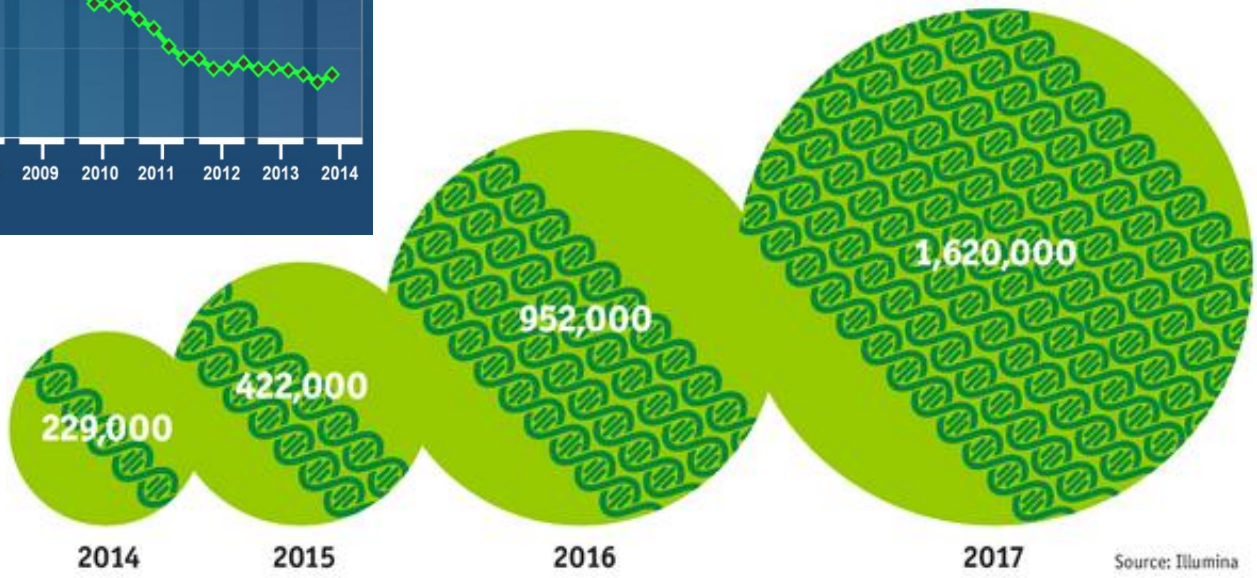
Almost-omics

# The genome sequencing pace

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



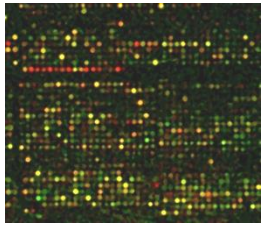

NGS is matching the cost of many conventional clinical tests




*http://www.economist.com/news/21631808-so-much-genetic-data-so-many-uses-genes-unzipped*

# Gene expression profiling. Historic perspective

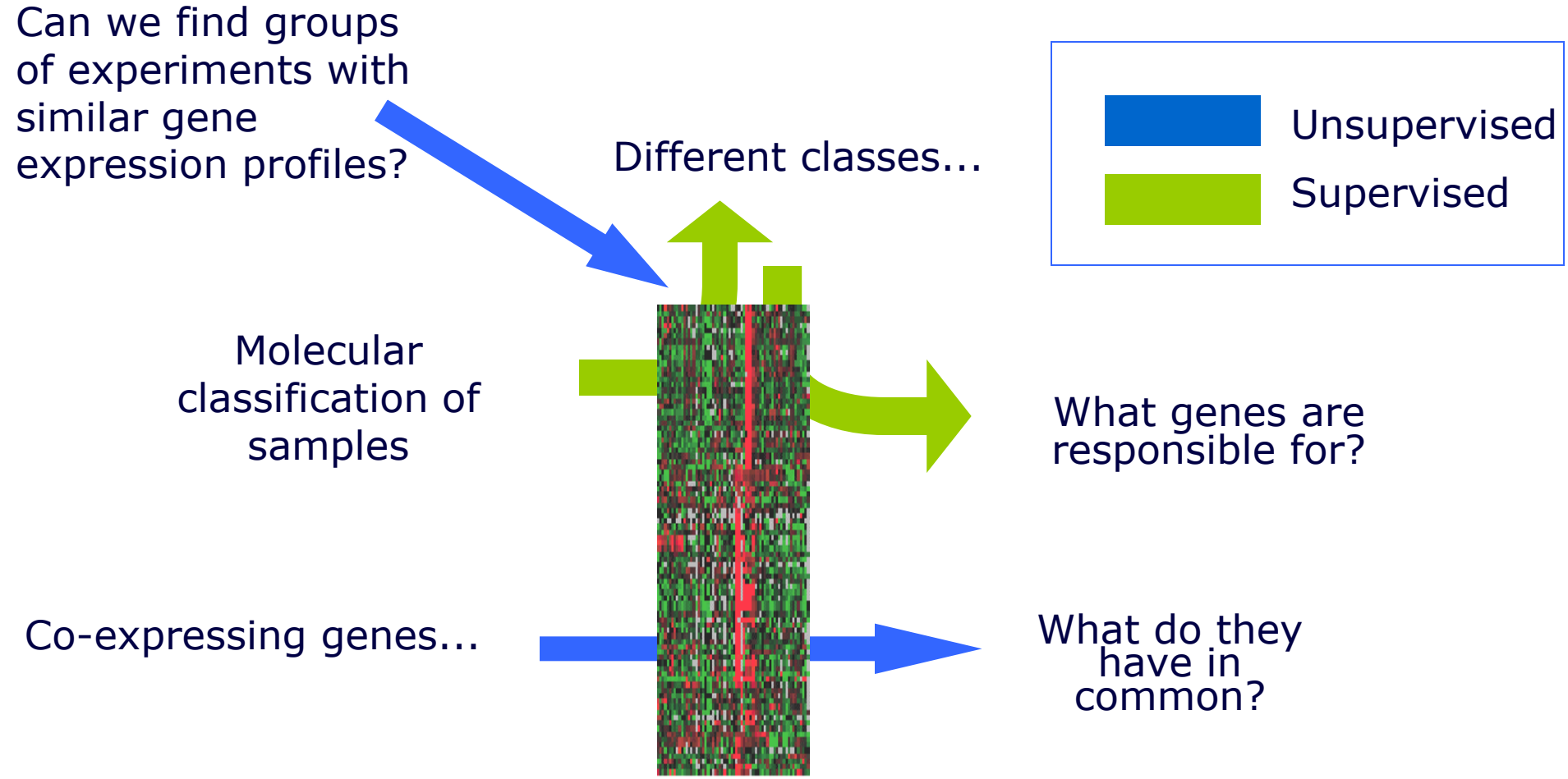
Differences at phenotype level are the visible cause of differences at molecular level which, in many cases, can be detected by measuring the levels of gene expression. The same holds for different experiments, treatments, strains, etc.



- Classification of phenotypes / experiments. Can we distinguish among classes (either known or unknown), values of variables, etc. using molecular gene expression data? (sensitivity)
- Selection of differentially expressed genes among the phenotypes / experiments. Did we select the relevant genes, all the relevant genes and nothing but the relevant genes? (specificity)
- Biological roles the genes are carrying out in the cell. What general biological roles are really represented in the set of relevant genes? (interpretation)

# Studies must be hypothesis driven.

What is the aim? Class discovery? sample classification? gene selection? ...

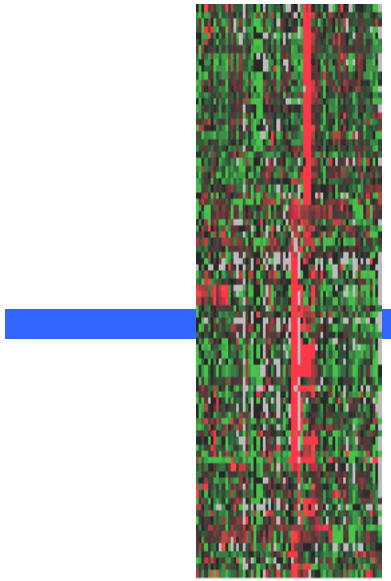


# Unsupervised problem: class discovery

Our interest is in discovering clusters of items (genes or experiments) which we do not know beforehand

Can we find groups of experiments with similar gene expression profiles?

Co-expressing genes...



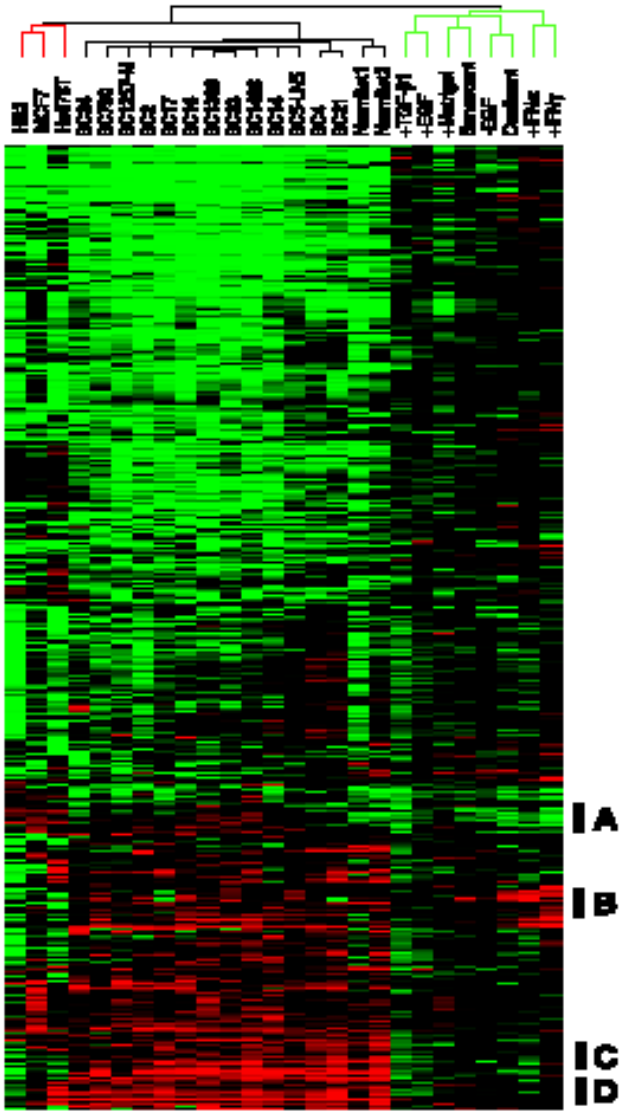
- What genes co-express?
- How many different expression patterns do we have?
- What do they have in common?
- Etc.

# Clustering of experiments: The rationale

If enough genes have their expression levels altered in the different experiments, we might be able of finding these classes by comparing gene expression profiles.

**Distinctive gene expression patterns in human mammary epithelial cells and breast cancers**

Overview of the combined *in vitro* and breast tissue specimen cluster diagram. A scaled-down representation of the 1,247-gene cluster diagram The black bars show the positions of the clusters discussed in the text: (*A*) proliferation-associated, (*B*) IFNregulated, (*C*) B lymphocytes, and (*D*) stromal cells.




*Perou et al., PNAS (1999)*

# Supervised problems.

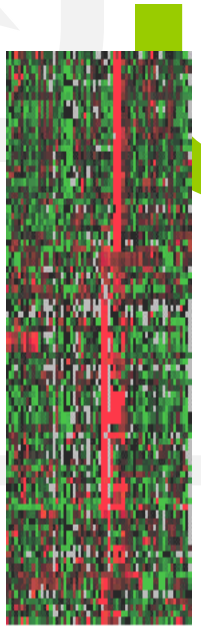
## Differential gene expression

Can we find groups of experiments with similar gene expression profiles?

Molecular classification of samples

Co-expressing genes...

**Different classes...**



**What genes are responsible for?**

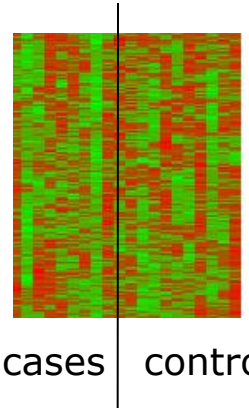
What do they have in common?

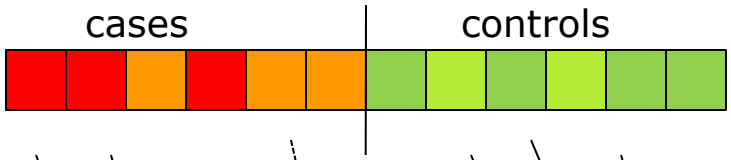
# Differential gene expression

The simplest way: univariant gene-by-gene.
Other multivariant approaches can be used

- **Two classes**
  - T-test
  - Limma
  - Fold-change
- **Multiclass**
  - Anova
  - Limma
- **Continuous variable (e.g. level of a metabolite)**
  - Pearson
  - Spearmam
  - Regression
- **Survival**
  - Cox model
- **Time Course**

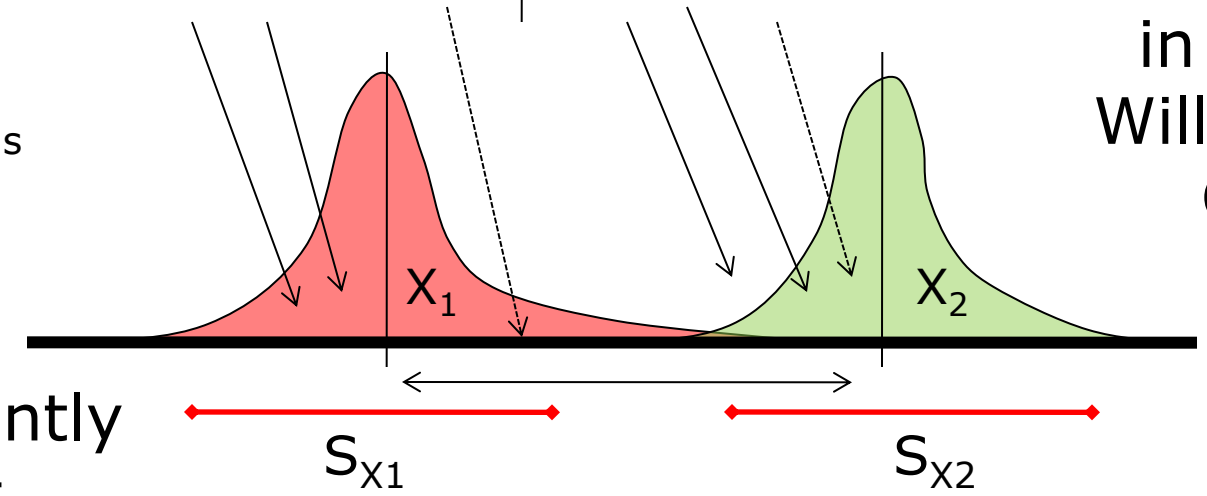
# Differential gene expression



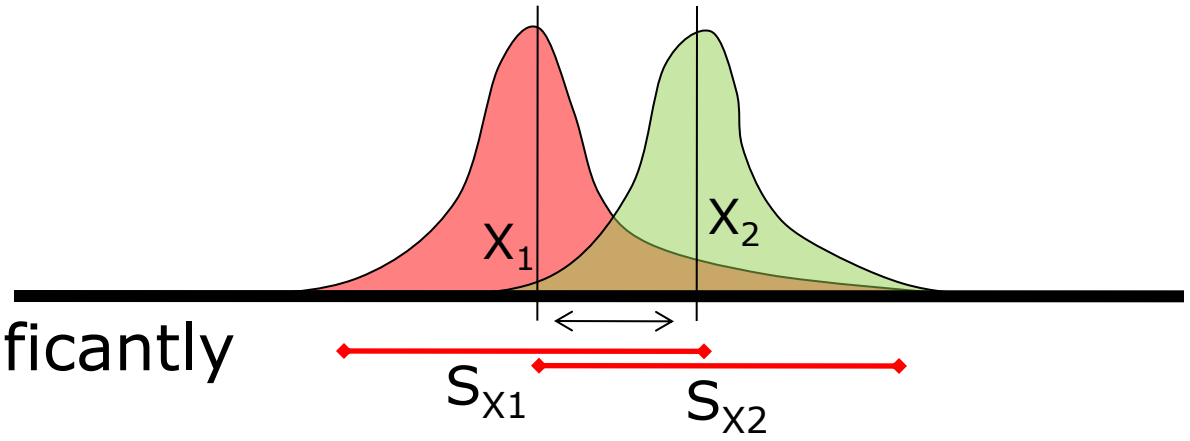




The t-statistic was introduced in 1908 by William Sealy Gosset

Significantly different




Non significantly different




$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}} \quad \text{being} \quad S_{X_1X_2} = \sqrt{\frac{S^2_{X_1} + S^2_{X_2}}{2}}.$$

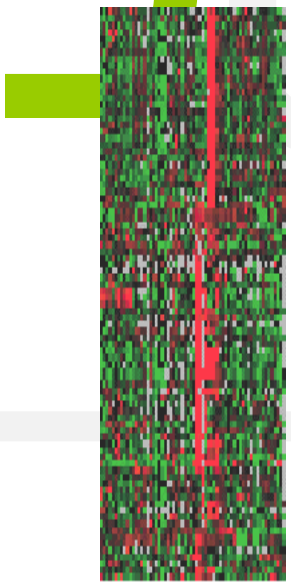
# Supervised problems.

## sample classification

Can we find groups of experiments with similar gene expression profiles?

Different classes...

Molecular classification of samples



What genes are responsible for?

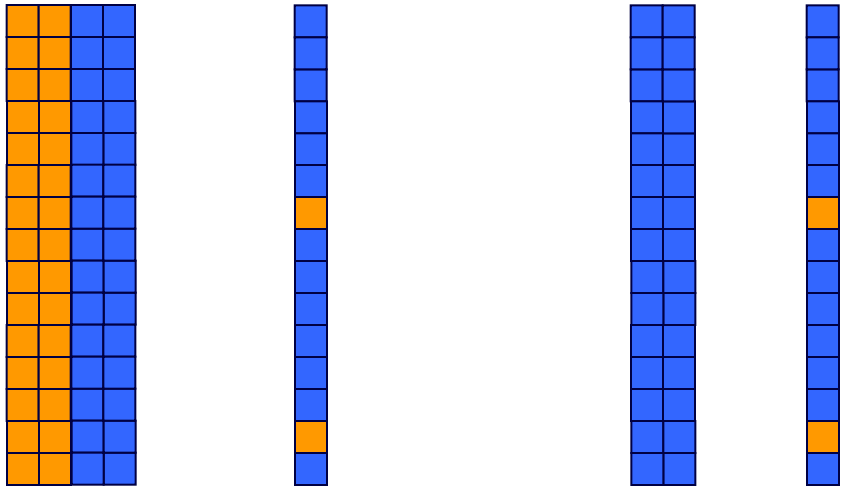
Co-expressing genes...

What do they have in common?

# Predictors and molecular signatures

What is a predictor?

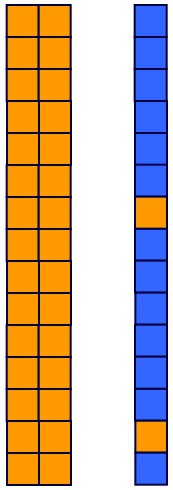
Intuitive notion:




Is X, A or B?



Diff (B, X) = 2



Diff (A, X) = 13

Most probably X belongs to class B

Algorithms: DLDA, KNN, SVM, random forests, PAM, etc.

# Predictor of clinical outcome in breast cancer

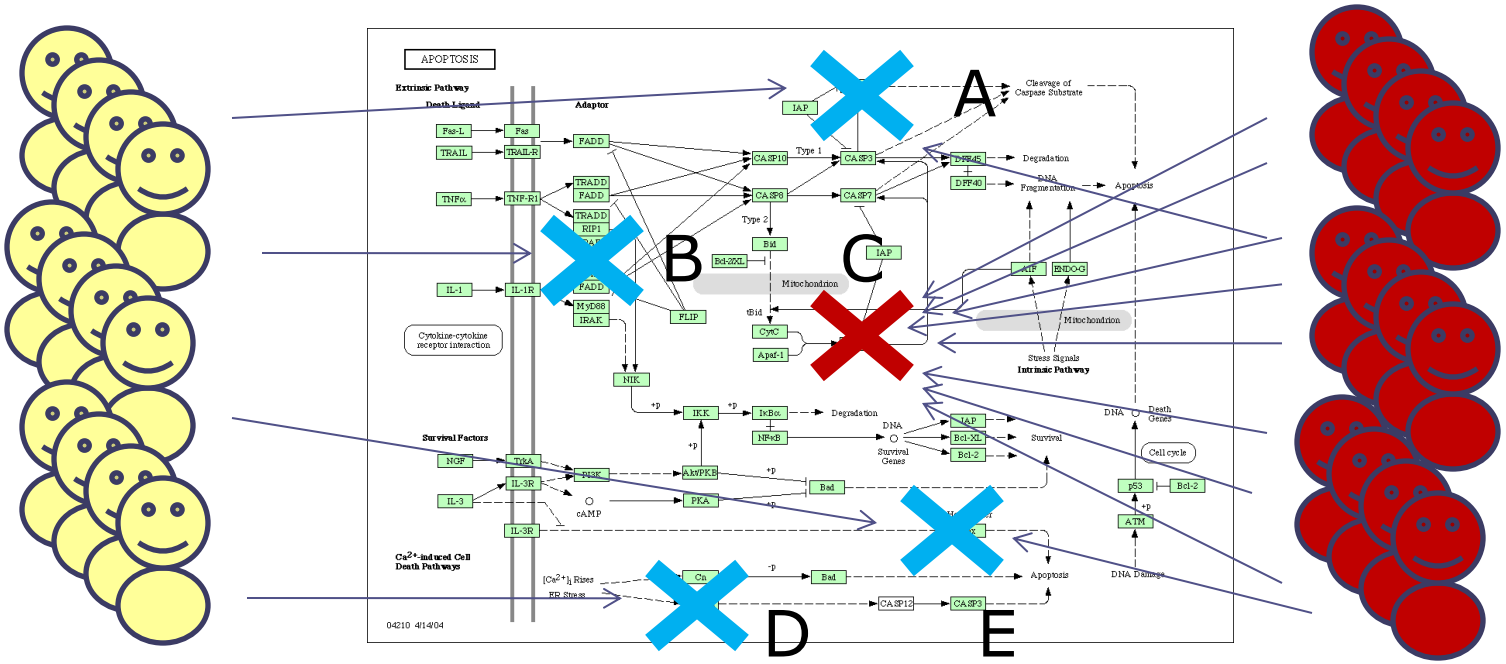

Genes are arranged to their correlation eith the pronostic groups

← Pronostic classifier with optimal accuracy

*van't Veer et al., Nature, 2002*

# Genotyping: finding mutations associated to diseases

The simplest case: monogenic disease

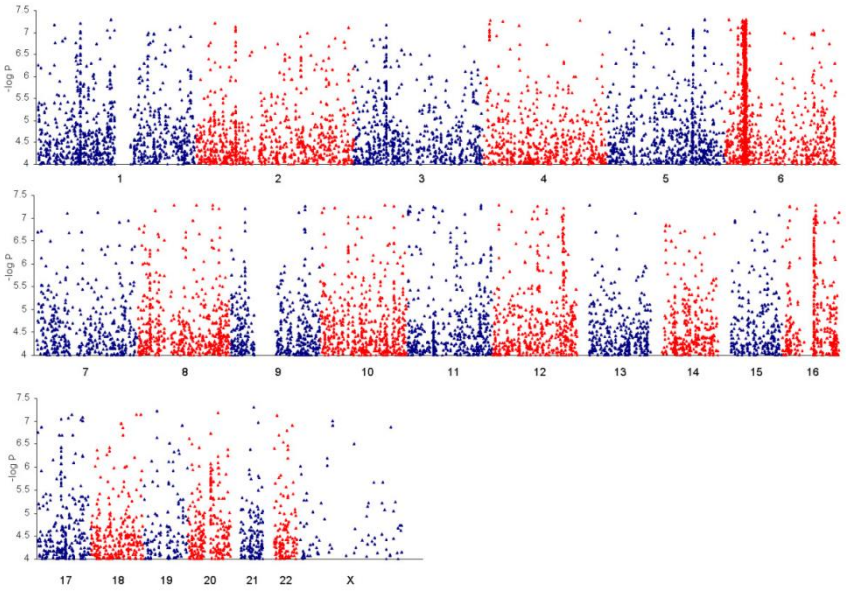


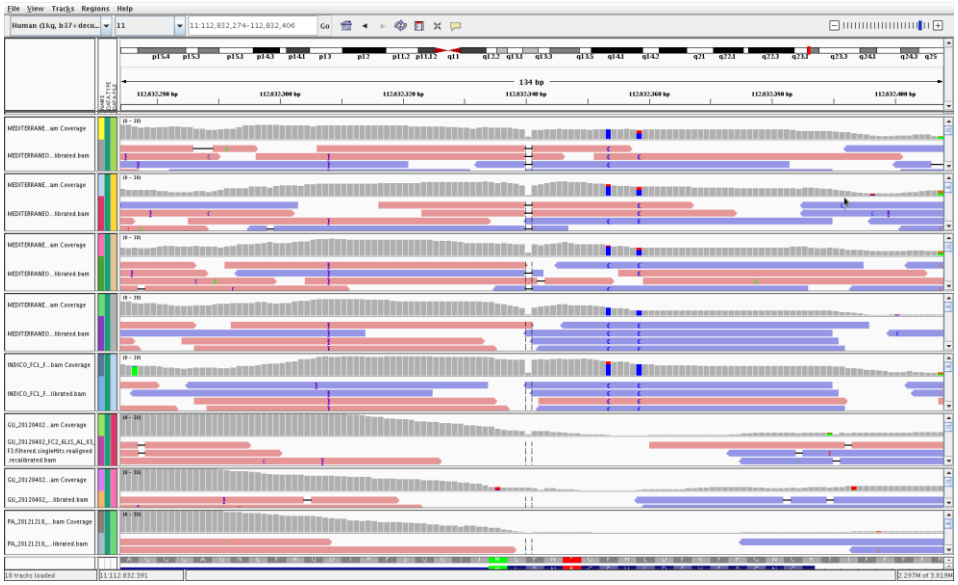

Controls

Cases

| | | |
|---|---|---|
| Gene A | 1 0 0 0 0 0 0 0 0 0 0 0 0 0 | 0 0 0 0 1 0 0 0 0 0 0 0 0 0 |
| Gene B | 0 0 0 1 0 0 0 0 0 0 0 0 0 0 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 |
| Gene C | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | 1 1 1 1 1 1 1 1 1 1 1 1 1 1 |
| Gene D | 0 0 0 0 0 0 0 0 0 0 1 0 0 0 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 |
| Gene E | 0 0 0 0 0 1 0 0 0 0 0 0 0 0 | 0 0 0 0 0 0 0 1 0 0 0 0 0 0 |

# Genotyping and nowadays NGS



# Primary data analysis tools

**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

**Variant calling analysis**

GATK and SAM mPileup HPC Implementation.
Statistics genomic tests

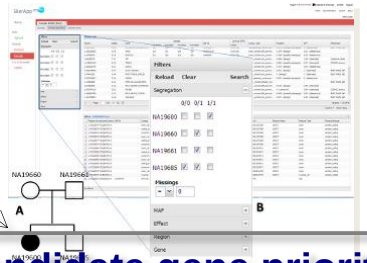
**VCF file**

**QC and preprocessing**

QC stats, filtering and preprocessing options

**Genome Maps. Variant VCF viewer**

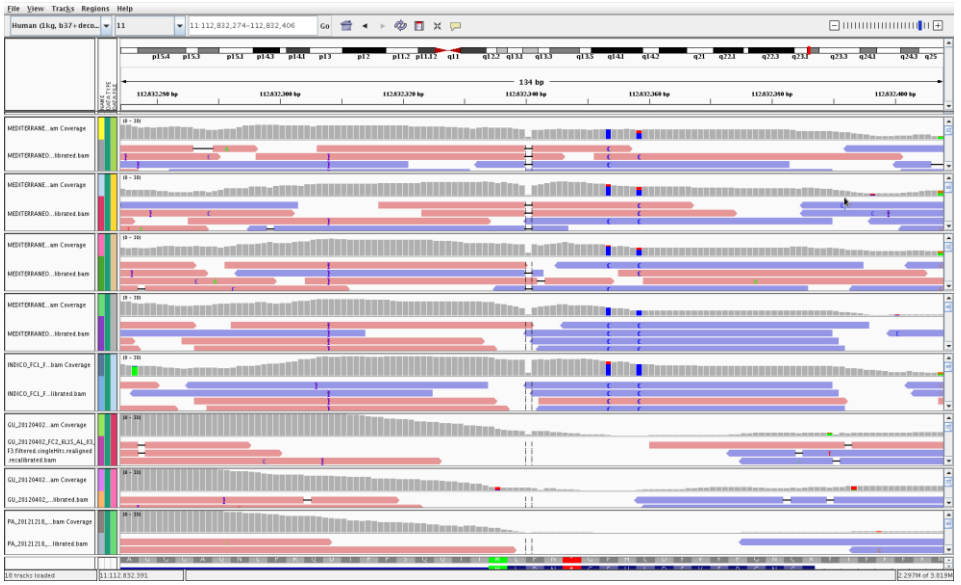
HTML5+SVG Web based viewer



**BiERApp candidate gene prioritization**

**Consequence type, pathogenicity, Population frequencies, etc**

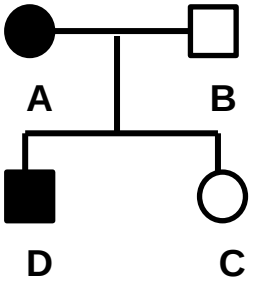
# The principle: comparison of patients to reference controls or segregation within families

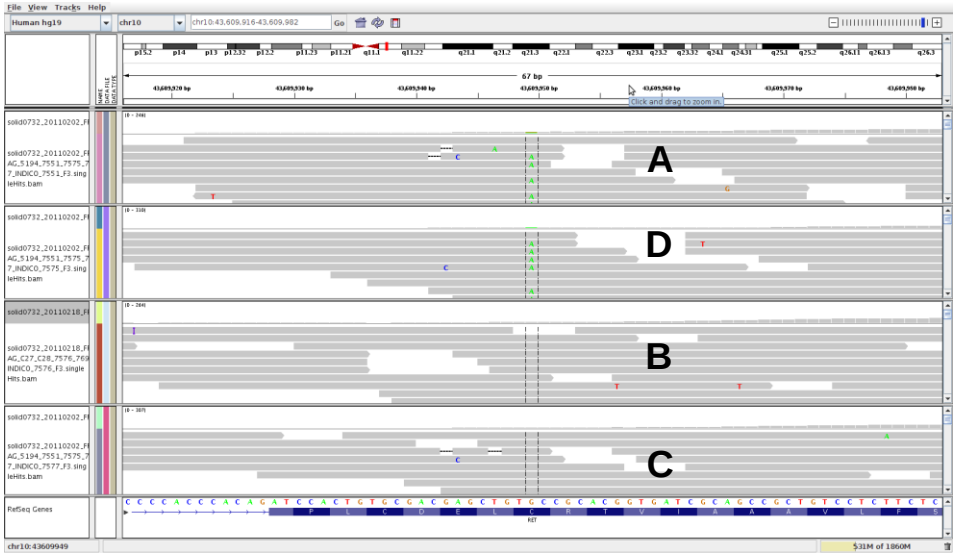


Cases

Controls

Segregation within a pedigree



# Variant/gene prioritization by successive filtering

## Variant level

Potential impact of the variant

Population frequencies

## Experimental design level

Family(es)
Trios
Case / control

## Functional (system) level

Gene set
Network analysis
Pathway analysis

Control of sequencing errors (missing values)

Testing strategies

# Pipeline of data analysis

| Initial QC | Mapping + QC | Variant calling + QC | Variant and gene prioritization + QC |
|---|---|---|---|
| Sequence cleansing<br>Base quality<br>Remove adapters<br>Remove duplicates<br><br>**FASTQ file** | Mapping (HPG)<br>Remove multiple mapping reads<br>Remove low quality mapping reads<br>Realigning<br>Base quality recalibrating<br><br>**BAM file** | Calling and labeling of missing values<br>Calling SNVs and indels (GATK) using 6 statistics based on QC, strand bias, consistence (poor QC callings are converted to missing values as well)<br>Create multiple VCF with missing, SNVs and indels<br><br>**VCF file** | Counts of sites with variants<br>Variant annotation (function, putative effect, conservation, etc.)<br>Inheritance analysis (including compound heterozygotes in recessive inheritance)<br>Filtering by frequency with external controls (**Spanish controls**, dbSNP, 1000g, 5500g) and annotation<br>Multi-family intersection of genes and variants<br>Network-based prioritization<br>**Report** |

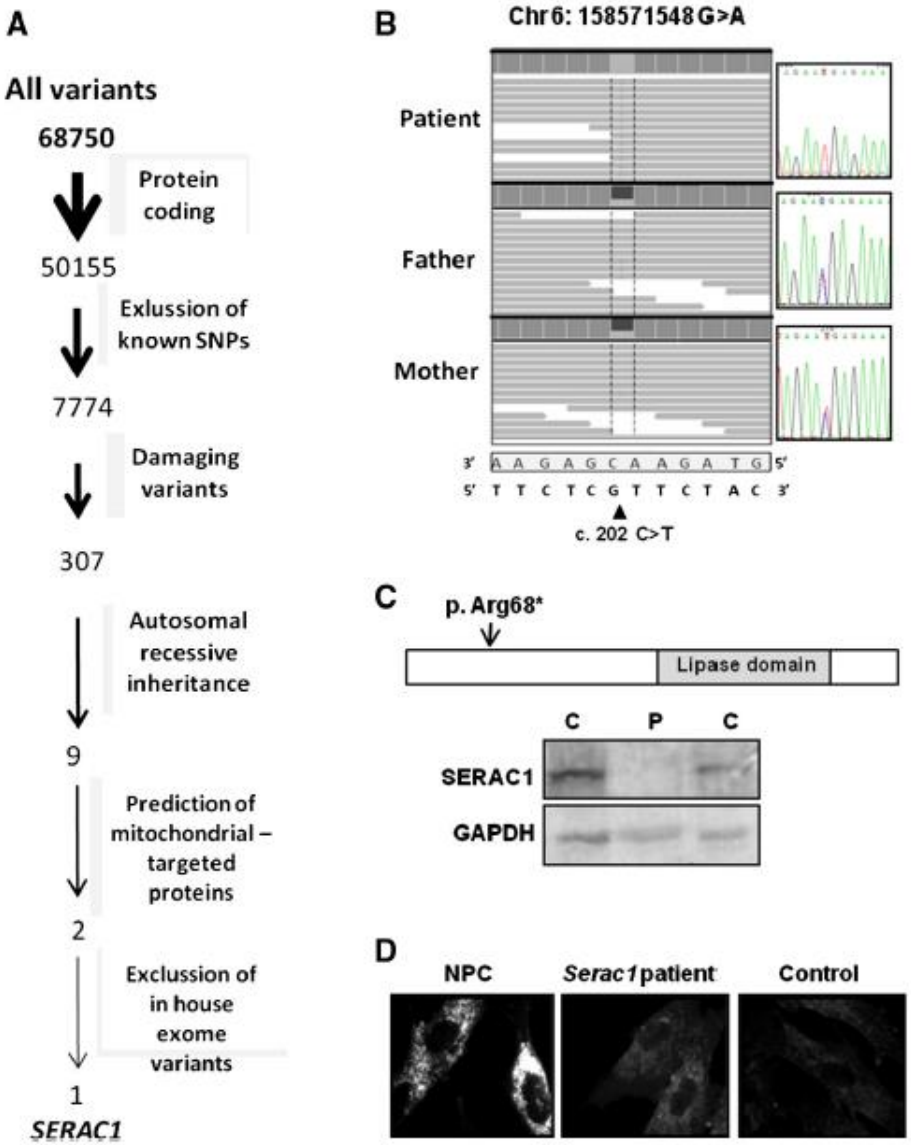
Primary analysis

Gene prioritization

# Successive Filtering approach

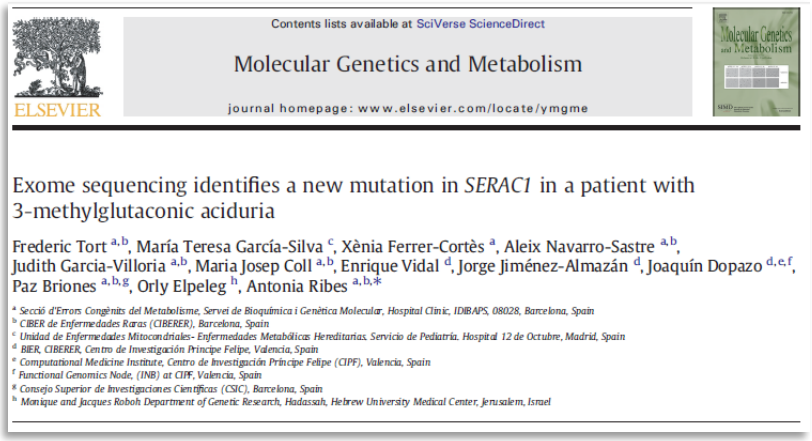
## An example with 3-Methylglutaconic aciduria syndrome

F. Tort et al. / Molecular Genetics and Metabolism xxx (2013) xxx–xxx




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

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

# Exome sequencing has been systematically used to identify Mendelian disease genes

ARTICLES

nature genetics

## Exome sequencing identifies the cause of a mendelian disorder

Sarah B Ng[1,10], Kati J Buckingham[2,10], Choli Lee[1], Abigail W Bigham[2], Holly K Tabor[2,3], Karin M Dent[4], Chad D Huff[5], Paul T Shannon[6], Ethylin Wang Jabs[7,8], Deborah A Nickerson[1], Jay Shendure[1] & Michael J Bamshad[1,2,9]

We demonstrate the first successful application of exome sequencing to discover the gene for a rare mendelian disorder of unknown cause, Miller syndrome (MIM%263750). For four affected individuals in three independent kindreds, we captured and sequenced coding regions to a mean coverage of 40× and sufficient depth to call variants at ~97% of each targeted exome. Filtering against public SNP databases and eight HapMap exomes for genes with two previously unknown variants in each of the four individuals identified a single candidate gene, *DHODH*, which encodes a key enzyme in the pyrimidine *de novo* biosynthesis pathway. Sanger sequencing confirmed the presence of *DHODH* mutations in three additional families with Miller syndrome. Exome sequencing of a small number of unrelated affected individuals is a powerful, efficient strategy for identifying the genes



REVIEWS

TRANSLATIONAL GENETICS

## Exome sequencing as a tool for Mendelian disease gene discovery

*Michael J. Bamshad*[*‡], *Sarah B. Ng*[‡], *Abigail W. Bigham*[*§], *Holly K. Tabor*[*||], *Mary J. Emond*[¶], *Deborah A. Nickerson*[‡] *and Jay Shendure*[‡]

Abstract | Exome sequencing — the targeted sequencing of the subset of the human genome that is protein coding — is a powerful and cost-effective new tool for dissecting the genetic basis of diseases and traits that have proved to be intractable to conventional gene-discovery strategies. Over the past 2 years, experimental and analytical approaches relating to exome sequencing have established a rich framework for discovering the genes underlying unsolved Mendelian disorders. Additionally, exome sequencing is being adapted to explore the extent to which rare alleles explain the heritability of complex diseases and health-related traits. These advances also set the stage for applying exome and whole-genome sequencing to facilitate clinical diagnosis and personalized disease-risk profiling.

PLoS GENETICS

## Whole-Exome Re-Sequencing in a Family Quartet Identifies *POP1* Mutations As the Cause of a Novel Skeletal Dysplasia

Evgeny A. Glazov[1☯]*, Andreas Zankl[2☯], Marina Donskoi[1], Tony J. Kenna[1], Gethin P. Thomas[1], Graeme R. Clark[1], Emma L. Duncan[1,3], Matthew A. Brown[1]*

1 University of Queensland Diamantina Institute, Princess Alexandra Hospital, Woolloongabba, Australia, 2 Centre for Clinical Research, The University of Queensland, Herston, Australia, 3 School of Medicine, Faculty of Health Sciences, The University of Queensland, Herston, Australia

**Abstract**

Recent advances in DNA sequencing have enabled mapping of genes for monogenic traits in families with small pedigrees and even in unrelated cases. We report the identification of disease-causing mutations in a rare, severe, skeletal dysplasia,

European Journal of Human Genetics (2011) 19, 115–117

www.nature.com/ejhg

SHORT REPORT

## Next generation sequencing in a family with autosomal recessive Kahrizi syndrome (OMIM 612713) reveals a homozygous frameshift mutation in *SRD5A3*

Kimia Kahrizi[1], Cougar Hao Hu[2], Masoud Garshasbi[2], Seyedeh Sedigheh Abedini[1], Shirin Ghadami[1], Roxana Kariminejad[1], Reinhard Ullmann[2], Wei Chen[2], H-Hilger Ropers[2], Andreas W Kuss[2], Hossein Najmabadi[1] and Andreas Tzschach[*,2]

As part of a large-scale, systematic effort to unravel the molecular causes of autosomal recessive mental retardation, we have previously described a novel syndrome consisting of mental retardation, coloboma, cataract and kyphosis (Kahrizi syndrome

*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



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

Cristina Méndez-Vidal,[1,2] María González-del Pozo,[1,2] Alicia Vela-Boza,[3] Javier Santoyo-López,[3] Francisco J. López-Domingo,[3] Carmen Vázquez-Marouschek,[4] Joaquin Dopazo,[3,5,6] 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/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*

# Where did the heritability go?

The missing heritability problem: individual genes cannot explain the heritability of traits

NEWS FEATURE PERSONAL GENOMES

## The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Vol 461|8 October 2009|doi:10.1038/nature08494 nature

REVIEWS

## Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

Genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases and traits, and have provided valuable insights into their genetic architecture. Most variants identified so far confer relatively small increments in risk, and explain only a small proportion of familial clustering, leading many to question how the remaining, 'missing' heritability can be explained. Here we examine potential sources of missing heritability and propose research strategies, including and extending beyond current genome-wide association approaches, to illuminate the genetics of complex diseases and enhance its potential to enable effective disease prevention or treatment.

**How to explain all this?**

Rare Variants, rare CNVs, epigenetics or.. epistatic effects?

**Table 1 | Estimates of heritability and number of loci for several complex traits**

| Disease | Number of loci | Proportion of heritability explained |
|---|---|---|
| Age-related macular degeneration[72] | 5 | 50% |
| Crohn's disease[21] | 32 | 20% |
| Systemic lupus erythematosus[73] | 6 | 15% |
| Type 2 diabetes[74] | 18 | 6% |
| HDL cholesterol[75] | 7 | 5.2% |
| Height[15] | 40 | 5% |
| Early onset myocardial infarction[76] | 9 | 2.8% |
| Fasting glucose[77] | 4 | 1.5% |

* Residual is after adjustment for age, gender, diabetes.

# Is the heritability missing or are we looking at the wrong place?

**How to explain missing heritability?**
Rare Variants, rare CNVs, epigenetics or.. **epistatic effects**?

**Table 1 | Estimates of heritability and number of loci for several complex traits**

| Disease | Number of loci | Proportion of heritability explained |
|---|---|---|
| Age-related macular degeneration[72] | 5 | 50% |
| Crohn's disease[21] | 32 | 20% |
| Systemic lupus erythematosus[73] | 6 | 15% |
| Type 2 diabetes[74] | 18 | 6% |
| HDL cholesterol[75] | 7 | 5.2% |
| Height[15] | 40 | 5% |
| Early onset myocardial infarction[76] | 9 | 3.8% |
| Fasting glucose[77] | 4 | 1.5% |

* Residual is after adjustment for age, gender, diabetes.


nature genetics

## Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

**SNPs discovered by genome-wide association studies (GWASs) account for only a small fraction of the genetic variation of complex traits in human populations. Where is the remaining heritability? We estimated the proportion of variance for human height explained by 294,831 SNPs genotyped on 3,925 unrelated individuals using a linear model analysis, and validated the estimation method with simulations based on the observed genotype data. We show that 45% of variance can be explained by considering all SNPs simultaneously. Thus, most of the heritability is not missing but has not previously been detected because the individual effects are too small to pass stringent significance tests. We provide evidence that the remaining heritability is due to incomplete linkage disequilibrium between causal variants and genotyped SNPs, exacerbated by causal variants having lower minor allele frequency than the SNPs explored to date.**

GWASs in human populations have discovered hundreds of SNPs

of variation that their effects do not reach stringent significance thresholds and/or the causal variants are not in complete linkage disequilibrium with the SNPs that have been genotyped [and] of complete [illegible], or [illegible] minor allele frequency (MAF) than genotyped SNPs. Here we test these two hypotheses and estimate the contribution of the heritability of height that is [illegible]

Height in humans is a classical quantitative trait, easy to measure and studied for well over a century as a model for investigating the genetic basis [illegible] of height has been estimated to be ~0.8 (refs. 9,11–13). Rare mutations that cause extreme short or tall stature have been found[14,15], but these do not explain much of the variation in the general population. Recent GWASs on tens of thousands of individuals have detected ~50 variants that are associated with height in the population, but these in total account for only ~5% of phenotypic variance[16–19].

Data from a GWAS that are collected to detect statistical associations between SNPs and complex traits are usually analyzed by testing each

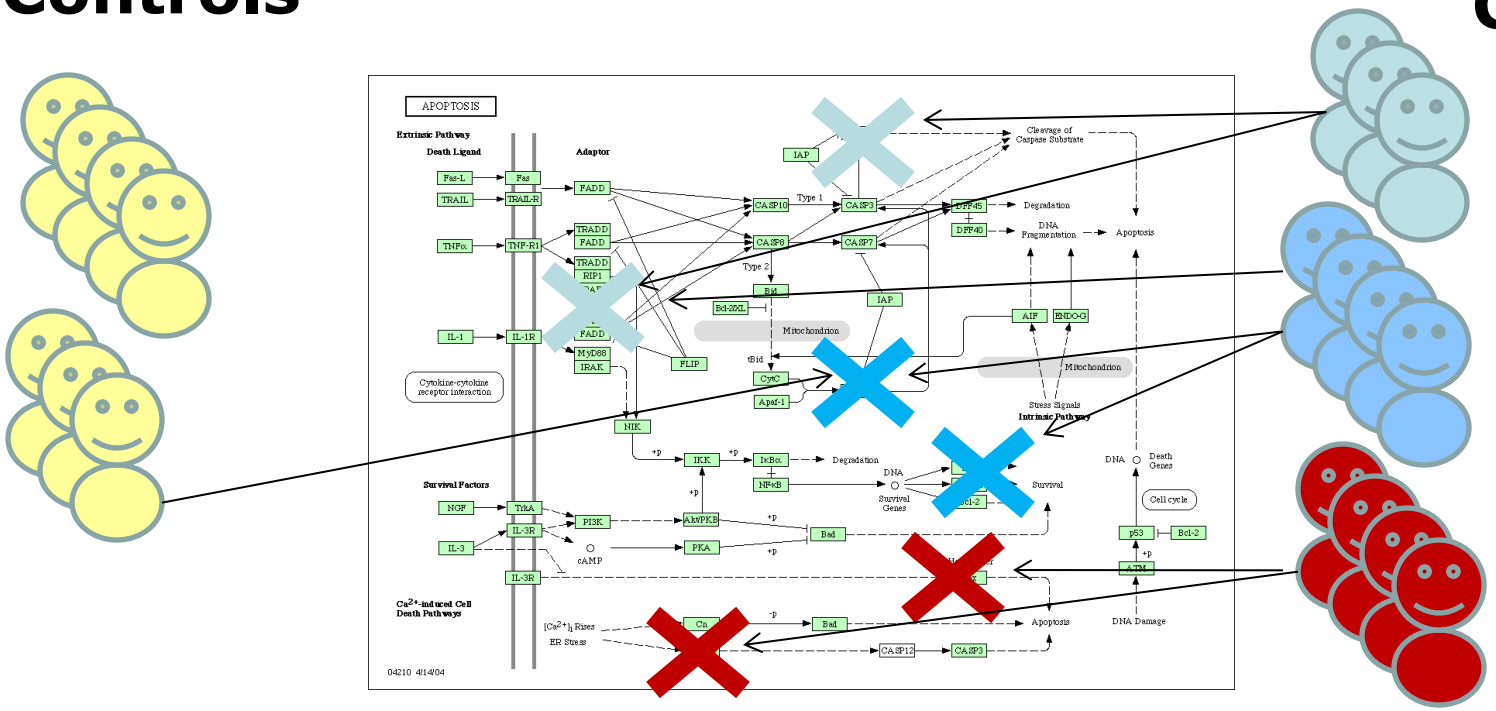


At the end, most of the heritability was there…

# An approach inspired on systems biology can help in detecting causal genes

**Controls**

**Cases**



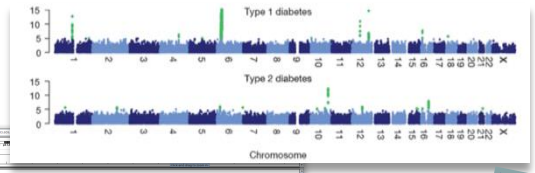

Affected **cases** in complex diseases will be a **heterogeneous** population with different mutations (or combinations).

Many cases and controls are needed to obtain significant associations.

The only **common element** is the (know or unknown) **pathway affected**.

**Disease understood as the failure of a functional module**

# From gene-based to function-based perspective



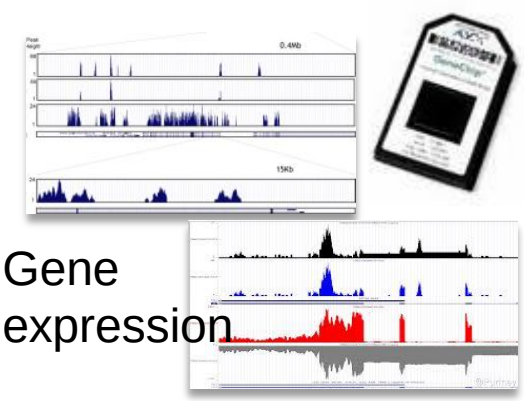

SNPs

WES/WGS

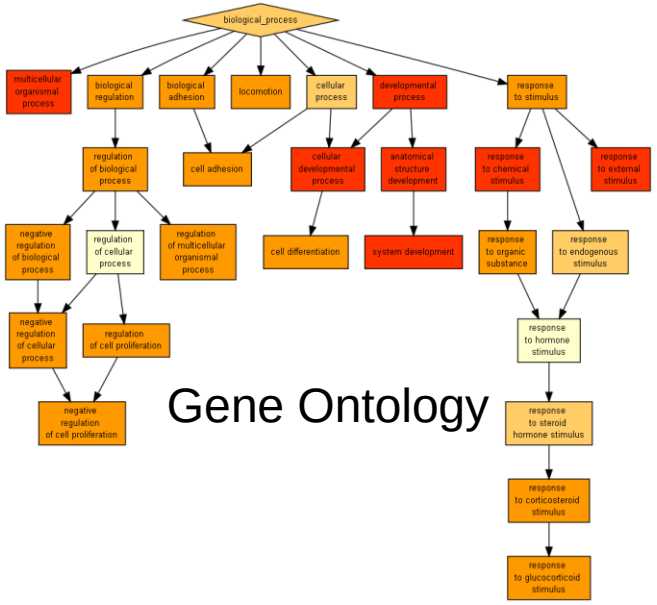
Gene integrity

AND/OR

Gene activity



Gene expression




Gene Ontology

**Gene Ontology** are **labels** to genes that describe, by means of a controlled vocabulary (ontology), the **functional role(s)** played by the genes in the cell.
A set of genes **sharing** a **GO** annotation can be considered a **functional module**.

# An example of GWAS

GWAS in Breast Cancer.

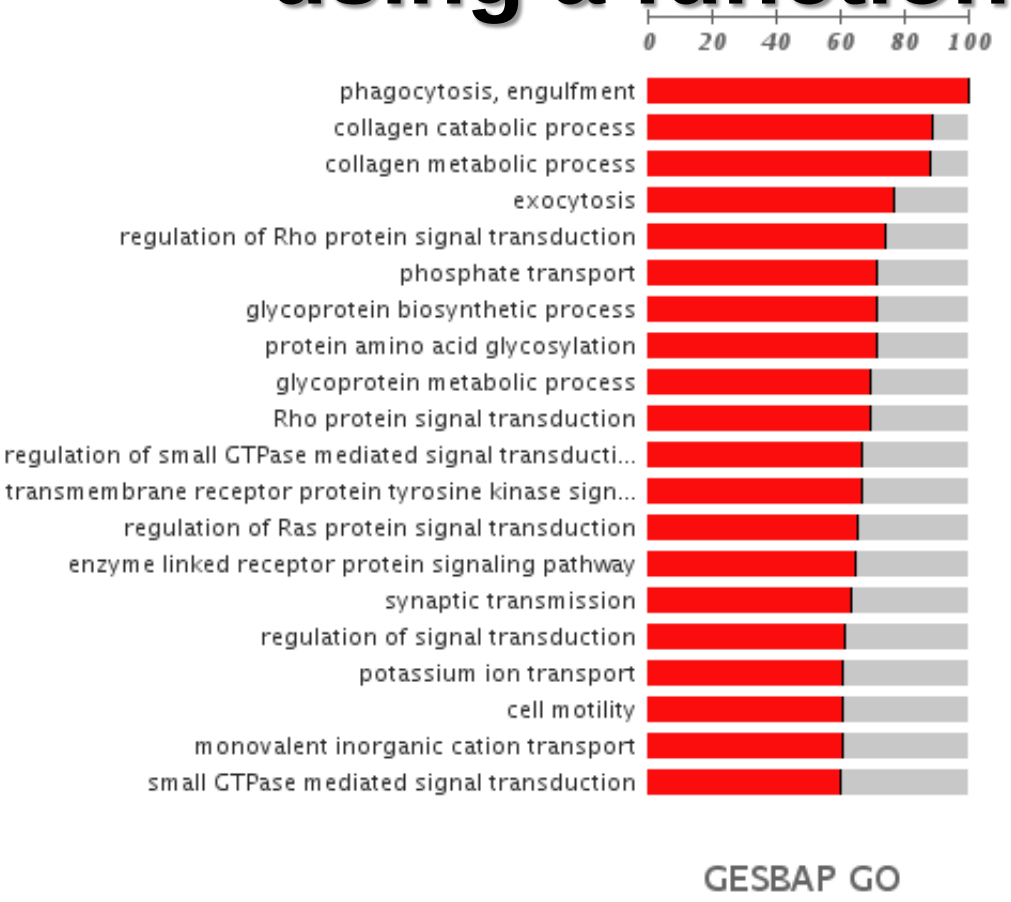
The CGEMS initiative. (Hunter et al. Nat Genet 2007)

1145 cases 1142 controls. Affy 500K

Conventional association test reports only 4 SNPs significantly mapping only on one gene: FGFR2

Conclusions: **conventional SNP-based** or **gene-based tests** are not providing much resolution.

# The same GWAS data re-analyzed using a function-based test




Breast Cancer

CGEMS initiative.
(Hunter et al. Nat Genet 2007)

1145 cases 1142 controls. Affy 500K

Only 4 SNPs were significantly associated, mapping only in one gene: FGFR2

PBA reveals 19 GO categories including *regulation of signal transduction* (FDR-adjusted p-value=$4.45 \times 10^{-03}$) in which FGFR2 is included.

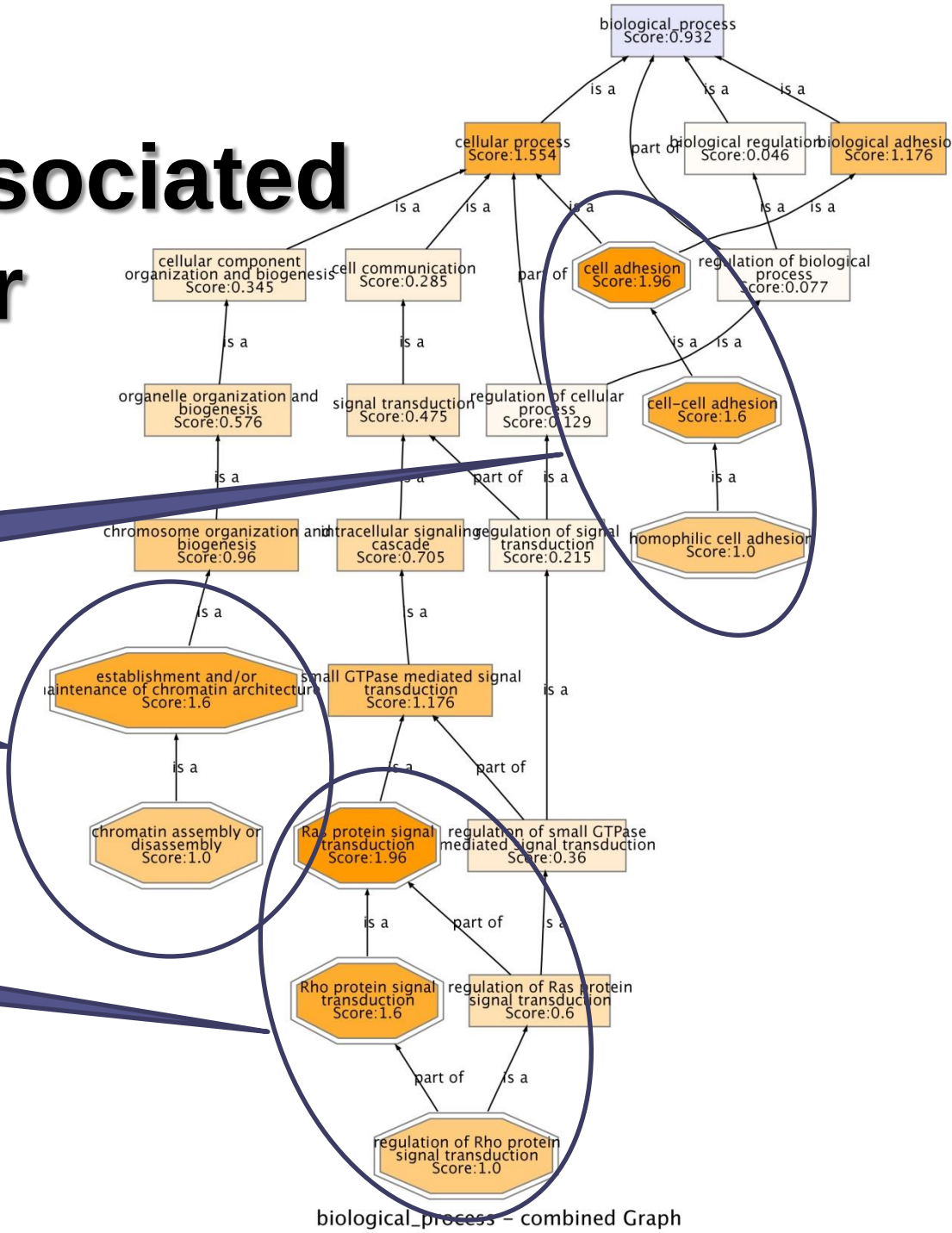
*Bonifaci et al., BMC Medical Genomics 2008; Medina et al., 2009 NAR*

# GO processes significantly associated to breast cancer

Metastasis

Chromosomal instability

Rho pathway

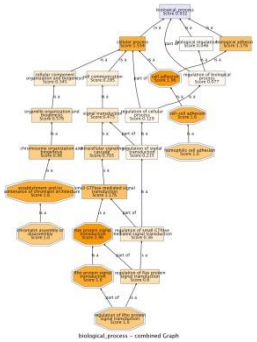


# From gene-based to function-based perspective

SNPs,
Gene expression

$Gene_1$
$Gene_2$
$Gene_3$
$Gene_4$
:
:
:
$Gene_{22000}$

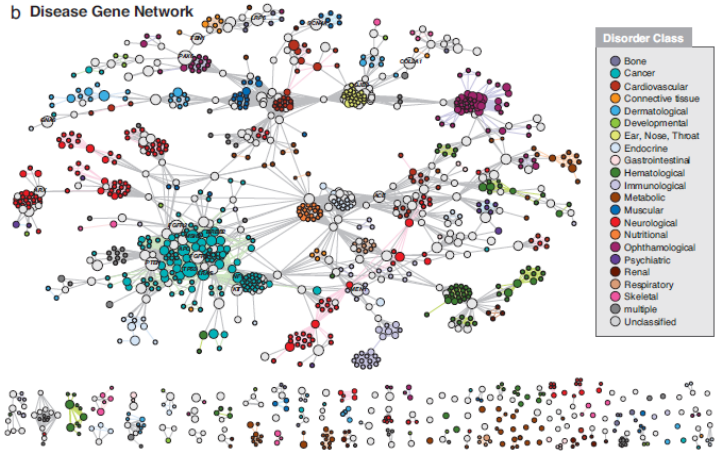
Gene Ontology



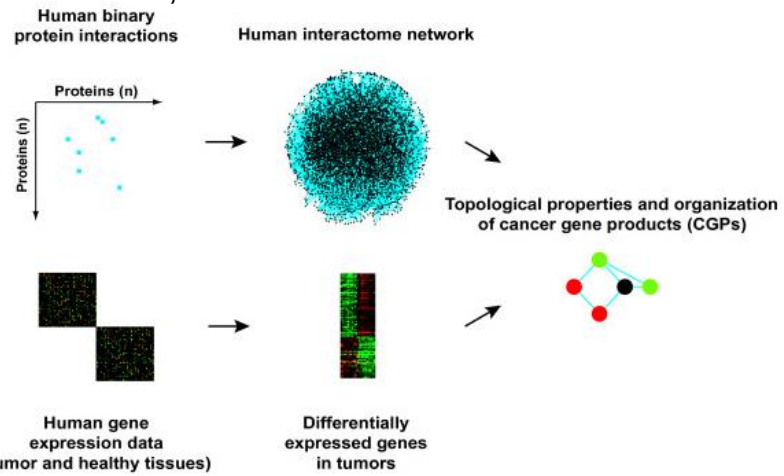
| | SNPs, gene exp. | GO |
|---|---|---|
| **Detection power** | Low (only very prevalent genes) | high |
| **Annotations available** | many | many |
| **Use** | Biomarker | Illustrative, give hints |

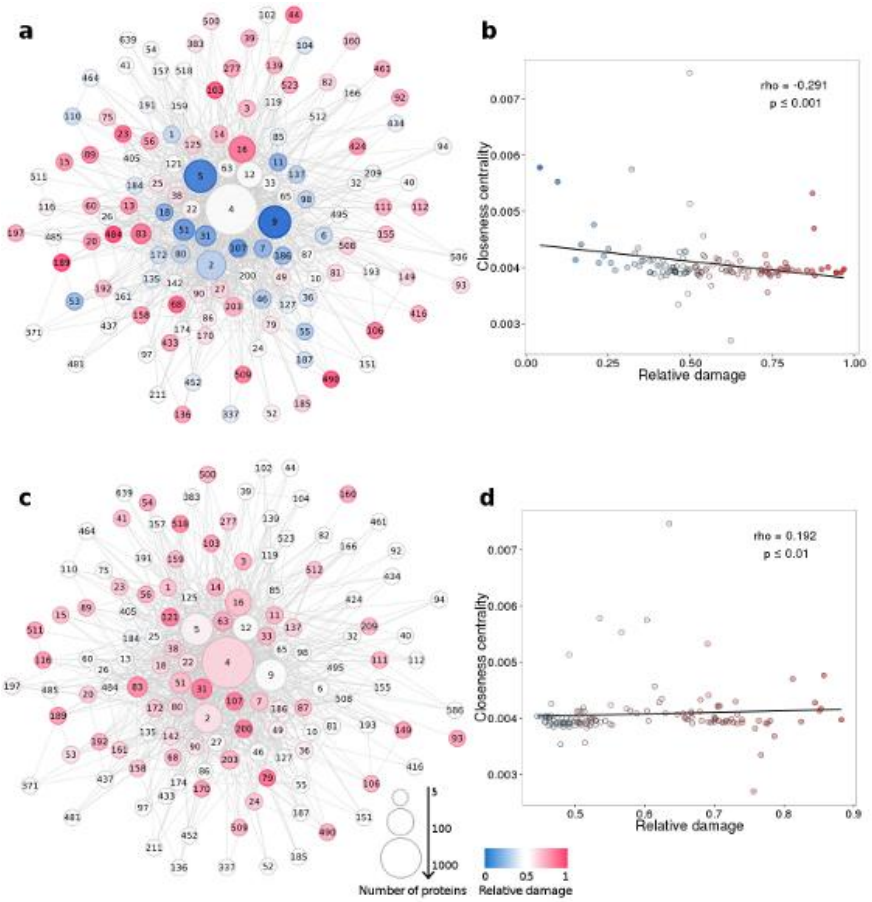
# Can the interactome help to find disease mutations?

Disease genes are close in the interactome
Goh 2007 PNAS



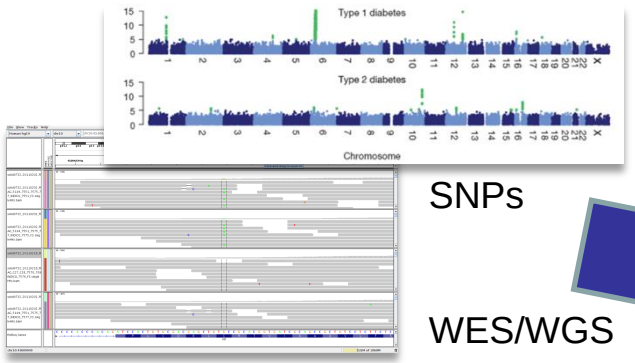

Cancer genes are central.
Hernandez, 2007 BMC Genomics




Deleterious mutations in 1000g (up) and somatic CLL deleterious mutations (down)
Garcia-Alonso 2014 Mol Syst Biol

# From gene-based to function-based perspective

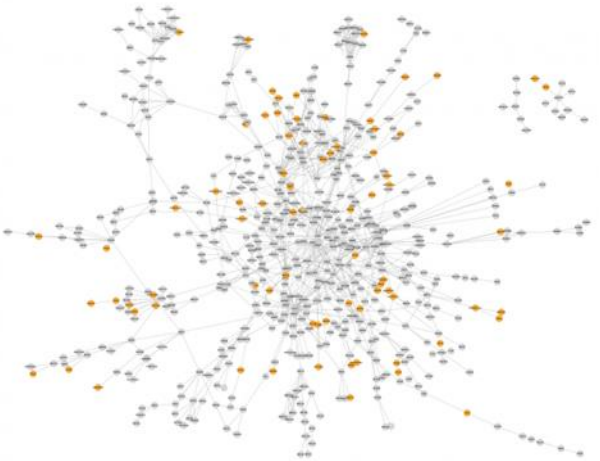



SNPs

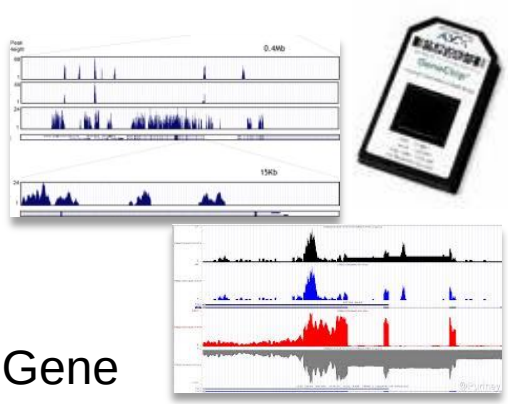
WES/WGS

Gene integrity

Using protein interaction networks as an scaffold to interpret the genomic data in a functionally-derived context



**AND/OR**



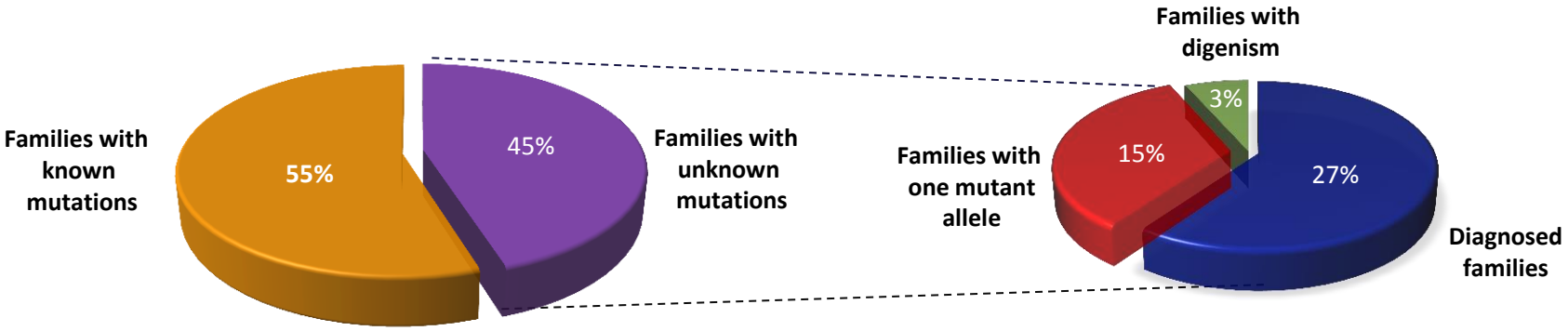
Gene expression

Gene activity

What part of the interactome is active and/or is damaged

# Example with Inherited Retinal Dystrophies

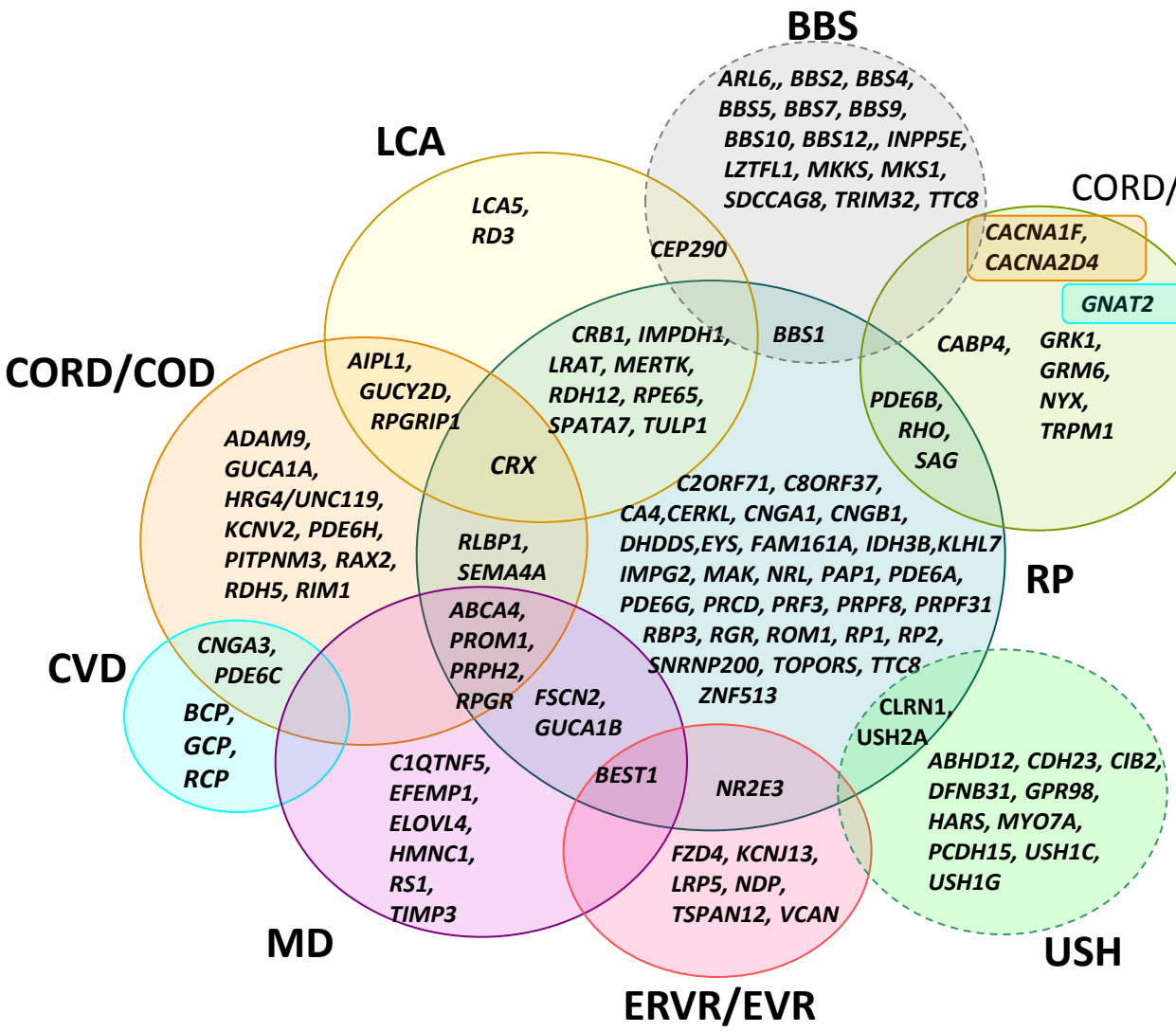
- Prevalence 1 in 3000
- Clinically and genetically very heterogeneous
- 190 GENES  account for aprox. 50% of IRDs.




**Novel variants and genes remain to be discovered**

# Network analysis.

## Is genetic overlapping among IRDs related to protein interaction?

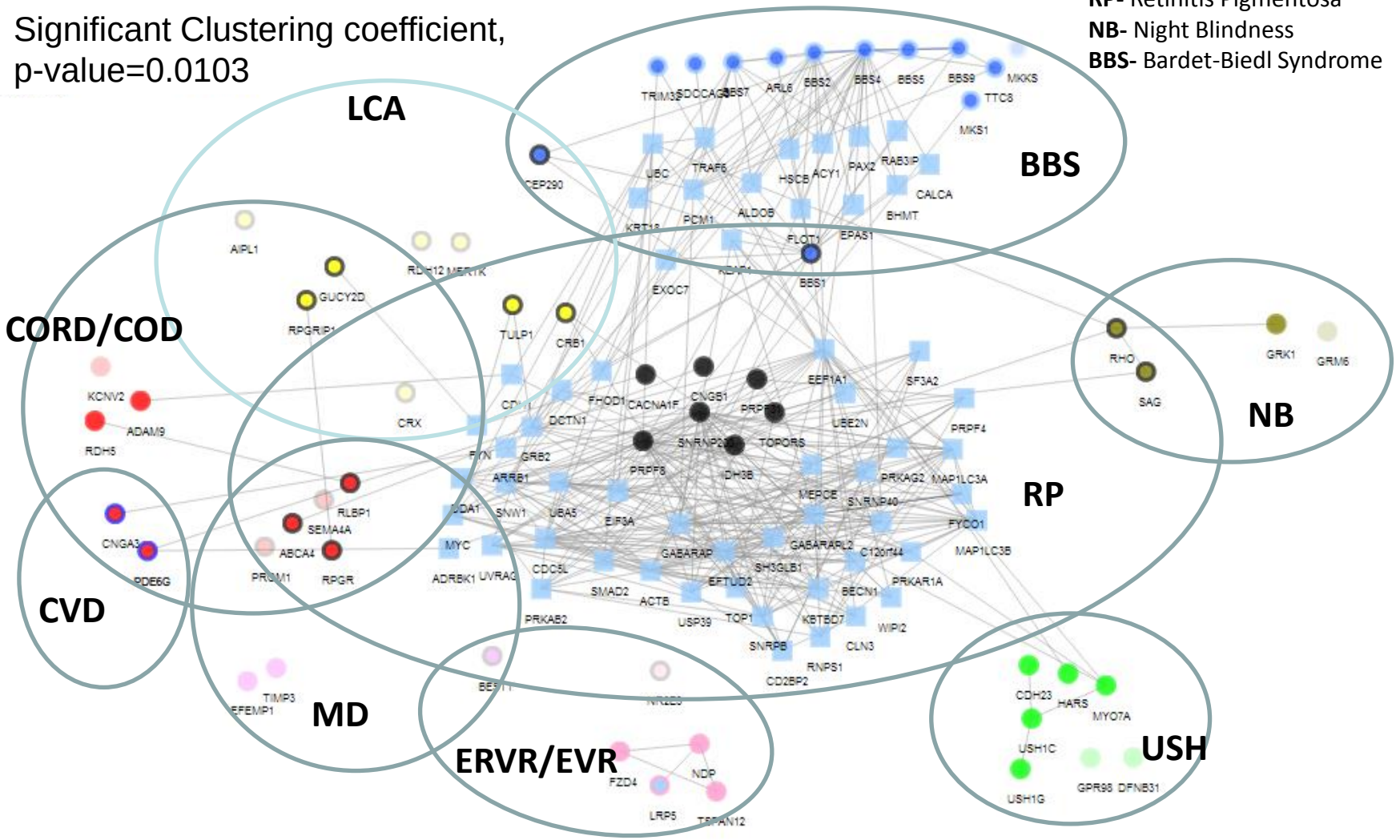



**LCA-**Leber Congenital Amaurosis
**CORD/COD-** Cone and cone-rod dystro.
**CVD-** Colour Vision Defects
**MD-** Macular Degeneration
**ERVR/EVR-** Erosive and Exudative Vitreoretinopathies
**USH-** Usher Syndrome
**RP-** Retinitis Pigmentosa
**NB-** Night Blindness
**BBS-** Bardet-Biedl Syndrome

# Connectivity among IRDs explain genetic overlap

**LCA-**Leber Congenital Amaurosis
**CORD/COD-** Cone and cone-rod dystro.
**CVD-** Colour Vision Defects
**MD-** Macular Degeneration
**ERVR/EVR-** Erosive and Exudative Vitreoretinopathies
**USH-** Usher Syndrome
**RP-** Retinitis Pigmentosa
**NB-** Night Blindness
**BBS-** Bardet-Biedl Syndrome

Significant Clustering coefficient, p-value=0.0103




SNOW Tool. *Minguez et al., NAR 2009 Implemented in Babelomics (http://www.babelomics.org)*

# Network analysis helps to find disease genes in complex diseases

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

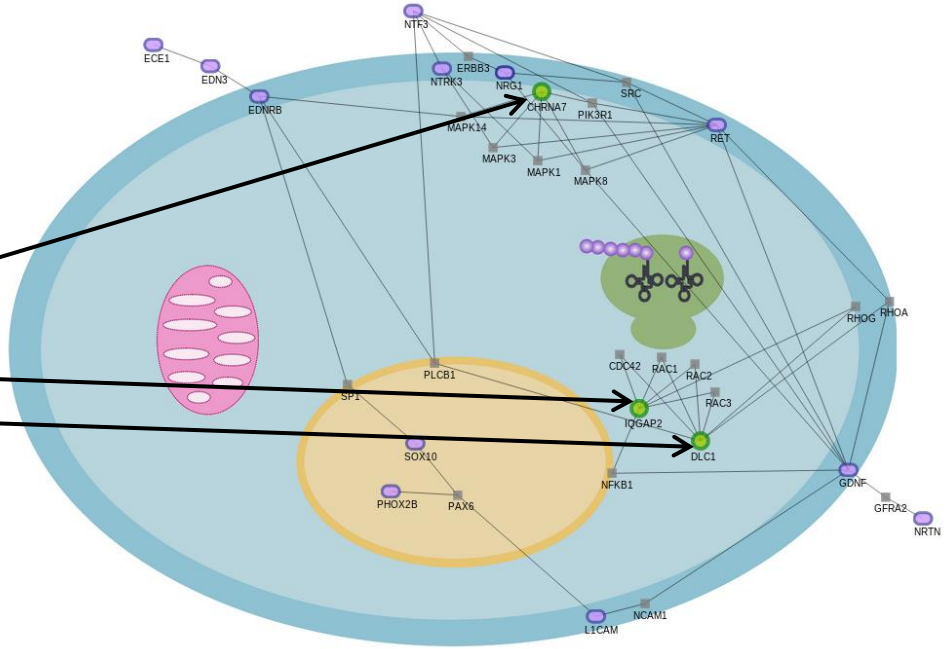
*Published online 27 July 2012* *Nucleic Acids Research, 2012, Vol. 40, No. 20 e158*
*doi:10.1093/nar/gks699*

**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)
*RASGEF1A** (rs1254964 p = $3.856 \times 10^{-05}$)
**no interactions known (yet)*

SNPs validated in independent cohorts

# Software that detects the most significant sub-network within a list of genes

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*

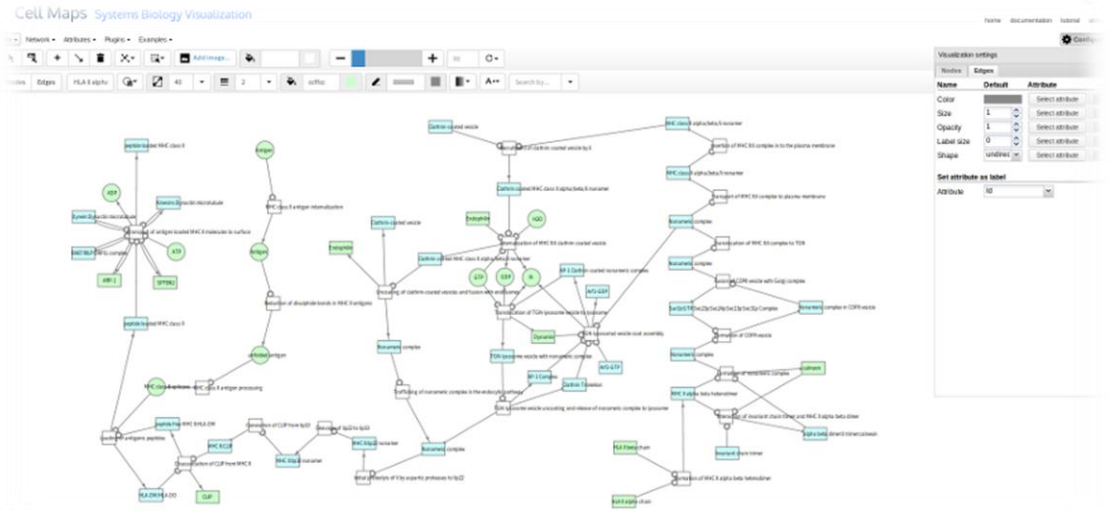
*Nucleic Acids Research, 2012, Vol. 40, No. 20 e158*
*doi:10.1093/nar/gks699*

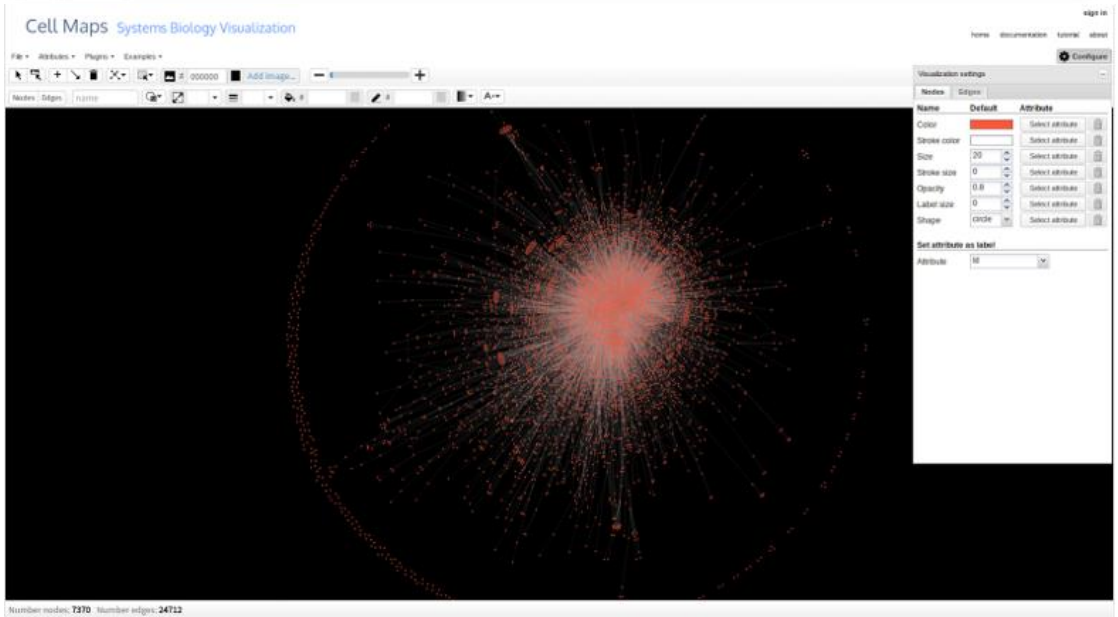
## Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments

**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




Network analysis and the interactive, web-based network viewer CellMaps

http://cellmaps.babelomics.org/

# From gene-based to function-based perspective

| | SNPs, gene expression, etc. | GO | Protein interaction networks |
|---|---|---|---|
| **Detection power** | Low (only very prevalent genes) | High | High |
| **Information coverage** | Almost all | Almost all | Less (~9000 genes in human) |
| **Use** | Biomarker | Illustrative, give hints | Biomarker* |

**Need of extra information (e.g. GO) to provide functional insights in the findings*

# Modeling pathways

Sebastian-Leon *et al. BMC Systems Biology* 2014, **8**:121
http://www.biomedcentral.com/1752-0509/8/121

BMC Systems Biology

METHODOLOGY ARTICLE Open Access

## Understanding disease mechanisms with models of signaling pathway activities

Patricia Sebastian-Leon[1], Enrique Vidal[1,2,3], Pablo Minguez[1,4], Ana Conesa[1], Sonia Tarazona[1], Alicia Amadoz[1], Carmen Armero[5], Francisco Salavert[1,2], Antonio Vidal-Puig[6], David Montaner[1] and Joaquín Dopazo[1,2,7*]

*Published online 8 June 2013* *Nucleic Acids Research, 2013, Vol. 41, Web Server issue* **W213–W217**
*doi:10.1093/nar/gkt451*

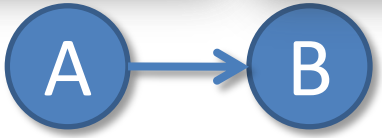
## Inferring the functional effect of gene expression changes in signaling pathways

**Patricia Sebastián-León[1], José Carbonell[1], Francisco Salavert[1,2], Rubén Sanchez[3], Ignacio Medina[1] and Joaquín Dopazo[1,2,4,*]**

[1]Department of Computational Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia 46012, Spain, [2]CIBER de Enfermedades Raras (CIBERER), Valencia 46012, Spain, [3]Genometra S.L., Valencia, Spain and [4]Functional Genomics Node (INB) at CIPF, Valencia 46012, Spain

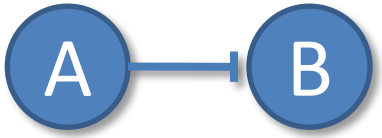
Received March 3, 2013; Revised April 18, 2013; Accepted May 2, 2013

Activation



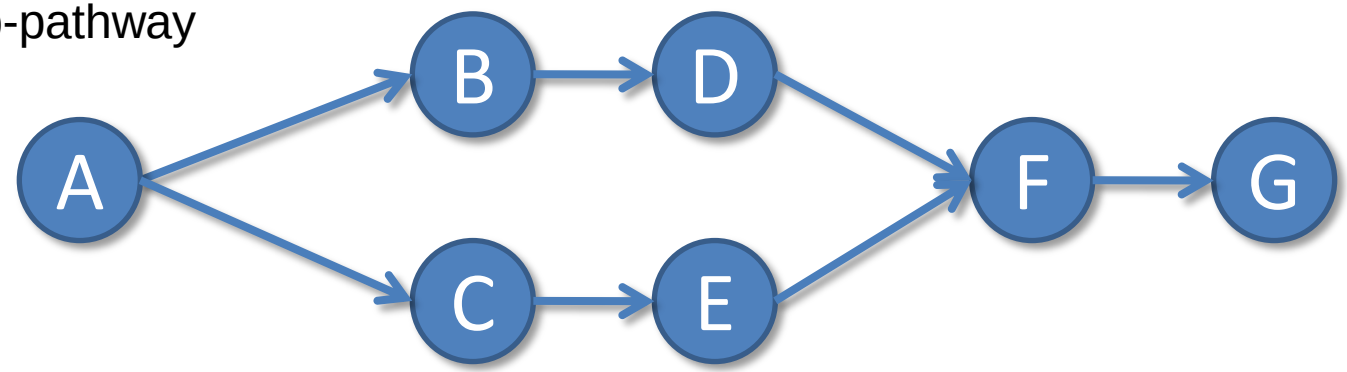

Prob. = P(A activated)P(B activated)

Inhibition




Prob. = [1-P(A activated)]P(B activated)

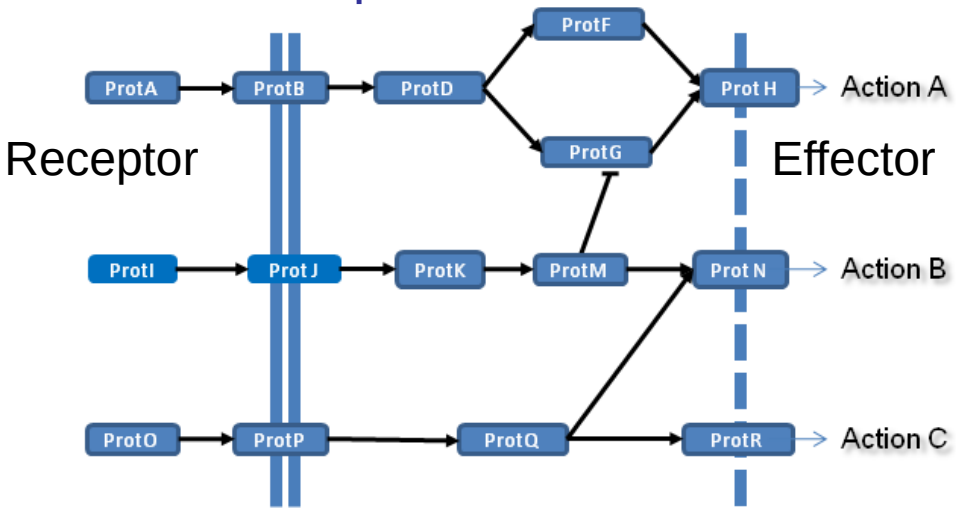
Sub-pathway



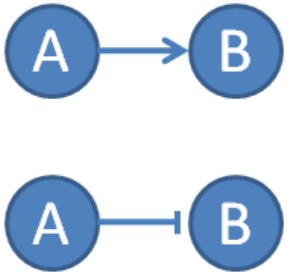

P(A→G activated) = P(A)P(B)P(D)P(F)P(G) + P(A)P(C)P(E)P(F)P(G) - P(A)P(F)P(G)P(B)P(C)P(D)P(E)

# From gene-based to mechanism-based perspective

**Transforming gene expression** values into another value that **accounts for a function**. Easiest example of modeling function: **signaling pathways**. Function: transmission of a signal from a receptor to an effector

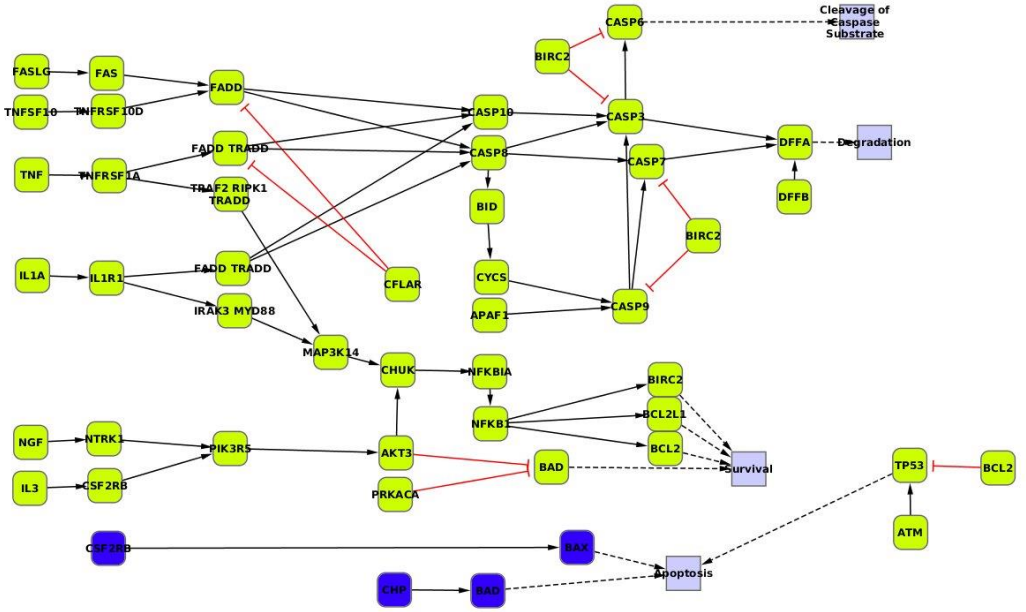



| | ProtH | ProtN | ProtR |
|---|---|---|---|
| ProtA | 1 | 0 | 0 |
| ProtI | 1 | 1 | 0 |
| ProtQ | 0 | 1 | 1 |
| function | Action A | Action B | Action C |

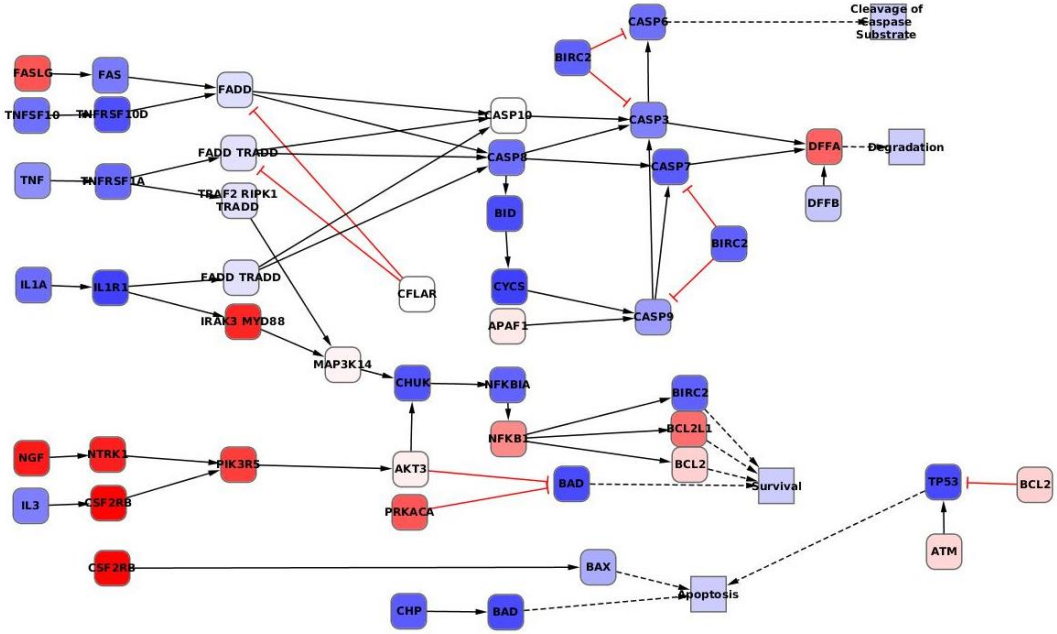



**Activations and repressions occur**

# Apoptosis inhibition is not obvious from gene expression




Two of the three possible sub-pathways leading to apoptosis are inhibited in colorectal cancer. Upper panel shows the inhibited sub-pathways in blue. Lower panel shows the actual gene up-regulations (red) and down-regulations (blue) that justify this change in the activity of the sub-pathways



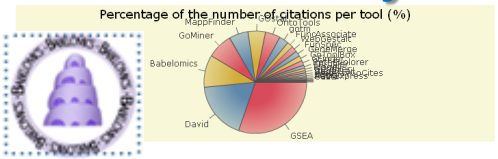
# From gene-based to function-based perspective

| | SNPs, gene expression, etc. | GO | Protein interaction networks | Models of cellular functions |
|---|---|---|---|---|
| **Detection power** | Low (only very prevalent genes) | High | High | Very high |
| **Information coverage** | Almost all | Almost all | Low (~9000 genes in human) | Low (~6700 genes in human)* |
| **Use** | Biomarker | Illustrative, give hints | Biomarker | Biomarker that explain disease mechanism |

*Only ~800 genes in human signaling pathways

# Software development

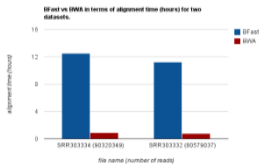
## Functional analysis




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

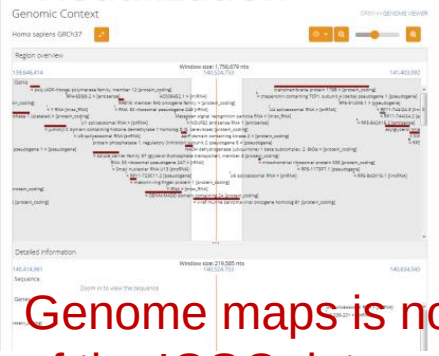

*See interactive map of for the last 24h use http://bioinfo.cipf.es/toolsusage*

## Mapping



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

## Visualization

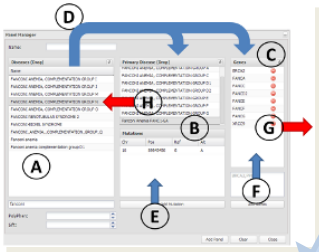


Ultrafast genome viewer with google technology

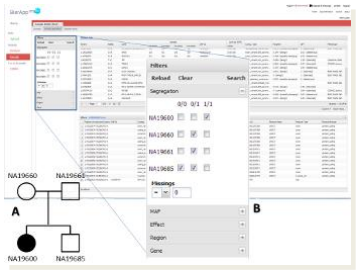
Genome maps is now part of the ICGC data portal

## Diagnostic

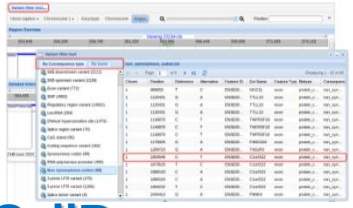
NGS panels



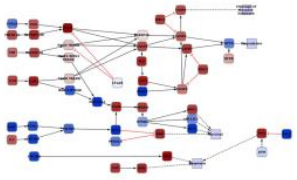
## Variant prioritization



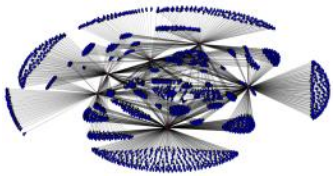
## Variant annotation



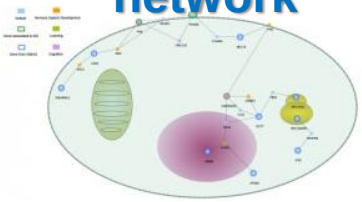
## Signaling network



## Regulatory network



## Interaction network



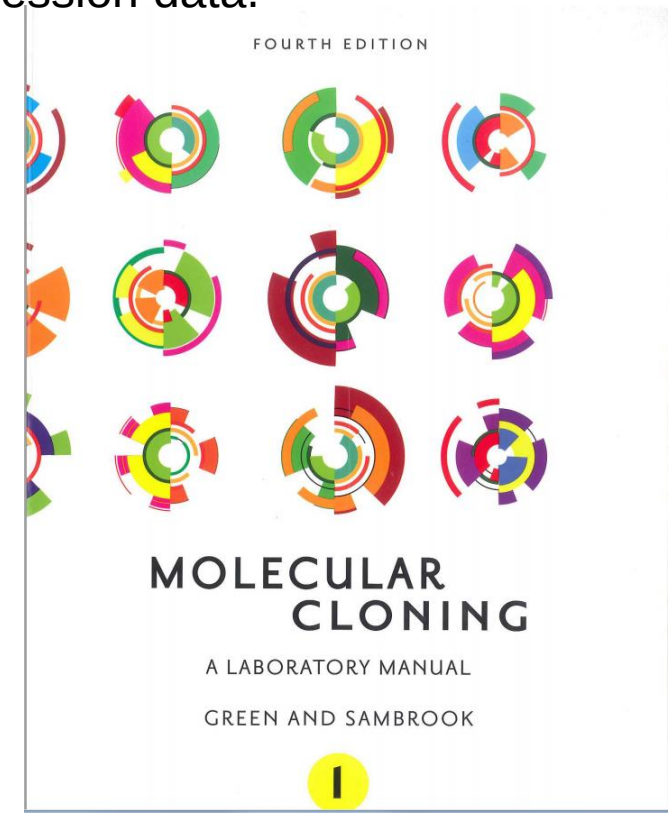
## CellBase

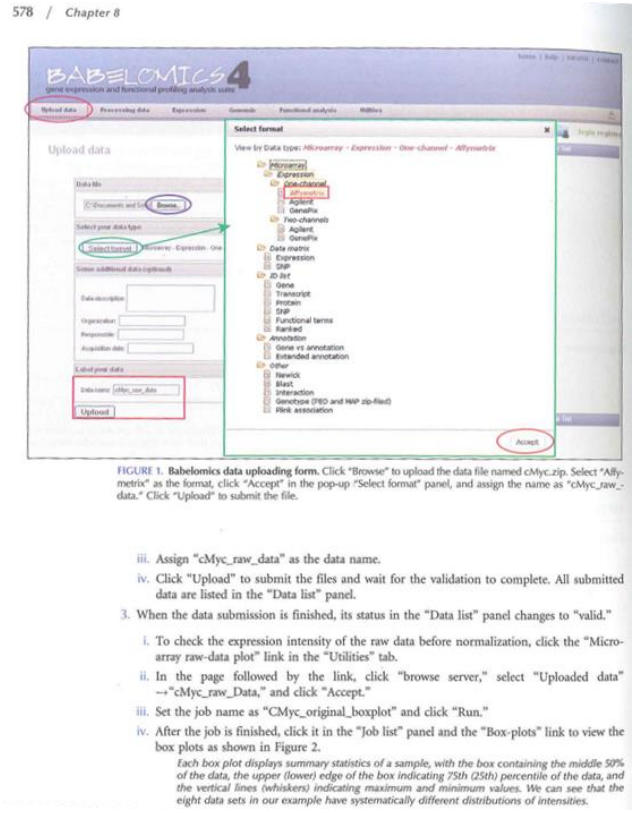
Knowledge database

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

# Babelomics in the Maniatis

The Babelomics suite of programs becomes a classic. Now is cited as a method in the last edition of **Molecular Cloning**, the popular **Maniatis**. 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.



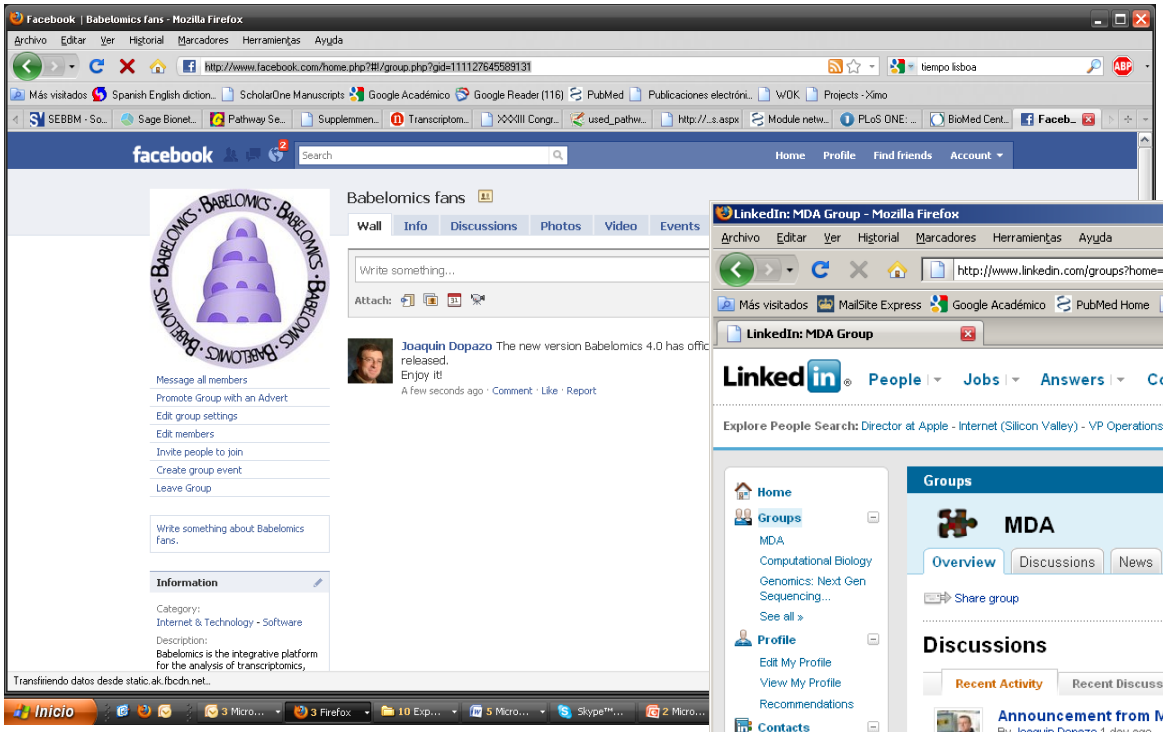



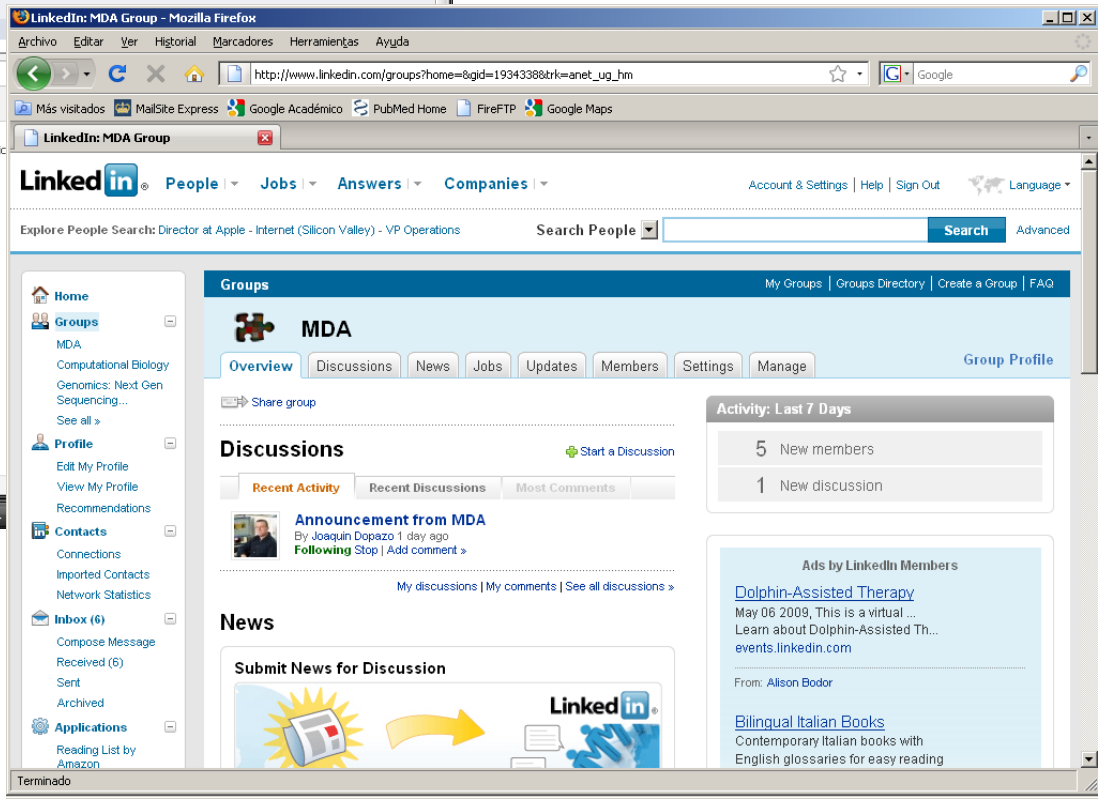

High impact developments

# SOCIAL:

## MDA group in Linked-in

## Babelomics group in Facebook and twitter



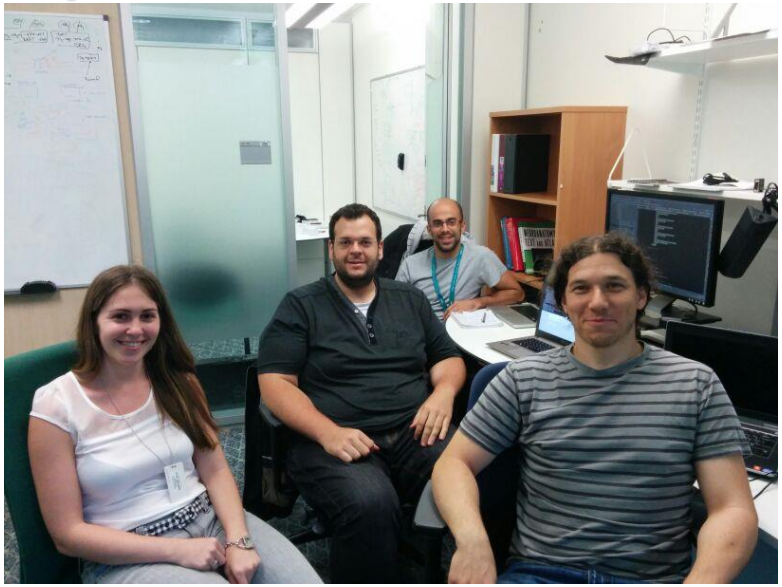




@babelomics

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

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

@xdopazo
@bioinfocipf



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

**… Ignacio Medina**
**Head of Computational Biology Lab, HPCS**
**University of Cambridge, UK**

**EMBL-EBI Scientific Collaborator**

Machine Learning for Personalized Medicine

Fundació La Marató de TV3

Proyectos PROMETEO Generalitat Valenciana

MGP Medical Genome Project

HPC4G

INB

er ciberer CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED DE ENFERMEDADES RARAS

MINISTERIO DE CIENCIA E INNOVACIÓN

F.E.D.E.R.

ciberer BIER

BULL STATegra

iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY
AFFILIATED REGIONAL INSTITUTE