Principles of genomic data analysis

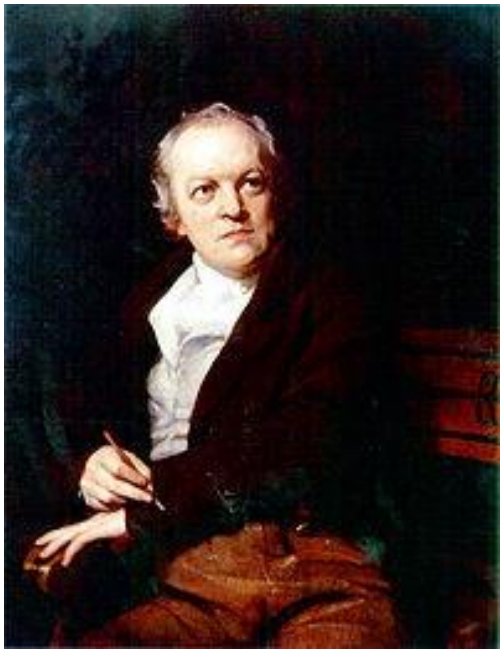
Joaquín Dopazo

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

MLPM Summer School, Paris, September18th, 2014

Background



The road of excess leads to the palace of wisdom

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

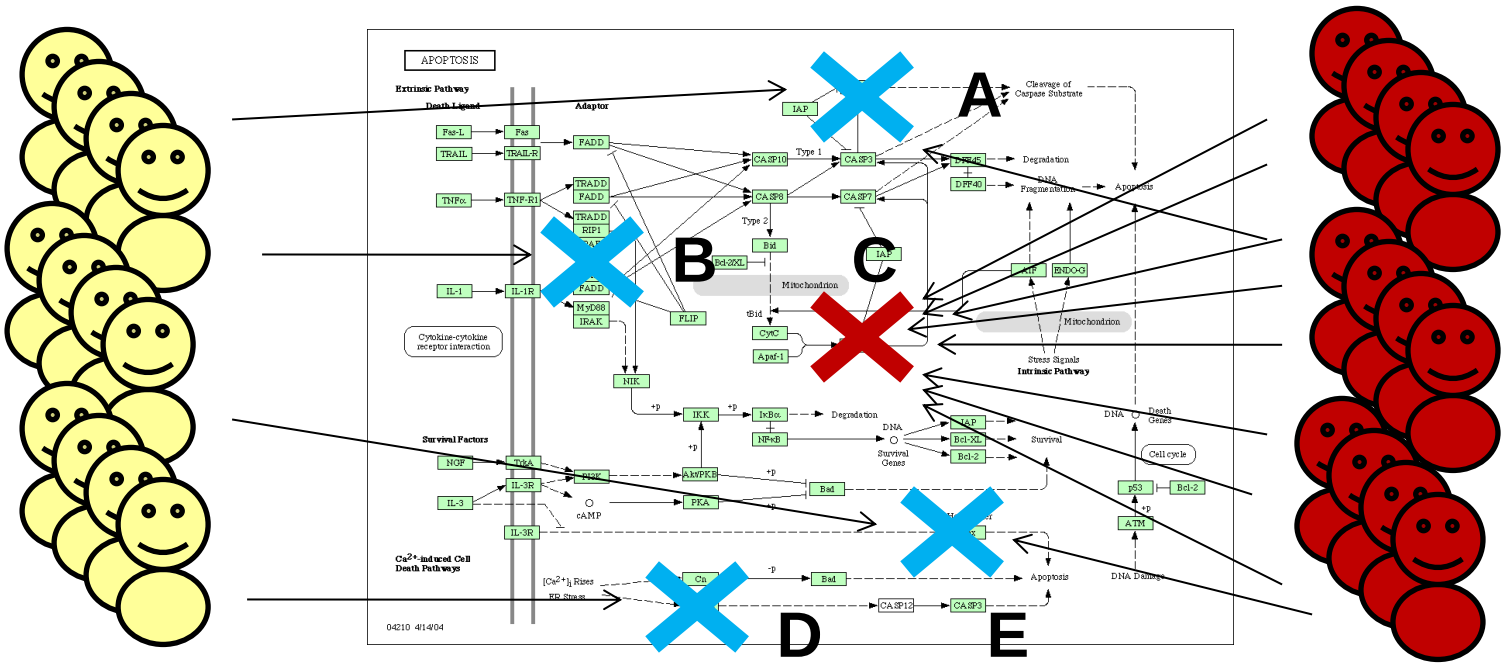
Biology has become a data-driven discipline. Biology is now bigger than physics, as measured by the size of budgets, by the size of the workforce, or by the output of major discoveries; and biology is likely to remain the biggest part of science through the twenty-first century.

The introduction and popularisation of high-throughput techniques offer the possibility of interrogating biological systems with an unprecedented level of detail.

But do we ask the proper questions…?

The paradigm

Looking for disease genes using a conventional testing schema: finding variants significantly over-represented in cases

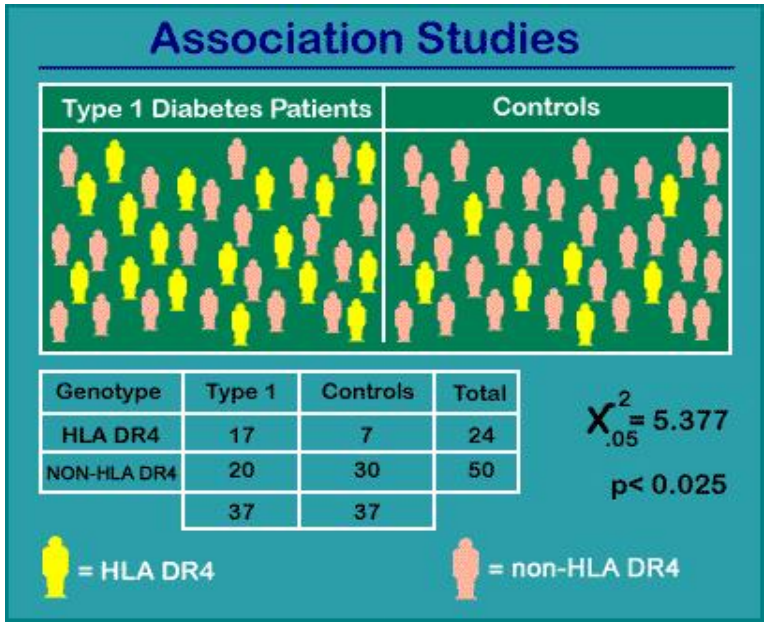


Controls

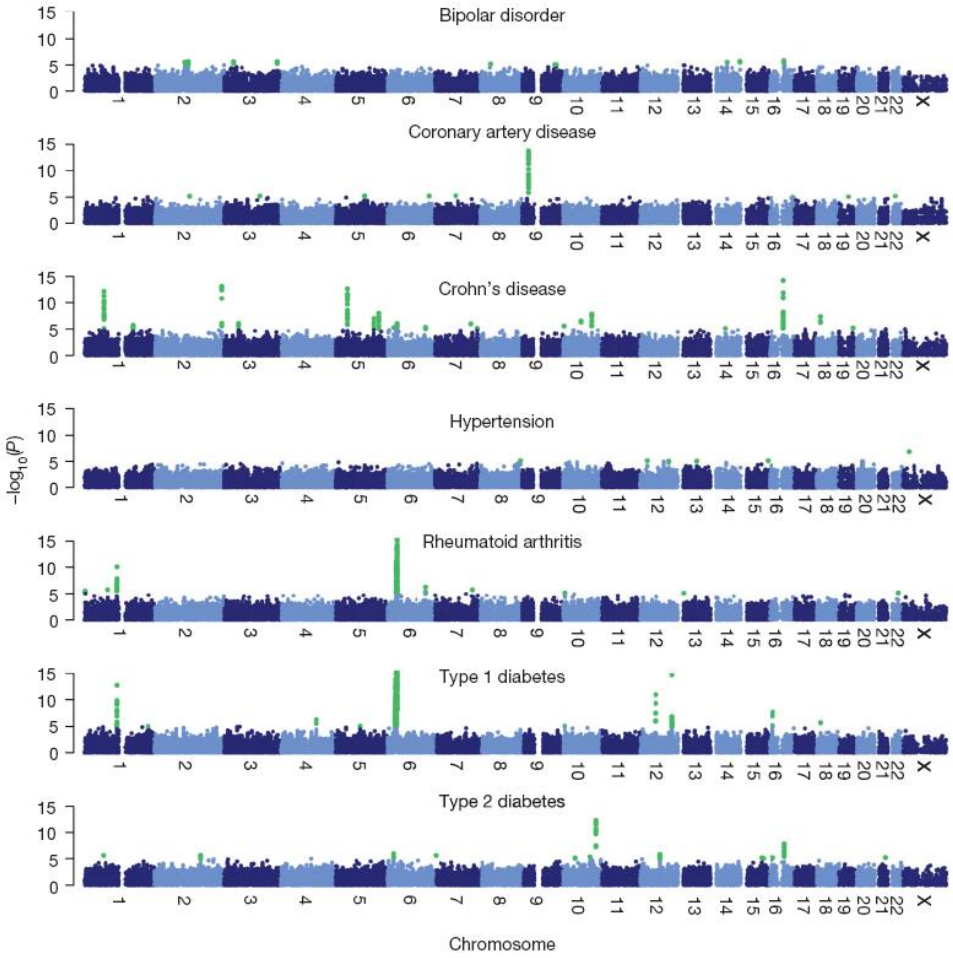
Cases

Candidate gene studies using GWAS

The promise in the 90s: "In 10 years we will unravel the genetic bases of complex diseases!!" ... And it seemed easy to fulfil...



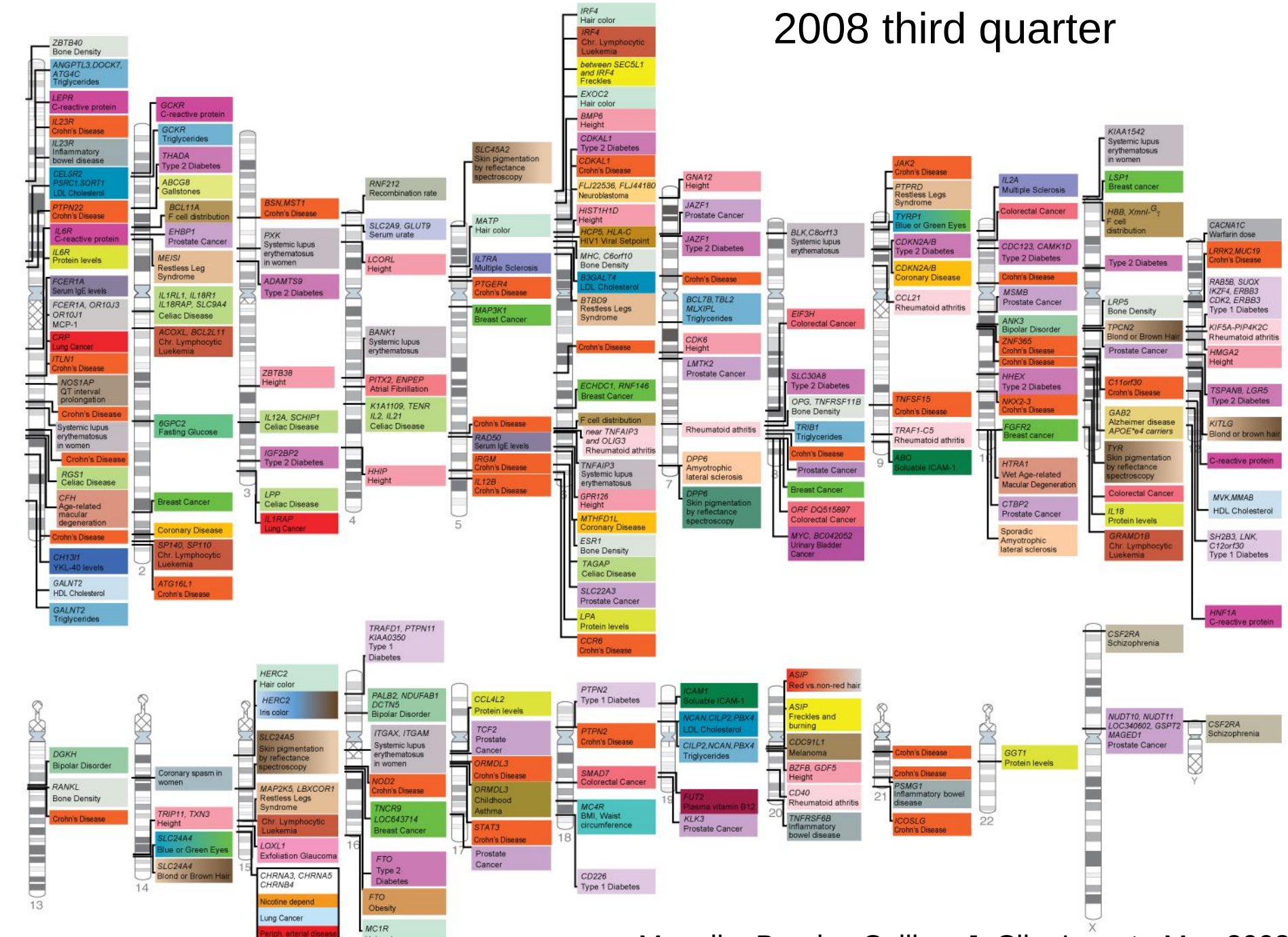
Odds Ratio: 3.6
95% CI = 1.3 to 10.4



WTCCC Home

The Wellcome Trust Case Control Consortium

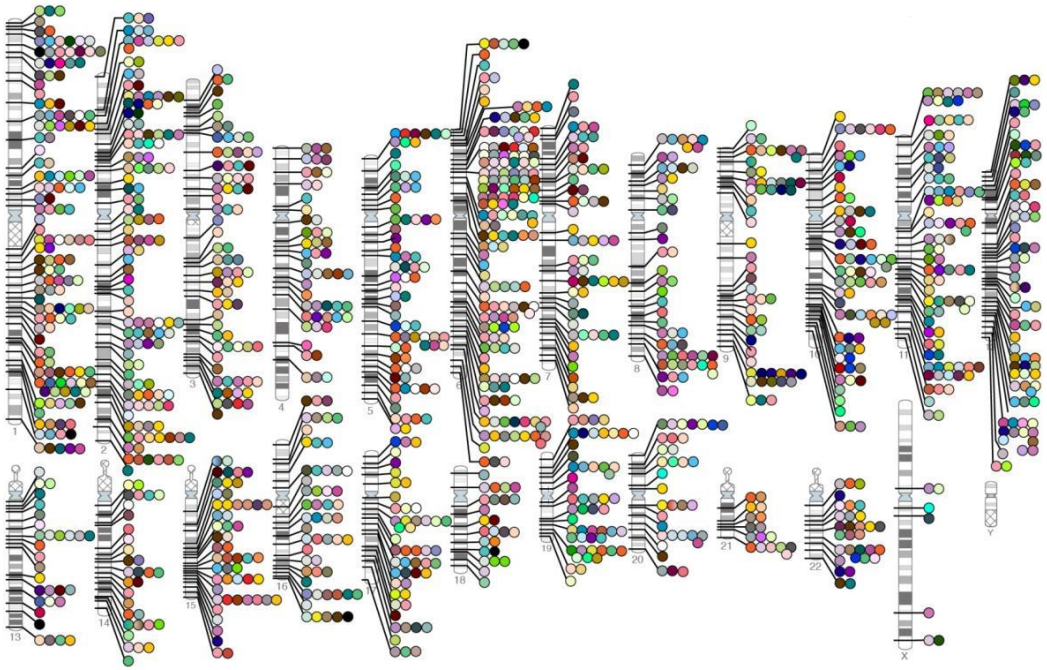
2008 third quarter



Manolio, Brooks, Collins, J. Clin. Invest., May 2008

Published Genome-Wide Associations

By the time of the completion of the human genome sequence, in **2005**, just a **few** genetic variants were known to be significantly associated to diseases.
When the first exhaustive catalogue of GWAS was compiled, in 2008, only three years later, more than **500** single nucleotide polymorphisms (SNPs) were associated to traits
Today such catalog had collected more than **1,900** papers reporting **14,012** SNPs significantly associated to more than **1,500** traits.



NHGRI GWA Catalog
www.genome.gov/GWAStudies

Lessons learned from GWAS

- **Many loci/variants** contribute to complex-trait variation
- There is evidence for **pleiotropy**, i.e., that the same **loci/variants** are associated with multiple traits.
- Much of the **heritability** of the trait **cannot be explained** by the **individual** loci/variants found associated to the trait.

Visscher et al, 2013 AJHG

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?

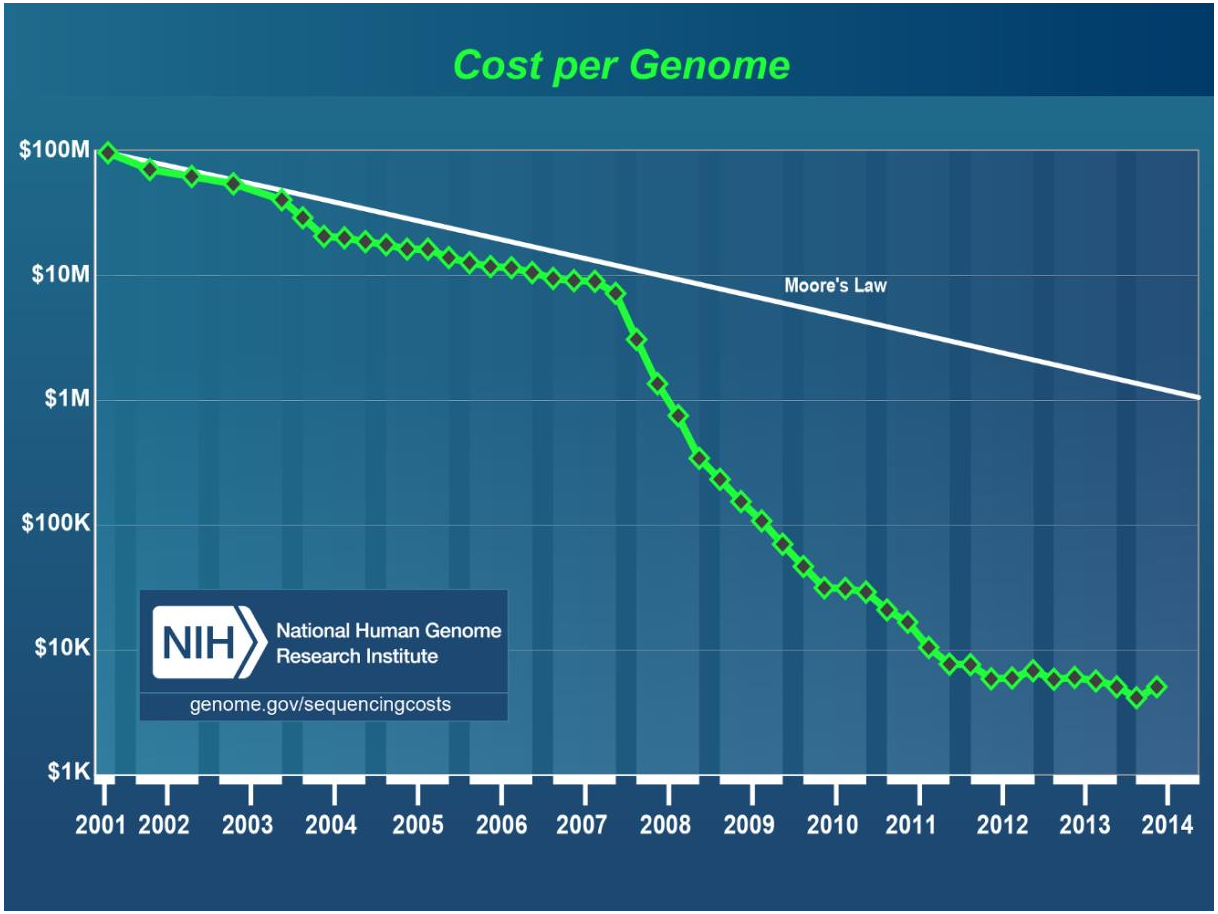
Rare Variants, rare CNVs, epigenetics or.. epistatic effects?

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.

If rare variants eluded detection because were under represented among the SNPs, genomic sequencing would reveal them.



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

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

c. All rights reserved.

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
© 2011 Macmillan Publishers Limited All rights reserved 1018-4813/11
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 OMIM 6 array-bas (c.203d interval. essential families and eye potential

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.

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

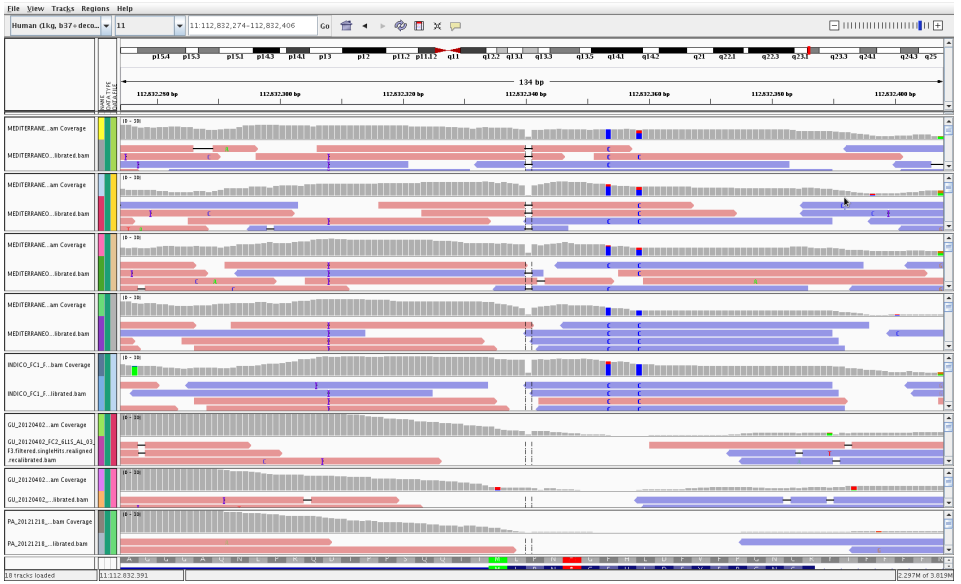
© 2013 Molecular Vision

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

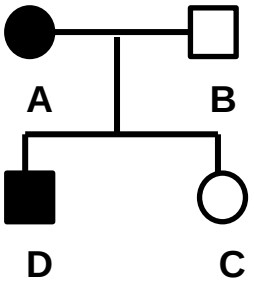
The principle: comparison of patients to reference controls or segregation within families

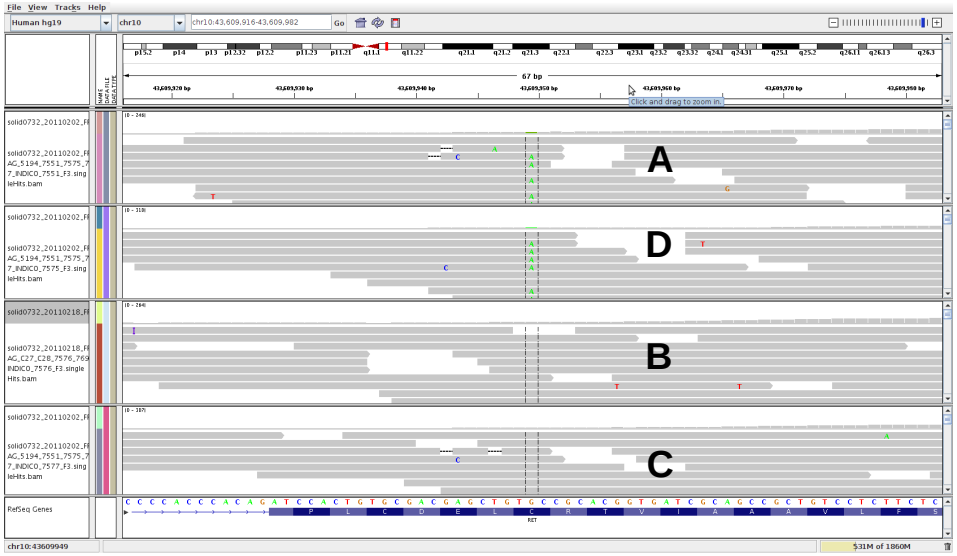


Cases

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

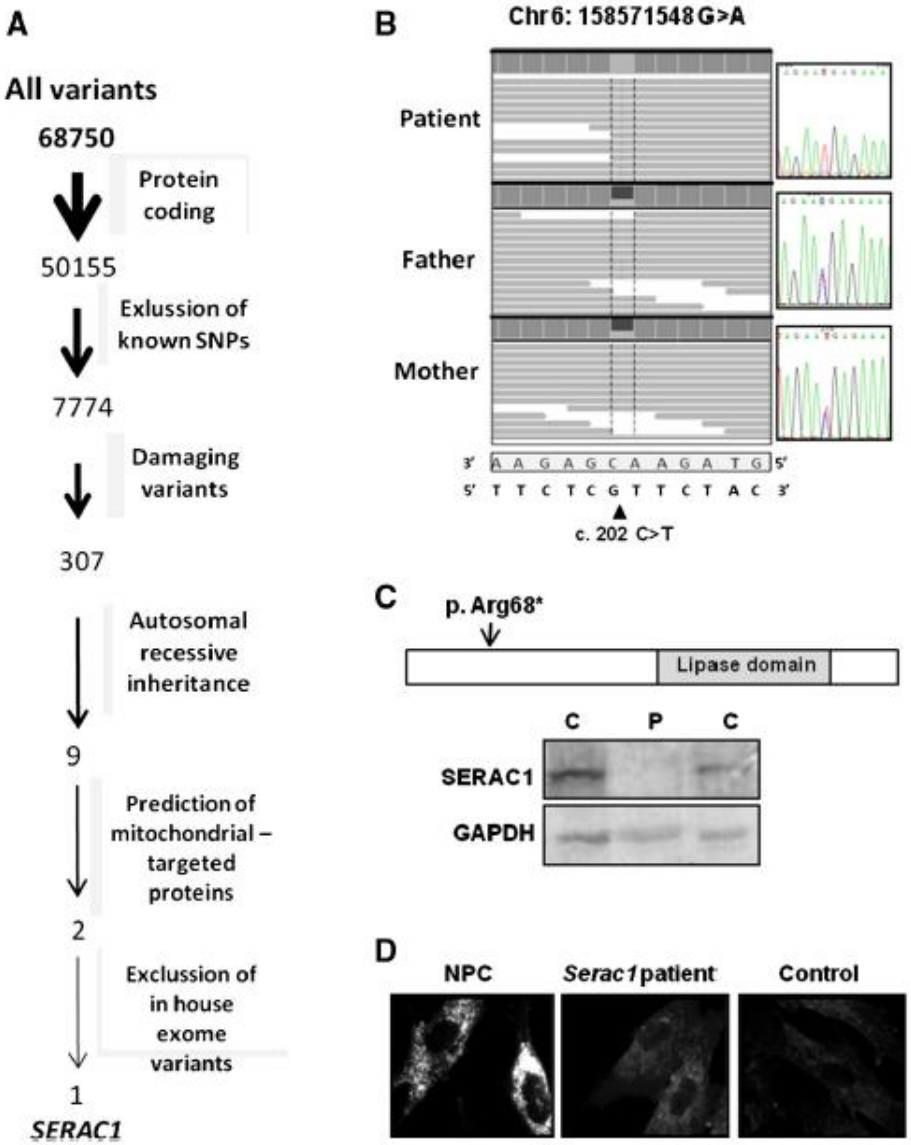
Control of sequencing errors (missing values)

Testing strategies

Successive Filtering approach

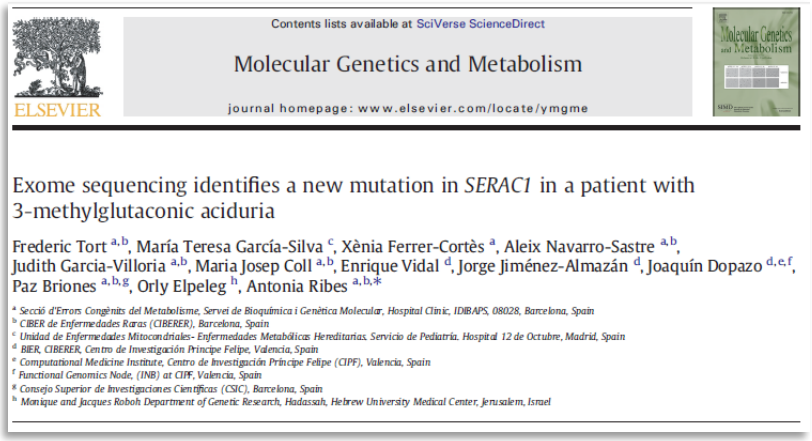
An example with 3-Methylglutaconic aciduria syndrome

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

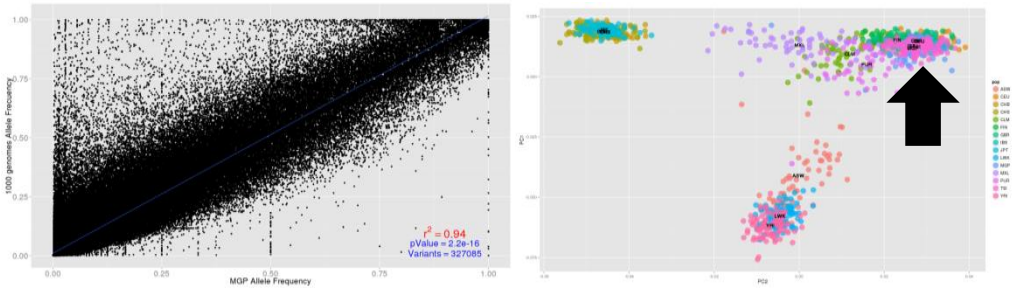
Frederic Tort [a,b], María Teresa García-Silva [c], Xènia Ferrer-Cortès [a], Aleix Navarro-Sastre [a,b], Judith Garcia-Villoria [a,b], Maria Josep Coll [a,b], Enrique Vidal [d], Jorge Jiménez-Almazán [d], Joaquín Dopazo [d,e,f], Paz Briones [a,b,g], Orly Elpeleg [h], Antonia Ribes [a,b,*]

[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

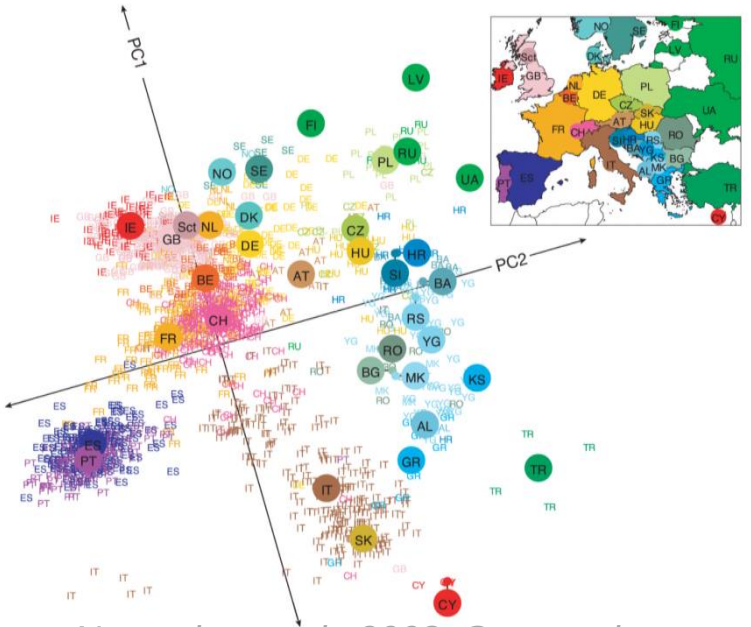
Use known variants and their population frequencies to filter out.

- Typically dbSNP, 1000 genomes and the 6515 exomes from the ESP are used as sources of population frequencies.
- We sequenced **300 healthy controls** (rigorously phenotyped) to add and extra filtering step to the analysis pipeline

How important do you think local information is to detect disease genes?



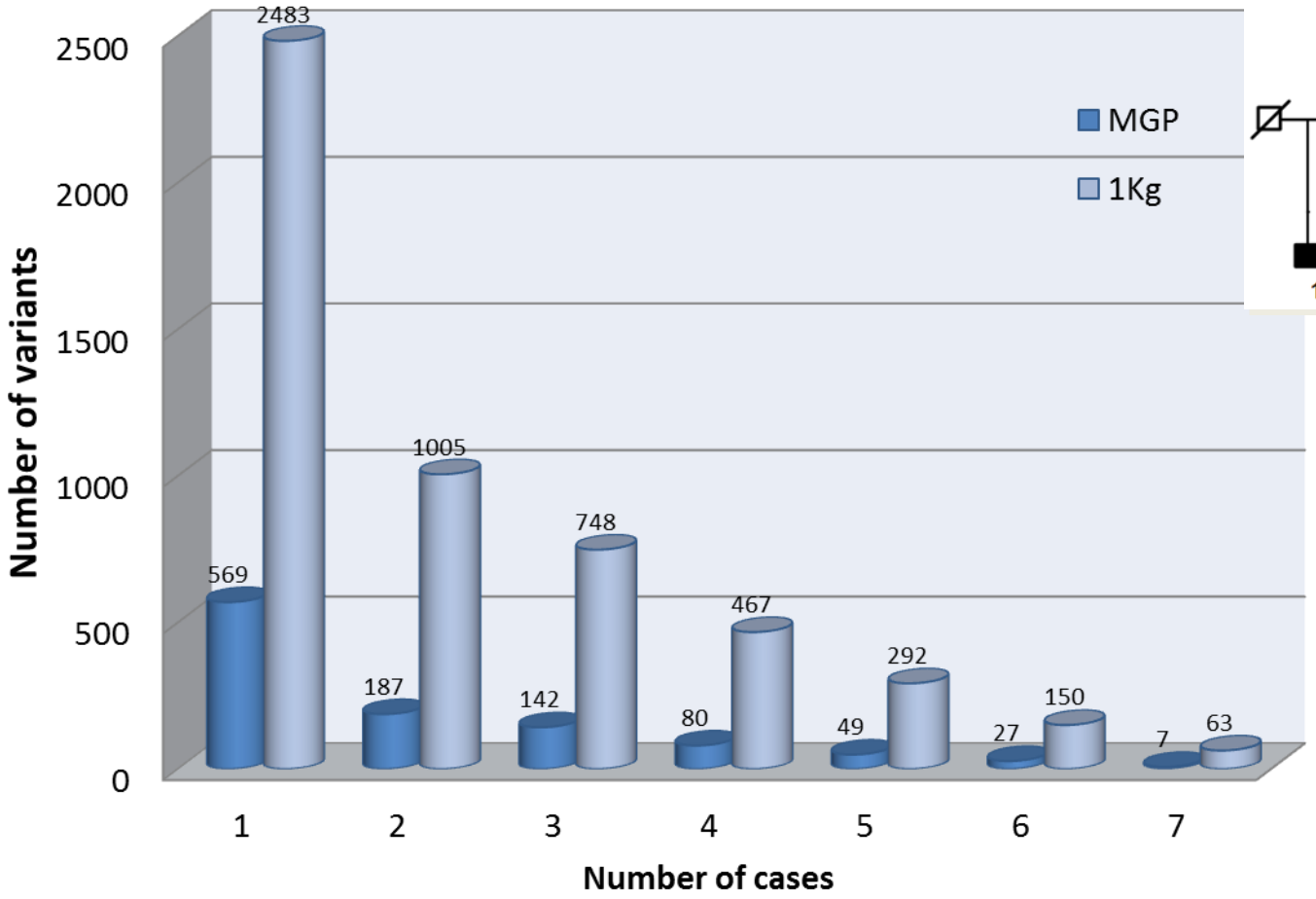
Comparison of MGP controls to 1000g

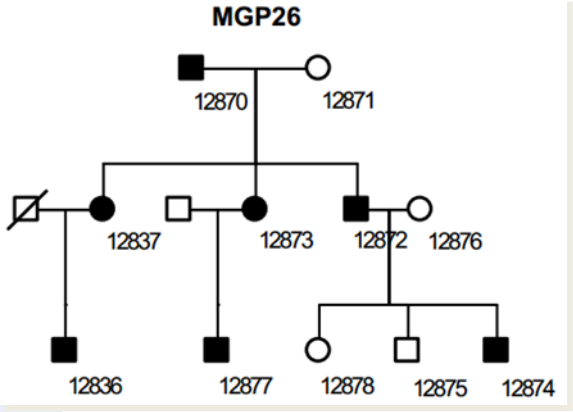


Novembre et al., 2008. Genes mirror geography within Europe. Nature

Filtering with or without local variants

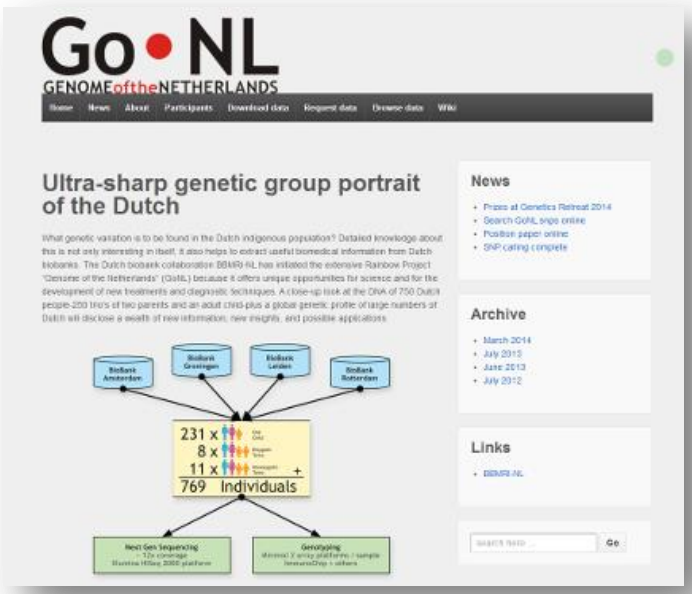
Number of genes as a function of individuals in the study of a dominant disease
Retinitis Pigmentosa autosomal dominant



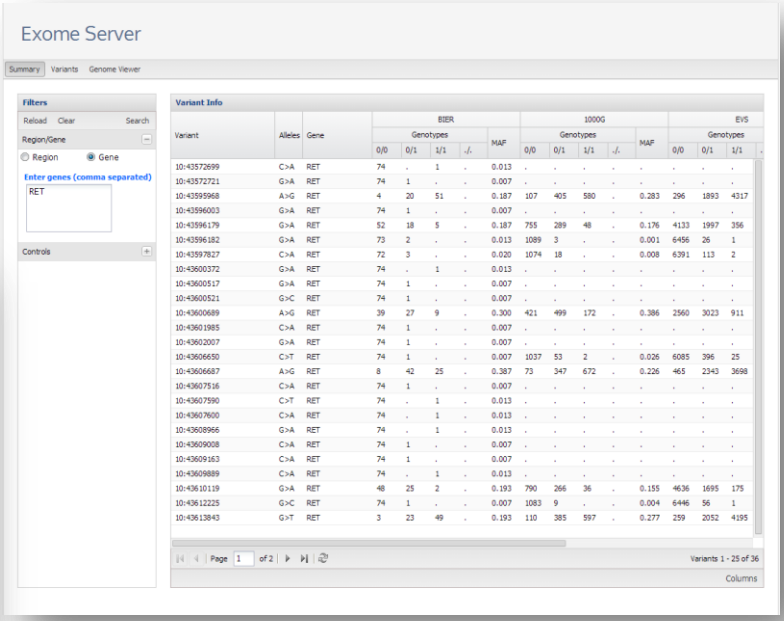


The use of local variants makes an enormous difference

The CIBERER Exome Server (CES): the first repository of variability of the Spanish population



Only another similar initiative exists: the GoNL http://www.nlgenome.nl/



EJHG Open

European Journal of Human Genetics (2014) 22, 221–227
© 2014 Macmillan Publishers Limited All rights reserved 1018-4813/14
www.nature.com/ejhg

ARTICLE

The Genome of the Netherlands: design, and project goals

Dorret I Boomsma*,[1,22], Cisca Wijmenga[2,22], Eline P Slagboom[3,22], Morris A Swertz[2,22], Lennart C Karssen[4], Abdel Abdellaoui[1], Kai Ye[3], Victor Guryev[5,6], Martijn Vermaat[7,8,9], Freerk van Dijk[2], Laurent C Francioli[10], Jouke Jan Hottenga[1], Jeroen FJ Laros[7,8,9], Qibin Li[11], Yingrui Li[11], Hongzhi Cao[11], Ruoyan Chen[11], Yuanping Du[11], Ning Li[12], Sujie Cao[12], Jessica van Setten[10], Androniki Menelaou[10], Sara L Pulit[10], Jayne Y Hehir-Kwa[15], Marian Beekman[16], Clara C Elbers[10], Heorhiy Byelas[2], Anton JM de Craen[16], Patrick Deelen[2], Martijn Dijkstra[2], Johan T den Dunnen[8,9], Peter de Knijff[8,9], Jeanine Houwing-Duistermaat[17], Vyacheslav Koval[18], Karol Estrada[18], Albert Hofman[4], Alexandros Kanterakis[2], David van Enckevort[7], Hailiang Mai[7], Mathijs Kattenberg[1], Elisabeth M van Leeuwen[4], Pieter BT Neerincx[2], Ben Oostra[19], Fernanodo Rivadeneira[18], Eka HD Suchiman[3], Andre G Uitterlinden[18], Gonneke Willemsen[1], Bruce H Wolffenbuttel[20], Jun Wang[11,13,14,22], Paul IW de Bakker[10,22], Gert-Jan van Ommen[21,22] and Cornelia M van Duijn*,[4,22]

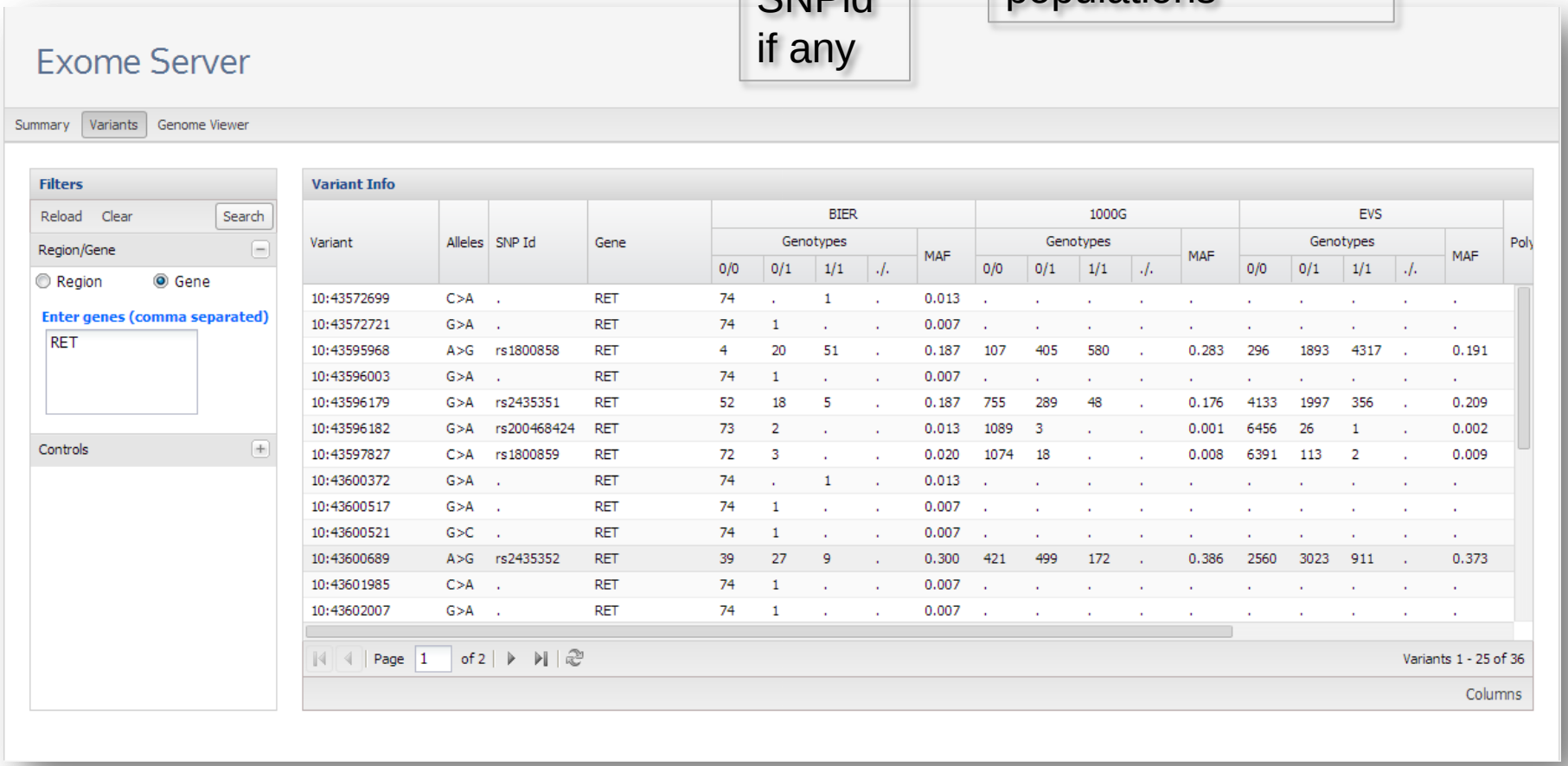
http://ciberer.es/bier/exome-server/

Information provided

Genomic coordinates, variation, and gene.

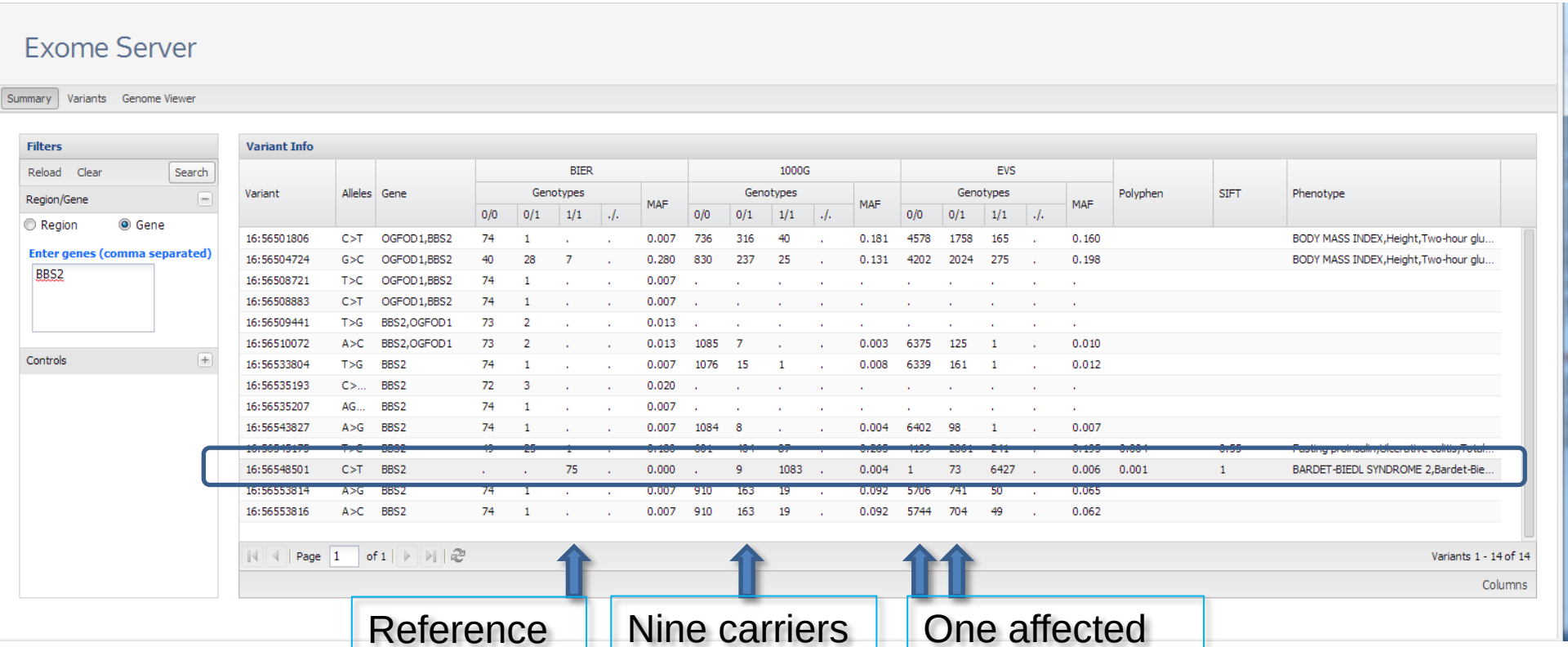
SNPid if any

Genotypes in the different reference populations



Variant	Alleles	SNP Id	Gene	BIER					1000G					EVS					Poly
				Genotypes				MAF	Genotypes				MAF	Genotypes				MAF	
				0/0	0/1	1/1	./.		0/0	0/1	1/1	./.		0/0	0/1	1/1	./.		
10:43572699	C>A	.	RET	74	.	1	.	0.013	.	.	.	.	.	.	.	.	.	.	
10:43572721	G>A	.	RET	74	1	.	.	0.007	.	.	.	.	.	.	.	.	.	.	
10:43595968	A>G	rs1800858	RET	4	20	51	.	0.187	107	405	580	.	0.283	296	1893	4317	.	0.191	
10:43596003	G>A	.	RET	74	1	.	.	0.007	.	.	.	.	.	.	.	.	.	.	
10:43596179	G>A	rs2435351	RET	52	18	5	.	0.187	755	289	48	.	0.176	4133	1997	356	.	0.209	
10:43596182	G>A	rs200468424	RET	73	2	.	.	0.013	1089	3	.	.	0.001	6456	26	1	.	0.002	
10:43597827	C>A	rs1800859	RET	72	3	.	.	0.020	1074	18	.	.	0.008	6391	113	2	.	0.009	
10:43600372	G>A	.	RET	74	.	1	.	0.013	.	.	.	.	.	.	.	.	.	.	
10:43600517	G>A	.	RET	74	1	.	.	0.007	.	.	.	.	.	.	.	.	.	.	
10:43600521	G>C	.	RET	74	1	.	.	0.007	.	.	.	.	.	.	.	.	.	.	
10:43600689	A>G	rs2435352	RET	39	27	9	.	0.300	421	499	172	.	0.386	2560	3023	911	.	0.373	
10:43601985	C>A	.	RET	74	1	.	.	0.007	.	.	.	.	.	.	.	.	.	.	
10:43602007	G>A	.	RET	74	1	.	.	0.007	.	.	.	.	.	.	.	.	.	.	

Occurrence of pathological variants in “normal” population

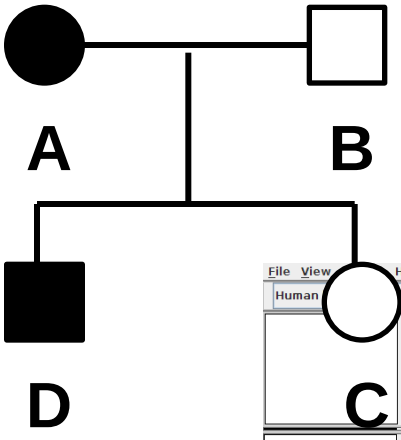


Reference genome is mutated

Nine carriers in 1000 genomes

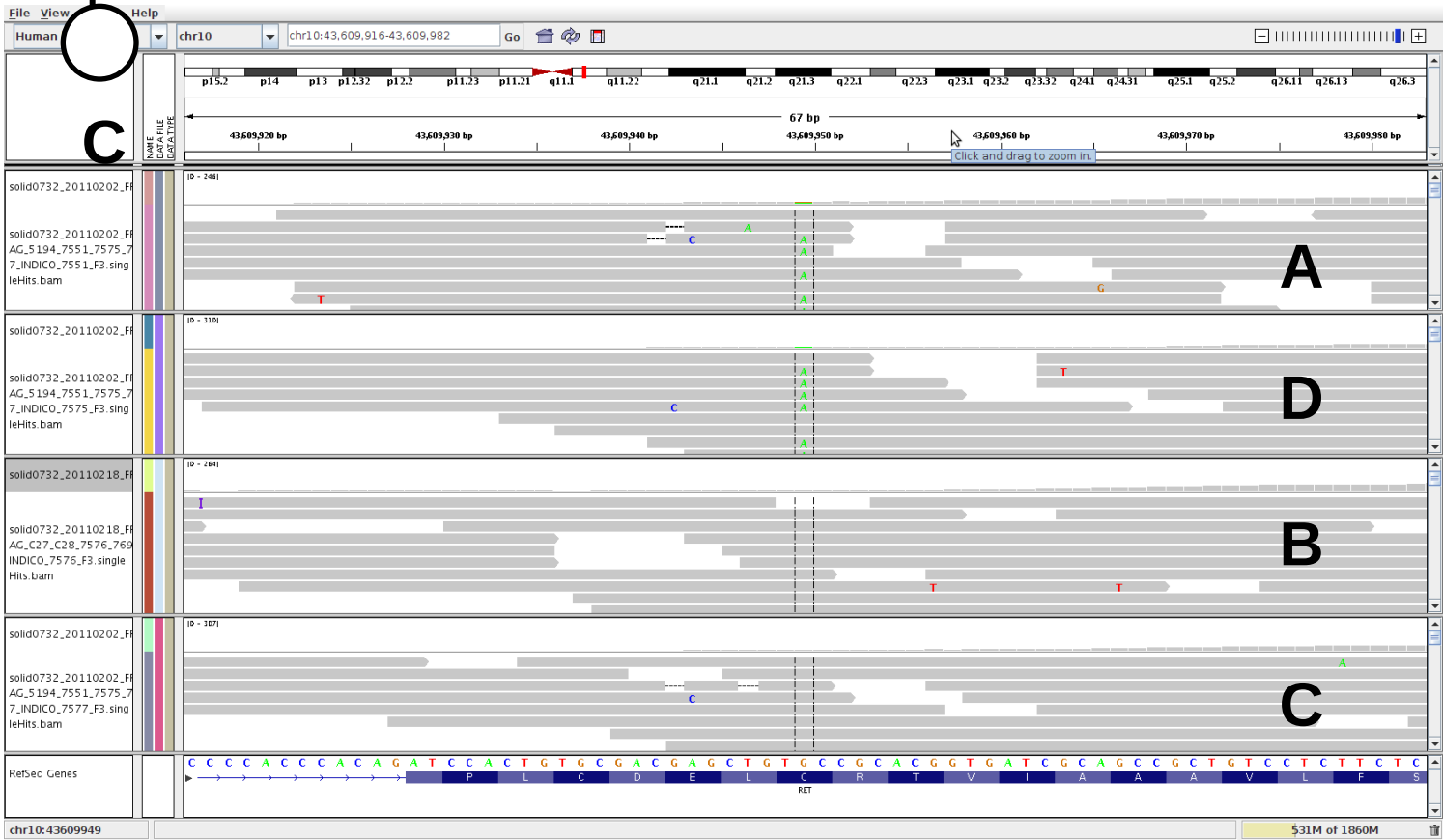
One affected and 73 carriers in EVS

Exome sequencing clearly finds mutations. Low rate of false negatives. An example with MTC

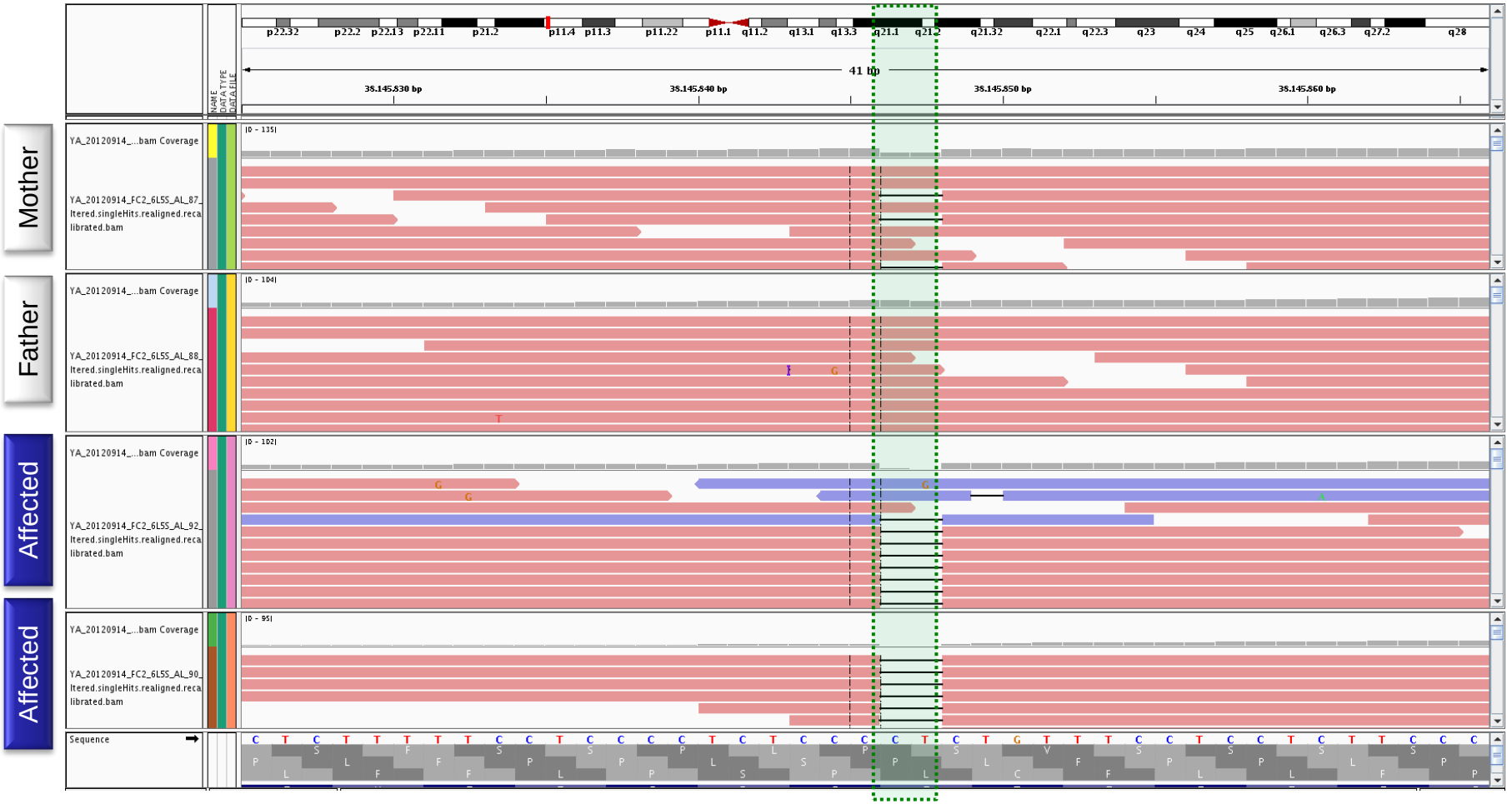


Dominant:
Heterozygotic in A and D
Homozygotic reference allele in B and C
Homozygotic reference allele in controls

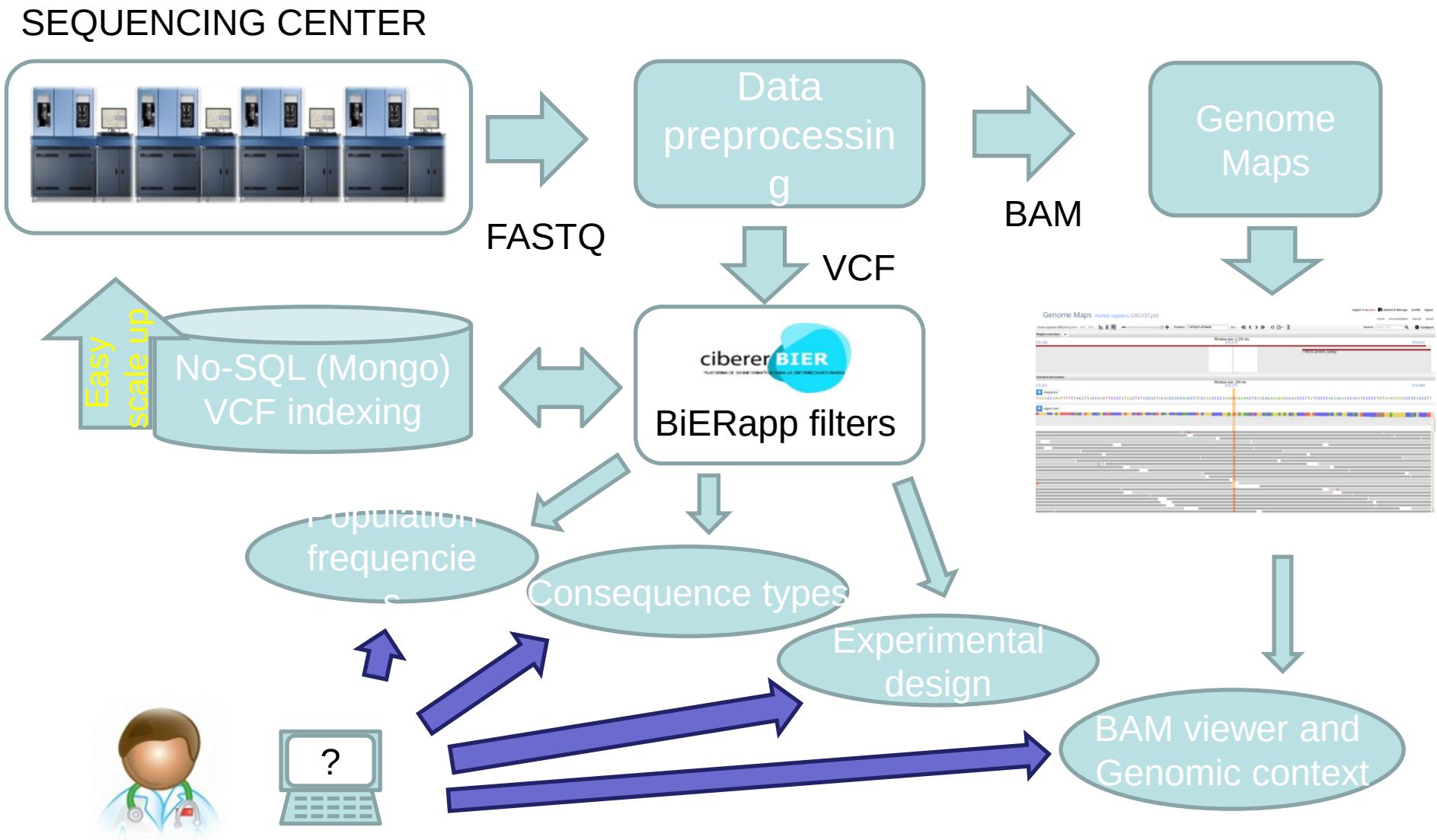
The codon 634 mutation

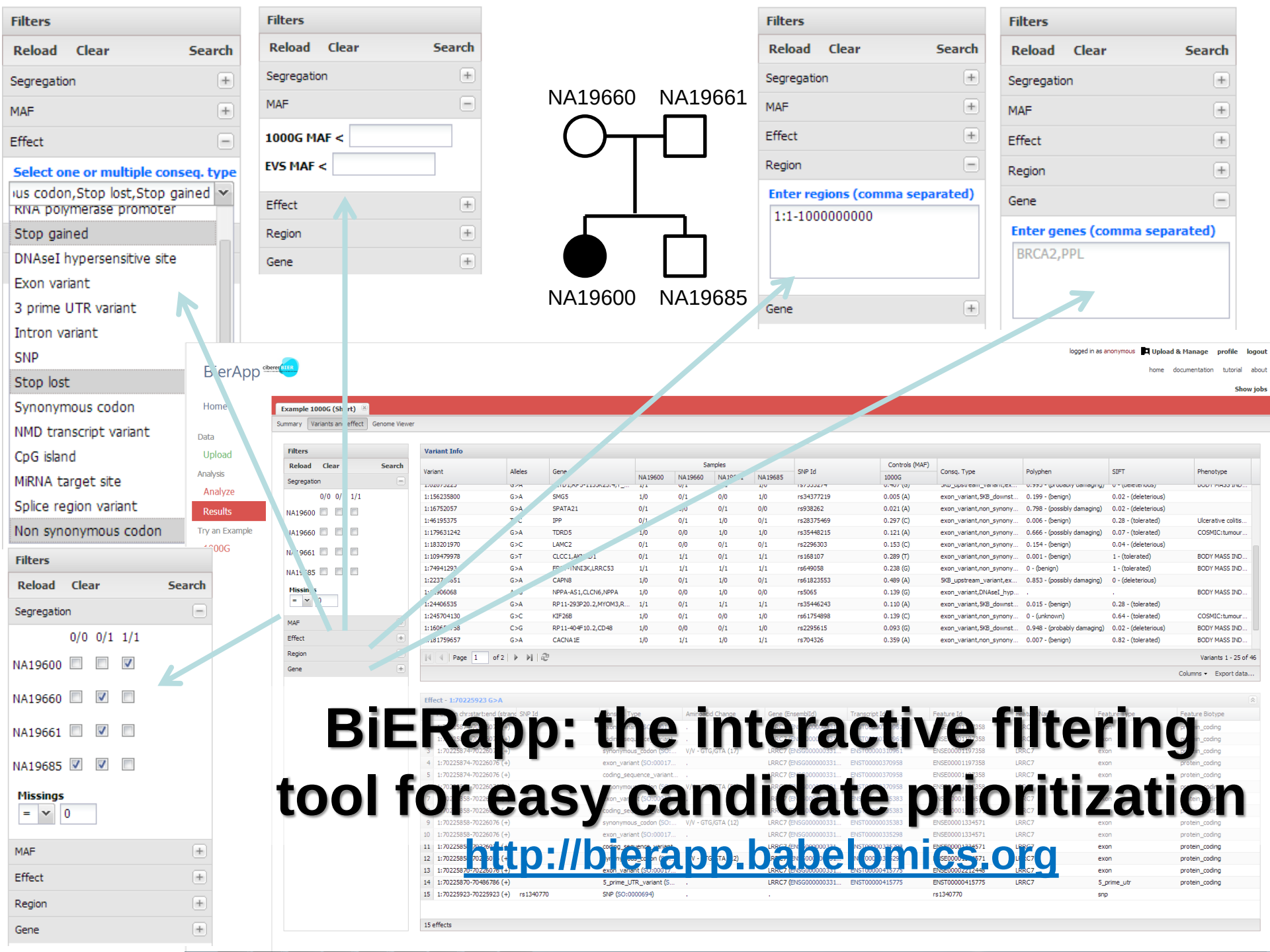


Another example: RP X-linked. Gene *RPGR*

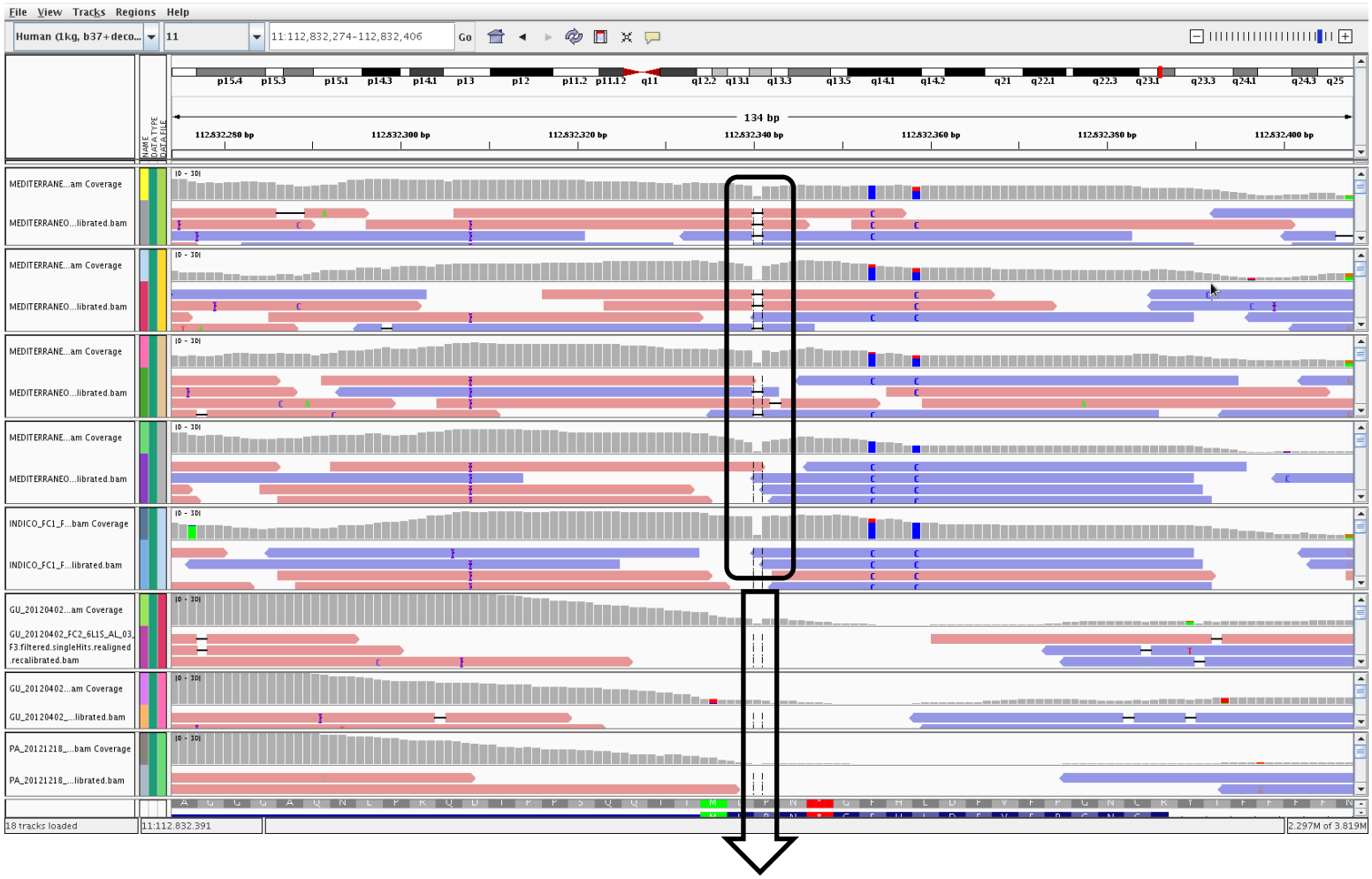


BiERapp: interactive web-based tool for easy candidate prioritization by successive filtering




Filters
Reload Clear Search
Segregation
MAF
Effect
Select one or multiple conseq. type
Stop gained
DNAseI hypersensitive site
Exon variant
3 prime UTR variant
Intron variant
SNP
Stop lost
Synonymous codon
NMD transcript variant
CpG island
MiRNA target site
Splice region variant
Non synonymous codon
1000G MAF <
EVS MAF <
Region
Gene
NA19660 NA19661
NA19600 NA19685
Enter regions (comma separated)
1:1-1000000000
Enter genes (comma separated)
BRCA2,PPL
0/0 0/1 1/1
NA19600
NA19660
NA19661
NA19685
Missings
BiERapp
Home
Data
Upload
Analysis
Analyze
Results
Try an Example
Variant Info
Effect - 1:70225923 G>A
BiERapp: the interactive filtering tool for easy candidate prioritization
http://bierapp.babelomics.org

However, WES finds more variants than you'd like. High rate of false positives



The promising variant (a frameshift present in all patients but not detected in controls) was nor real. It was not properly covered by reads in controls.

Consecutive filters approach. Exome sequencing produces many false positives

Table 1 | **Mean number of coding variants in two populations**

Variant type	Mean number of variants (± sd) in African Americans	Mean number of variants (± sd) in European Americans
Novel variants		
Missense	303 (± 32)	192 (± 21)
Nonsense	5 (± 2)	5 (± 2)
Synonymous	209 (± 26)	109 (± 16)
Splice	2 (± 1)	2 (± 1)
Total	520 (± 53)	307 (± 33)
Non-novel variants		
Missense	10,828 (± 342)	9,319 (± 233)
Nonsense	98 (± 8)	89 (± 6)
Synonymous	12,567 (± 416)	10,536 (± 280)
Splice	36 (± 4)	32 (± 3)
Total	23,529 (± 751)	19,976 (± 505)
Total variants		
Missense	11,131 (± 364)	9,511 (± 244)
Nonsense	103(± 8)	93 (± 6)
Synonymous	12,776 (± 434)	10,645 (± 286)
Splice	38 (± 5)	34 (± 4)
Total	24,049 (± 791)	20,283 (± 523)

The table lists the mean number (± standard deviation (sd)) of novel and non-novel coding single nucleotide variants from 100 sampled African Americans and 100 European Americans. Non-novel variants refer to those found in dbSNP131 or in 200 other control exomes. Capture was performed using the Nimblegen V2 target. The analysis pipeline consisted of: alignment using the Burrows–Wheeler alignment tool; recalibration; realignment around insertion–deletions and merging with the Genome Analysis Toolkit (GATK)[91]; and removal of duplicates with PICARD. Variants were called using the following parameters: quality score > 50, allele balance ratio < 0.75; homopolymer run > 3; and quality by depth < 8. Variants were called from a RefSeq37.2 target (35,804,408 bp).

Bamshad et al., 2011, Nat.Rev.Genet.

Average values obtained per exome (>800)

After filtering by:	SNVs
Conventional filter QC, coverage…	60,000
Mapping and haplotype coherence, missing sites…	30,000
Nonsynonymous (nonsense and missense)	5,000
Unknown (not present in controls)	150-300
Segregate with the families	< 100

We <u>can</u> detect the disease mutation(s)… along with many other unrelated variants

And there are many real variants with potential phenotypic effect

Findings:
20.000 total variants
1000 new variants
300-500 LOF variants (>50 homozygous)
100 known variants associated to disease

A report must contain:
1) **Diagnostic variants**
2) **Therapy-related variants**
3) **Susceptibility variants**
4) **Incidental findings with risk for the patient**

My first exome…

List of variants

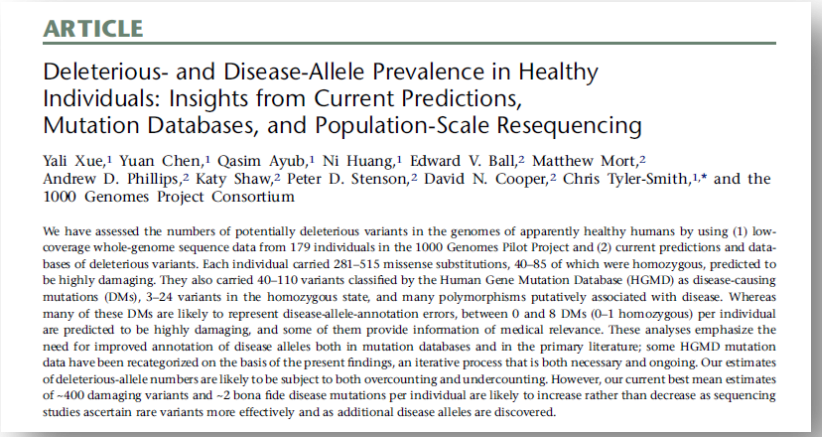
none	1	2116429	C	missense	0	PRKCZ,LOC10
none	1	2116429	C	missense	0	PRKCZ,LOC10
none	1	2116429	C	missense	0	PRKCZ,LOC10
none	1	2116429	C	utr-3	0	PRKCZ,LOC10
none	1	2318893	C	missense	0	MORN1
none	1	2452167	C	missense	0	PANK4
dbSNP_1000Genomes	1	2452569	T	coding-synonymous	2985862	PANK4
none	1	3680294	A	missense	0	CCDC27
none	1	3745852	T	missense	0	KIAA0562
none	1	3746432	G	missense	0	KIAA0562
dbSNP_1000Genomes	1	3755675	T	coding-synonymous	1891941	KIAA0562
none	1	6029181	G	missense	0	NPHP4
none	1	6101899	A	missense	0	KCNAB2
none	1	6101899	A	intron	0	KCNAB2
none	1	6132842	C	coding-synonymous	0	KCNAB2
none	1	6132842	C	coding-synonymous	0	KCNAB2
none	1	6535559	T	missense	0	PLEKHG5
none	1	6535559	T	missense	0	PLEKHG5
none	1	6535559	T	missense	0	PLEKHG5
none	1	6535559	T	missense	0	PLEKHG5
none	1	6535559	T	missense	0	PLEKHG5
none	1	6647590	A	missense	0	ZBTB48
none	1	6694129	T	missense	0	THAP3
none	1	6695719	T	utr-3	0	DNAJC11
none	1	6704720	C	missense	0	DNAJC11
none	1	6711636	C	coding-synonymous	0	DNAJC11
dbSNP_1000Genomes	1	7889941	C	coding-synonymous	2640908	PER3
dbSNP_1000Genomes	1	7890117	T	missense	2640909	PER3
dbSNP_1000Genomes	1	8425900	T	coding-synonymous	3753275	RERE
dbSNP_1000Genomes	1	8425900	T	utr-5	3753275	RERE
dbSNP_1000Genomes	1	8425900	T	coding-synonymous	3753275	RERE
none	1	9086361	C	missense	0	SLC2A7
none	1	9117600	A	missense	0	SLC2A5
none	1	9117600	A	missense	0	SLC2A5
none	1	9129619	C	utr-5	0	SLC2A5
none	1	9129619	C	utr-5	0	SLC2A5
none	1	9770594	C	coding-synonymous	0	PIK3CD
none	1	10042458	A	missense	0	NMNAT1

```
Known snps phenotypic effect
-----------------------------

Hits     Description
8        Amyotrophic Lateral Sclerosis (ALS)
7        Parkinson's disease
6        Rheumatoid Arthritis
5        common polymorphism
4        Multiple complex diseases-Crohn's disease, combined control
      dataset
3        Alzheimer's Disease
3        LDL cholesterol
3        Skin pigmentation
3        Type 1 diabetes
2        Coronary Artery Disease
2        in allele DQB1*0501 and allele DQB1*0502
2        Multiple complex diseases-Bipolar disorder
2        Multiple complex diseases-Coronary Artery Disease, gender
     differentiated
2        Multiple complex diseases-Crohn's disease, combined control
      dataset, gender differentiated
2        Multiple complex diseases-Type I Diabetes, combined control
      dataset
2        Multiple complex diseases-Type II Diabetes Mellitus,
     combined control dataset
2        Systemic Lupus Erythematosus (SLE), gender differentiated
     in women
2        Triglycerides
2        Type I Diabetes
1        893Ser-expressing (ABCB1:2677G>T (Ala893Ser)) cells showed
     47% lower intracellular digo...
1        A study in 336 recipients of hematopoietic-cell transplants
```

A high level of deleterious variability exists in the human genome

- Variants predicted to severely affect the function of human protein coding genes known as loss-of-function (LOF) variants were thought:
  - To have a potential deleterious effect
  - To be associated to severe Mendelian disease



ARTICLE

Deleterious- and Disease-Allele Prevalence in Healthy Individuals: Insights from Current Predictions, Mutation Databases, and Population-Scale Resequencing

Yali Xue,[1] Yuan Chen,[1] Qasim Ayub,[1] Ni Huang,[1] Edward V. Ball,[2] Matthew Mort,[2] Andrew D. Phillips,[2] Katy Shaw,[2] Peter D. Stenson,[2] David N. Cooper,[2] Chris Tyler-Smith,[1,*] and the 1000 Genomes Project Consortium

We have assessed the numbers of potentially deleterious variants in the genomes of apparently healthy humans by using (1) low-coverage whole-genome sequence data from 179 individuals in the 1000 Genomes Pilot Project and (2) current predictions and databases of deleterious variants. Each individual carried 281–515 missense substitutions, 40–85 of which were homozygous, predicted to be highly damaging. They also carried 40–110 variants classified by the Human Gene Mutation Database (HGMD) as disease-causing mutations (DMs), 3–24 variants in the homozygous state, and many polymorphisms putatively associated with disease. Whereas many of these DMs are likely to represent disease-allele-annotation errors, between 0 and 8 DMs (0–1 homozygous) per individual are predicted to be highly damaging, and some of them provide information of medical relevance. These analyses emphasize the need for improved annotation of disease alleles both in mutation databases and in the primary literature; some HGMD mutation data have been recategorized on the basis of the present findings, an iterative process that is both necessary and ongoing. Our estimates of deleterious-allele numbers are likely to be subject to both overcounting and undercounting. However, our current best mean estimates of ~400 damaging variants and ~2 bona fide disease mutations per individual are likely to increase rather than decrease as sequencing studies ascertain rare variants more effectively and as additional disease alleles are discovered.

- However, an unexpectedly large number of LOF variants have been found in the genomes of apparently healthy individuals: 281-515 missense substitutions per individual, 40-85 of them in homozygous state and predicted to be highly damaging.

- A similar proportion was observed in miRNAs and possibly affect to any functional element in the genome

Genome Medicine

Search Genome Medicine for

Home | Articles | Authors | Reviewers | About this journal | My Genome Medicine | Subscriptions

Research

Highly accessed | Open Access

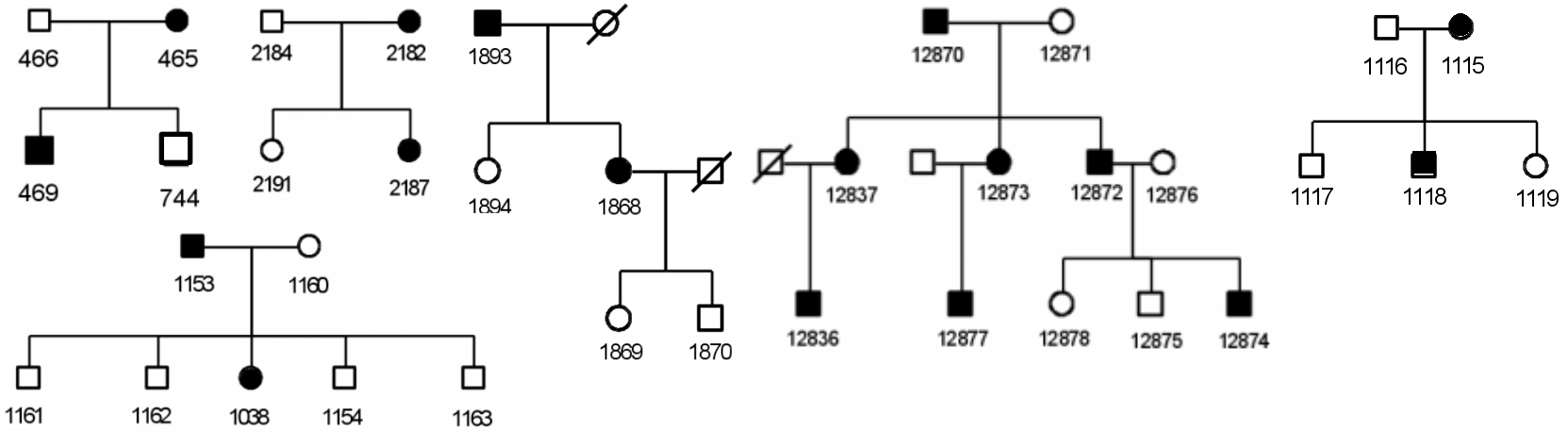
A map of human microRNA variation uncovers unexpectedly high levels of variability

Jose Carbonell, Eva Alloza, Pablo Arce, Salud Borrego, Javier Santoyo, Macarena Ruiz-Ferrer, Ignacio Medina, Jorge Jimenez-Almazan, Cristina Mendez-Vidal, Maria Gonzalez-del Pozo, Alicia Vela, Shomi S Bhattacharya, Guillermo Antinolo and Joaquin Dopazo

Such apparently deleterious mutation must be first detected and then distinguished from real pathological mutations

Moreover, even Mendelian genes can be elusive.

Intuitive belief: multiple family information should help



	Families					
	1	2	3	4	5	6
Variants	3403	82	4	0	0	0
Genes	2560	331	35	8	1	0

Observation: this is not always true, not even in cases of Mendelian diseases

Is the single-gene approach realistic? Can we easily detect disease-related variants?

There are several problems:

a) Interrogating 60Mb sites (3000 Mb in genomes) produces too many variants. A large number of these segregating with our experimental design

b) There is a non-negligible amount of **apparently deleterious** variants that (apparently) has no pathologic effect

c) In many cases we are not targeting rare but **common** variants (which occur in normal population)

d) In many cases only one variant does not explain the disease but rather a **combination** of them (epistasis)

e) Consequently, the few individual variants found associated to the disease usually account for a **small portion** of the trait **heritability**

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 [illegible]

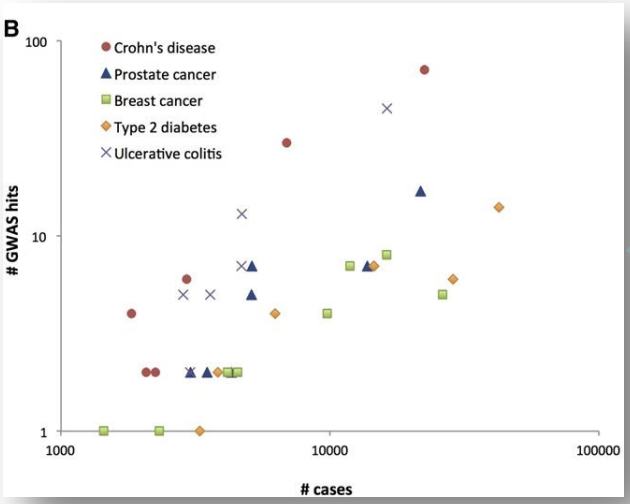
Height in humans is a classical quantitative trait, easy to measure and studied for well over a century as a model for investigating the genetic basis [illegible] of height has been estimated to be ~0.8 (refs. 9,11–13). Rare mutations that cause extreme short or tall stature have been found[14,15], but 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

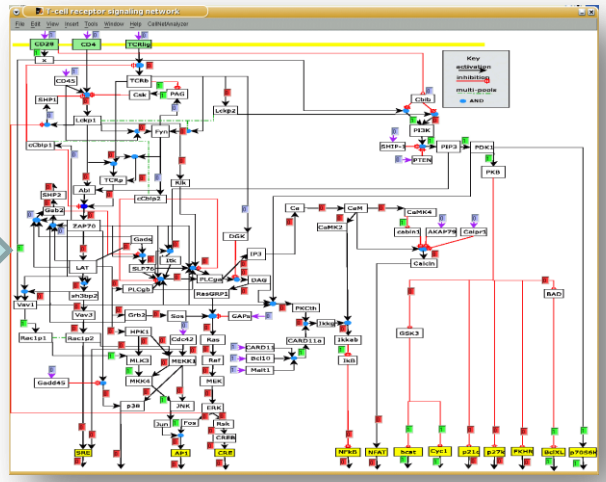
2010 Nature America, Inc. All rights reserved.

At the end, most of the heritability was there…

At the crossroad: how detection power of genomic technologies can be increased?



There are two (non mutually exclusive) ways



Scaling up: by increasing sample size.

It is known that larger size allows detecting more individual gene (biomarker) associations.

Limitations: Budget, patients availability and the own nature of the disease.

Changing the perspective: systems approach to understand variation

Interactions, multigenicity can be better detected and the role of variants understood in the context of disease mechanism.

Limitations: Available information

Modular nature of human genetic diseases

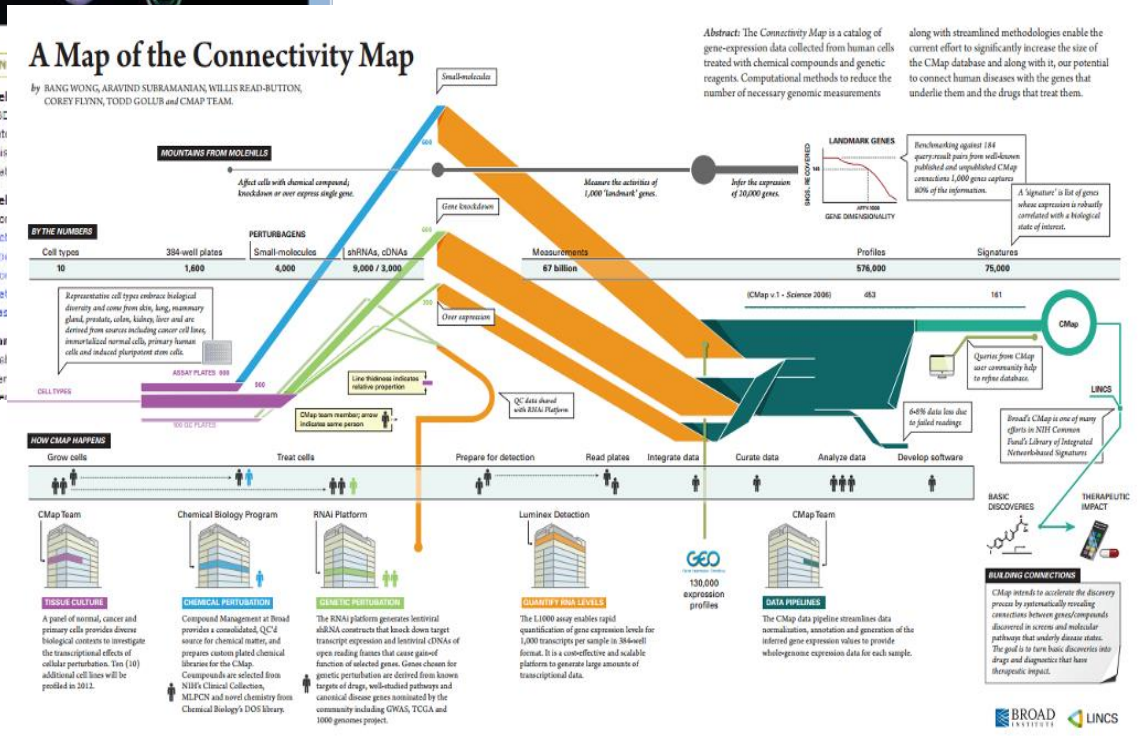
- With the development of **systems biology**, studies have shown that phenotypically similar diseases are often caused by **functionally related genes**, being referred to as the **modular nature of human genetic diseases** (Oti and Brunner, 2007; Oti et al, 2008).
- This modularity suggests that **causative genes** for the same or phenotypically similar diseases may generally reside in the same **biological module**, either a **protein complex** (Lage et al, 2007), a **sub-network** of protein interactions (Lim et al, 2006) , or a **pathway** (Wood et al, 2007)

Initiatives to explore beyond single gene biomarkers

LINCS aims to create a **network-based** understanding of biology by cataloging changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbing agents, and by using **computational** tools to integrate this diverse information into a comprehensive view of normal and disease states that can be applied for the development of **new biomarkers** and **therapeutics**.

LINCS aims to create a network-based understanding of biology by cataloging changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbing agents, and by using computational tools to integrate this diverse information into a comprehensive view of normal and disease states that can be applied for the development of new biomarkers and therapeutics. By generating and making public data that indicates how cells respond to various genetic and environmental stressors, the LINCS project will help us gain a more detailed understanding of cell pathways and aid efforts to develop therapies that might restore perturbed pathways and networks to their normal states.

This website is a source of information for the research community and general public about the LINCS project. It contains information about the experiments conducted, as well as links to participating LINCS centers' websites, data releases from LINCS centers, and tools that can be used for analyzing the data.

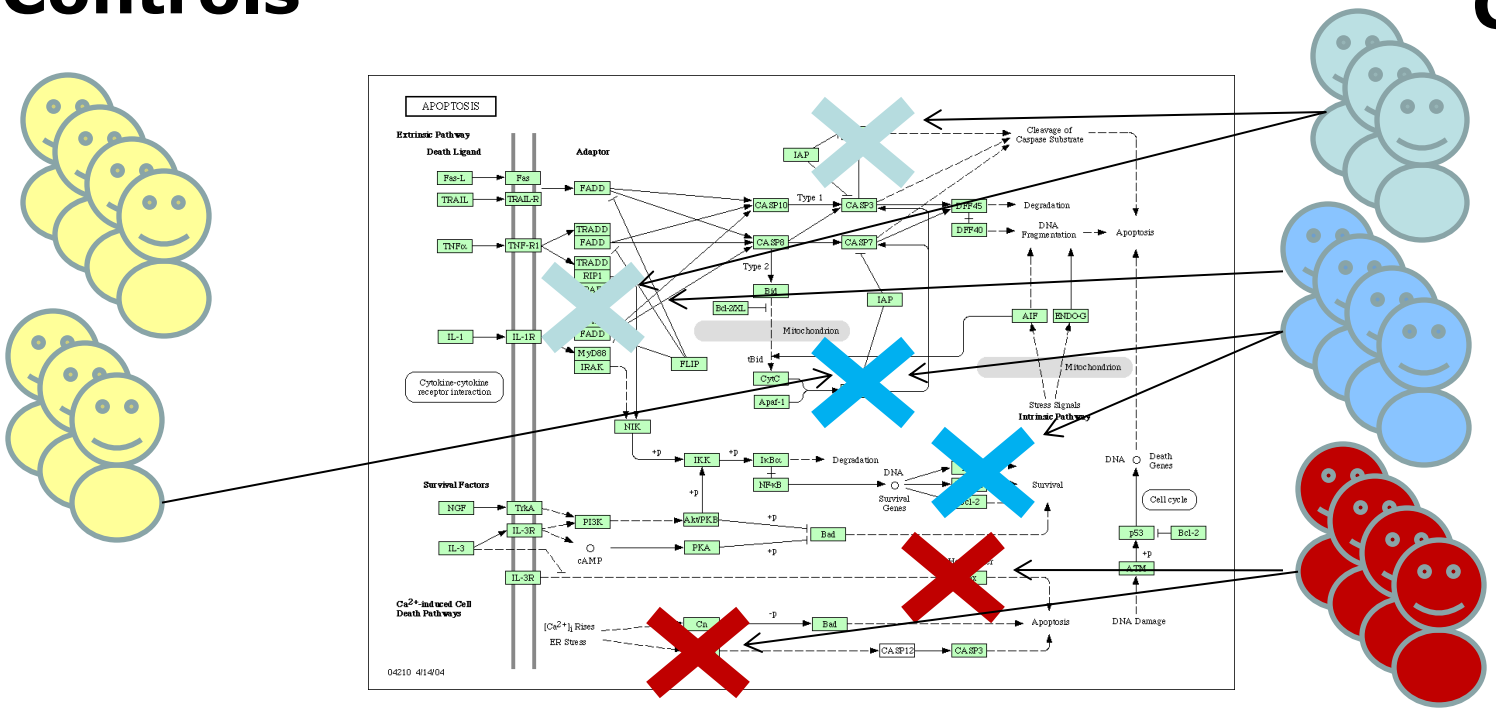


And the connectivity map, a pioneer initiative

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

Controls

Cases



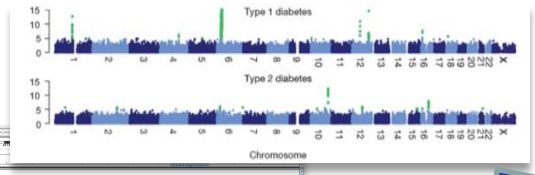
Affected **cases** in complex diseases will be a **heterogeneous** population with different mutations (or combinations).

Many cases and controls are needed to obtain significant associations.

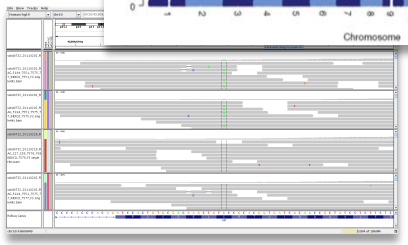
The only **common element** is the (know or unknown) **pathway affected**.

Disease understood as the failure of a functional module

From gene-based to function-based perspective



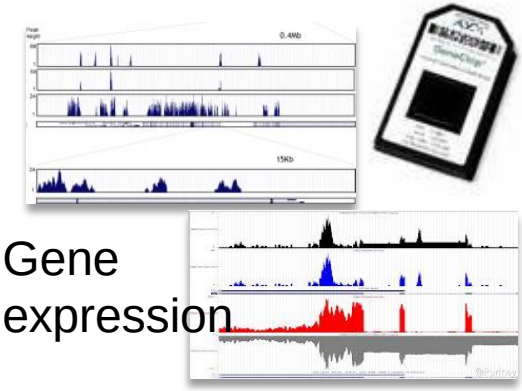
SNPs



WES/WGS

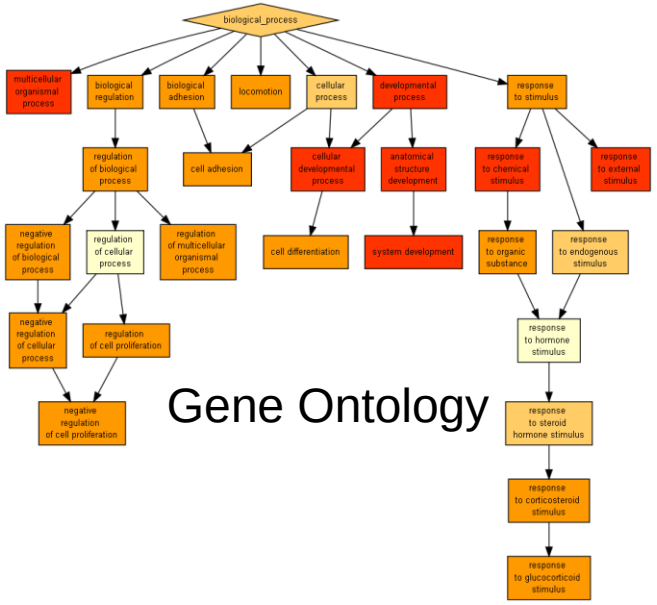
Gene integrity

AND/OR



Gene expression

Gene activity



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.

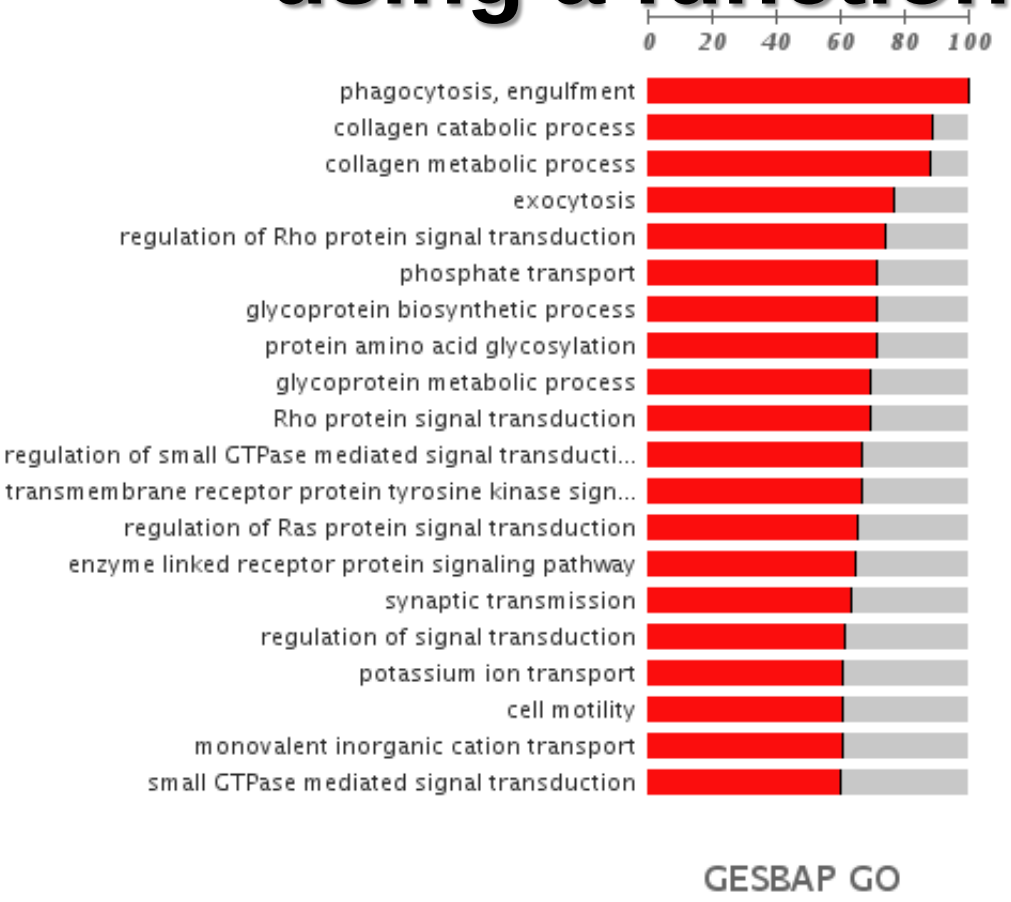
The CGEMS initiative. (Hunter et al. Nat Genet 2007)

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=4.45×10^{-03}) in which FGFR2 is included.

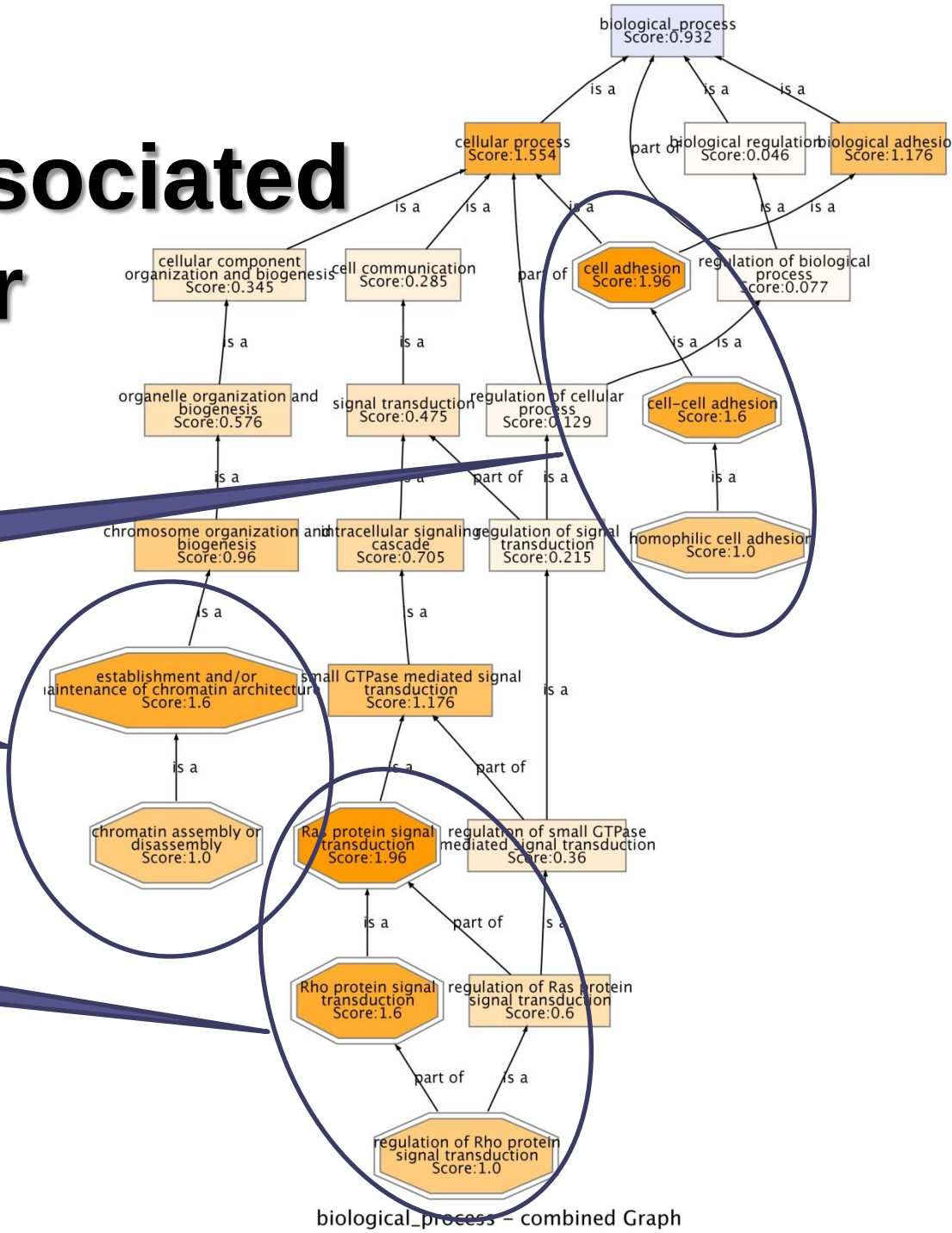
Bonifaci et al., BMC Medical Genomics 2008; Medina et al., 2009 NAR

GO processes significantly associated to breast cancer

Metastasis

Chromosomal instability

Rho pathway

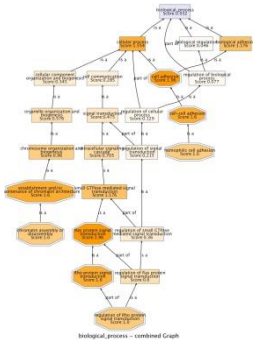


From gene-based to function-based perspective

SNPs,
Gene expression

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

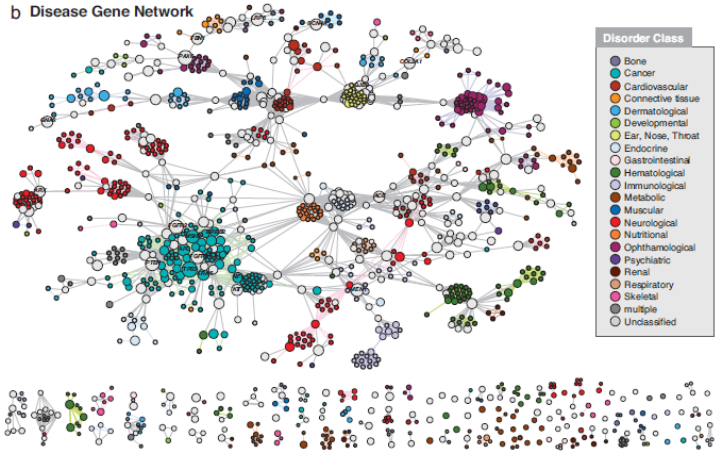
Gene Ontology



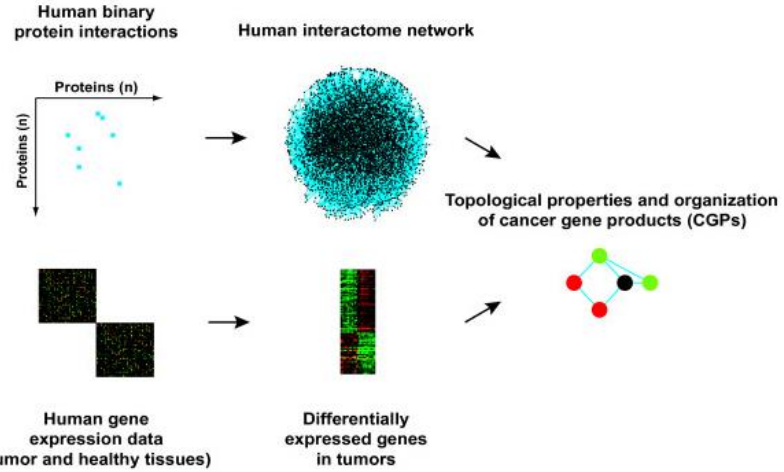
	SNPs, gene exp.	GO
Detection power	Low (only very prevalent genes)	high
Annotations available	many	many
Use	Biomarker	Illustrative, give hints

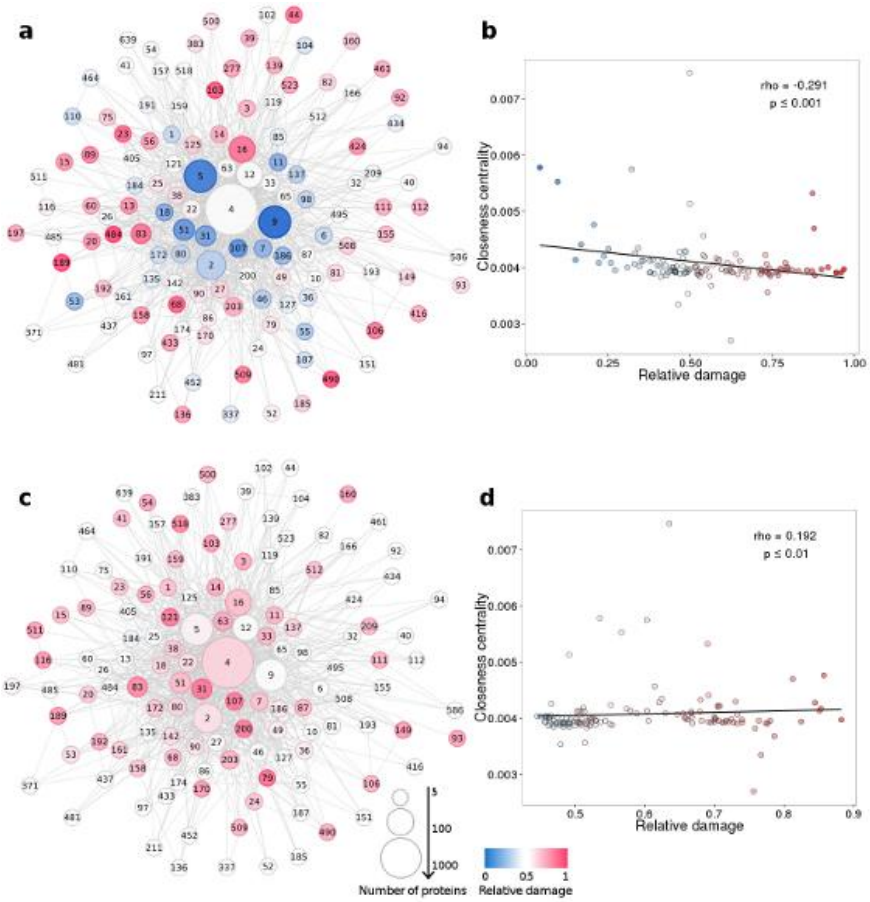
Can the interactome help to find disease mutations?

Disease genes are close in the interactome
Goh 2007 PNAS



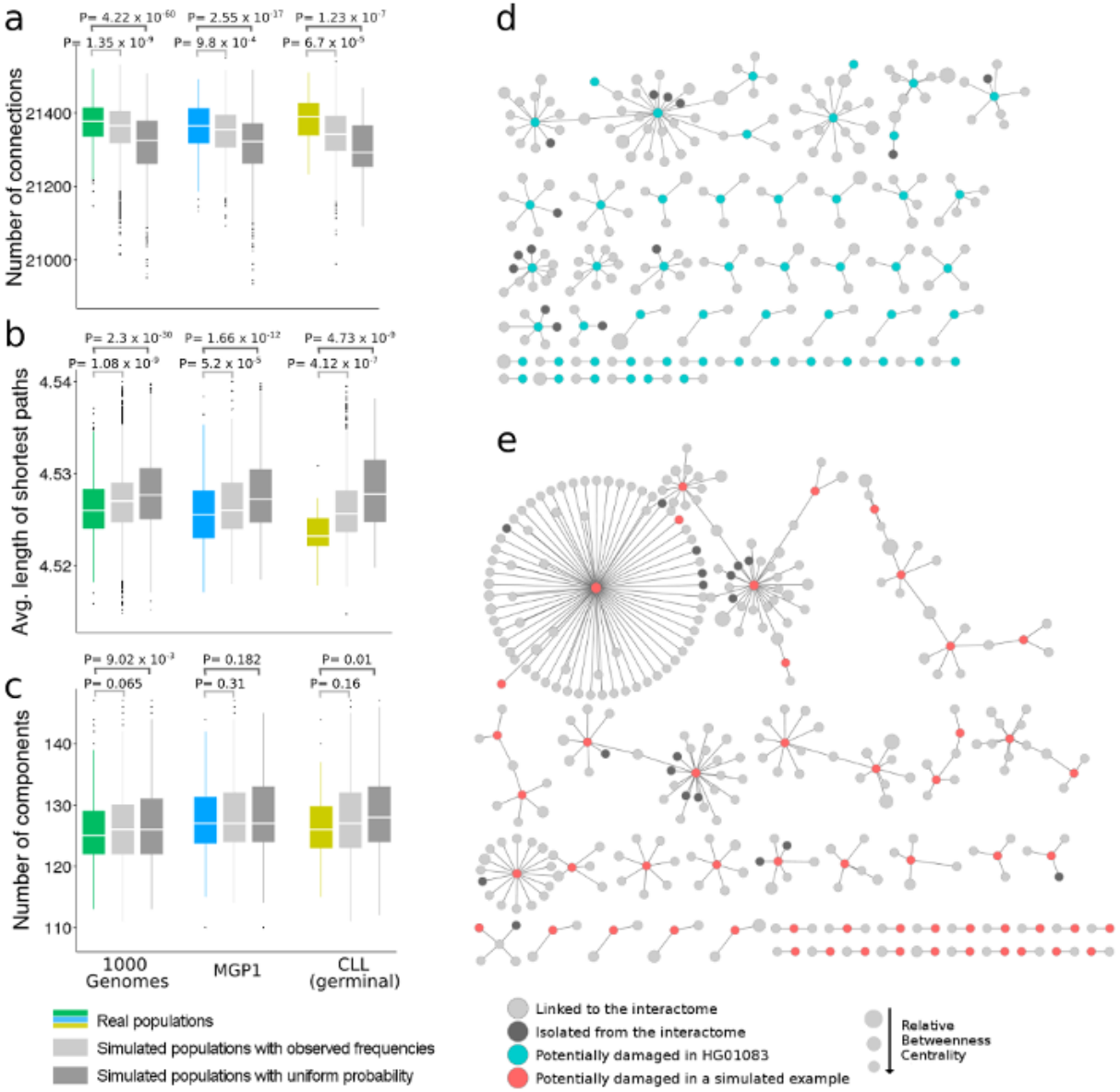
Cancer genes are central.
Hernandez, 2007 BMC Genomics





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

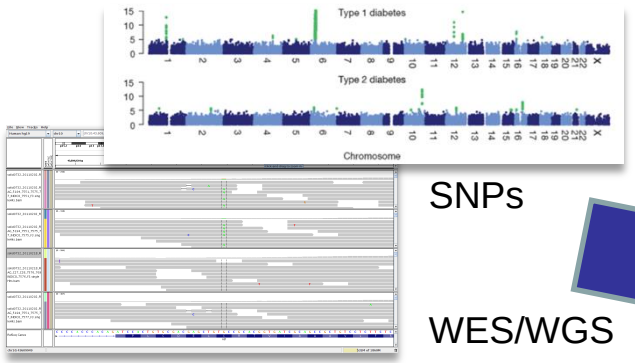
The role of interactome in buffering the deleteriousness of LoF mutations



Comparison of the interactome damage between real and random individuals after removing the nodes corresponding to proteins containing deleterious variants in both alleles (homozygote). Two different scenarios are simulated: Simulated populations with **uniform probability**, where proteins are randomly removed, and Simulated populations with **observed frequencies**, where proteins are removed with a probability proportional to the frequency of variation in the 1000 genomes population

Garcia-Alonso 2014 Mol Syst Biol

From gene-based to function-based perspective

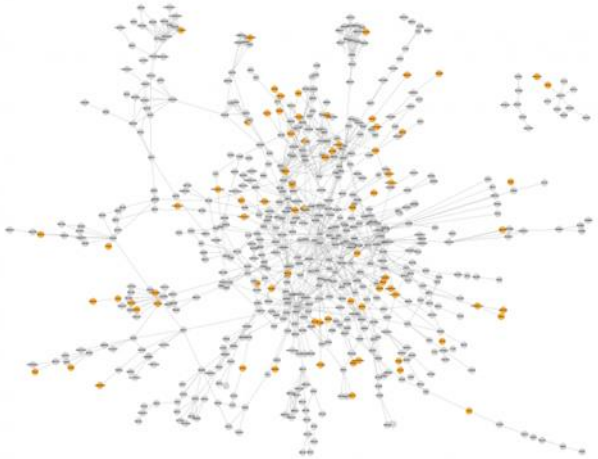


SNPs

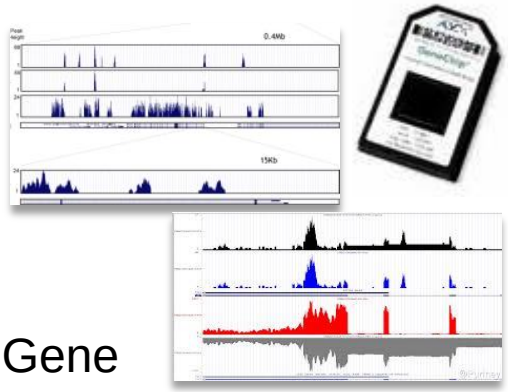
WES/WGS

Gene integrity

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



AND/OR



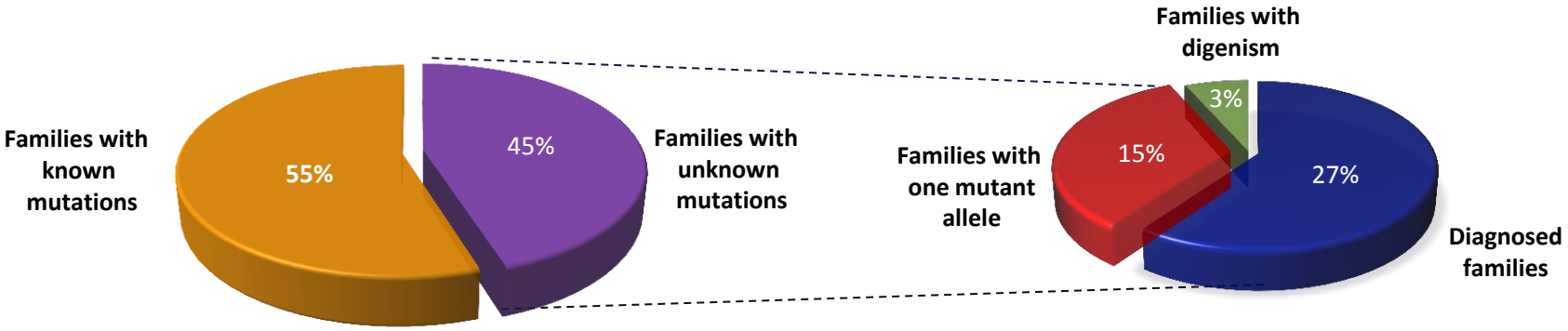
Gene expression

Gene activity

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

Example with Inherited Retinal Dystrophies

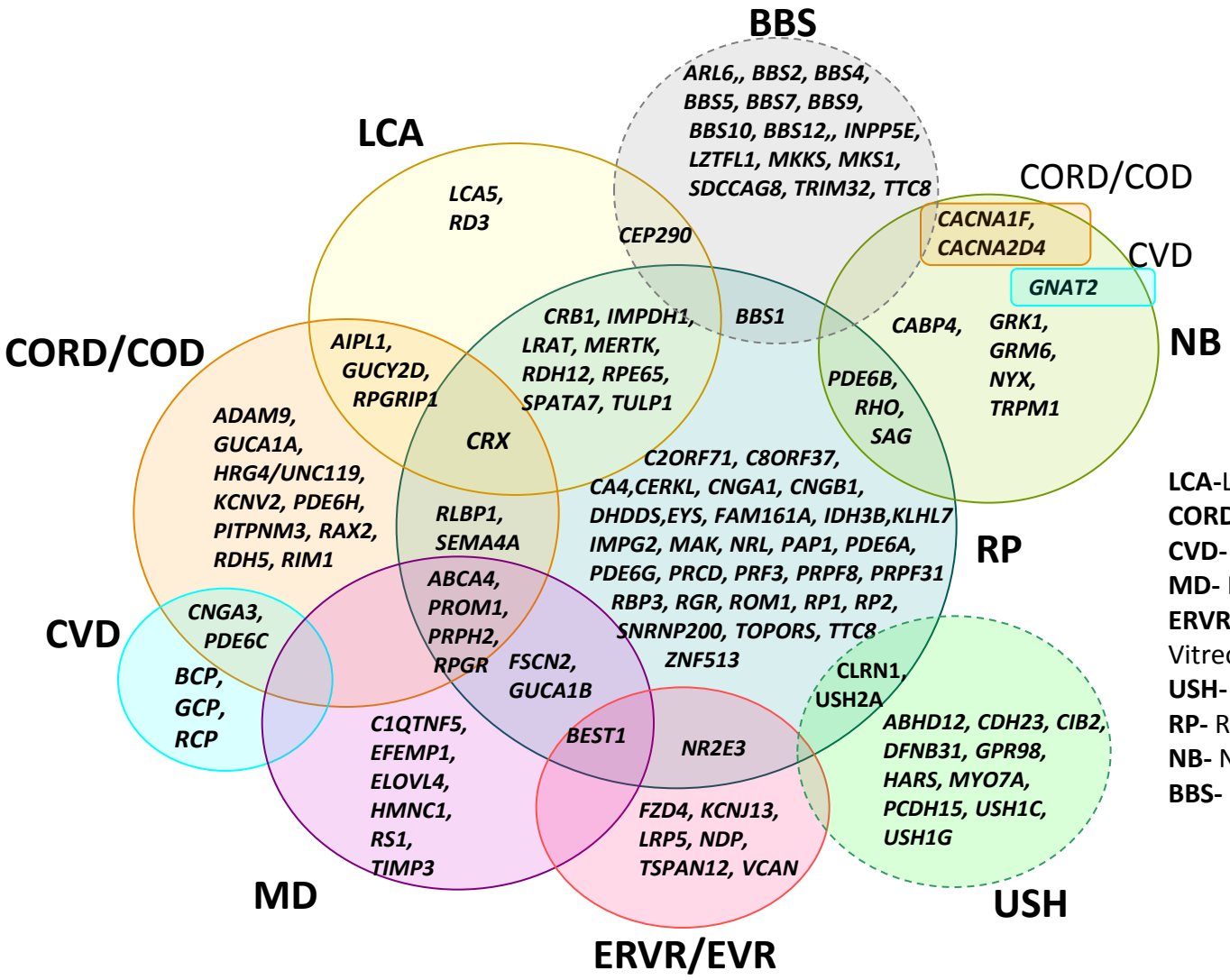
- Prevalence 1 in 3000
- Clinically and genetically very heterogeneous
- 190 GENES account for aprox. 50% of IRDs.



Novel variants and genes remain to be discovered

Network analysis.

Is genetic overlapping among IRDs related to protein interaction?

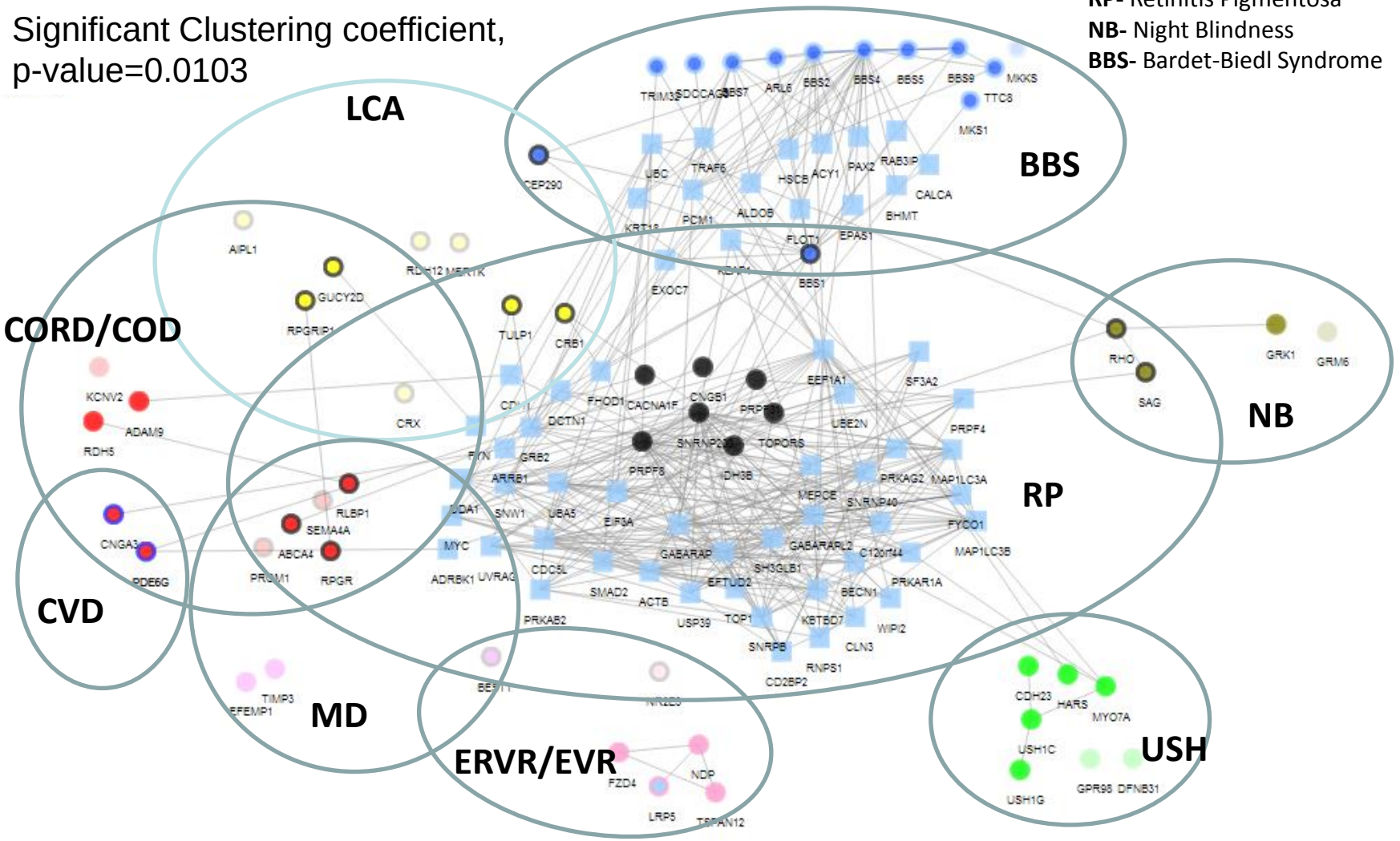


LCA-Leber Congenital Amaurosis
CORD/COD- Cone and cone-rod dystro.
CVD- Colour Vision Defects
MD- Macular Degeneration
ERVR/EVR- Erosive and Exudative Vitreoretinopathies
USH- Usher Syndrome
RP- Retinitis Pigmentosa
NB- Night Blindness
BBS- Bardet-Biedl Syndrome

Connectivity among IRDs explain genetic overlap

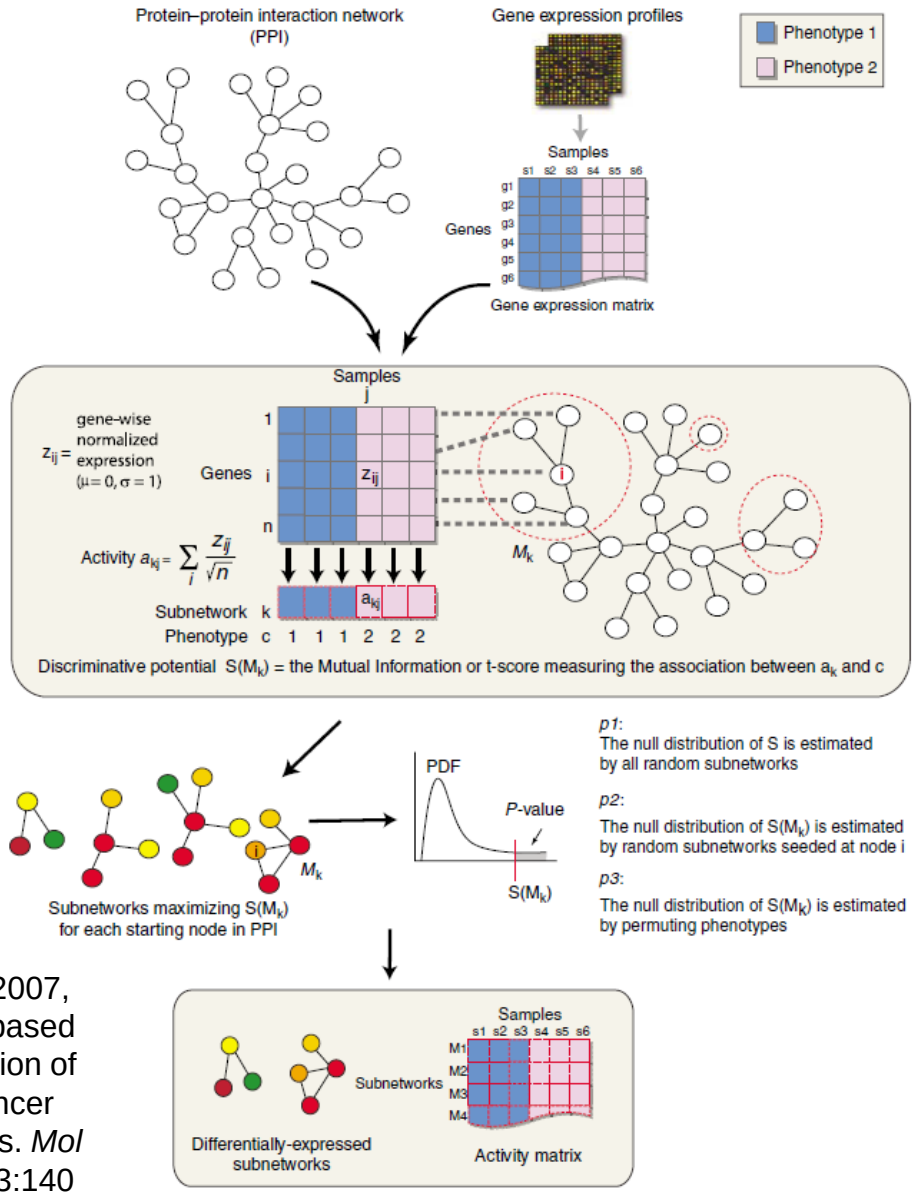
LCA-Leber Congenital Amaurosis
CORD/COD- Cone and cone-rod dystro.
CVD- Colour Vision Defects
MD- Macular Degeneration
ERVR/EVR- Erosive and Exudative Vitreoretinopathies
USH- Usher Syndrome
RP- Retinitis Pigmentosa
NB- Night Blindness
BBS- Bardet-Biedl Syndrome

Significant Clustering coefficient, p-value=0.0103



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

Network enrichment methods

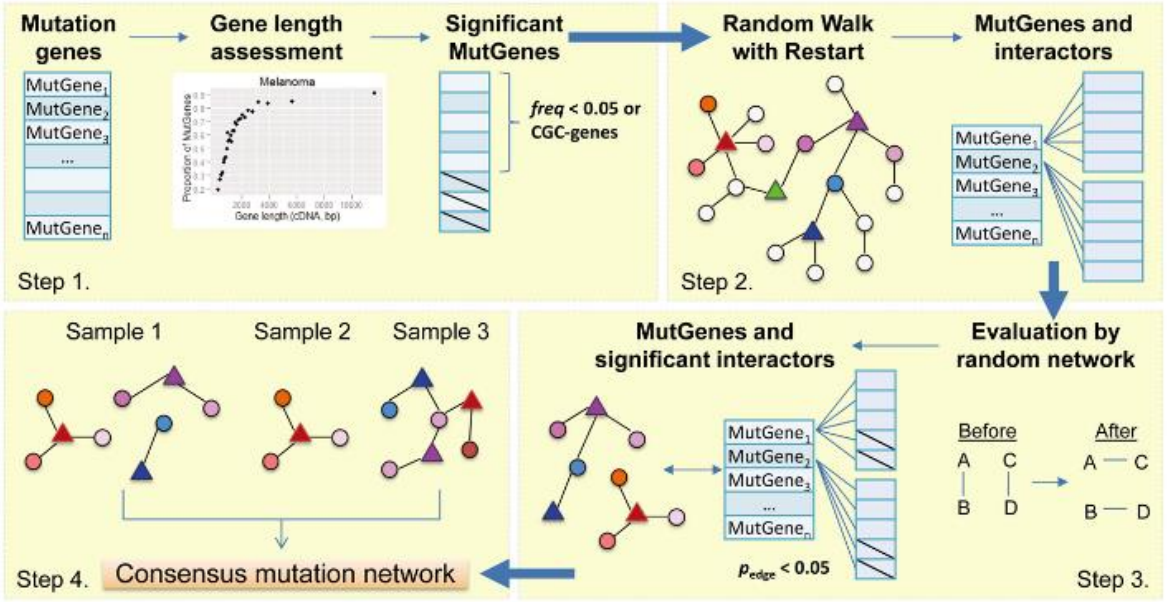


Chuang, 2007, Network-based classification of breast cancer metastasis. *Mol Syst Biol.*3:140

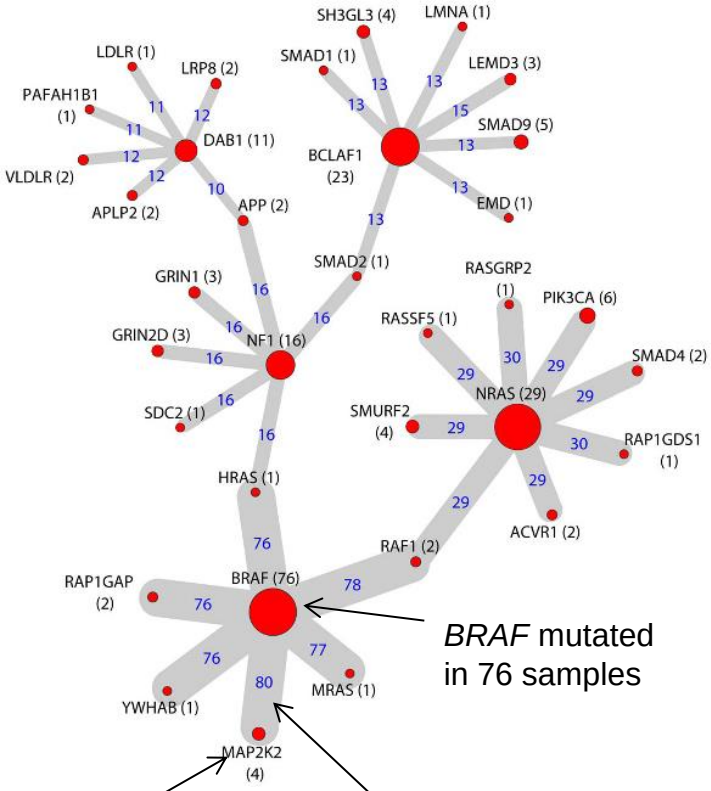
Network analysis in one of its first proposals.

An objective function (discrimination potential) is used to evaluate a growing network built up around a seed (known genes, most differentially expressed genes, etc.)

Mutation network analysis



Subgraphs in the melanoma consensus mutation network



- This approach prioritizes well-known, infrequently mutated genes, which are shown to interact with highly recurrently mutated genes yet have been ignored by conventional single-gene-based approaches.
- These observations suggest cancer genes could mutate in a broad range of frequency spectrums, making it difficult for the frequency-based filtering approach to be effective.

Jia 2013 PLoS Comp. Biol.

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

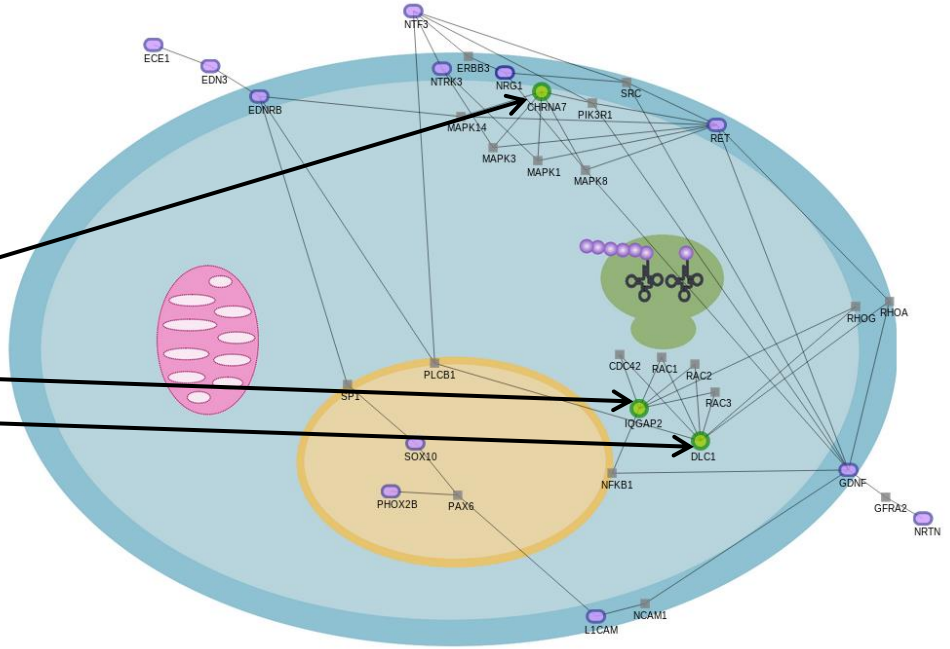
Published online 27 July 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 20 e158*
doi:10.1093/nar/gks699

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 = 3.856×10^{-05})
**no interactions known (yet)*

SNPs validated in independent cohorts

Programs to detect 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

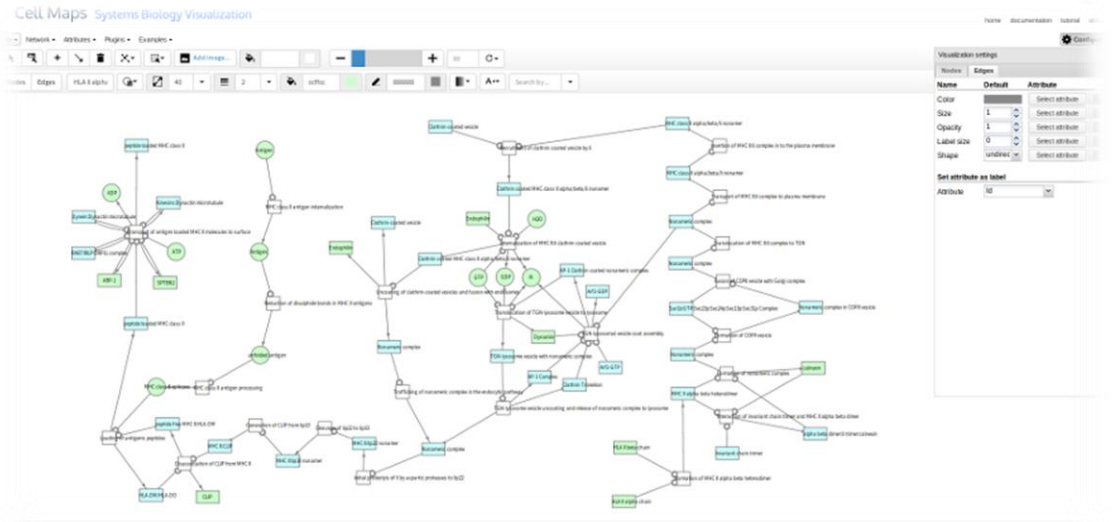
Nucleic Acids Research, 2012, Vol. 40, No. 20 ***e158***
doi:10.1093/nar/gks699

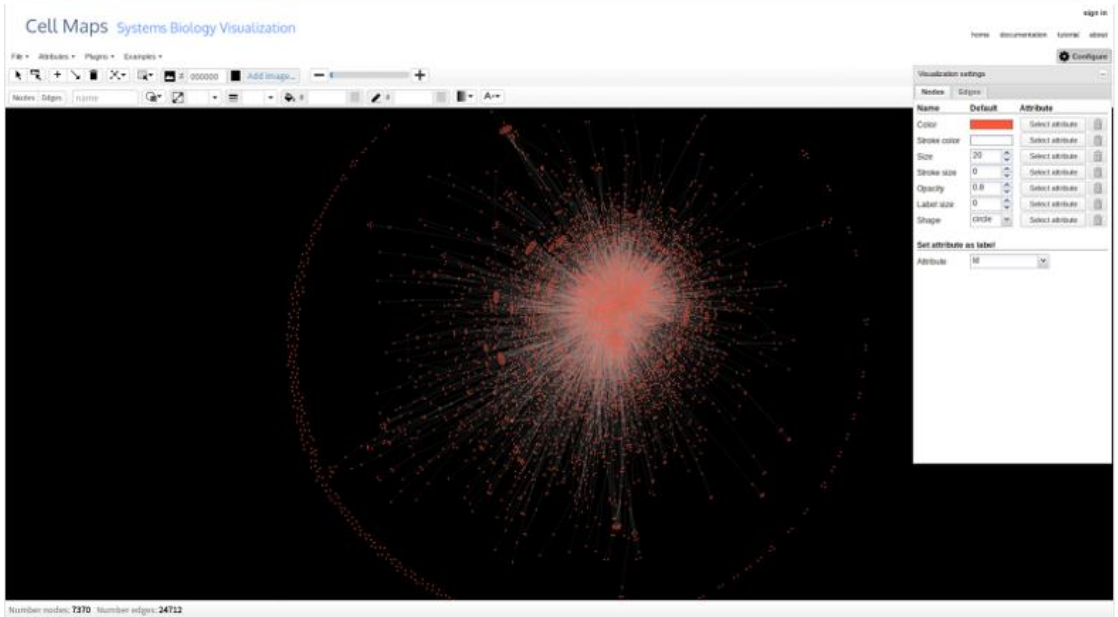
Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments

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/

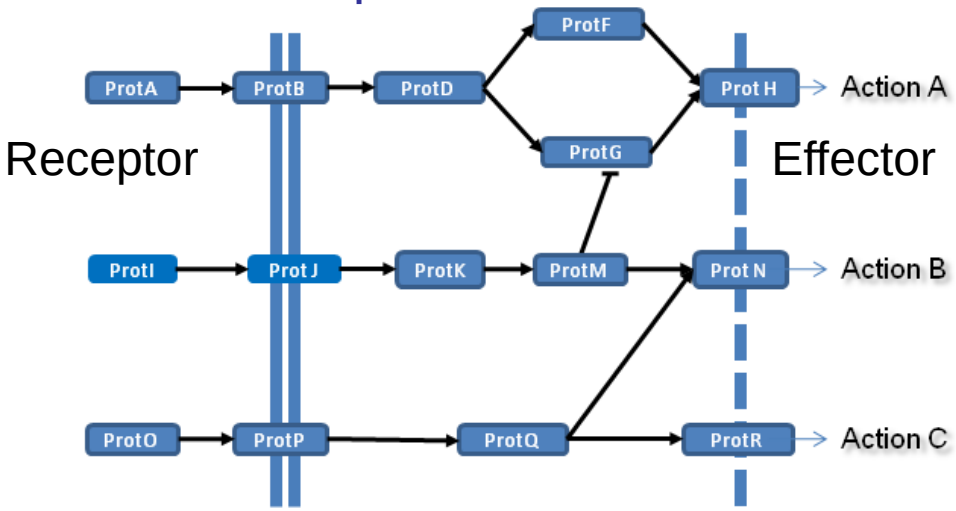
From gene-based to function-based perspective

	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*

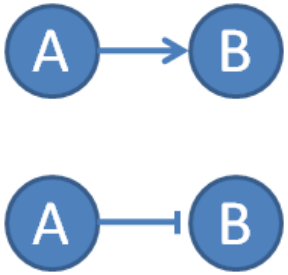
**Need of extra information (e.g. GO) to provide functional insights in the findings*

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

Modeling pathways

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

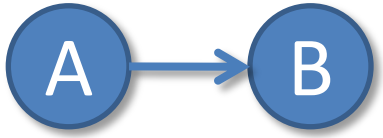
Inferring the functional effect of gene expression changes in signaling pathways

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

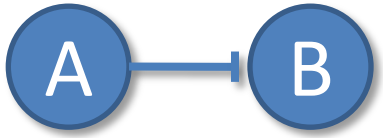
Received March 3, 2013; Revised April 18, 2013; Accepted May 2, 2013

Activation



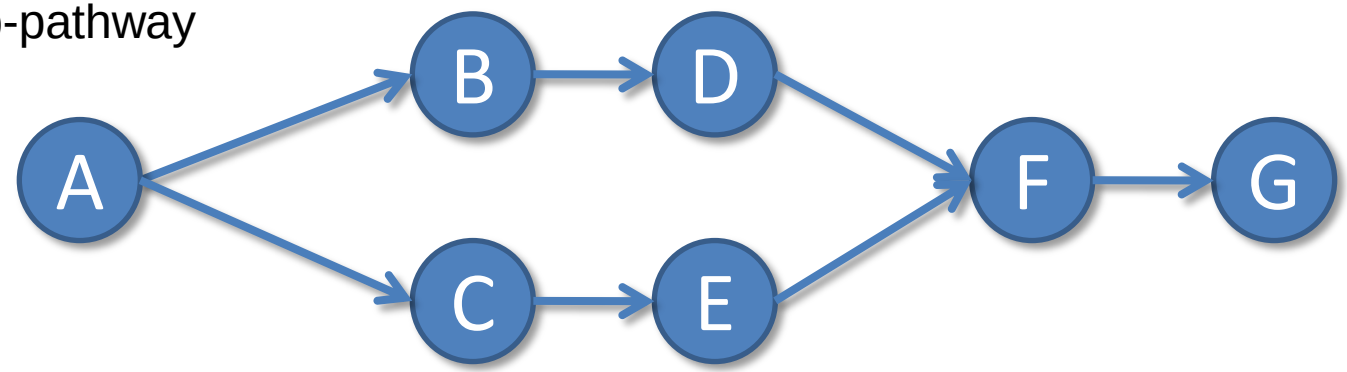
Prob. = P(A activated)P(B activated)

Inhibition



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

Sub-pathway



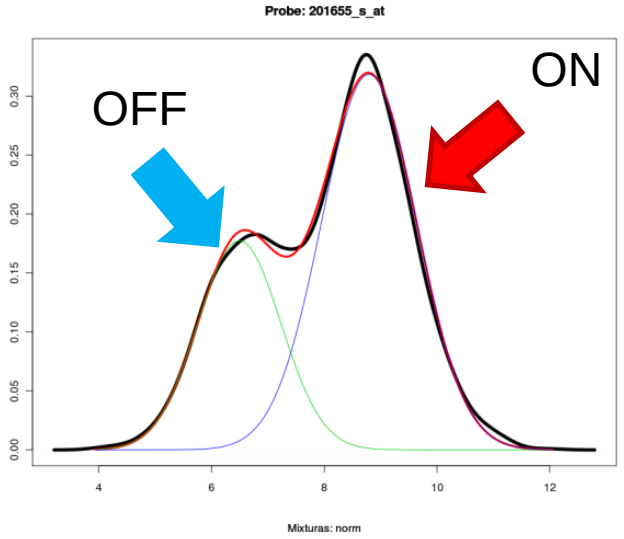
P(A→G activated) = P(A)P(B)P(D)P(F)P(G) + P(A)P(C)P(E)P(F)P(G) - P(A)P(F)P(G)P(B)P(C)P(D)P(E)

Modeling pathways

We only need to estimate the probabilities of **gene activation** and then calculate the probability for each **circuit** of being **active**.

Gene expression

A large dataset of affymetrix microarrays (10,000) is used to adjust a mixture of distributions of **gene activity for all the genes**.



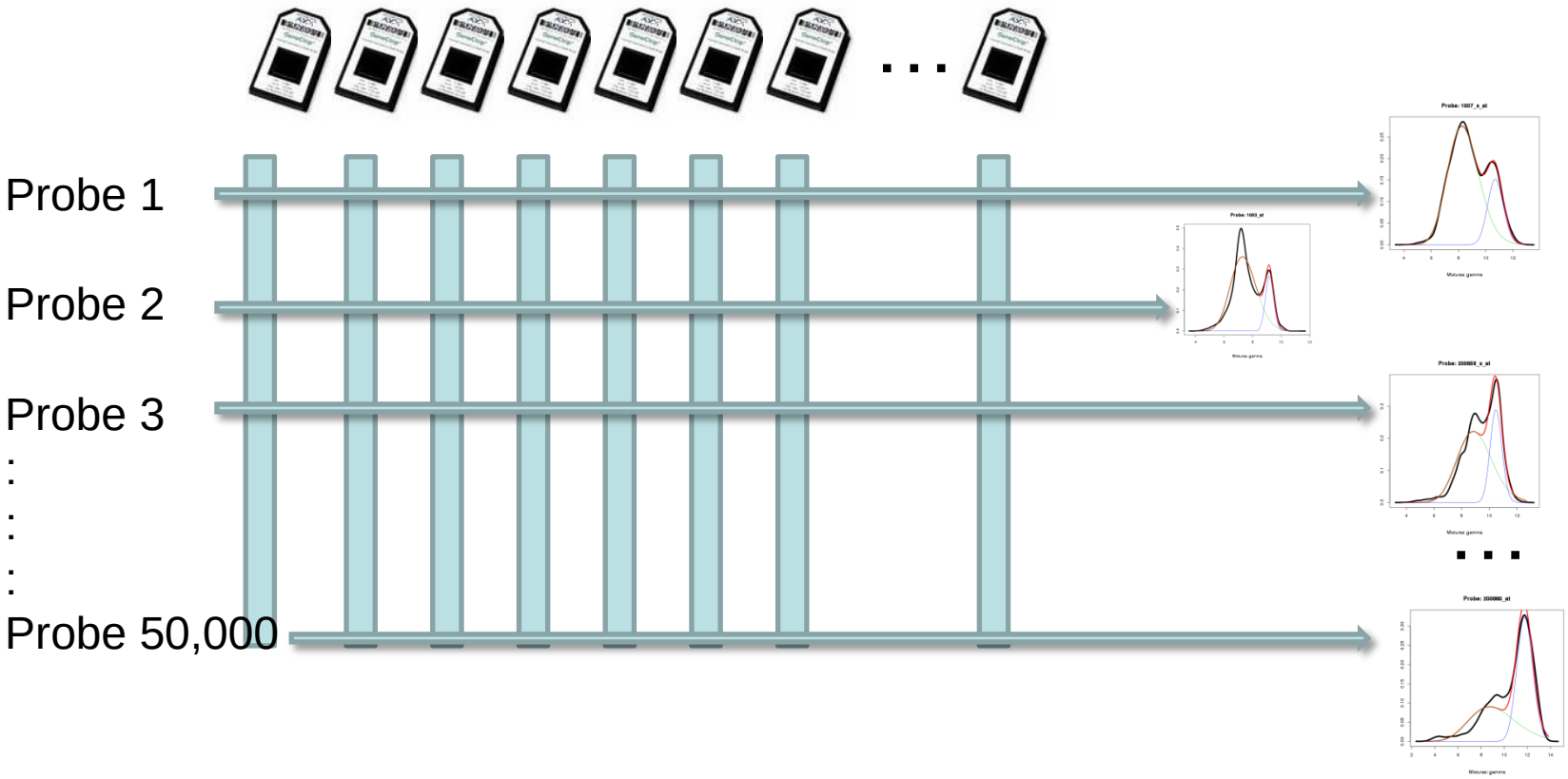
Then, the activation state of any gene from a **new** microarray can be calculated as a probability:

$$P\left(x, Activated\right) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}$$

And the probability for each circuit of the pathway of being ***active*** can be calculated as well:

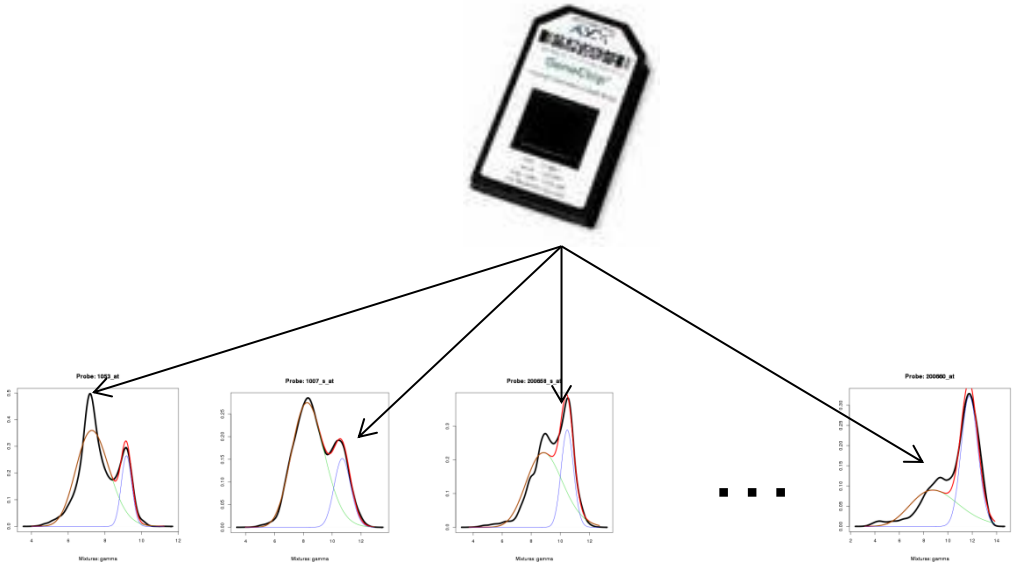
$$P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P\left(A_k\right) - \sum_{i<j} P\left(A_i \cap A_j\right) + \sum_{i<j<k} P\left(A_i \cap A_j \cap A_k\right) + \ldots + (-1)^{n+1} P\left(\bigcap_{k=1}^{n} A_k\right)$$

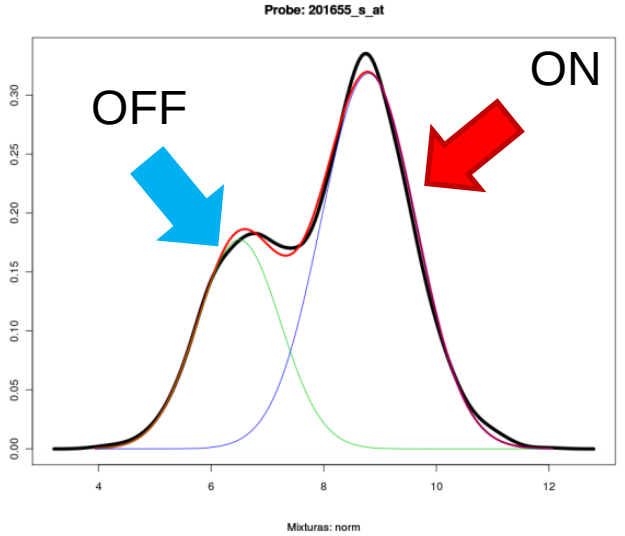
Obtaining probability distributions for ALL the probes in the microarray



A large dataset of affymetrix microarrays (10,000) is used to adjust a mixture of distributions of gene activity for all the probes.

Using probability distributions to estimate gene activation probabilities





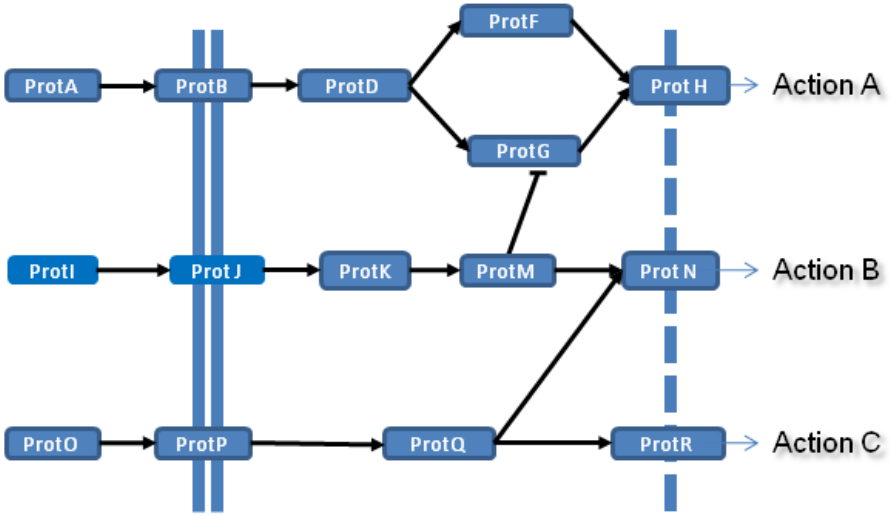
P_{probe} 0,001 0.89 0.5 …… 0.4

Then, the activation state of any probe from a new microarray can be calculated as a probability:

$$P\left(x, Activated\right) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}$$

Finally, gene activation probabilities are summarized from their corresponding probes as the 90% percentil value (to avoid outliers)

Gene activation probabilities are transformed into signal transduction probabilities



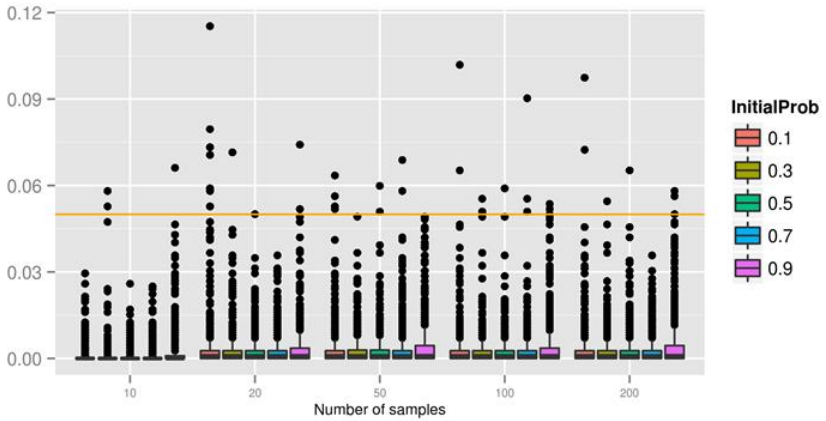
And the probability of being ***active*** for each circuit of each pathway can be calculated as well:

$$P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P\left(A_k\right) - \sum_{i<j} P\left(A_i \cap A_j\right) + \sum_{i<j<k} P\left(A_i \cap A_j \cap A_k\right) + \ldots + (-1)^{n+1} P\left(\bigcap_{k=1}^{n} A_k\right)$$

We have transformed a physical genomic measure (gene expression) into a value that accounts for cell functionality

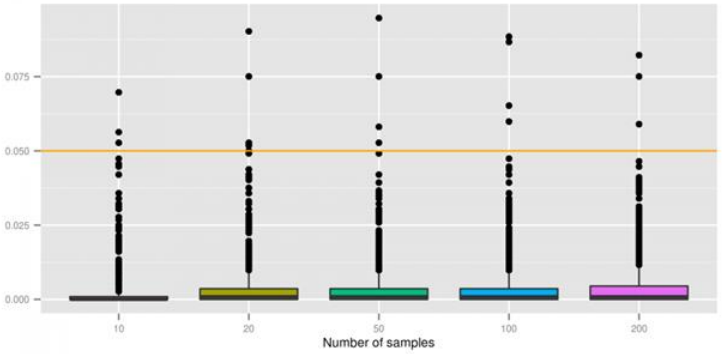
Sensitivity and specificity of model results

Simulated datasets with equal random gene expression values with noise added



Absence of false positives (low type I error)

Circuit activities used for disease class prediction. Accuracy is evaluated by 10-fold cross-validation. A good prediction is a proxy for sensitivity (low type II error)



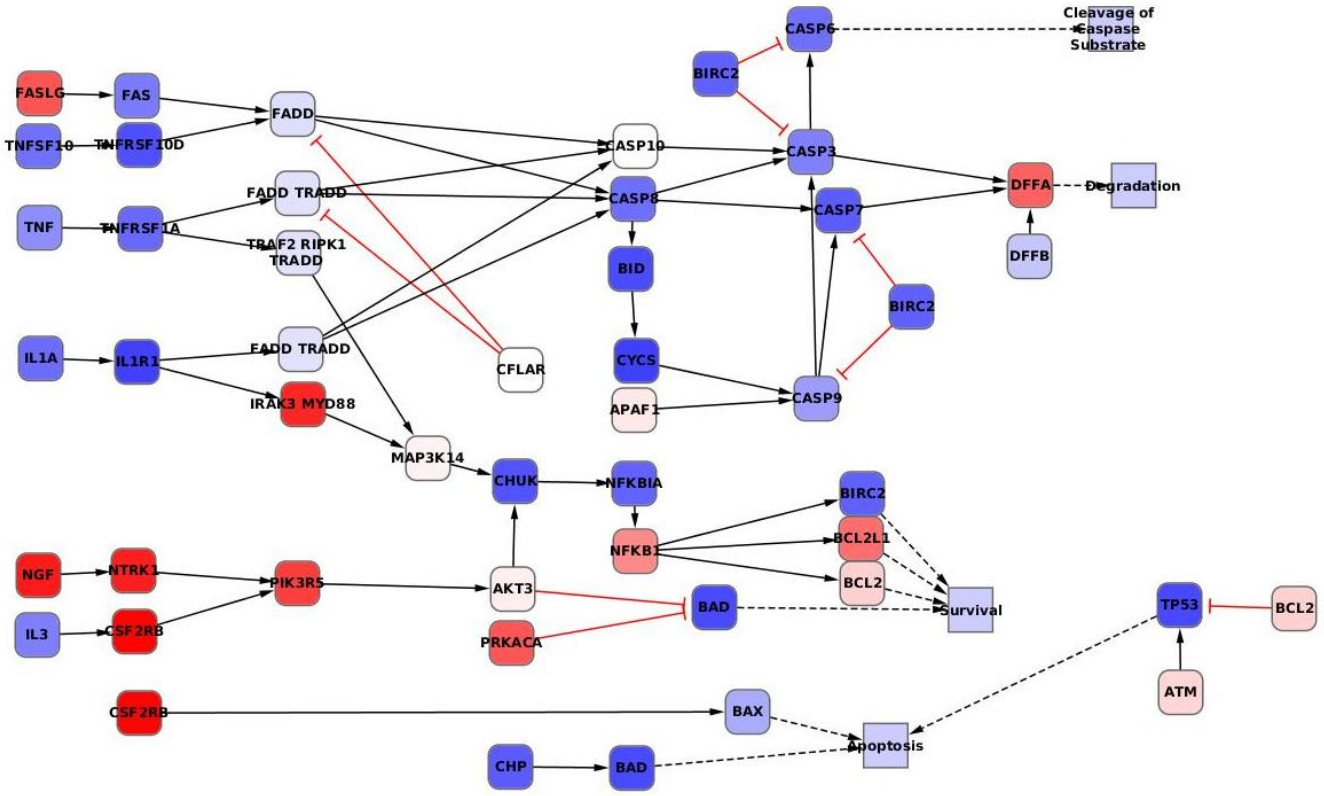
Real dataset of pediatric acute myeloid leukemia with gene expression microarray data of 237 samples

Dataset		KNN	RF	SVM
	Accuracy	0.86-0.89	0.90-0.92	0.99
Breast cancer	MCC	0.74-0.79	0.81-0.83	0.98
	RMSC	0.29-0.32	0.31	0.04
	AUC	0.97-0.98	0.98	0.99
	Accuracy	0.88-0.89	0.92-0.95	0.96
AML	MCC	0.76-0.78	0.84-0.90	0.92-0.93
	RMSC	0.27-0.30	0.31	0.10-0.11
	AUC	0.94	0.98	0.96

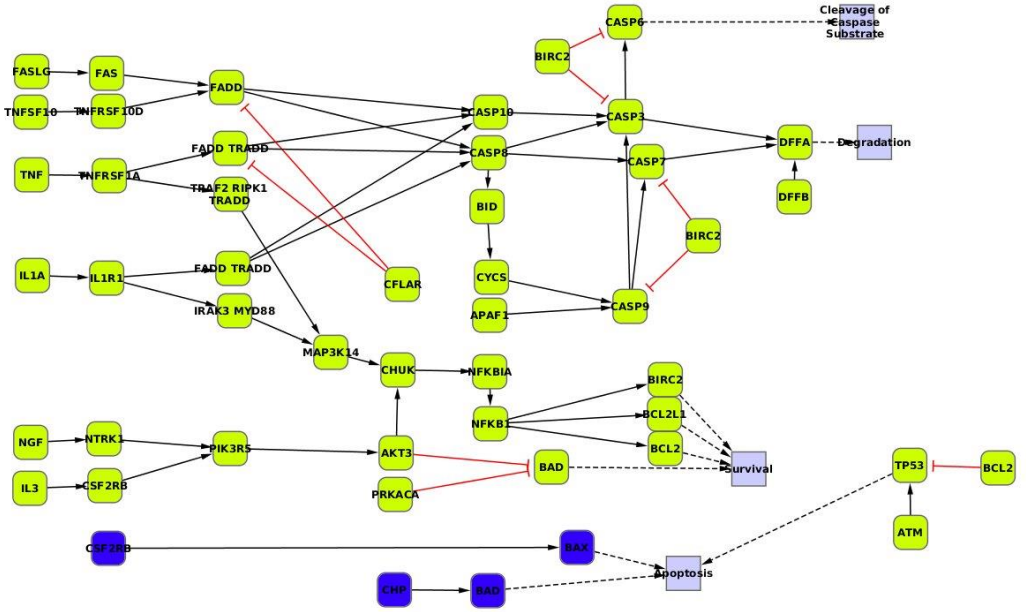
The effects of changes in gene activity are not obvious

What would you predict about the consequences of gene activity changes in the apoptosis pathway in a case control experiment of colorectal cancer?

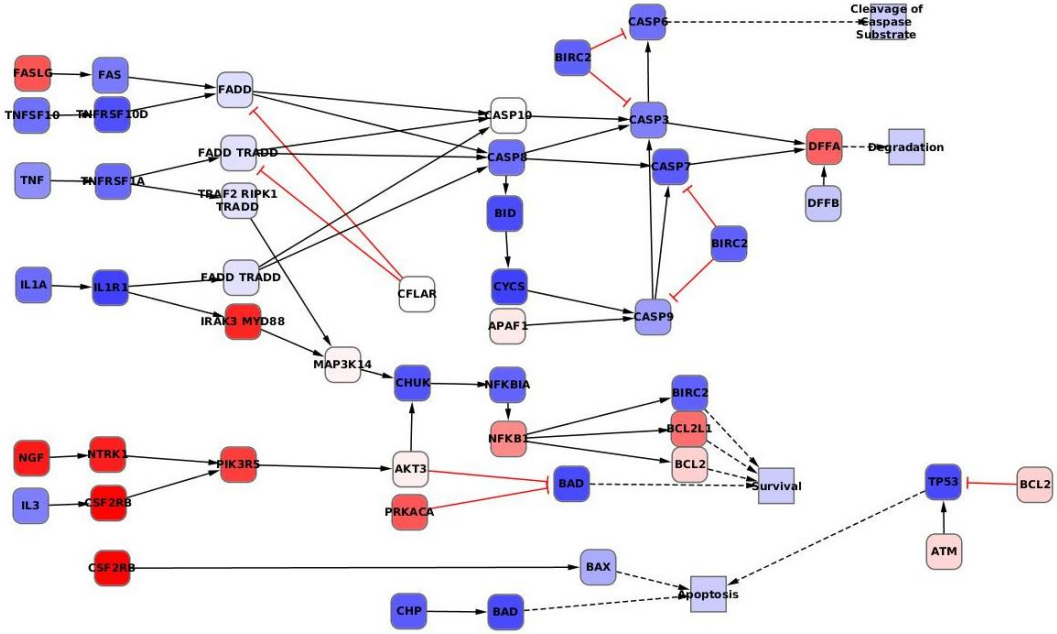
The figure shows the gene up-regulations (red) and down-regulations (blue)



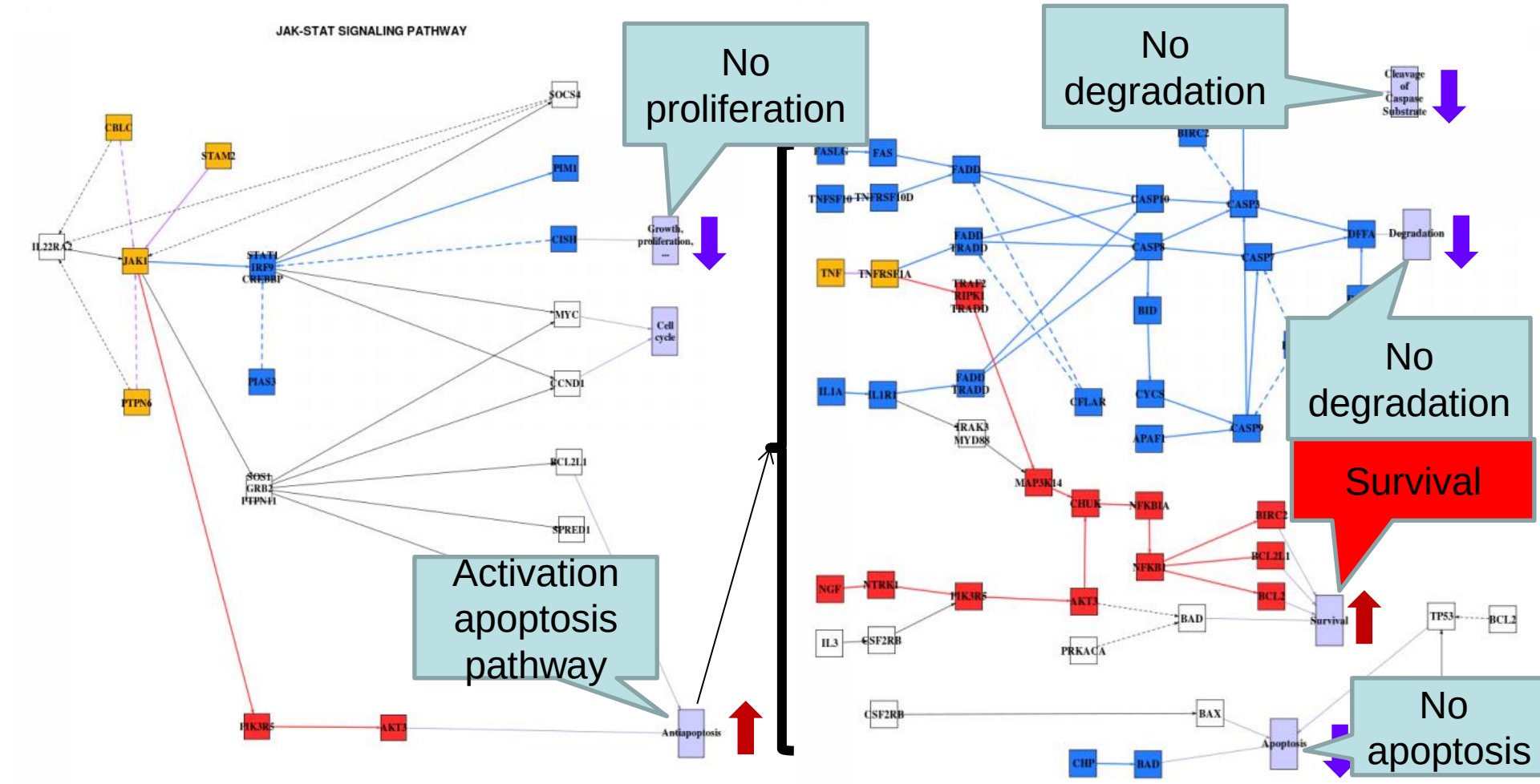
Apoptosis inhibition is not obvious from gene expression



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

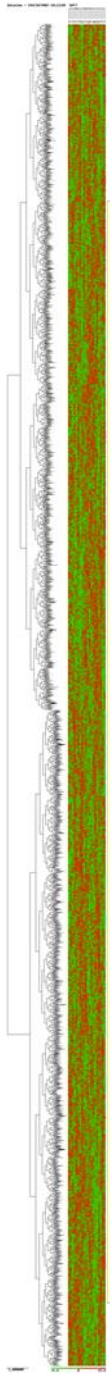


Different pathways cross-talk to deregulate programmed death in Fanconi anemia



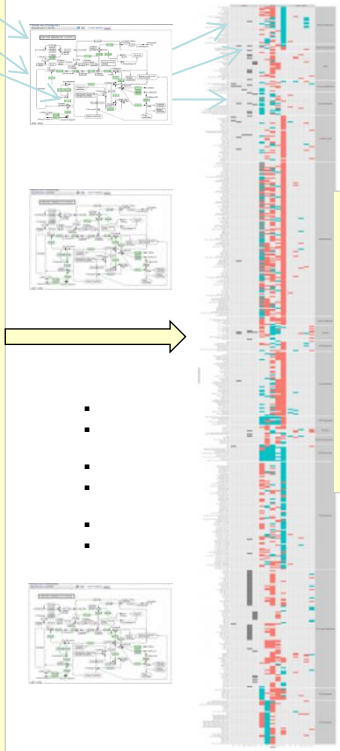
FA is a rare chromosome instability syndrome characterized by aplastic anemia and cancer and leukemia susceptibility. It has been proposed that disruption of the apoptotic control, a hallmark of FA, accounts for part of the phenotype of the disease.

Mechanisms-based biomarkers can also be used to predict features



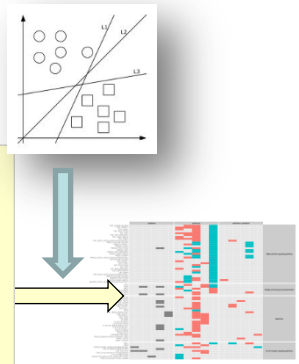
Gene$_1$
Gene$_2$
Gene$_3$
Gene$_4$
:
Gene$_{22000}$

Raw measurement

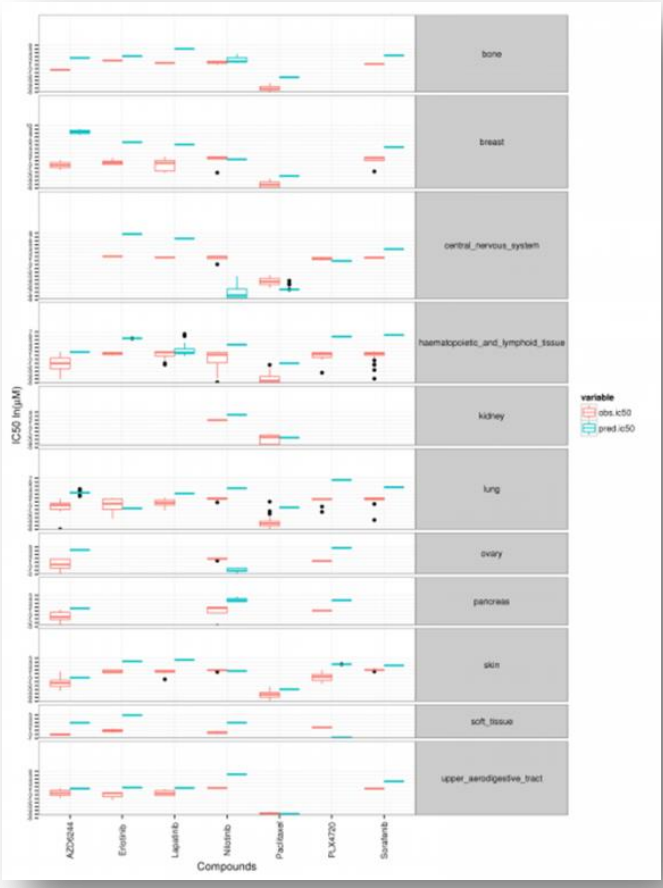


Transformed measurement: Mechanism-based biomarker

Circuit$_1$
Circuit$_2$
Circuit$_3$
Circuit$_4$
:
Circuit$_{800}$

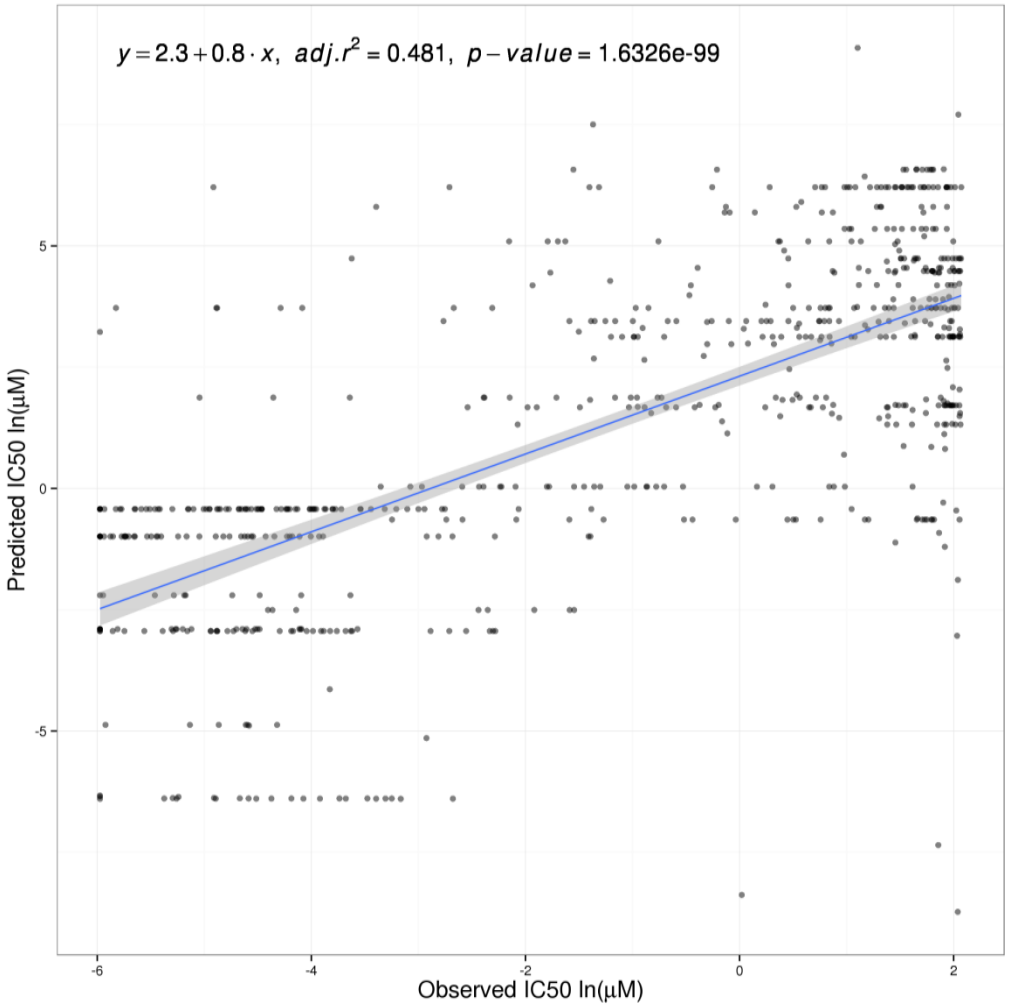


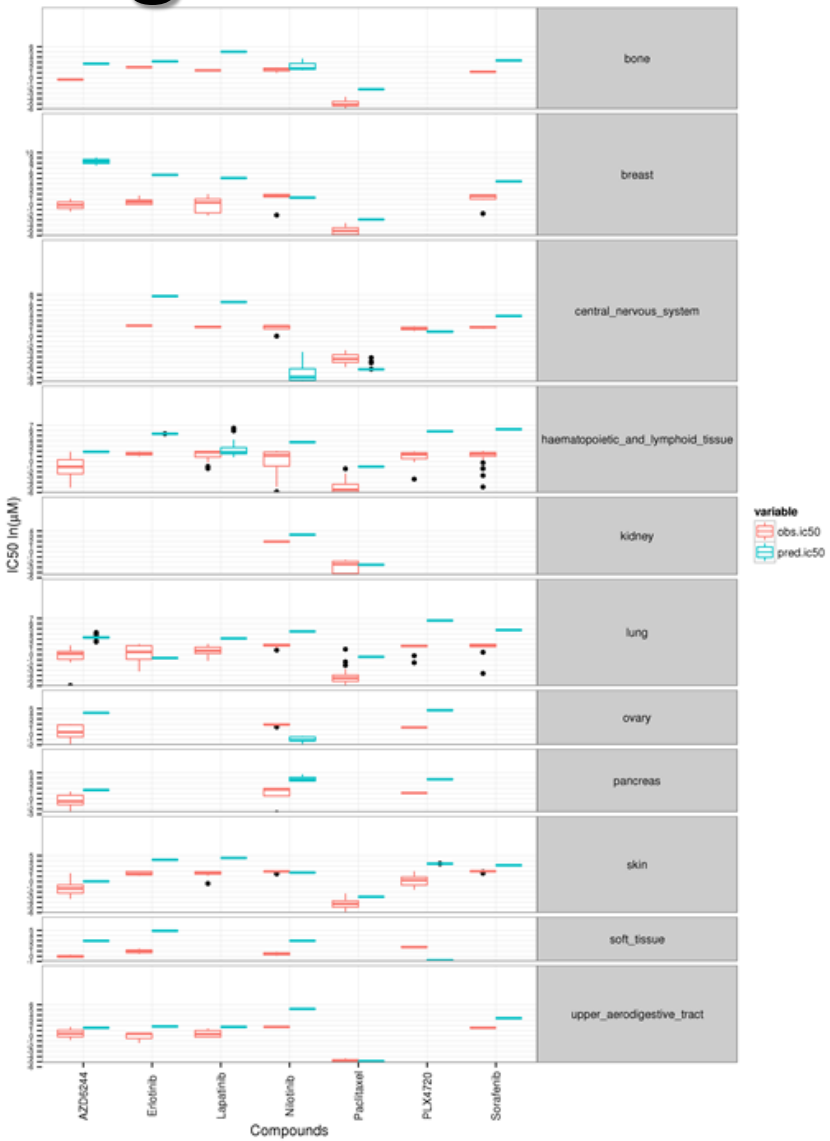
Sub-selection of highly discriminative mechanism-based biomarkers



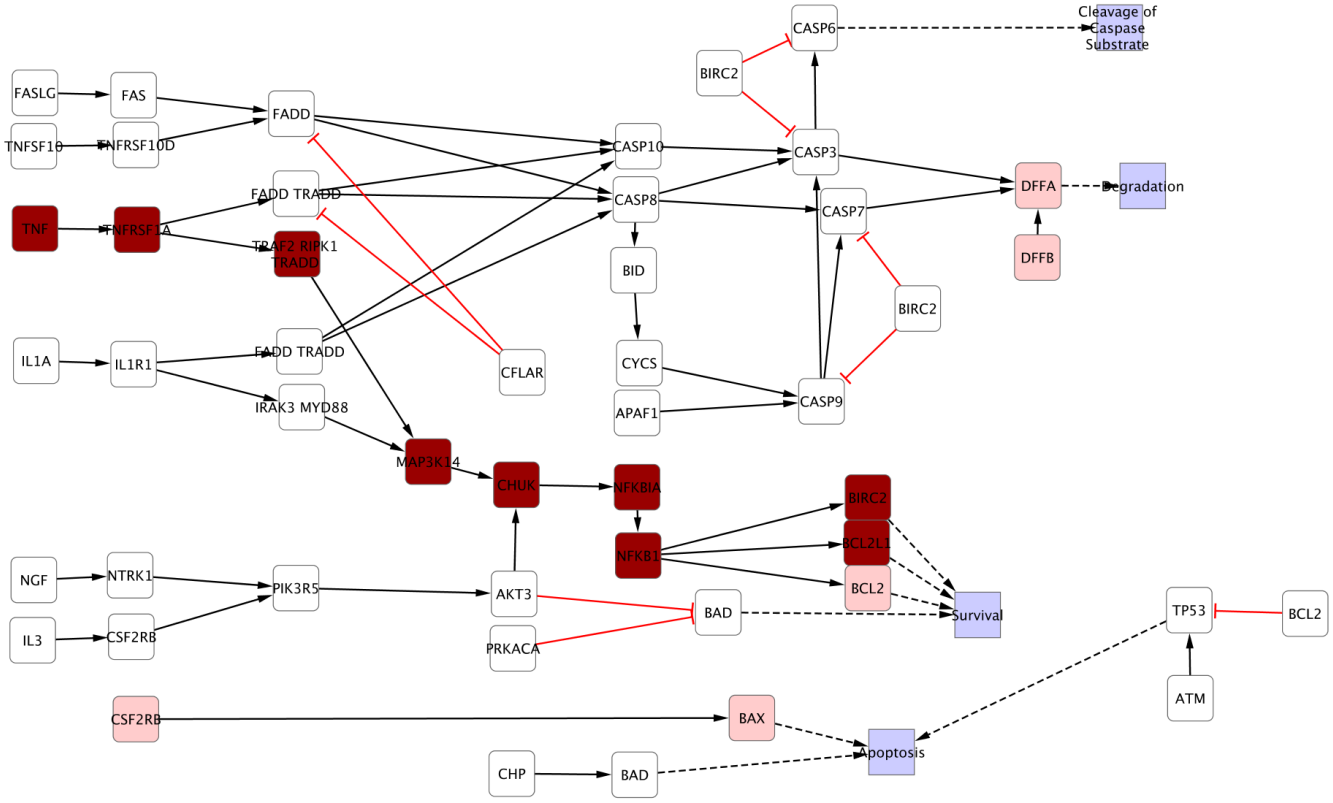
Mechanism-based biomarkers (signaling circuits here) can be used to predict phenotypes that can be discrete (e.g. disease subtype) or even continuous (e.g. the IC50 of drugs, model animals from cell lines, etc.).

Prediction of IC_{50} values from the activity of signaling circuits





Mechanism-based biomarkers have meaning by themselves



Unlike in single gene biomarkers, the selected mechanism-based biomarkers (probability of circuit activation) have meaning by itself. Survival activation in the apoptosis pathway is one of the predictive mechanism-based biomarkers of breast cancer.

From gene-based to function-based perspective

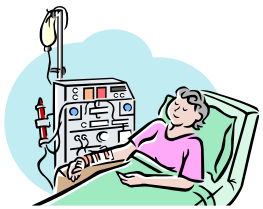
	SNPs, gene expression, etc.	GO	Protein interaction networks	Models of cellular functions
Detection power	Low (only very prevalent genes)	High	High	Very high
Information coverage	Almost all	Almost all	Low (~9000 genes in human)	Low (~6700 genes in human)*
Use	Biomarker	Illustrative, give hints	Biomarker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

*Only ~800 genes in human signaling pathways

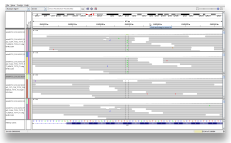
Is individualized treatment a realistic option today?

An integrated view

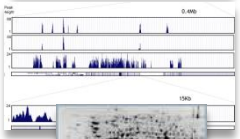
Patient



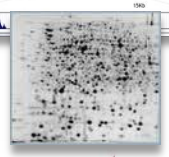
Patient's omic data



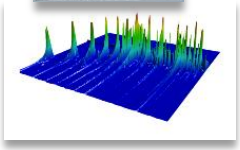
Genomics and transcriptomics



Epigenomics

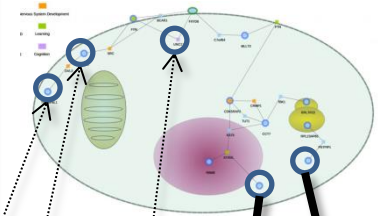


Proteomics

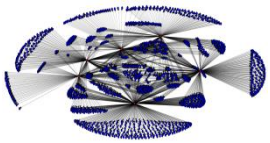


Metabolomics

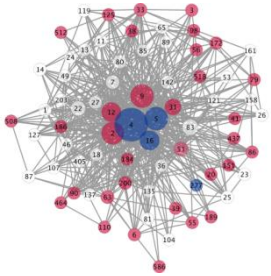
Systems biology computational models



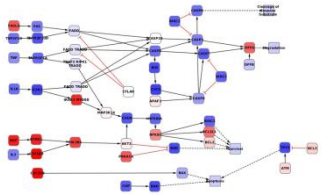
Biological knowledge



Regulation

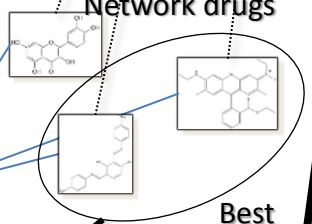


Interaction



Function

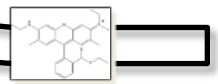
Network drugs



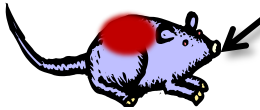
Best combination

Cell culture

Drug treatment



Personalized therapy



Xenograft model

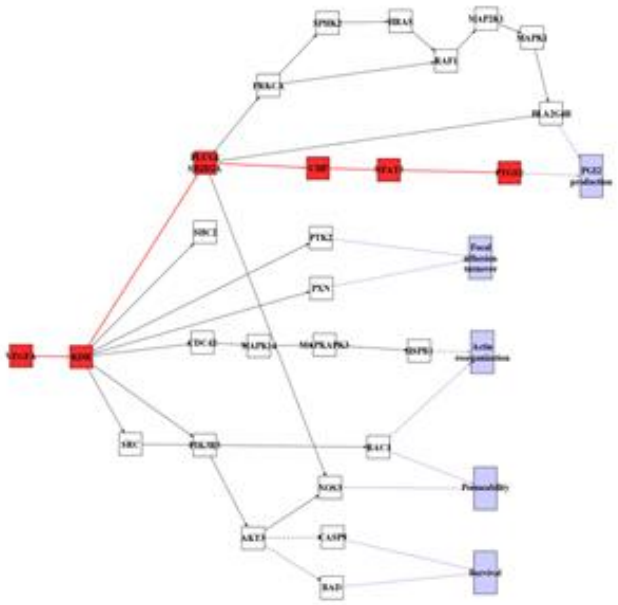
Therapeutic targets

Diagnostic biomarkers
Personalized medicine

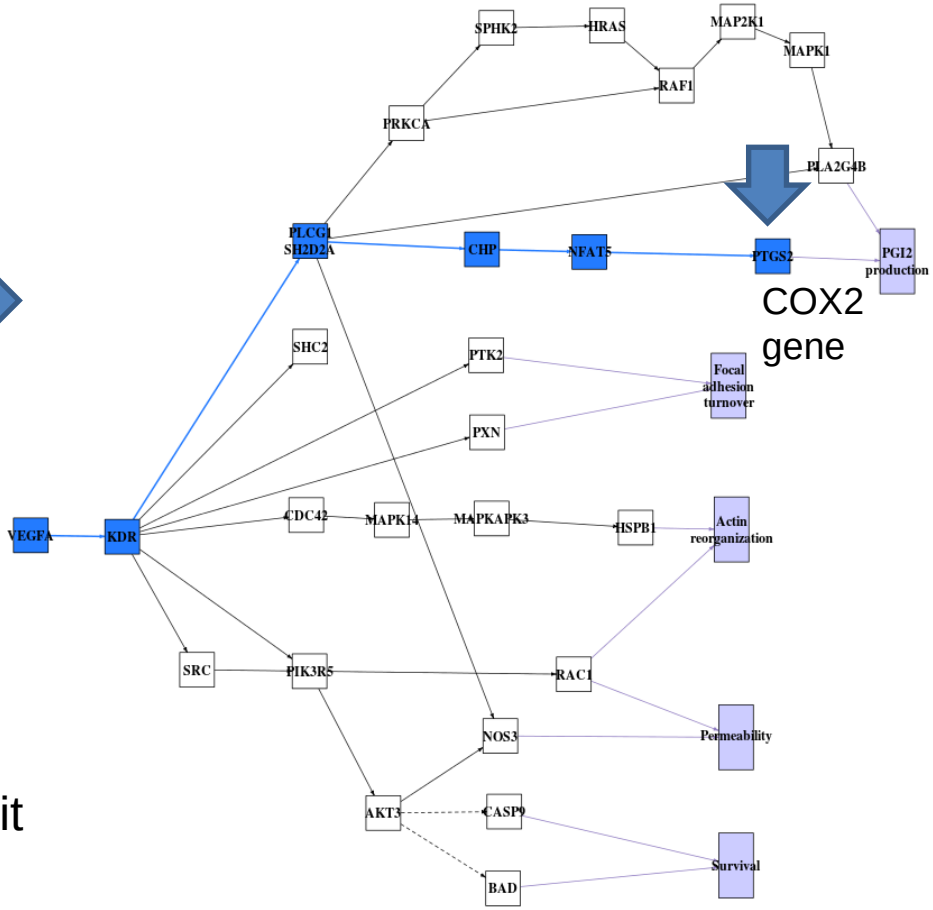
Dopazo, 2003, Drug Discovery Today

In silico KOs or over-expressions

If we can calculate the effect of gene expression over signaling, we can simulate KOs (or over-expressions)



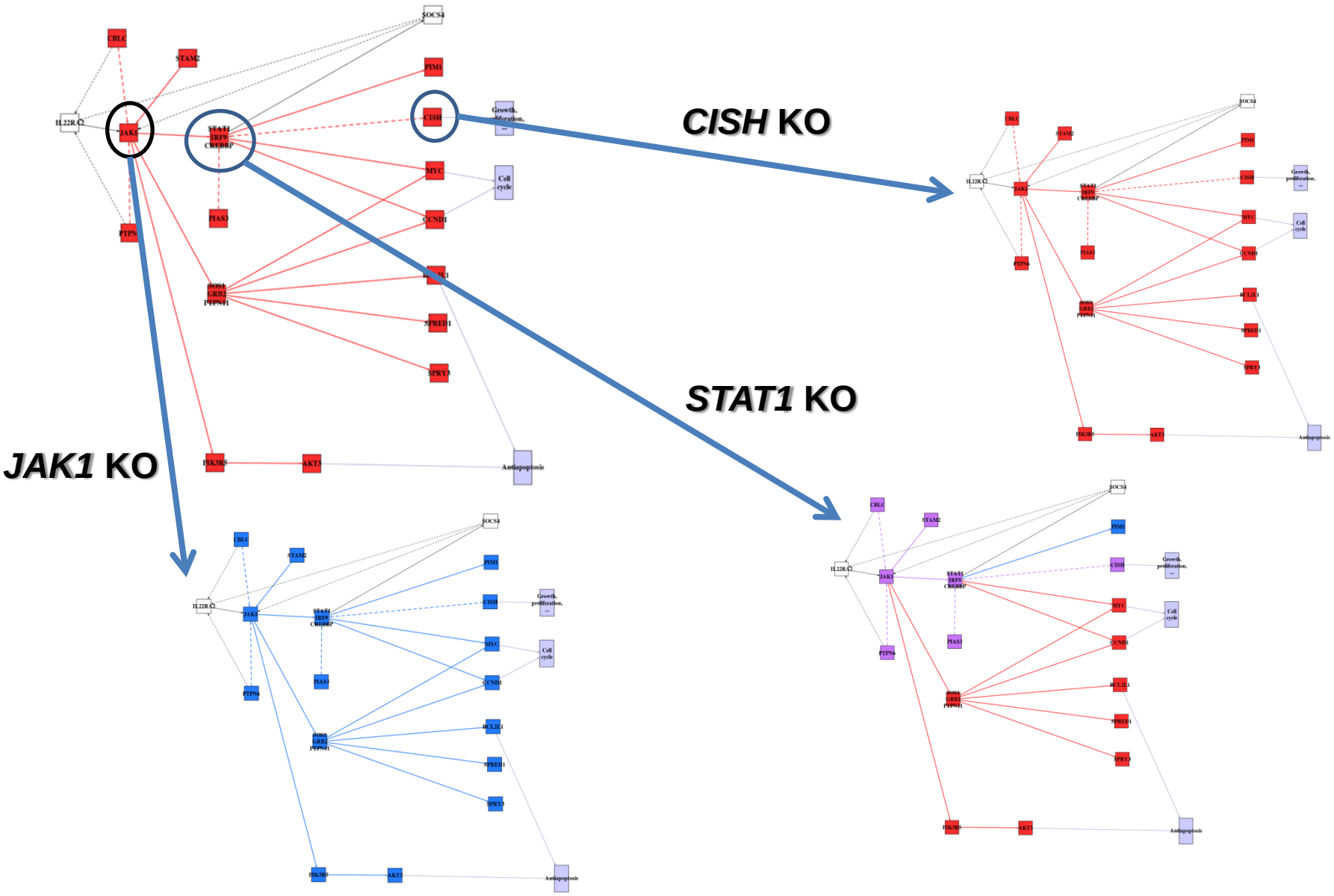
KO



Colorectal cancer activates a signaling circuit of VEGF pathway that produces PGI2.

Virtual KO of COX2 interrupts the circuit (known therapeutic inhibitor in CGR

Different virtual KOs predict different effects



Distribution of Human Genetic Diseases

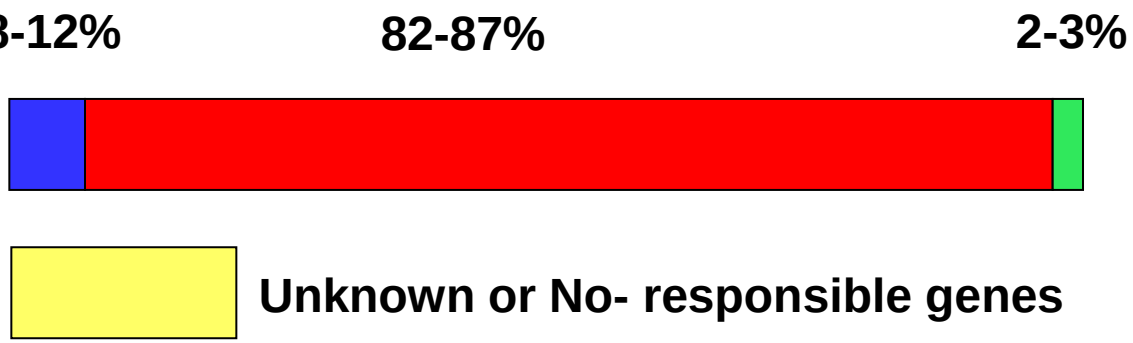
Hospital Universitario La Paz

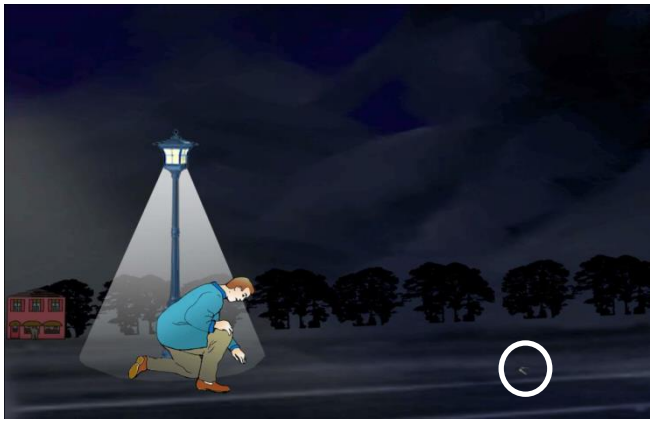
Pablo Lapunzina, Personal communication

5 Kb- ? Mb **1 bp- 200 bp** **No dosage changes**

GENOMICS **GENETICS** **EPIGENETICS**

All genetic/genomic or epigenetic diseases with known cause:
~ 5000 disorders



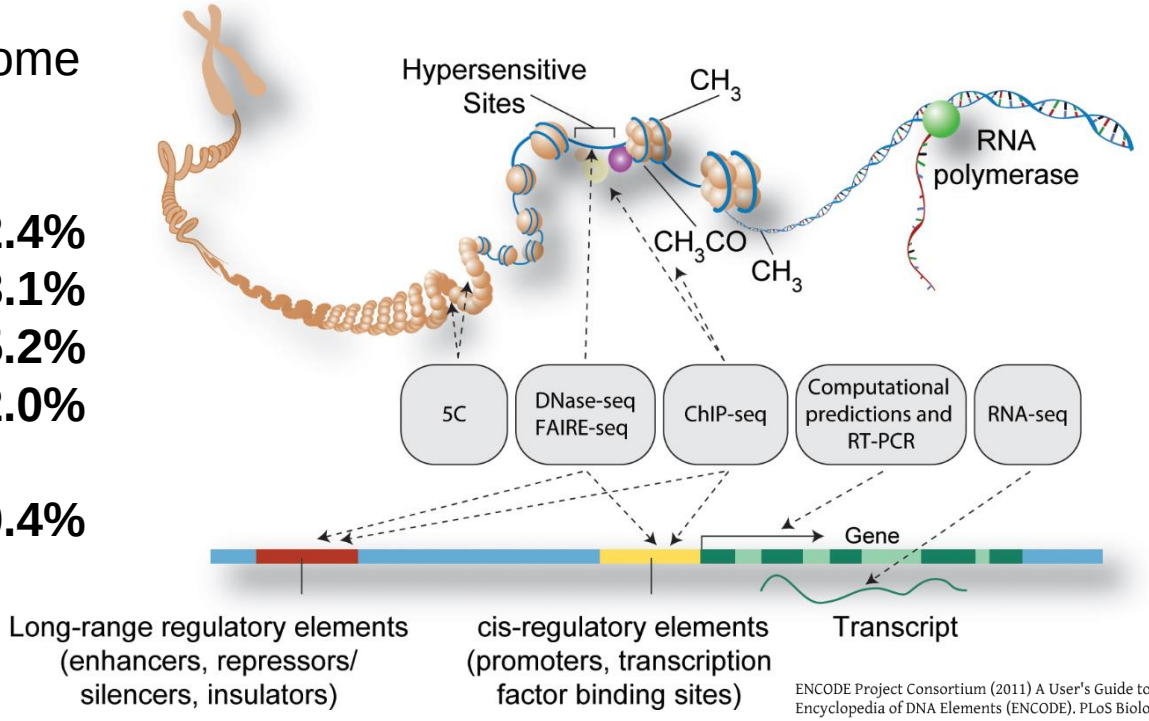


Fact: exons represent a comparatively small part of the complete genome
Other fact: there is still a lot of missing heritability

The ENCODE project suggests a functional role for a large fraction of the genome

Which percentage of the genome is occupied by:

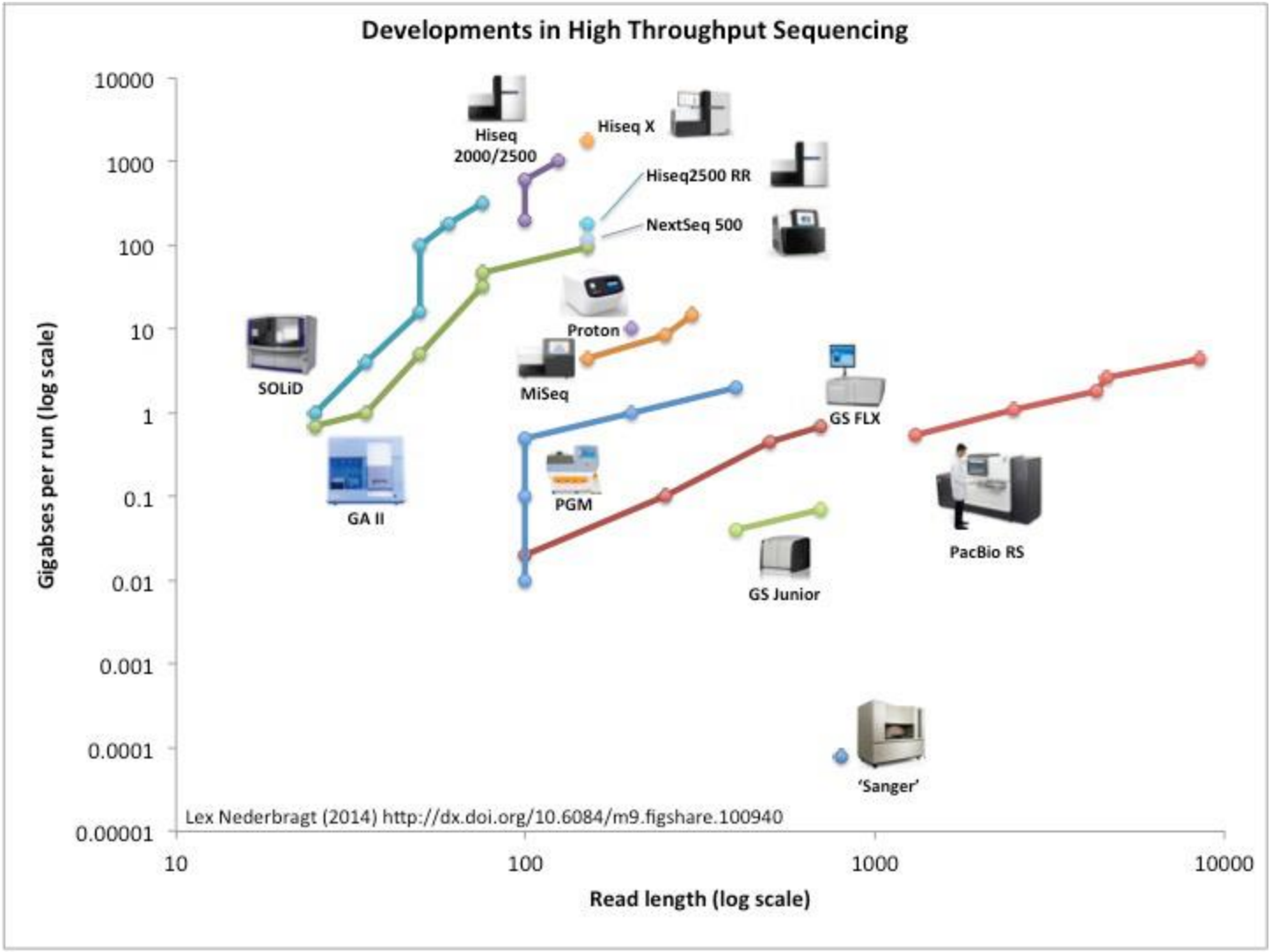
Coding genes:	**2.4%**
TFBSs	**8.1%**
Open chromatin regions	**15.2%**
Different RNA types	**62.0%**
Total annotated elements:	**80.4%**



ENCODE Project Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biology 9: 1-24

Exomes are only covering a small fraction of the potential functionality of the genome (2.4%).
Is the **missing heritability** hidden in the remaining 78%?
If so, what type of variant should be expect to discover? SNVs? SVs?

Future prospects



http://flxlexblog.wordpress.com/2014/06/11/developments-in-next-generation-sequencing-june-2014-edition/

Future prospects

We need to efficiently query all the information contained in the genome, including all the epigenomic signatures.

This means **data integration** and "**epistatic**" queries

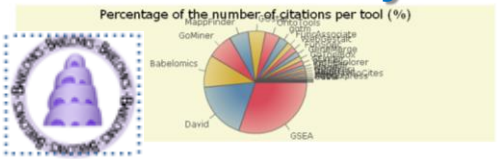
We need to prepare our **health systems** to deal with all the genomic data flood

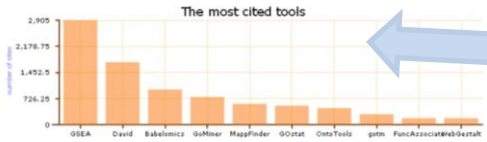
Information about variations	Processed	Raw
Genome variant information (VCF)	150 MB	250 GB
Epigenome	150 MB	250 GB
Each transcriptome	20 MB	80 GB
Individual complete variability	400 MB	525 GB
Hospital (100.000 patients)	40 TB	50 PB

There are **technical** problems and **conceptual** problems on how genomic information is managed that must be addressed in the near future.

Software development

Functional analysis

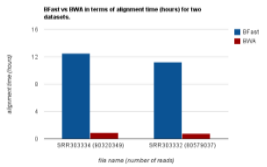




Babelomics is the third most cited tool for functional analysis. Includes more than 30 tools for advanced, systems-biology based data analysis

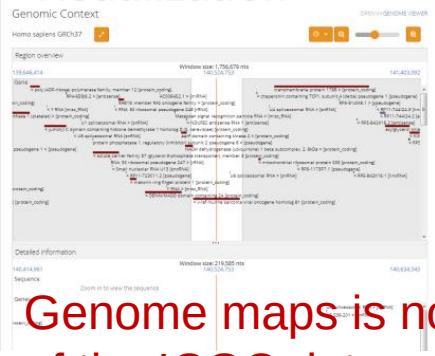
See interactive map of for the last 24h use http://bioinfo.cipf.es/toolsusage

Mapping



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

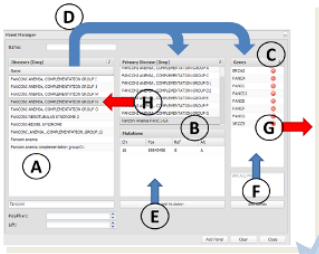
Visualization



Ultrafast genome viewer with google technology

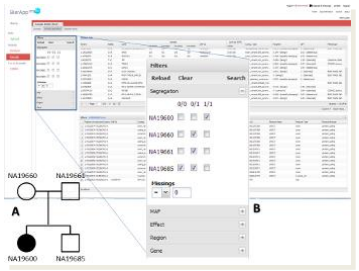
Genome maps is now part of the ICGC data portal

Diagnostic

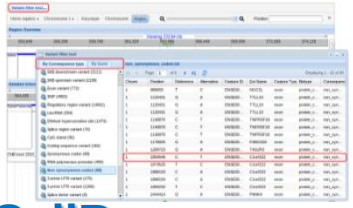


NGS panels

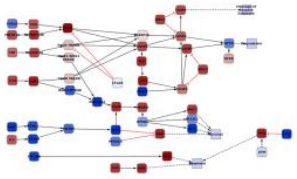
Variant prioritization



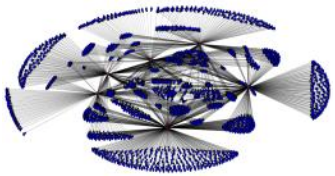
Variant annotation



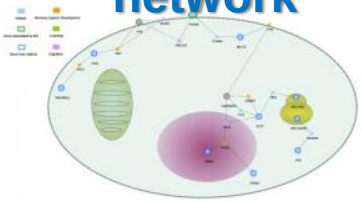
Signaling network



Regulatory network



Interaction network



CellBase

Knowledge database

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

From data to knowledge

Contribution from the computational biology side

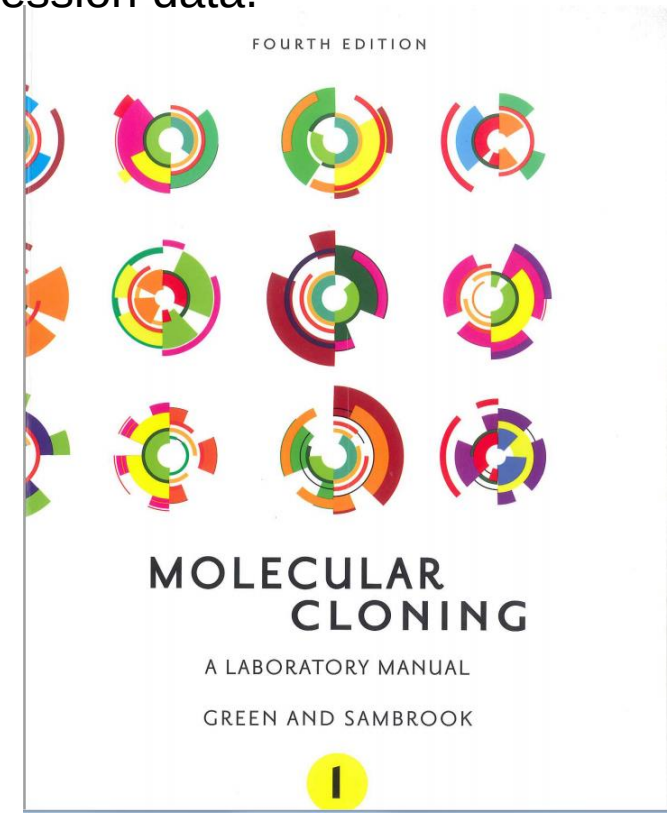
(from molecular biology to genomic/computational biology)

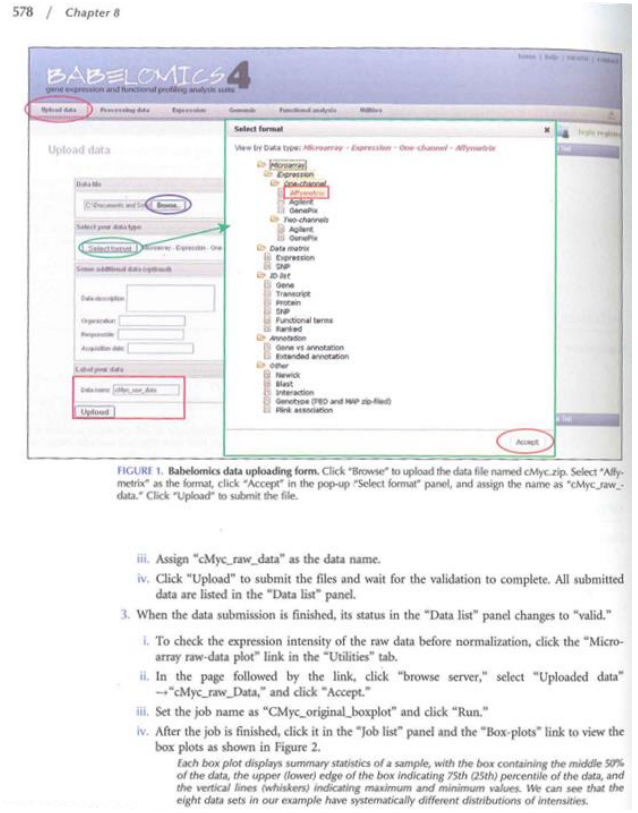
Scientific relevance

- **Technical contribution**: huge datasets need to be managed by computers. Technically useful, scientifically irrelevant
- **Data analysis contribution**: bioinformatics is necessary to properly analyze *big data*. Of course p-values are necessary, but do not provide novelty. No *big data* vision, only analysis
- **Science from the computational biology perspective**:
  - Do not use genomic data to simply scaling up conventional approaches. New ideas emerge when data are seen from the proper perspective.
  - *Big data*-driven questions and hypotheses can only be formulated when such data can be easily managed
  - Niche disciplines: genomics and systems biology
  - Most challenging problems: Complex diseases, disease mechanisms, biological processes (e.g. metabolism). drug action mechanisms, drug repositioning…

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.





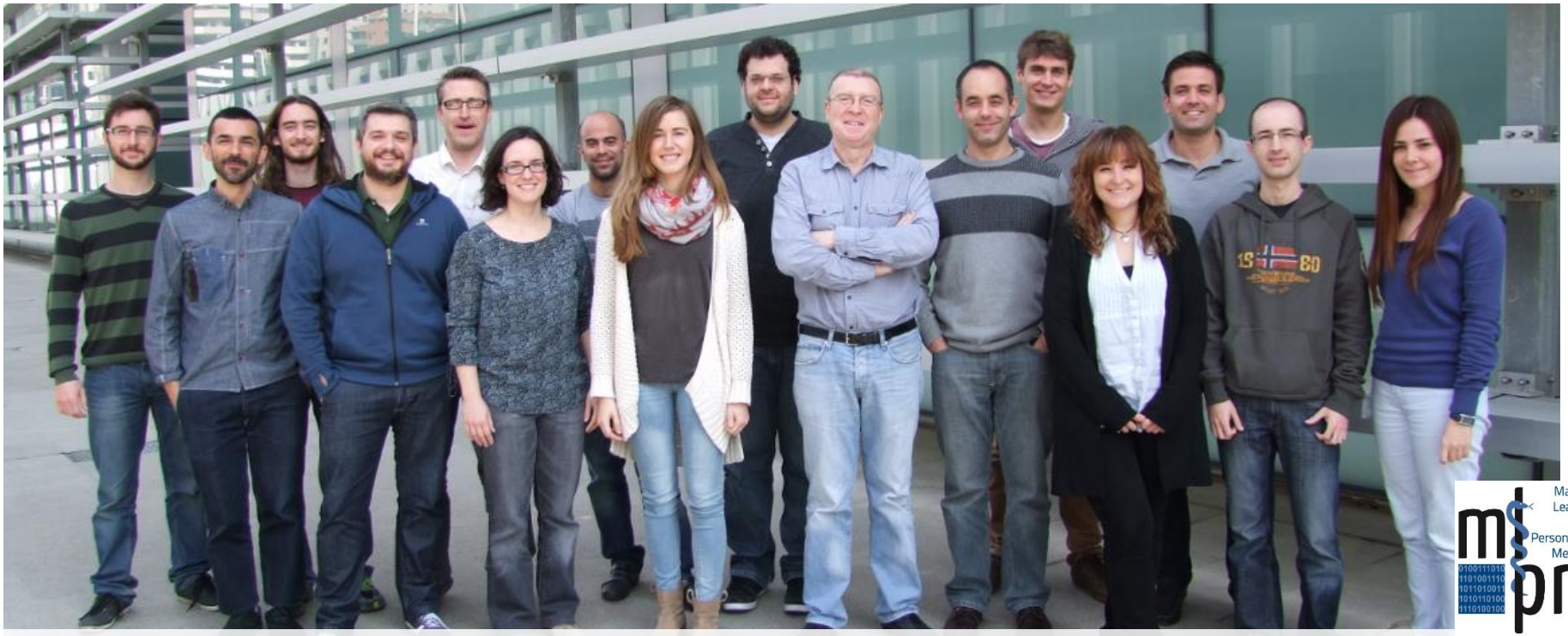
High impact developments

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

@xdopazo
@bioinfocipf

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 CIBERER Network of Centers for Rare Diseases.



FUTURE CLINIC Medicina Genómica Personalizada

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