

Variant prioritization

Joaquín Dopazo

Computational Genomics Department,
Centro de Investigación Príncipe Felipe (CIPF),
Functional Genomics Node, (INB),
Bioinformatics in Rare Diseases (BiER-CIBERER),
Valencia, Spain.

<http://bioinfo.cipf.es>
[http://www.babelomics.org](http://wwwbabelomics.org)
<http://www.hpc4g.org>
 @xdopazo

University of Cambridge, 17-18 June, 2015

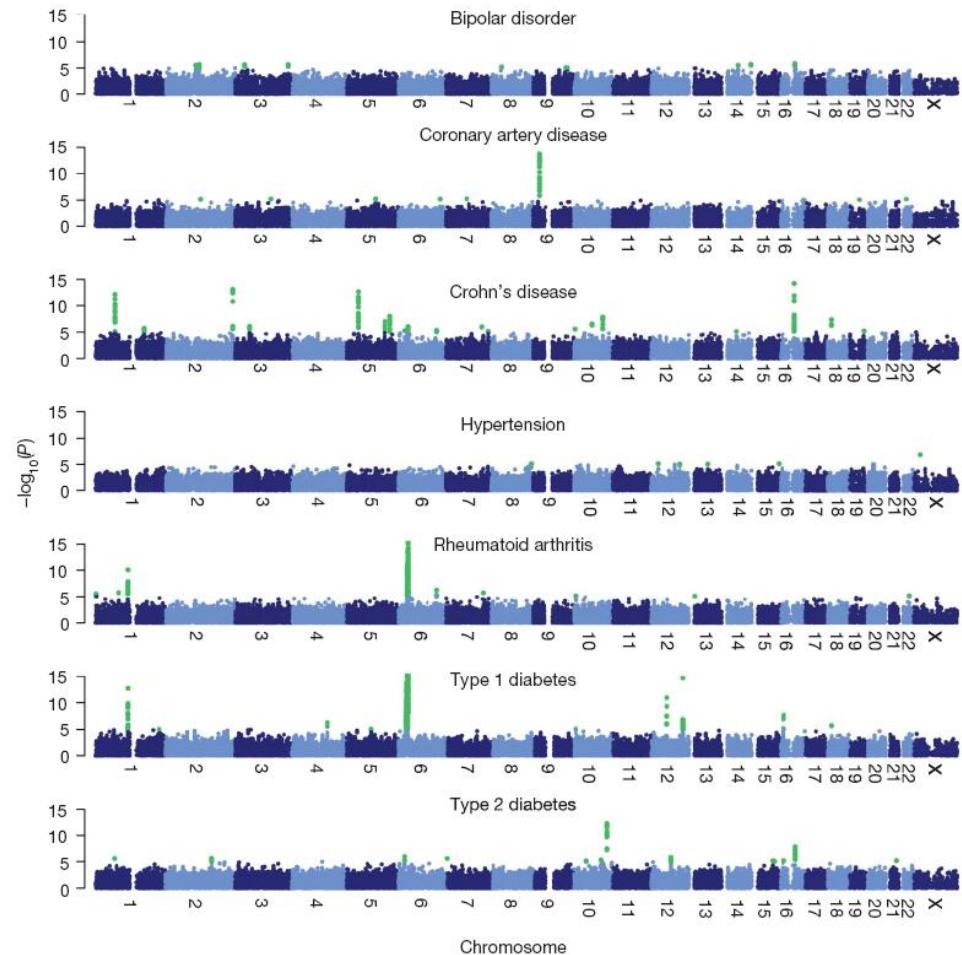
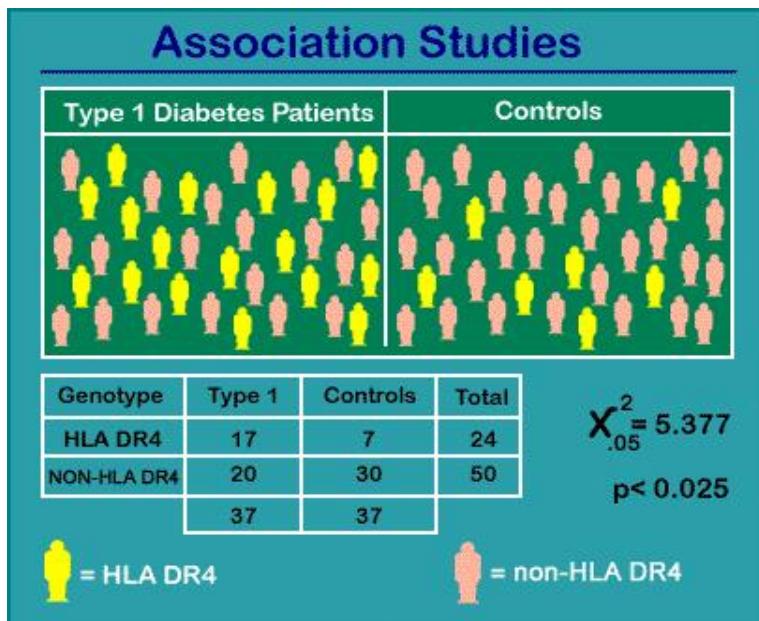


Fundació
La Marató de TV3

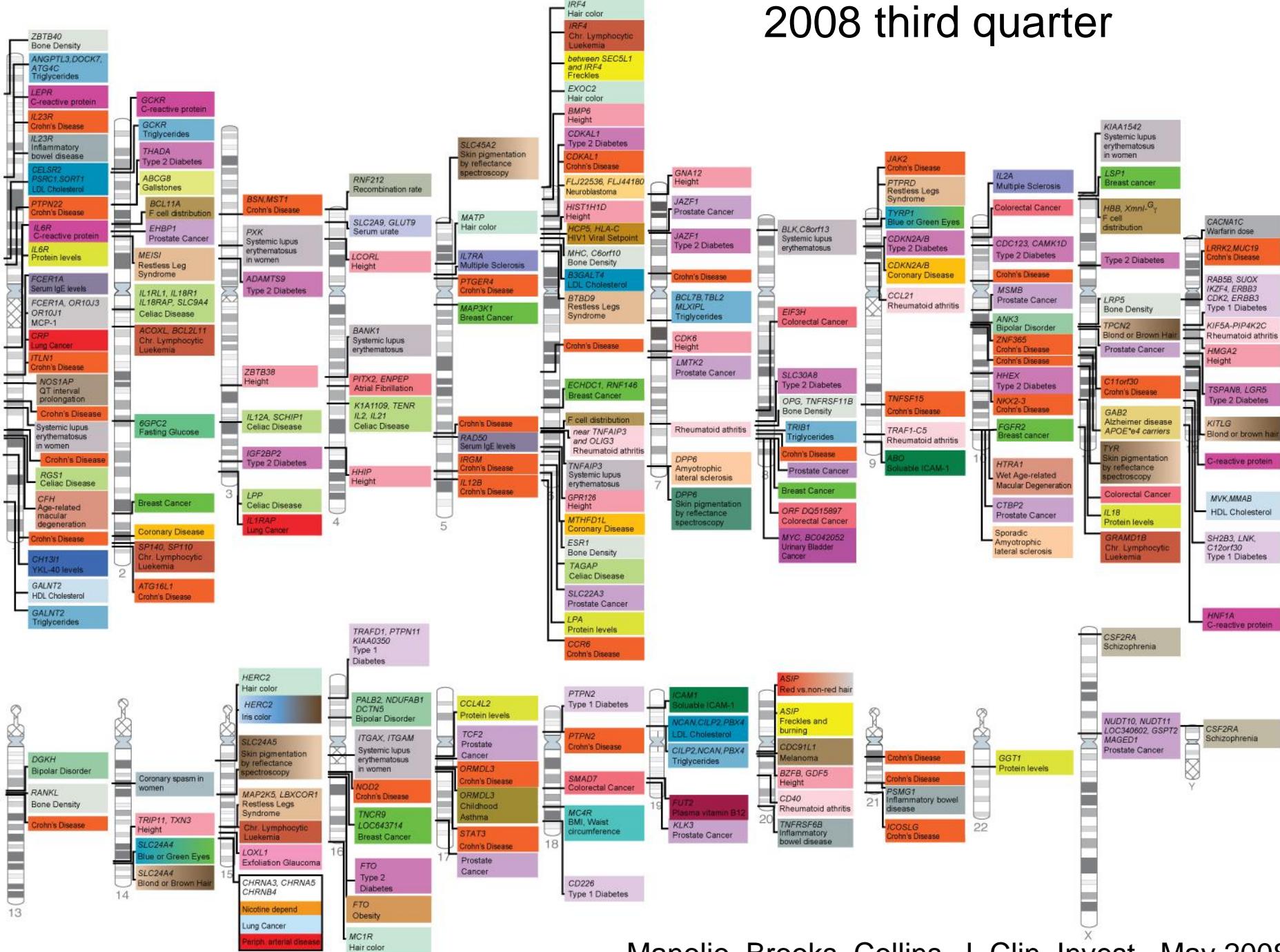


The dawn of genomic data production

Candidate gene studies using GWAS



2008 third quarter

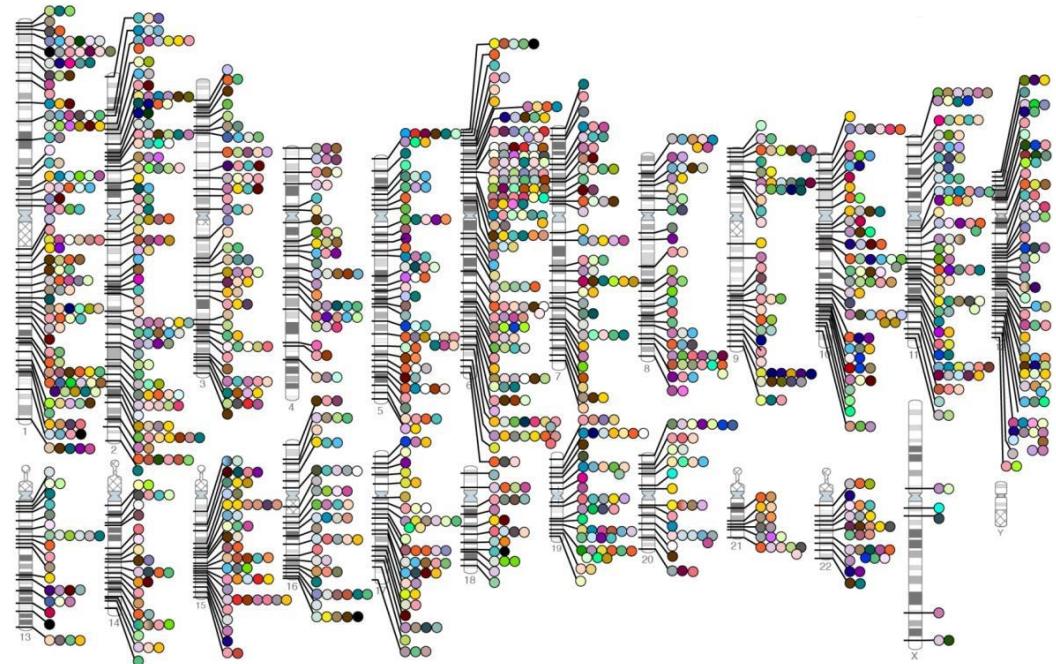


Published Genome-Wide Associations

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, the catalog has collected more than 1,900 papers reporting **14,012** SNPs significantly associated to more than **1,500** traits.



NHGRI GWA Catalog
www.genome.gov/GWASStudies

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.

Where did the heritability go?

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

NEWS FEATURE PERSONAL GENOMES NATURE/Vol 456/November 2008



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. Manolio¹, Francis S. Collins², Nancy J. Cox³, David B. Goldstein⁴, Lucia A. Hindorff⁵, David J. Hunter⁶, Mark I. McCarthy⁷, Erin M. Ramos⁵, Lon R. Cardon⁸, Aravinda Chakravarti⁹, Judy H. Cho¹⁰, Alan E. Guttmacher¹, Augustine Kong¹¹, Leonid Kruglyak¹², Elaine Mardis¹³, Charles N. Rotimi¹⁴, Montgomery Slatkin¹⁵, David Valle⁹, Alice S. Whittemore¹⁶, Michael Boehnke¹⁷, Andrew G. Clark¹⁸, Evan E. Eichler¹⁹, Greg Gibson²⁰, Jonathan L. Haines²¹, Trudy F. C. Mackay²², Steven A. McCarroll²³ & Peter M. Visscher²⁴

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 this problem?

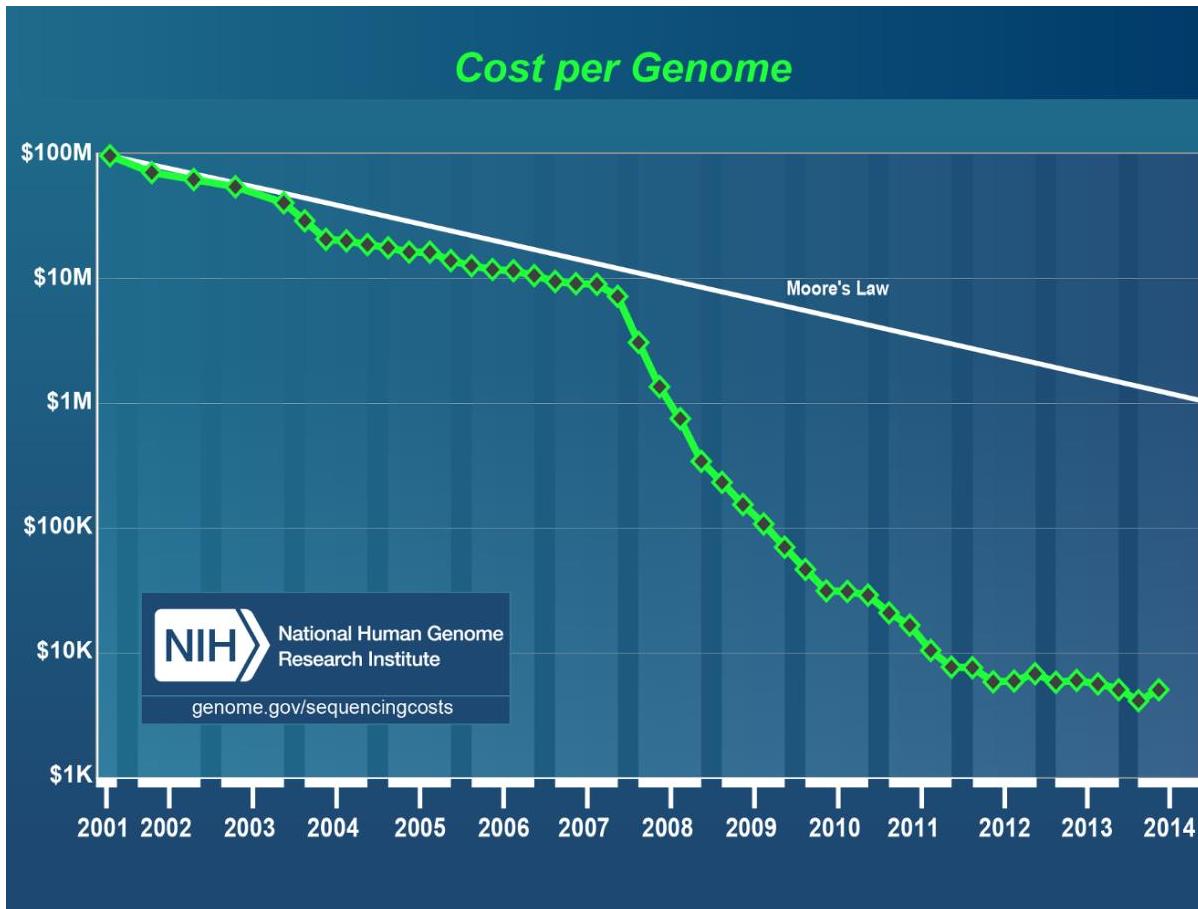
Rare Variants, rare CNVs, epigenetics or.. epistatic effects?

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.



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 Lee¹, Abigail W Bigham², Holly K Tabor^{2,3}, Karin M Dent⁴, Chad D Huff⁵, Paul T Shannon⁶, Ethylin Wang Jabs^{7,8}, Deborah A Nickerson¹, Jay Shendure¹ & 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 (OMIM 263750). For four affected individuals in three independent kindreds, we captured and sequenced coding regions to a mean coverage of 40x, and sufficient depth to call variants at ~97% of each targeted exon. 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.

OPEN ACCESS Freely available online

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^{2,3}, Marina Donskoi¹, Tony J. Kenna¹, Gethin P. Thomas¹, Graeme R. Clark¹, 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, St. Lucia, Australia, ³ Murdoch Childrens Research Institute, Melbourne, Victoria, Australia

European Journal of Human Genetics (2011) 19, 115–117
© 2011 Macmillan Publishers Limited All rights reserved 1088-4813/11
www.nature.com/ejhg



small pedigrees
skeletal dysplasia,
dysplasia, brachydactyly.
The two
a rare form of
sequencing.
encodes a core
the *RMRP* RNA
and activity of
by which *POP1*

? Mutations As the
sense, which permits

SHORT REPORT

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

Kimia Kahrizi¹, Cougar Hao Hu², Masoud Garshabi², Seyedeh Sedigheh Abedini¹, Shirin Ghadami¹, Roxana Kariminejad³, Reinhard Ullmann⁴, Wei Chen², H-Hilger Ropers², Andreas W Kuss², Hossein Najmabadi¹ and Andreas Tschach^{*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 612713
array-based
(c.203del
interval.
essential
families
and eye
potential
European

Keywords:
consanguinity

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

© 2013 Molecular Vision

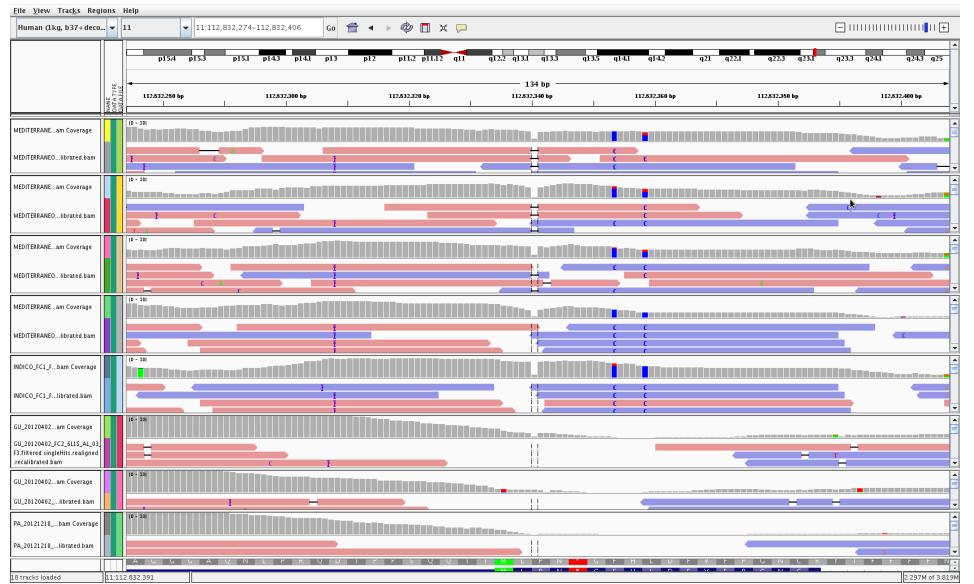
Whole-exome sequencing identifies novel compound heterozygous mutations in *USH2A* in Spanish patients with autosomal recessive retinitis pigmentosa

Cristina Méndez-Vidal,^{1,2} María González-del Pozo,^{1,2} Alicia Vela-Boza,³ Javier Santoyo-López,³ Francisco J. López-Domínguez,³ Carmen Vázquez-Marouschek,⁴ Joaquín Dopazo,^{3,5,6} Salud Borrego,^{1,2} Guillermo António,^{1,2,3}

¹Department of Genetics, Reproduction and Fetal Medicine, Institute of Biomedicine of Seville, University Hospital Virgen del Rocío/CSIC/University of Seville, Seville, Spain; ²Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Seville, Spain; ³Medical Genome Project, Genomics and Bioinformatics Platform of Andalucía (GBPA), Seville, Spain;

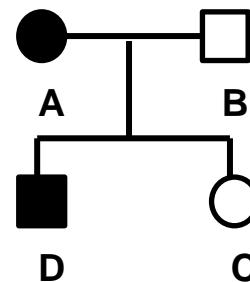
⁴Department of Ophthalmology, University Hospital Virgen del Rocío, Seville, Spain; ⁵Department of Bioinformatics, Centro de Investigación Príncipe Felipe, Valencia, Spain; ⁶Functional Genomics Node (INB), Centro de Investigación Príncipe Felipe, Valencia, Spain

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

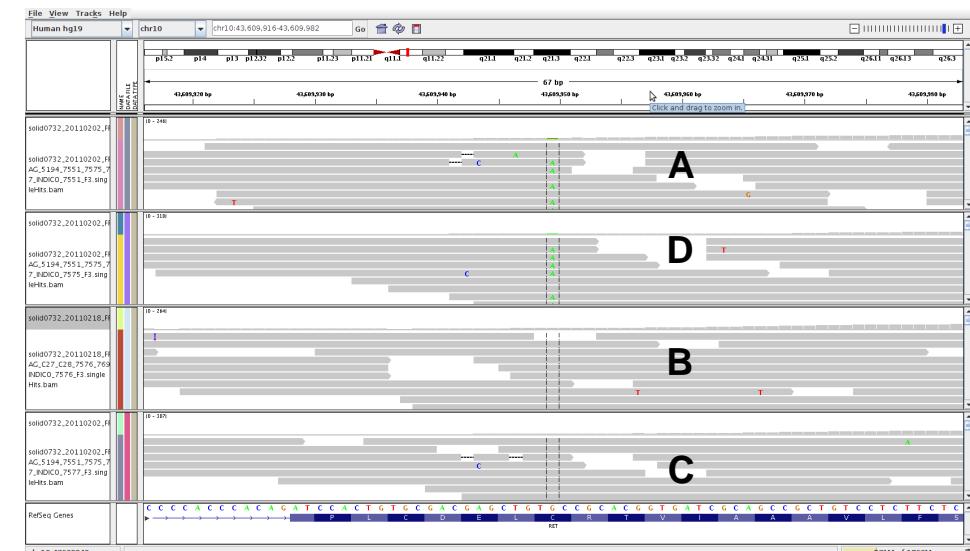


Cases

Controls



Segregation
within a
pedigree



Variant/gene prioritization by heuristic filtering



Variant level

Potential impact of the variant

Population frequencies

Experimental design level

Family(es)
Trios
Case / control

Functional (system) level

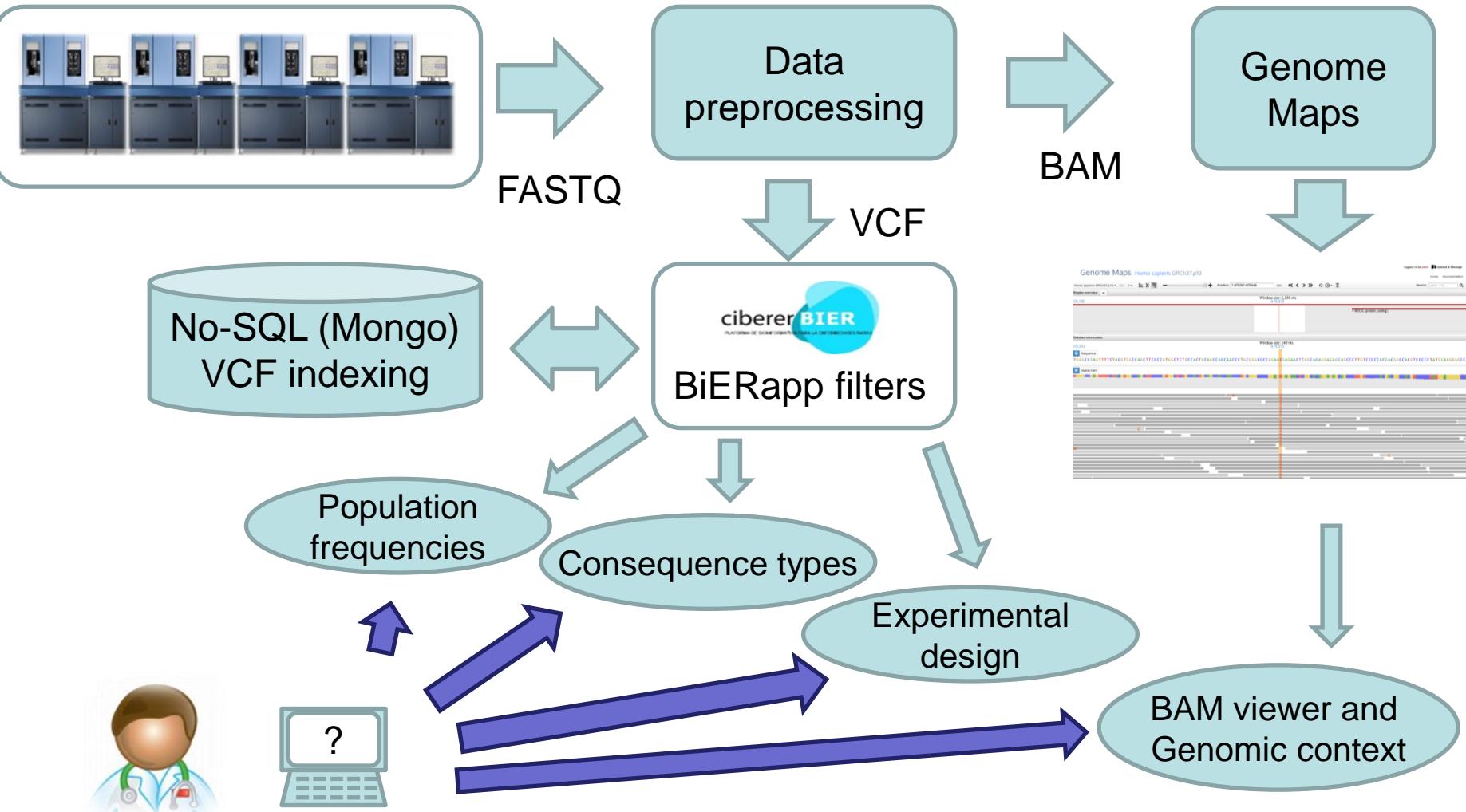
Gene set
Network analysis
Pathway analysis

Control of sequencing errors (missing values)

Testing strategies

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

SEQUENCING CENTER



BiERapp: interactive heuristic filtering tool for candidate gene prioritization

Menu BierApp ciberBIEB anonymous logout upload & manage profile jobs

Example 1000G (Short) Variant Browser Variants 1-10 of 22

Filter

Clear	Submit		
Aggregation			
0/0	0/1	1/1	1..
A19600:	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
A19600:	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A19601:	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
A19685:	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
MAF			
1000G MAF <:	<input type="text"/>		
EVS MAF <:	<input type="text"/>		
100G Populations			
African MAF <:	<input type="text"/>		
American MAF <:	<input type="text"/>		
Asian MAF <:	<input type="text"/>		
European MAF <:	<input type="text"/>		
Position			
NP id:	<input type="text"/>		
Chromosomal Location:			
1:1-1000000;2:1-1000000			
Gene / Transcript:			
Consequence Type			
<input type="checkbox"/> 5KB_downstream_variant			
<input type="checkbox"/> coding_sequence_variant			
<input type="checkbox"/> RNA_polymerase_promoter			
<input checked="" type="checkbox"/> stop_gained			
<input type="checkbox"/> DNase1_hypersensitive_site			
<input type="checkbox"/> exon_variant			
<input type="checkbox"/> 3_prime_UTR_variant			
<input type="checkbox"/> intron_variant			
<input type="checkbox"/> SNP			
<input checked="" type="checkbox"/> stop_lost			

Variant Browser

Variant	Alleles	Gene	Samples				SNP Id	Controls (MAF)		Conseq_Type	Polyphen	SIFT	Phenotype
			NA19600	NA19660	NA19661	NA19685		1000G	EVS				
4:172735897	C>G	GALNTL6	1/0	1/1	0/1	1/0	0.245 (C)	0.252	exon_variant,non_sy...	.	.	.	
17:8167600	T>C	PFAS	1/0	1/1	1/0	1/0	0.456 (T)	0.310	exon_variant,non_sy...	0.967 - (probably da...	.	.	
4:86844835	A>G	ARHGAP24	0/1	1/1	0/1	0/1	0.350 (G)	0.317	5KB_upstream_vari...	1 - (probably damage...	.	.	
4:169299528	T>C	DDX60L	0/1	1/1	1/0	0/1	0.474 (C)	0.428	exon_variant,non_sy...	0.001 - (benign)	0.63 - (tolerated)	.	
17:39240504	A>G	KRTAP4-9.KRTAP4-7	1/0	1/1	1/0	0/1	0.458 (A)	.	exon_variant,DNAs...	0 - (unknown)	0.75 - (tolerated)	.	
11:308314	T>C	IFITM2	0/1	1/1	0/1	0/1	0.406 (T)	0.454	5KB_upstream_vari...	0.306 - (benign)	0.63 - (tolerated)	.	
15:65916527	A>T	SLC24A1	1/0	1/1	1/0	1/0	0.297 (T)	0.160	exon_variant,5KB_d...	0.862 - (possibly da...	.	.	
6:42666061	T>C	PRPH2	0/1	1/1	0/1	0/1	0.219 (T)	0.223	exon_variant,5KB_d...	0.001 - (benign)	0.37 - (tolerated)	.	
5:90518792	T>G	CTD-221E18.1.RI...	0/1	1/1	0/1	0/1	0.477 (T)	0.481	5KB_upstream_vari...	.	.	.	
17:11420375	G>A	PRR3	1/0	1/1	1/0	1/0	0.199 (G)	0.117	exon variant 5KR d...	0 - (unknown)	0.4 - (tolerated)	.	

Variant Data

Genomic Context Effect & Annotation Study Summary

Position: 15:65916466-65916588 Window size: 6,149nts Region overview

65,916,453 65,916,527 65,916,588

65,916,466 65,916,527 65,916,588

65,916,466 65,916,527 65,916,588

Gene SNP

Powered by [Genome Maps](#)

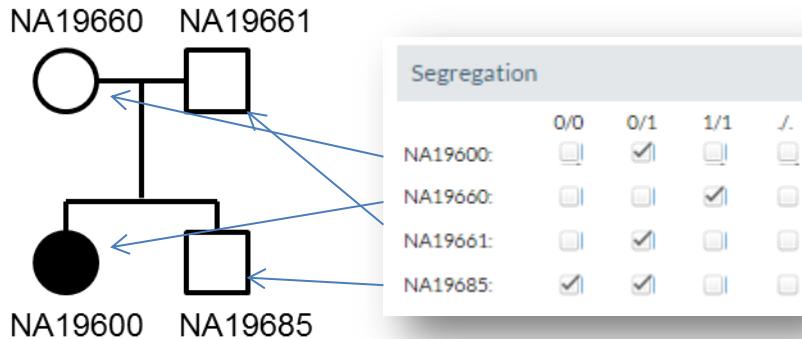
NA19660 NA19661

NA19600 NA19685

```

graph TD
    NA19660((NA19660)) --- NA19661((NA19661))
    NA19661 --- NA19600((NA19600))
    NA19661 --- NA19685((NA19685))
    NA19600 --- NA19685
  
```

BiERapp interactive filters



Any pedigree with any inheritance model can easily be defined

MAF

1000G MAF < :	<input type="text"/>
EVS MAF < :	<input type="text"/>
1000G Populations	
African MAF < :	<input type="text"/>
American MAF < :	<input type="text"/>
Asian MAF < :	<input type="text"/>
European MAF < :	<input type="text"/>

Different population frequencies can be used (including local population)

Consequence Type

- 5KB_downstream_variant
- coding_sequence_variant
- RNA_polymerase_promoter
- 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

Position

SNP id:	<input type="text"/>
Chromosomal Location:	1:1-1000000;2:1-1000000
Gene / Transcript:	<input type="text"/>

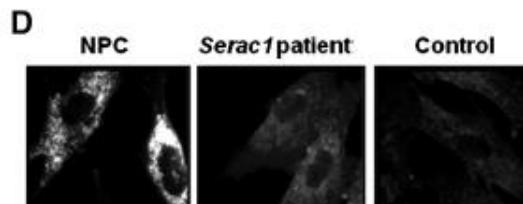
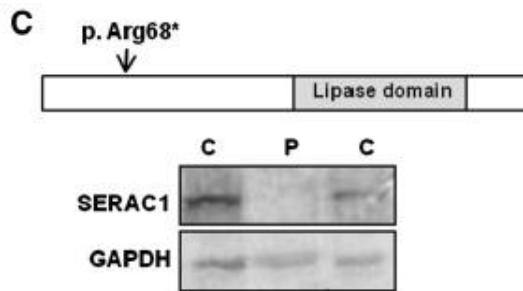
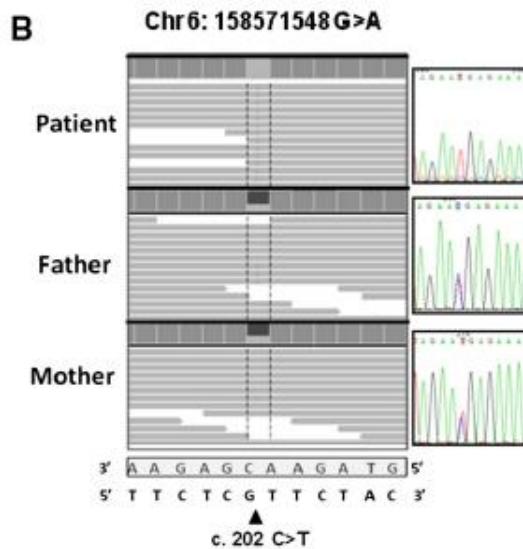
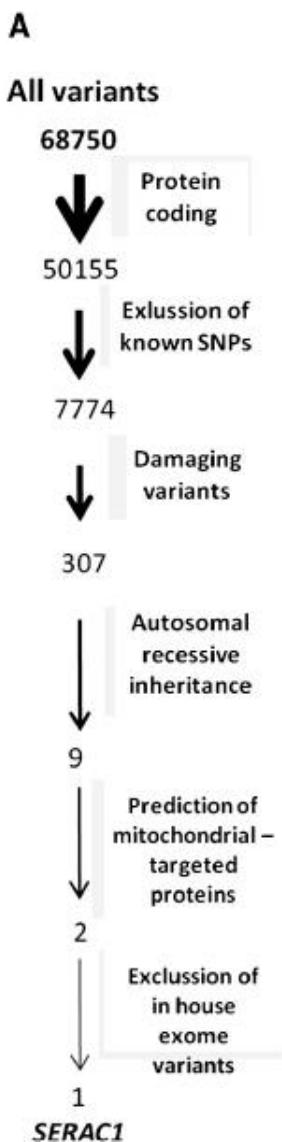
Regions, genes or known SNPs can be selected

Different consequence types can be selected. Also SIFT, PolyPhen and PhastCons can be used

Heuristic 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-MGAuria) 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.



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 García-Villoria ^{a,b}, María 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 Mitochondriales- Enfermedades Metabólicas Hereditarias, Servicio de Pediatría, Hospital 12 de Octubre, Madrid, Spain

^d BIER, CIBERER, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain

^e Computational Medicinal Institute, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain

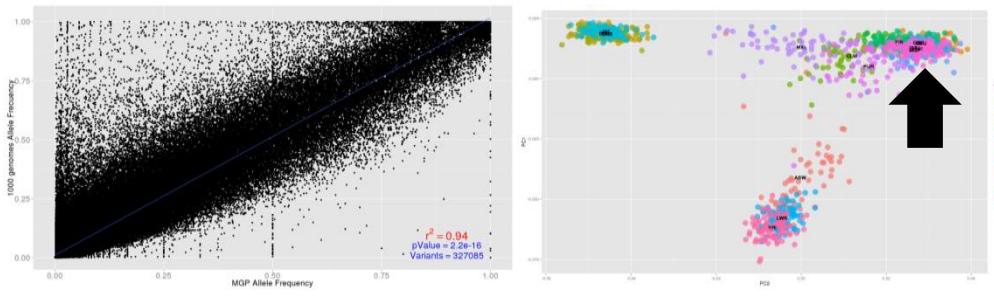
^f Functional Genomics Node, (INB) at CIPF, Valencia, Spain

^g Consejo Superior de Investigaciones Científicas (CSIC), Barcelona, Spain

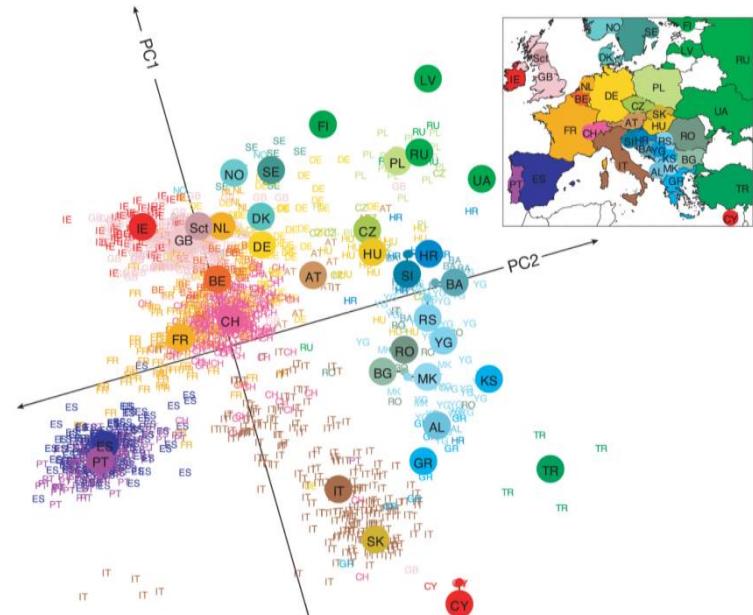
^h Monique and Jacques Roboh Department of Genetic Research, Hadassah, Hebrew University Medical Center, Jerusalem, Israel

Use known variants and their population frequencies to filter out false candidates

- 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 an extra filtering step to the analysis pipeline



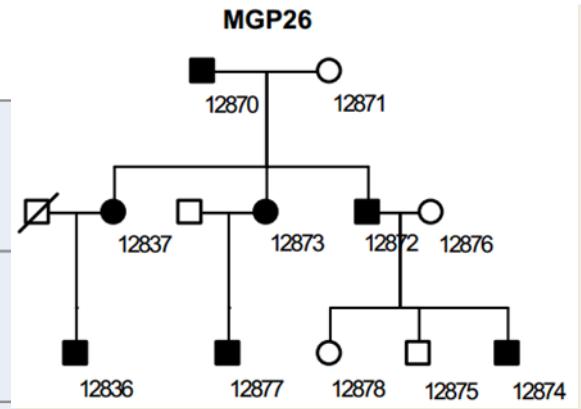
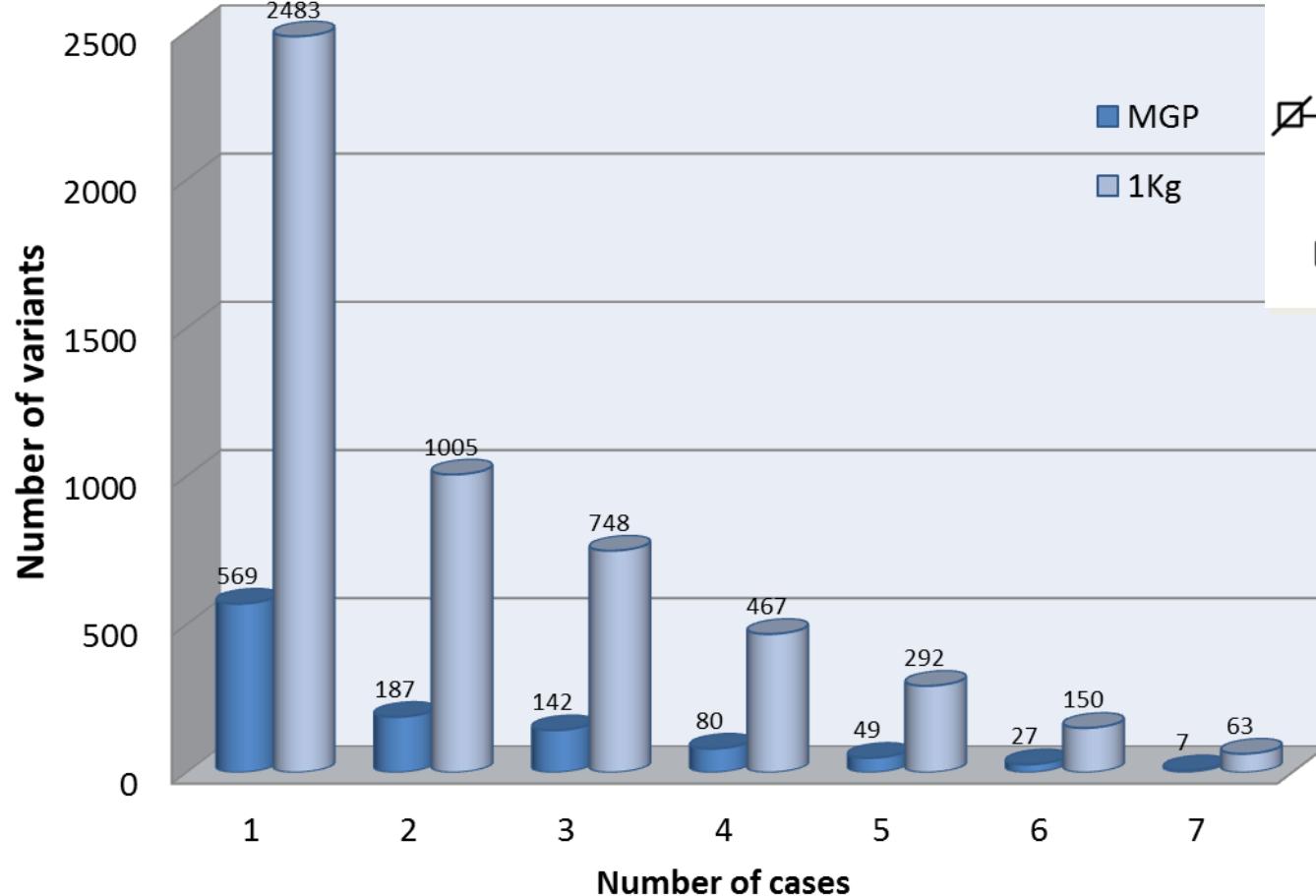
How important do you think is local information to detect disease genes?



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

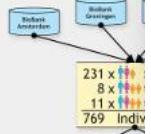
The CIBERER Spanish Variant Server (CSVs): the first repository of variability of the Spanish population



Home News About Participants Download data Request data Browse data Wiki

Ultra-sharp genetic group portrait of the Dutch

What genetic variation is to be found in the Dutch and this is not only interesting in itself, it also helps to establish biobanks. The Dutch biobank collaborative GO NL ("Genome of the Netherlands" (GoNL)) because it offers development of new biomarkers and diagnostic tools for people across the country and an atlas of the Dutch genome will disclose a wealth of new information.



ARTICLES

• Prize at Genetics Retreat 2014
• Search GoNL single online

nature genetics

Whole-genome sequence variation, population structure and demographic history of the Dutch population

The Genome of the Netherlands Consortium*

Whole-genome sequencing enables complete characterization of genetic variation, but geographic clustering of rare alleles demands many diverse populations be studied. Here we describe the Genome of the Netherlands (GoNL) Project, in which we sequenced 231 individuals from 8 Dutch families, and performed whole-genome resequencing of 11 additional families. We identified 1.1 million single-nucleotide variants and 1.2 million insertions and deletions. The intermediate coverage (~1x) and trio design enabled extensive characterization of strain-level variation, including meiotic events (0.5–50 bp) previously catalogued and de novo characterized. Population-level analysis revealed a high degree of differentiation between the Dutch and other European populations, with lower frequency alleles. Population genetic analyses demonstrate fine-scale structure across the country and support single nucleotide polymorphism (SNP)-based clustering at the level of provinces. The GoNL Project illustrates how single population whole-genome sequencing can provide detailed characterization of genetic variation and may guide the design of future population studies.

OPEN ACCESS freely available online

PLOS GENETICS

Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population

Elaine T. Lim^{1,2,3,4}, Peter Wurz^{2,6,7}, Aki S. Havulainen^{5,6}, Pärt Pälta^{1,3,8}, Taru Tukainen^{1,3,9}, Karoliina Päkinen⁵, Tuomas Esko^{2,3,9,10}, Reetiki Mägi⁷, Michael Inouye⁹, Tuuli Lappalainen^{1,3,11}, Yingleong Chan^{1,4,9}, Rany M. Salem^{1,10}, Monkol Lek^{1,2,4}, Jason Flannick^{1,2}, Xueling Sim^{1,14}, Alisa Manning^{1,2}, Class Lagat^{1,2}, Mikael Maksimov^{1,15}, Marjaana Jylhä^{1,16}, Benjamin Horne², Ruth F. Anuj Goyal^{1,2}, Martin Farre^{1,17}, Sekar Kathiresan¹⁰, Stacey Levy^{1,18}, Leif Grönberg^{1,19}, Jaakko Kaprio^{1,20}, David M. Altshuler^{2,3}, Cedric Nelson B. Freimer^{1,21}, Tanja Richard Durbin^{1,2,22}, Daniel G. Palotie^{1,2,5,44,22}, Aarno Palotie^{1,2,5,44,22}, for the Focus on Genomes of Icelanders

ARTICLES

Large-scale whole-genome sequencing of the Icelandic population

Daniel F. Gudbjartsson^{1,2,31}, Hannes Helgason^{1,2,32}, Sigrun Ólafsdóttir¹, Florian Zink¹, Arnarudur Oddsson¹, Arnarudur Gylfason¹, Sonja Beschorner², Gisli Massonsson¹, Björn V. Halldorsson^{3,4}, Eirikur Hjartarson⁵, Gunnar Ólafsson⁶, Sverrir N. Sverrisson⁷, Michael I. Frigge⁸, Kristanna Helgadóttir⁹, Jonna Saemundsdóttir¹⁰, Hafdi Þ. Hafðadóttir¹¹, Hrefna Jónsdóttir¹², Gísli Sigurðardóttir¹³, Guðrún Gudmundsdóttir¹⁴, Ólafur D. Svartsson¹⁵, Solveig Gretarsdóttir¹⁶, Þorunn Rafnur¹⁷, Björn Þrifjóldóttir¹⁸, Ólafur Ólafsson¹⁹, Einar S. Björnsson^{20,21}, Sigríður Ólafsdóttir²², Hildur Þórharsdóttir²³, Þorsteinn Steingrimsdóttir^{21,24}, Þorsteinn S. Gudbjartsson²⁵, Ágúst Þorgeirsson²⁶, Gíður Jónsdóttir²⁷, Ágúst Sigurðsson²⁸, Glyða Björnsdóttir²⁹, Þóra Þ. Ólafsson³⁰, Ólafur Ólafsson²⁹, Þorsteinn Ólafsson²², Ólafur Ólafsson³¹, Ólafur Ólafsson³², Ólafur Ólafsson³³, Ólafur Ólafsson³⁴, David O. Arman^{2,29}, Ólafur Th. Magnússon², Ágústine Kong^{2,22}, Gisli Masson¹, Unnur Thorsteinsdóttir^{1,2}, Ágúst Helgason^{2,29}, Patrick Solem¹, Kari Stefansson^{1,2}

FOCUS ON GENOMES OF ICELANDERS

ARTICLES

CIBERER Spanish Variant Server 1.0.1

Search Stats

BIER

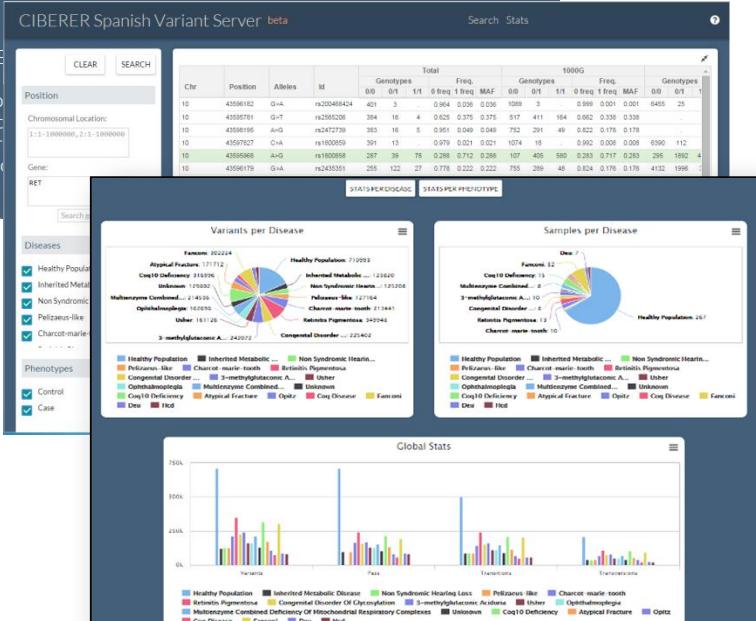
Try it now

Overview

Welcome to the CIBERER Spanish Variant Server (CSVs). We provide information about the variability of the Spanish population for scientific/medical purposes. Local variations in the CSVs currently stored. We accept submissions of samples.

Variants per Disease

Disease	Famconi	Healthy Population	Unrelated Individual	Non-Syndromic Hearing Loss	Ocular Disease	Chloracne-like	Unspecified	Other	3-methylglutaconic Aciduria	Cholesteryl Ester储积症	Unrelated Individual	Unknown	Optic	Global Stats
Famconi	202244	177172	133120	123000	112054	101208	9171	8741	844	809	7609	7089	6809	109000
Healthy Population														



Only another similar initiative exists, the Dutch genome, and more recently Finnish and Icelander populations have also been sequenced

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

The CSVs can be accessed via web

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

CIBERER Spanish Variant Server 1.0.1 ?

Gene or location

Case & control

Healthy population + different diseases (pseudocontrols) uncheck similar

Variant Annotations

Data provided:
Chromosomal location, change, genotypes, genotypes in other populations, pathogenicity indexes, phenotype, etc.

Chr	Position	Alleles	Id	Total					1000G					EVS					SIFT	POLYPHEN	PhastCons	phyloP	Phenotype				
				Genotypes	Freq.	0/0	0/1	1/1	0 freq	1 freq	MAF	Genotypes	Freq.	0/0	0/1	1/1	0 freq	1 freq						MAF	0/0	0/1	1/1
10	43596179	G>A	rs2435351	255	122	27	0.782	0.218	0.218	755	289	48	0.824	0.176	0.176	4132	1998	355	0.791	0.209	0.209	0.066	0.533				
10	43596182	G>A	rs200468424	401	3	.	0.998	0.004	0.004	1089	3	.	0.998	0.001	0.001	6455	25	.	0.998	0.002	0.002	0.005	-1.268				
10	43597827	C>A	rs1800859	391	13	.	0.984	0.016	0.016	1074	18	.	0.992	0.008	0.008	6390	112	1	0.991	0.009	0.009	0.772	0.533				
10	43597873	G>A		403	1	.	0.999	0.001	0.001	0.86 (tolerated)	0.004 (benign)	0.899	-0.313			
10	43597874	G>A		403	1	.	0.999	0.001	0.001	0.46 (tolerated)	0.071 (benign)	0.899	0.655			
10	43598195	A>G	rs2472739	383	16	5	0.968	0.032	0.032	752	291	49	0.822	0.178	0.178	0.004	0.533				
10	43600359	G>A		403	1	.	0.999	0.001	0.001	0.000	-0.485					
10	4369	C>G		403	1	.	0.999	0.001	0.001	0.001	0.374					
10	4372	G>A		403	1	.	0.998	0.002	0.002	0.003	0.456					
10	43517	G>A		403	1	.	0.999	0.001	0.001	0.14 (tolerated)	0.443 (benign)	0.096	0.550				

Page 2 of 10 Displaying 11 - 20 of 92

Genomic Context **Variant Annotations** Population Frequencies Phenotypes

Gene Name	Ensembl Gene Id	Ensembl Transcript Id	Conseq. type	Relative Position	Codon	Strand	Biotype	cDNA Position	cds Position	AA Position	AA Change	Sift	Polyphen
RET	ENSG00000165731	ENST000003565710	synonymous variant		gtC/gtA		protein_coding	607		375		-	-
RET	ENSG00000165731	ENST00000498820	intron variant				protein_coding					-	-
RET	ENSG00000165731	ENST00000340058	synonymous variant		gtC/gtA		protein_coding	555		375		-	-
RET	ENSG00000165731	ENST00000479913	upstream_gene_variant				retained_intron					-	-
			regulatory_region_variant									-	-

of 1 Displaying 1 - 5 of 5

The CSVs can be accessed via web

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

CIBERER Spanish Variant Server 10.1

Search Stats ?

Position

Chromosomal Location: 1:1-1000000, 2:1-1000000

Gene: RET

Diseases

- Healthy Population
- Inherited Metabolic Disease
- Non Syndromic Hearing Loss
- Pelizaeus-like
- Charcot-marie-tooth
- Retinitis Pigmentosa
- Congenital Disorder Of Glycosylation
- 3-methylglutaconic Aciduria
- Usher
- Ophthalmoplegia
- Multienzyme Combined

Phenotypes

Genomic Context Variant Annotations Population Frequencies Phenotypes

Gene Name Ensembl Gene Id Ensembl Transcript Id Conseq. type Rel. Position Codon Strand Biotype cDNA Position cds Position AA Position AA Change Sift Polyphen

RET	ENSG00000165731	ENST00000355710	synonymous variant		gtCgtA	protein_coding	607	375	-	-
RET	ENSG00000165731	ENST00000408820	intron variant		gtCgtA	protein_coding	-	-	-	-
RET	ENSG00000165731	ENST00000340058	synonymous variant		gtCgtA	protein_coding	555	375	-	-
RET	ENSG00000165731									

Study Population SuperPopulation Ref. Allele Alt. Allele Ref. Allele Freq. Alt. Allele Freq. MAF 0/0 0/1 1/1

ESP_6500	European_American	European_American	C	A	0.988	0.012	0.012	0	0	0
ESP_6500	African_American	African_American	C	A	0.997	0.003	0.003	0	0	0
1000GENOMES	phase_1_AMR	phase_1_AMR	C	A	0.980	0.020	0.020	0	0	0
1000GENOMES	phase_1_EUR	phase_1_EUR	C	A	0.980	0.020	0.020	0	0	0
					0.990	0.010	0.010	0	0	0
					1.000	0.000	0.000	0	0	0
					1.000	0.000	0.000	0	0	0

Displaying 1 - 7 of 7

Region overview Window size: 141 Position: 10:43622062-43622202 Go! << < > >>

Detailed information Window size: 141 nts

SNP

Occurrence of pathological variants in “normal” population

Exome Server

Summary Variants Genome Viewer

Filters
Reload Clear Search
Region/Genome
 Region Gene
Enter genes (comma separated)
BBS2

Controls +

Variant	Alleles	Gene	BIER				1000G				EVS				Polyphen	SIFT	Phenotype				
			Genotypes		MAF	Genotypes		MAF	Genotypes		MAF	Genotypes		MAF							
			0/0	0/1		1/1	./.		0/0	0/1		1/1	./.								
16:56501806	C>T	OGFOD1,BBS2	74	1	.	.	.	0.007	736	316	40	.	0.181	4578	1758	165	.	0.160	BODY MASS INDEX,Height,Two-hour glu...		
16:56504724	G>C	OGFOD1,BBS2	40	28	7	.	.	0.280	830	237	25	.	0.131	4202	2024	275	.	0.198	BODY MASS INDEX,Height,Two-hour glu...		
16:56508721	T>C	OGFOD1,BBS2	74	1	.	.	.	0.007		
16:56508883	C>T	OGFOD1,BBS2	74	1	.	.	.	0.007		
16:56509441	T>G	BBS2,OGFOD1	73	2	.	.	.	0.013		
16:56510072	A>C	BBS2,OGFOD1	73	2	.	.	.	0.013	1085	7	.	.	0.003	6375	125	1	.	0.010			
16:56533804	T>G	BBS2	74	1	.	.	.	0.007	1076	15	1	.	0.008	6339	161	1	.	0.012			
16:56535193	C>T	BBS2	72	3	.	.	.	0.020			
16:56535207	AG...>...	BBS2	74	1	.	.	.	0.007			
16:56543827	A>G	BBS2	74	1	.	.	.	0.007	1084	8	.	.	0.004	6402	98	1	.	0.007			
16:56545175	T>C	BBS2	19	23	1	.	.	0.100	601	101	67	.	0.265	1193	2801	241	.	0.195	Fasting proinsulin/secretive cotts,Tot...		
16:56548501	C>T	BBS2	.	.	75	.	.	0.000	.	9	1083	.	0.004	1	73	6427	.	0.006	0.001	1	BARDET-BIEDL SYNDROME 2,Bardet-Bie...
16:56553814	A>G	BBS2	74	1	.	.	.	0.007	910	163	19	.	0.092	5706	741	50	.	0.065			
16:56553816	A>C	BBS2	74	1	.	.	.	0.007	910	163	19	.	0.092	5744	704	49	.	0.062			

Page 1 of 1 | << << >> >> | Columns | Variants 1 - 14 of 14

Reference genome is mutated

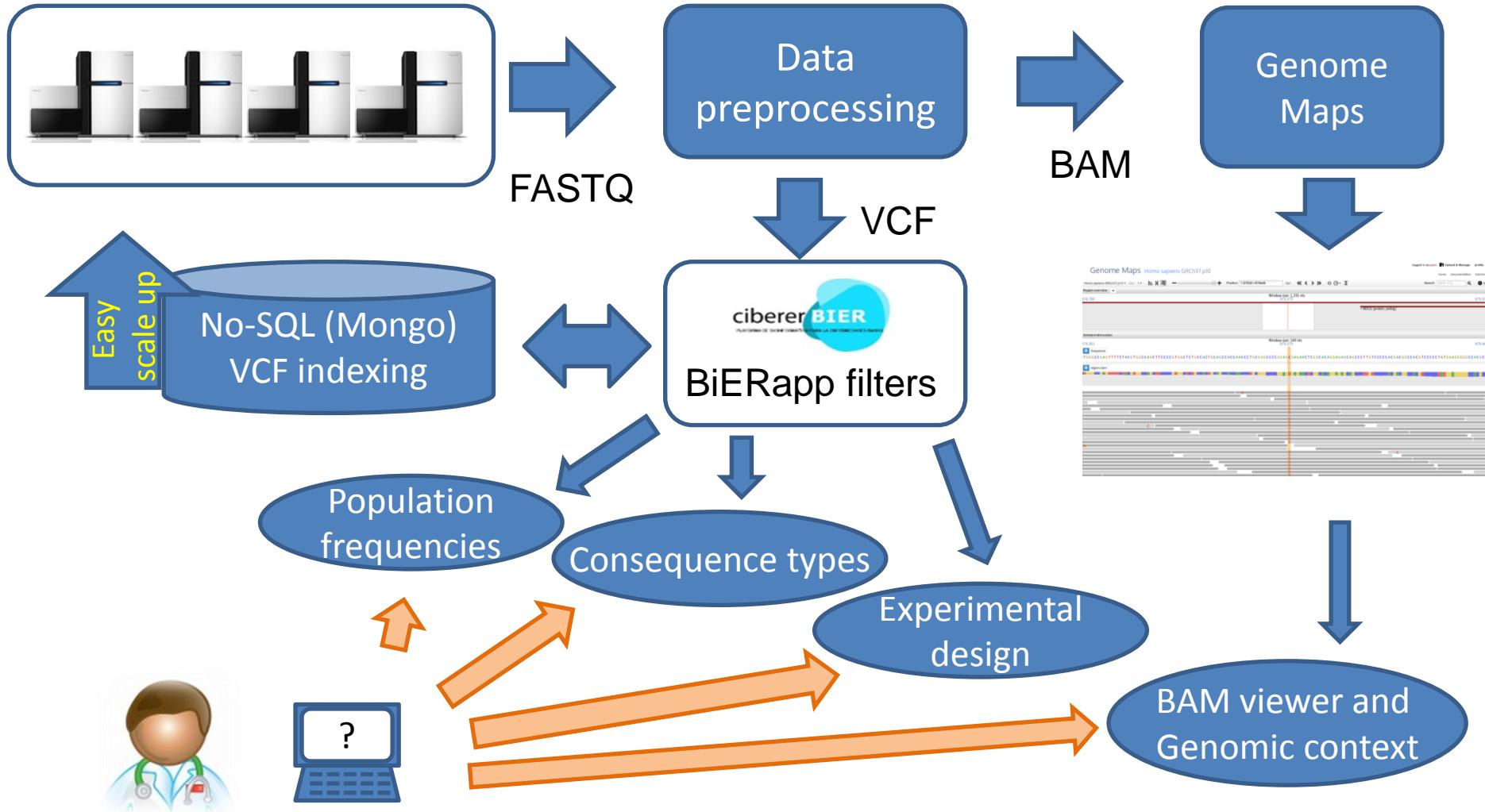
Nine carriers in 1000 genomes

One affected and 73 carriers in EVS

An example of end user's tool

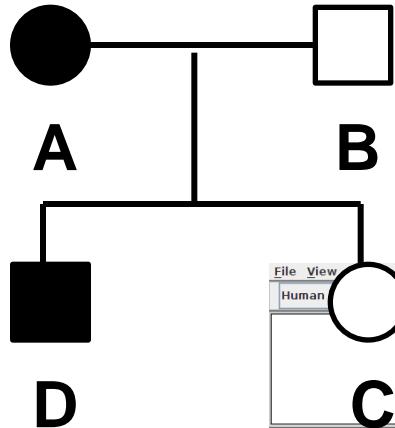
BiERapp: interactive web-based tool for easy candidate prioritization by heuristic filtering

SEQUENCING CENTER

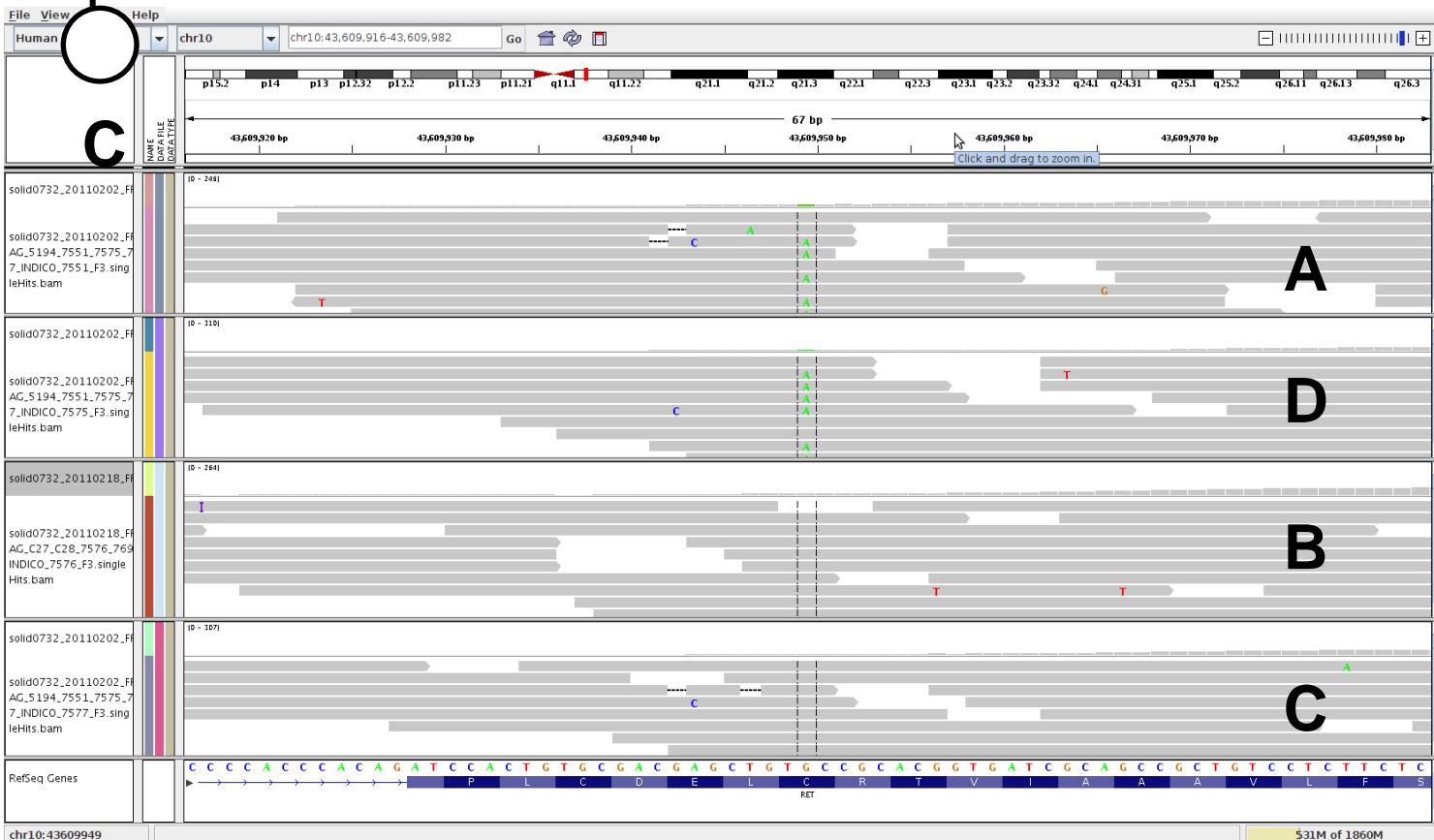


How efficient is exome/genome sequencing?

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



Heuristic filtering approach. Exome sequencing produces many false positives

Table 1 | Mean number of coding variants in two populations

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

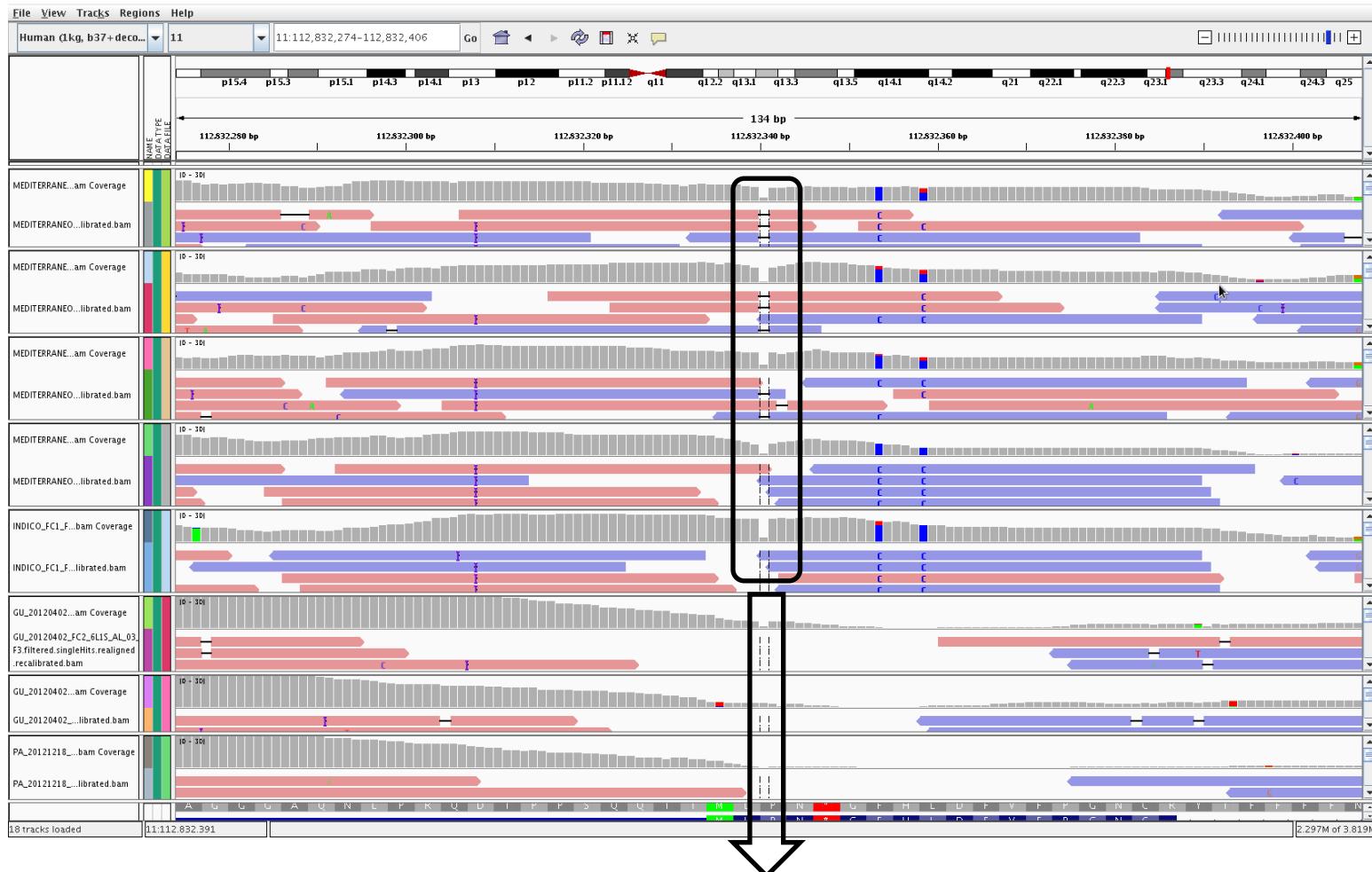
The table lists the mean number (\pm standard deviation (sd)) of novel and non-novel coding single nucleotide variants from 100 sampled African Americans and 100 European Americans. Non-novel variants refer to those found in dbSNP131 or in 200 other control exomes. Capture was performed using the Nimblegen 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).

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 can detect the disease mutation(s)... along with many other unrelated variants

Some false positives are errors that can easily be avoided. E.g. missing positions



The promising variant (a frameshift present in all patients but not detected in controls) was not real. It was not properly covered by reads in controls.

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

Known snps phenotypic effect									
							Hits	Description	
none	1	2116429	C	missense	0	PRKCZ,LOC10	8	Amyotrophic Lateral Sclerosis (ALS)	
none	1	2116429	C	missense	0	PRKCZ,LOC10	7	Parkinson's disease	
none	1	2116429	C	missense	0	PRKCZ,LOC10	6	Rheumatoid Arthritis	
none	1	2116429	C	utr-3	0	PRKCZ,LOC10	5	common polymorphism	
none	1	2318893	C	missense	0	MORN1	4	Multiple complex diseases—Crohn's disease , combined control dataset	
none	1	2452167	C	missense	0	PANK4	3	Alzheimer's Disease	
dbSNP_1000Genomes	1	2452569	T	coding-synonymous	2985862	PANK4	2	LDL cholesterol	
none	1	3680294	A	missense	0	CCDC27	2	Skin pigmentation	
none	1	3745852	T	missense	0	KIAA0562	2	Type 1 diabetes	
none	1	3746432	G	missense	0	KIAA0562	2	Coronary Artery Disease	
dbSNP_1000Genomes	1	3755675	T	coding-synonymous	1891941	KIAA0562	2	in allele DQB1*0501 and allele DQB1*0502	
none	1	6029181	G	missense	0	NPHP4	2	Multiple complex diseases—Bipolar disorder	
none	1	6101899	A	missense	0	KCNAB2	2	Multiple complex diseases—Coronary Artery Disease , gender differentiated	
none	1	6101899	A	intron	0	KCNAB2	2	Multiple complex diseases—Crohn's disease , combined control dataset , gender differentiated	
none	1	6132842	C	coding-synonymous	0	KCNAB2	2	Type I Diabetes	
none	1	6132842	C	coding-synonymous	0	KCNAB2	2	Systemic Lupus Erythematosus (SLE) , gender differentiated in women	
none	1	6535559	T	missense	0	PLEKHG5	2	Triglycerides	
none	1	6535559	T	missense	0	PLEKHG5	2	Type I Diabetes	
none	1	6535559	T	missense	0	PLEKHG5	1	893Ser-expressing (ABCB1:2677G>T (Ala893Ser)) cells showed 47% lower intracellular digo...	
none	1	6535559	T	missense	0	PLEKHG5	1	A study in 336 recipients of hematopoietic-cell transplants	
none	1	6535559	T	missense	0	ZBTB48			
none	1	6647590	A	missense	0	THAP3			
none	1	6694129	T	missense	0	DNAJC11			
none	1	6695719	T	utr-3	0	DNAJC11			
none	1	6704720	C	missense	0	DNAJC11			
none	1	6711636	C	coding-synonymous	0	PER3			
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	1002458	A	missense	0	NNM1AT1			

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
- 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

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.

The screenshot shows the homepage of the **Genome Medicine** journal. At the top, there is a search bar with the placeholder "Search Genome Medicine for". Below the search bar, there is a navigation menu with links for "Home", "Articles", "Authors", "Reviewers", "About this journal", "My Genome Medicine", and "Subscriptions".

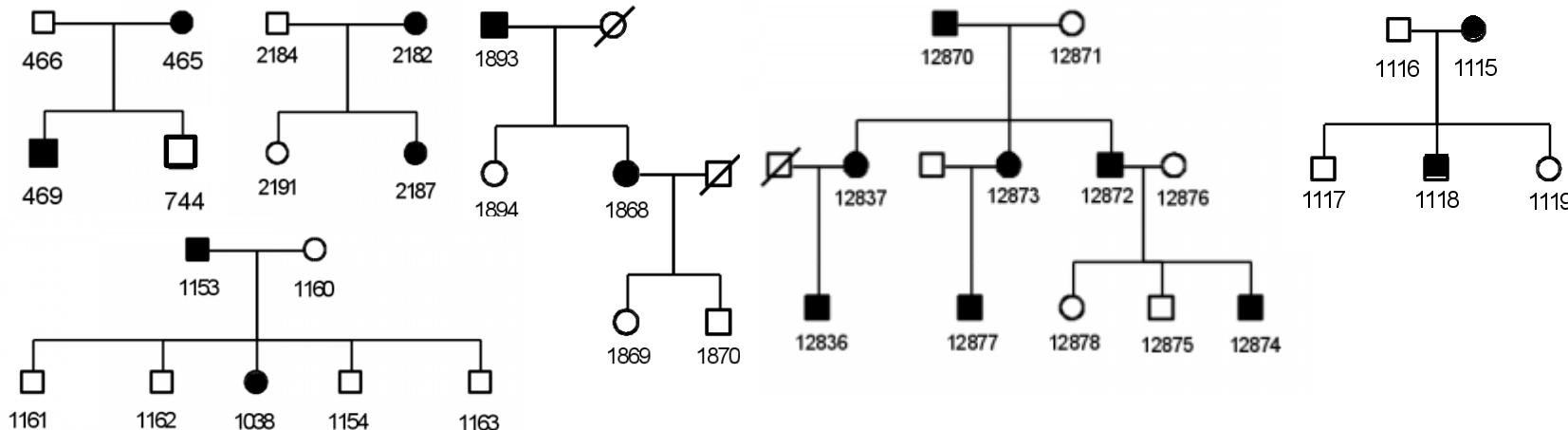
The main content area features a research article titled "A map of human microRNA variation uncovers unexpectedly high levels of variability". The article is marked as "Highly accessed" and "Open Access". The authors listed are Jose Carbonell, Eva Alloza, Pablo Arce, Salud Borrego, Javier Santoyo, Macarena Ruiz-Ferrer, Ignacio Medina, Jorge Jimenez-Almazan, Cristina Mendez-Vidal, Maria Gonzalez-del Pozo, Alicia Vela, Shomi S Bhattacharya, Guillermo Antinolo and Joaquin Dopazo.

The overall layout is clean and professional, typical of a scientific journal website.

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	1.9%
Fasting glucose ⁷⁷	4	1.5%

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

Human
genetics

NEWS FEATURE PERSONAL GENOMES NATURE Vol 456 November 2008



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.

2010 Nature America, Inc. All rights reserved.

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

Jian Yang¹, Beben Benyamin¹, Brian P McEvoy¹, Scott Gordon¹, Anjali K Henders¹, Dale R Nyholt¹, Pamela A Madden², Andrew C Heath², Nicholas G Martin¹, Grant W Montgomery¹, Michael E Goddard³ & Peter M Visscher¹

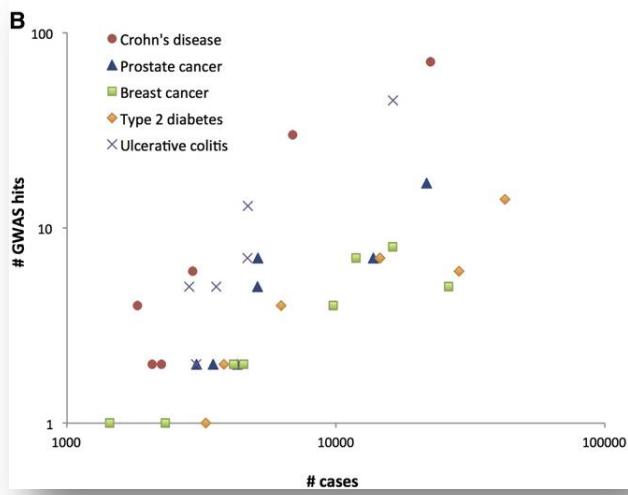
SNPs discovered by genome-wide association studies (GWASs) account for only a small fraction of the genetic variation of complex traits in human populations. Where is the remaining heritability? We estimated the proportion of variance for human height explained by 294,831 SNPs genotyped in 3,925 unrelated 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.

At the end, most of the heritability was there... variation that their effects do not reach stringent significance thresholds and/or the causal variants are not in complete linkage disequilibrium (LD) with the SNPs that have been genotyped. Lack of complete LD might, for example, occur if causal variants have a minor allele frequency (MAF) than genotyped SNPs. Here we test these two hypotheses by estimating the contribution of each to the heritability of human height and comparing it with the

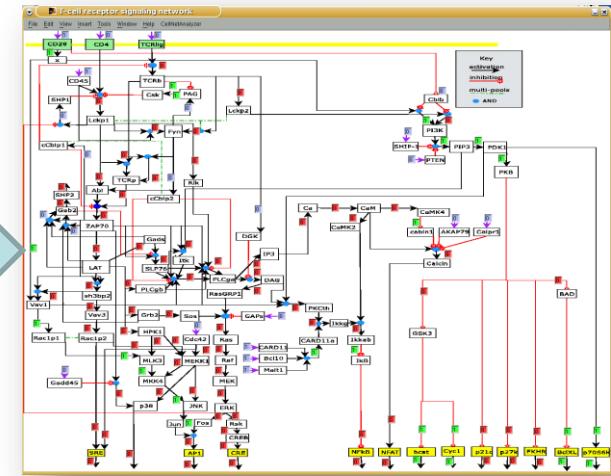
Height in humans is a classical quantitative trait, easy to measure and studied for well over a century as a model for investigating the genetic basis of complex traits. The heritability of height has been estimated to be ~0.8 (refs. 9,11–13). Rare mutations that cause extreme short or tall stature have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

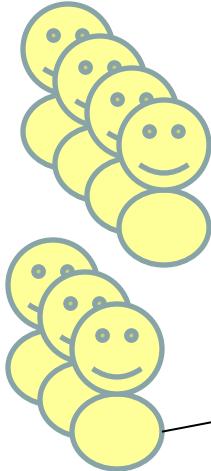
Limitations: Available information

Modular nature of human genetic diseases

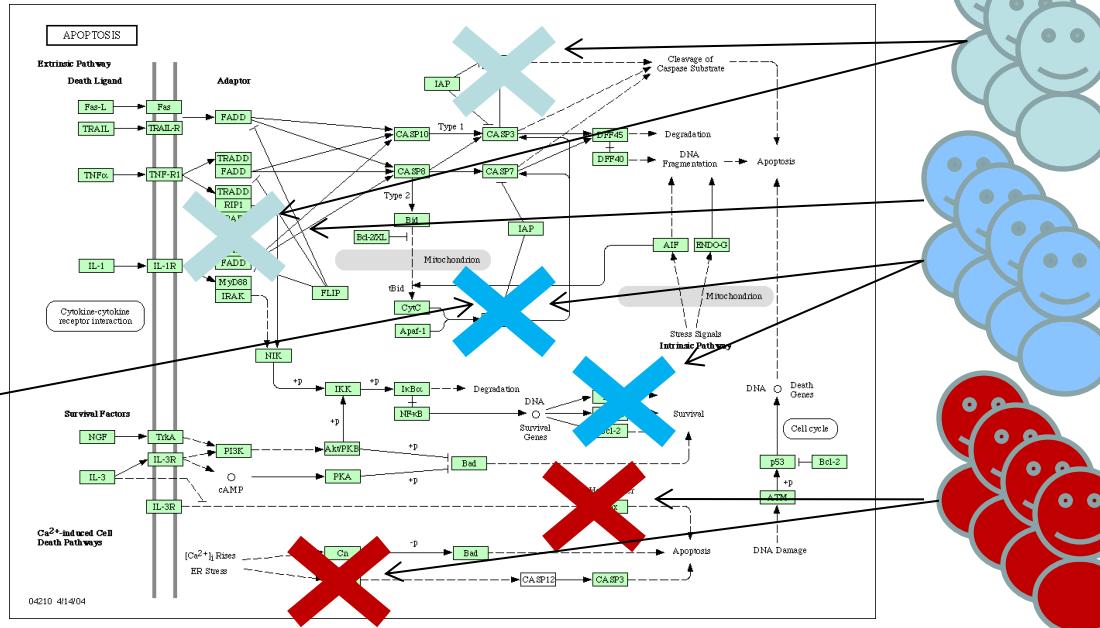
- With the development of **systems biology**, studies have shown that phenotypically similar diseases are often caused by **functionally related genes**, being referred to as the **modular nature of human genetic diseases** (Oti and Brunner, 2007; Oti et al, 2008).
- This modularity suggests that **causative 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)

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

Controls



Cases



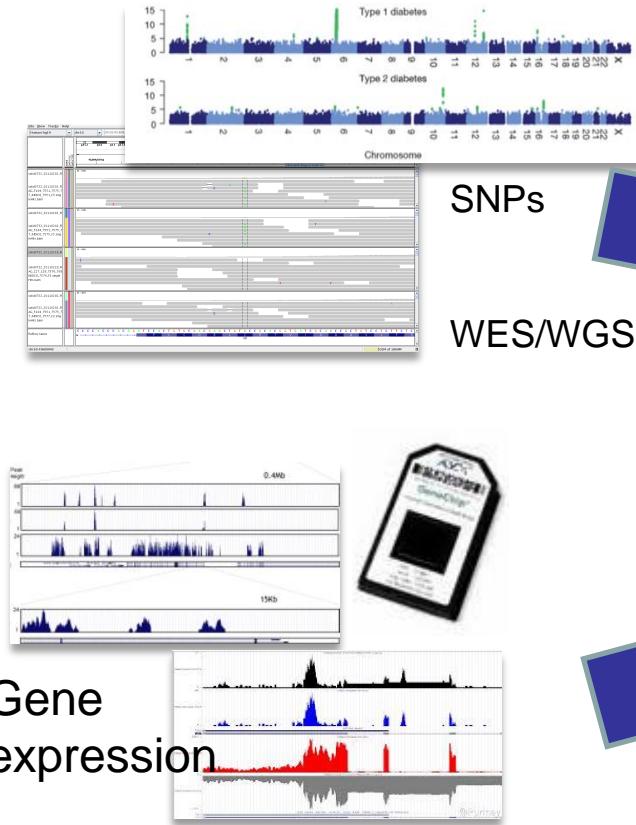
Affected **cases** in complex diseases will be a **heterogeneous** population with different mutations (or combinations).

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

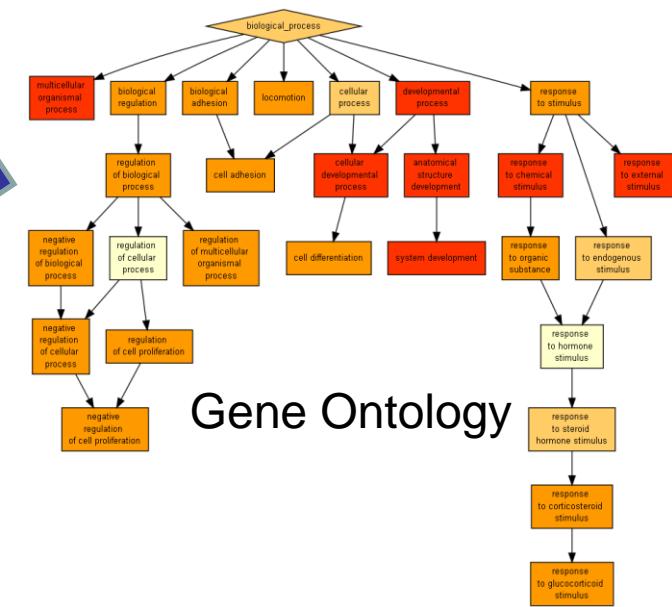
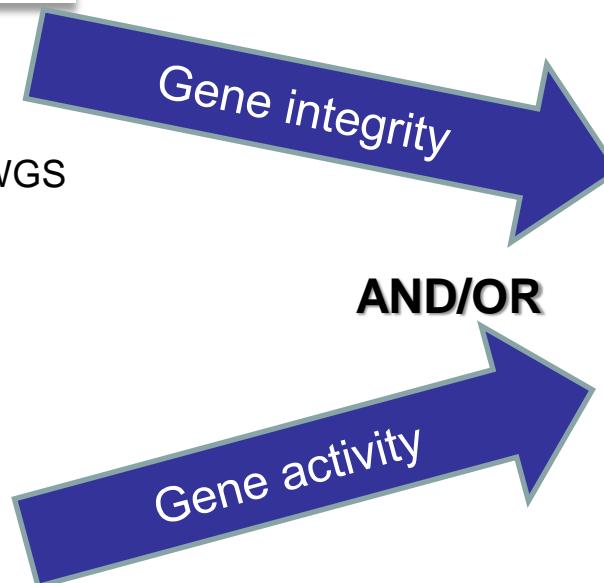
From gene-based to function-based perspective



SNPs

WES/WGS

Gene expression



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.

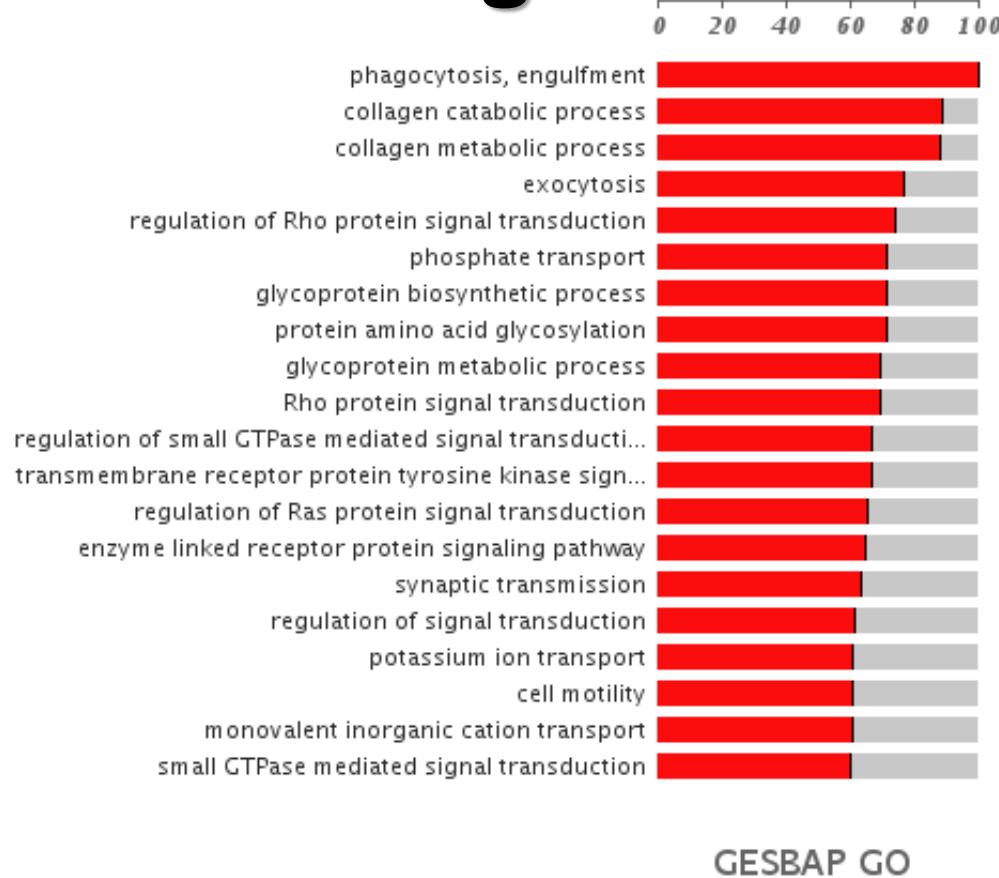
The CGEMS initiative. (Hunter et al. Nat Genet 2007)

1145 cases 1142 controls. Affy 500K

Conventional association test reports only 4 SNPs
significantly mapping 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

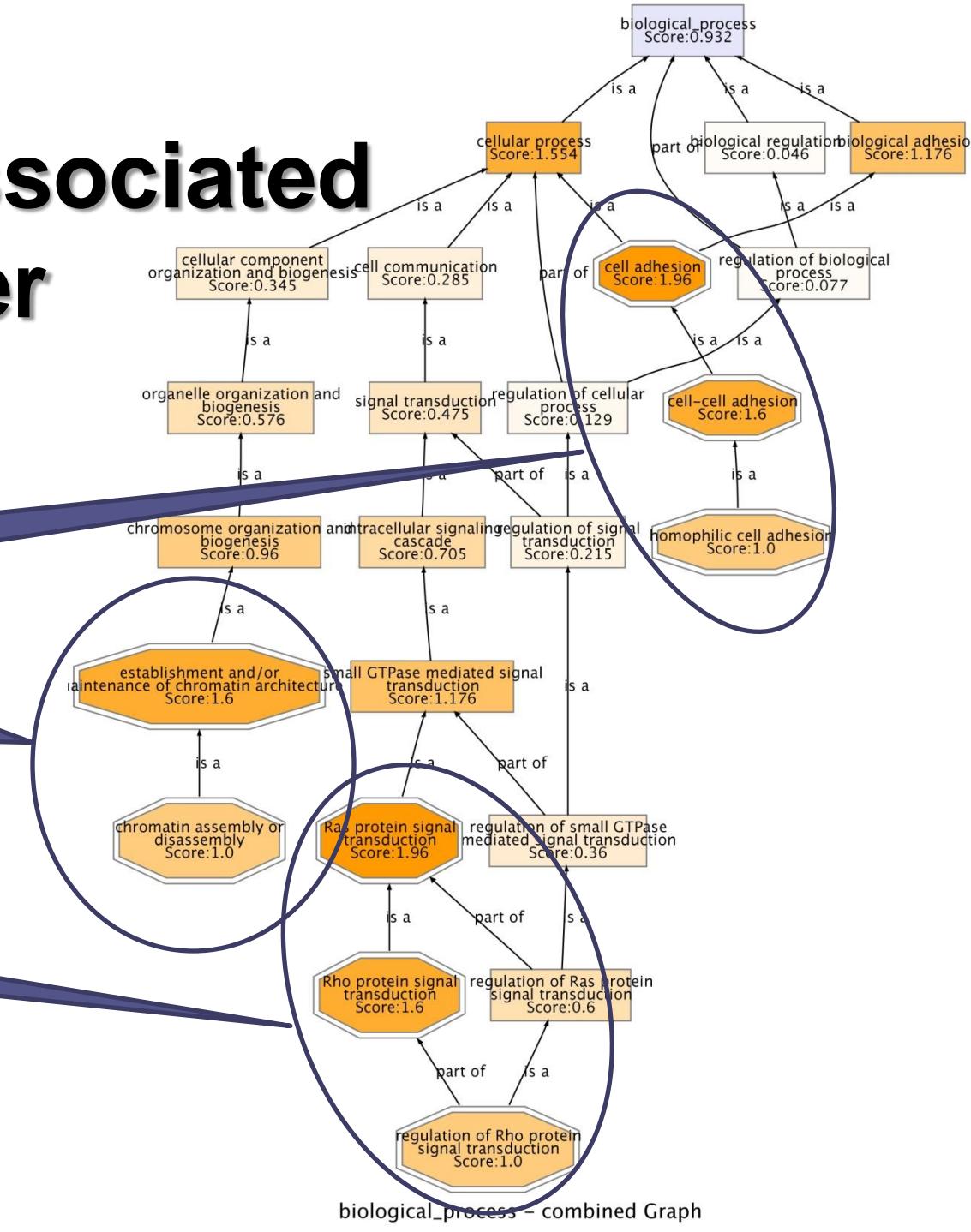
PBA reveals 19 GO categories including *regulation of signal transduction* (FDR-adjusted p-value=4.45x10⁻⁰³) in which FGFR2 is included.

GO processes significantly associated to breast cancer

Metastasis

Chromosomal instability

Rho pathway



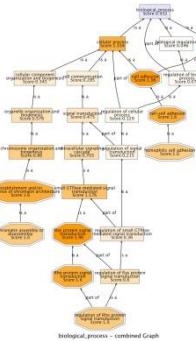
From gene-based to function-based perspective

SNPs,
Gene expression

Gene₁
Gene₂
Gene₃
Gene₄
:
:
:
Gene₂₂₀₀₀



Gene
Ontology

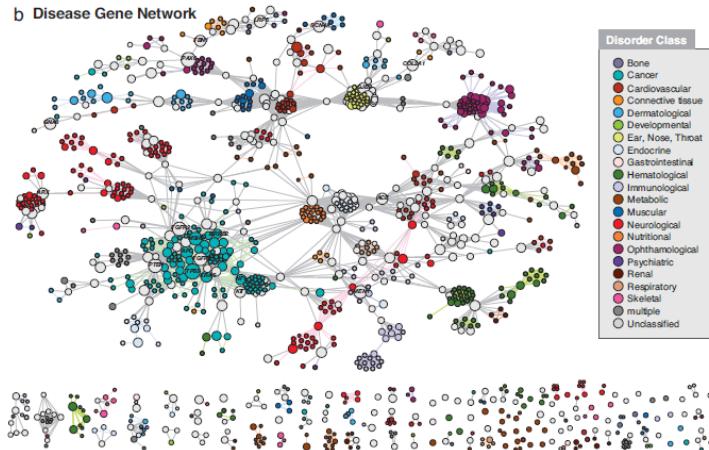


	SNPs, gene exp.	GO
Detection power	Low (only very prevalent genes)	high
Annotations available	many	many
Use	Biomarker	Illustrative, give hints

Can the interactome help to find disease mutations?

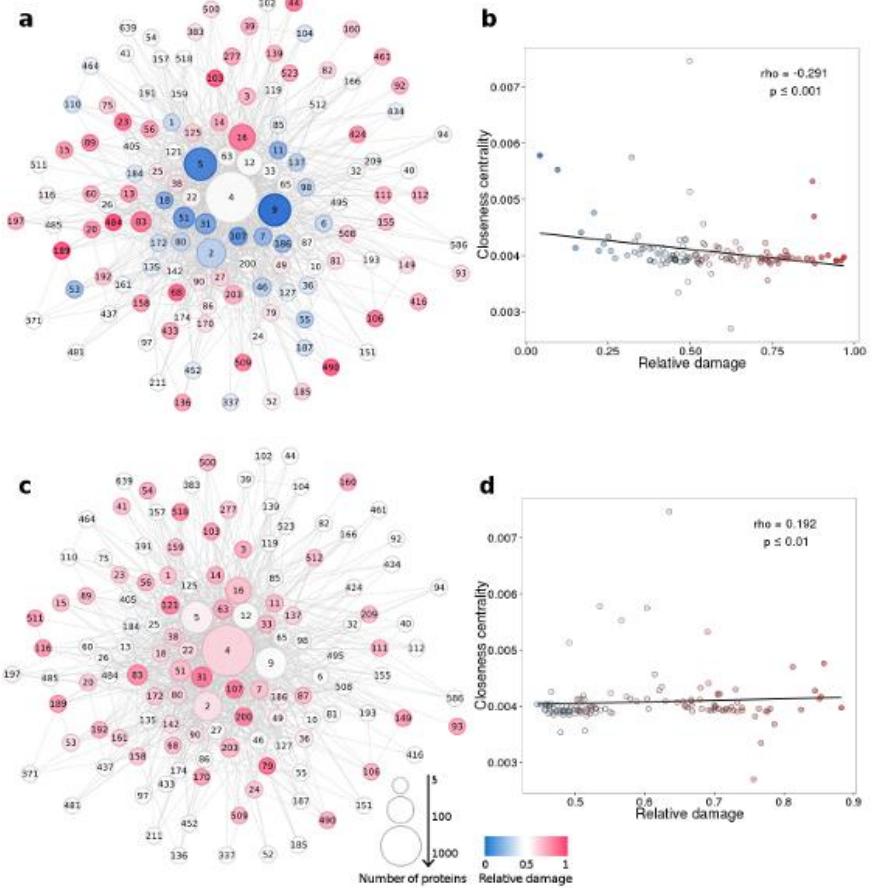
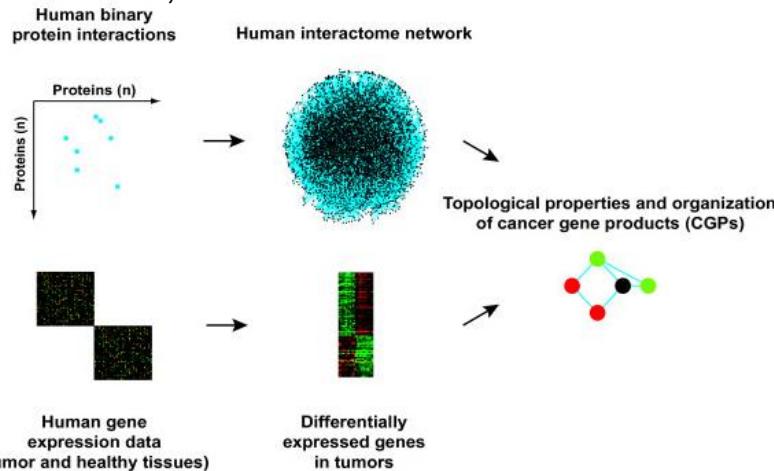
Disease genes are close in the interactome

Goh 2007 PNAS



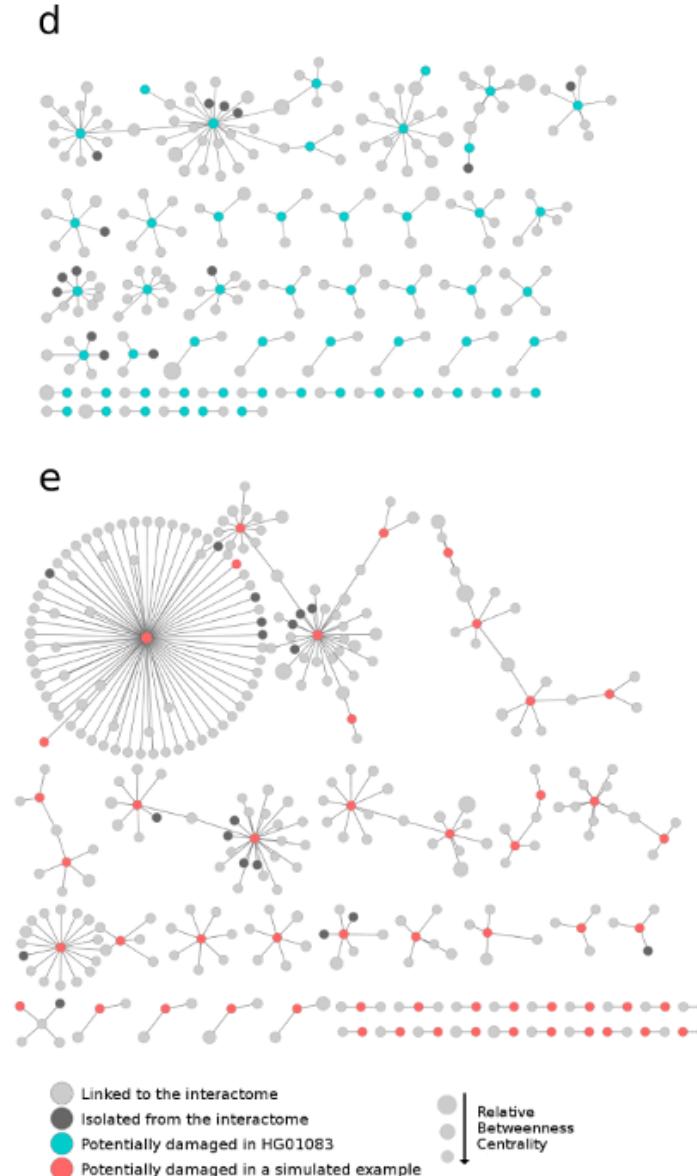
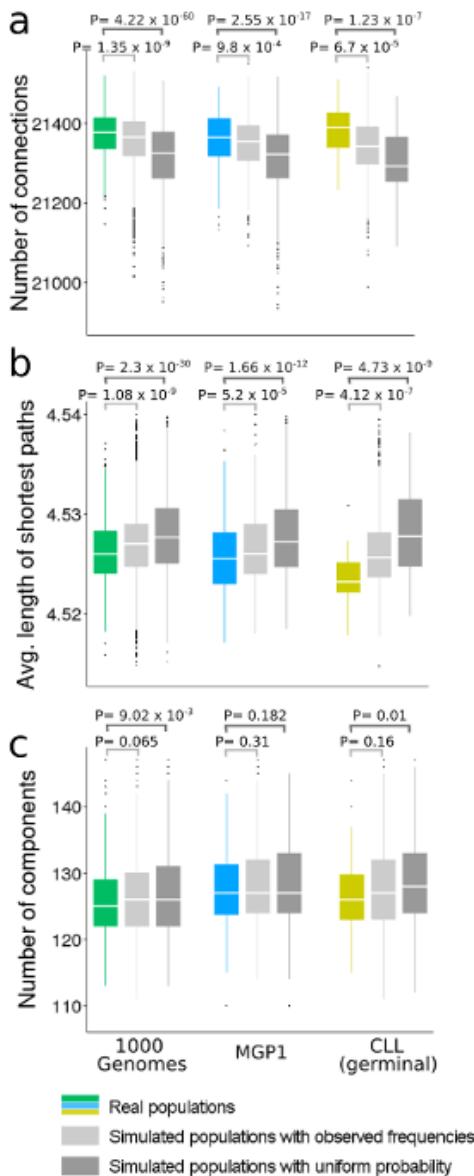
Cancer genes are central.

Hernandez, 2007 BMC Genomics



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

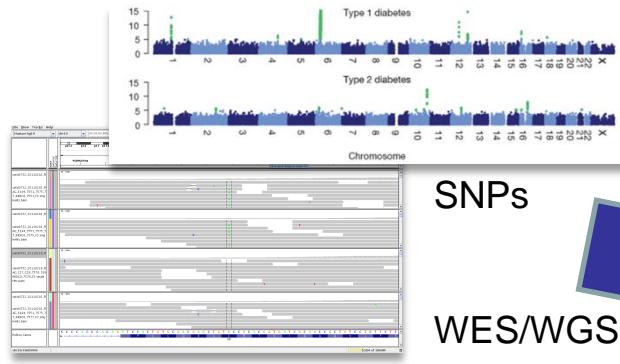
The role of interactome in buffering the deleteriousness of LoF mutations



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

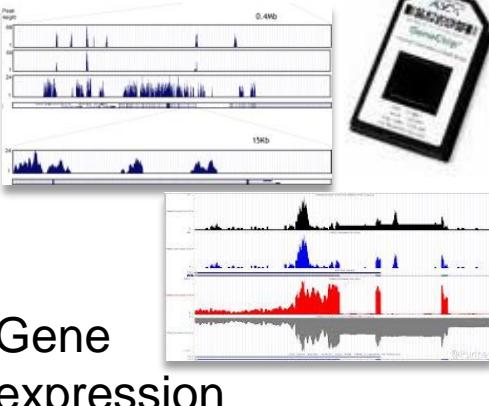
Garcia-Alonso 2014 Mol Syst Biol

From gene-based to function-based perspective



SNPs

WES/WGS



Gene expression

Gene integrity

AND/OR

Gene activity

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



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

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-Alonso³, Ana Torroglosa^{1,2}, Martina Marbà^{3,4}, Ma Valle Enguix-Riego^{1,2}, David Montaner³, Guillermo Antiñolo^{1,2}, Joaquín Dopazo^{2,3,4*} and Salud Borrego^{1,2*}

* Corresponding authors: Joaquín Dopazo idopazo@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

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-Alonso¹, Roberto Alonso¹, Enrique Vidal¹, Alicia Amadoz¹, Alejandro de María¹, Pablo Minguez², Ignacio Medina^{1,3} and Joaquín Dopazo^{1,3,4,*}

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

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

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

SNPs validated in independent cohorts

Nucleic Acids Research Advance Access published May 19, 2009

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

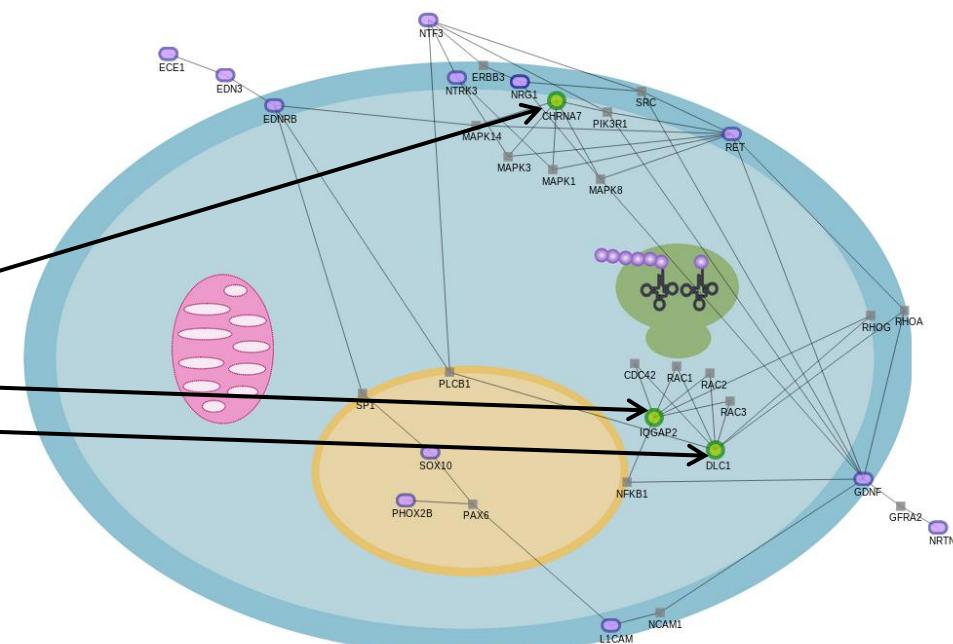
SNOW, a web-based tool for the statistical analysis of protein-protein interaction networks

Pablo Minguez¹, Stefan Götz^{1,2}, David Montaner¹, Fatima Al-Shahrour¹ and Joaquin Dopazo^{1,2,3,*}

¹Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF),

²CIBER de Enfermedades Raras (CIBERER) and ³Functional Genomics Node (INB) at CIPF, Valencia, Spain

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



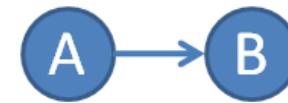
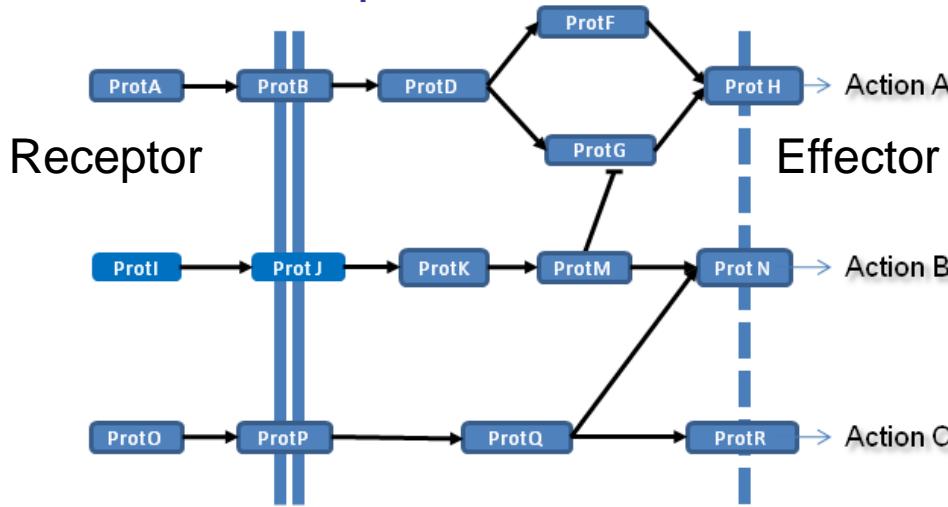
From gene-based to function-based perspective

	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



**Activations
and
repressions
occur**

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

Modeling pathways

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



METHODOLOGY ARTICLE

Open Access

Understanding disease mechanisms with models of signaling pathway activities

Patricia Sebastian-Leon¹, Enrique Vidal^{1,2,3}, Pablo Minguez^{1,4}, Ana Conesa¹, Sonia Tarazona¹, Alicia Amadorz¹, Carmen Armero⁵, Francisco Salavert^{1,2}, Antonio Vidal-Puig⁶, David Montaner¹ 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

Inferred the functional effect of gene expression changes in signaling pathways

Patricia Sebastián-León¹, José Carbonell¹, Francisco Salavert^{1,2}, Rubén Sanchez³, Ignacio Medina¹ and Joaquín Dopazo^{1,2,4,*}

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

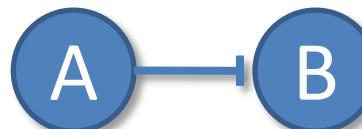
Received March 3, 2013; Revised April 18, 2013; Accepted May 2, 2013

Activation



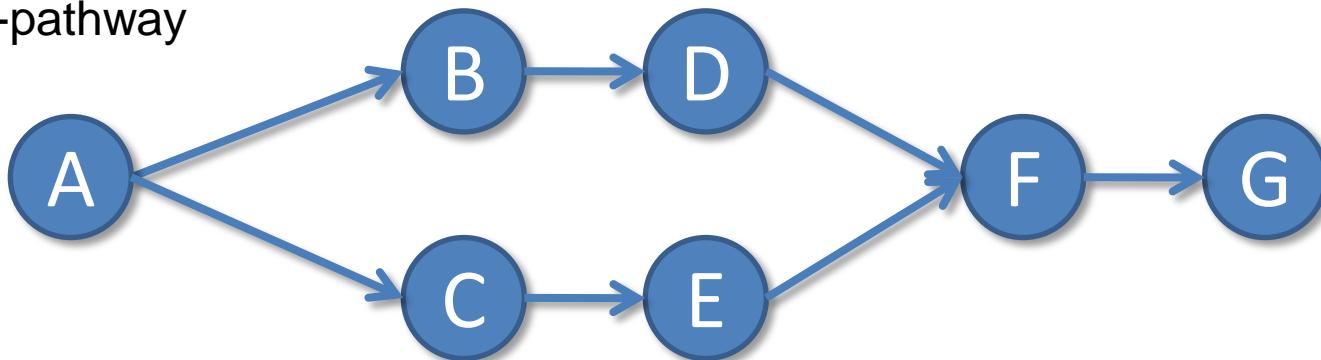
$$\text{Prob.} = P(\text{A activated})P(\text{B activated})$$

Inhibition



$$\text{Prob.} = [1 - P(\text{A activated})]P(\text{B activated})$$

Sub-pathway

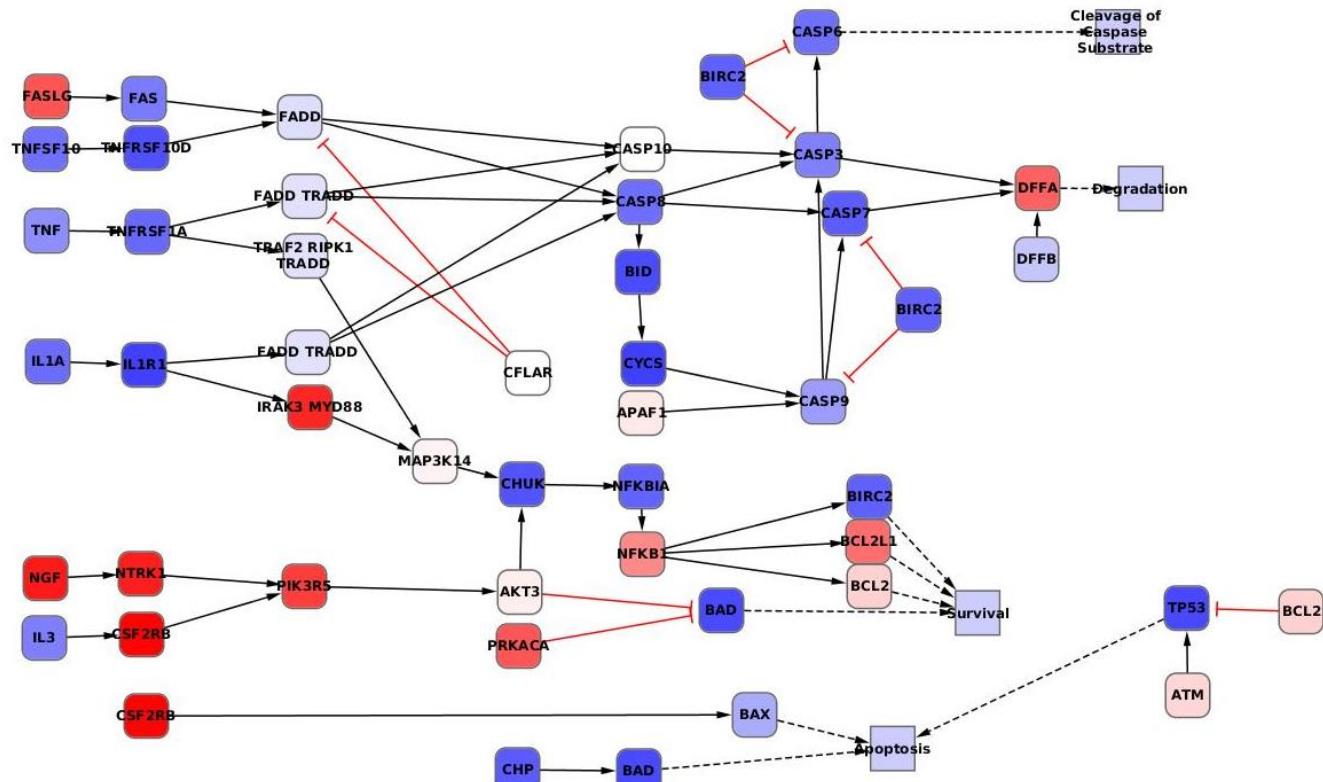


$$P(\text{A} \rightarrow \text{G activated}) = P(\text{A})P(\text{B})P(\text{D})P(\text{F})P(\text{G}) + P(\text{A})P(\text{C})P(\text{E})P(\text{F})P(\text{G}) - P(\text{A})P(\text{F})P(\text{G})P(\text{B})P(\text{C})P(\text{D})P(\text{E})$$

The effects of changes in gene activity are not obvious

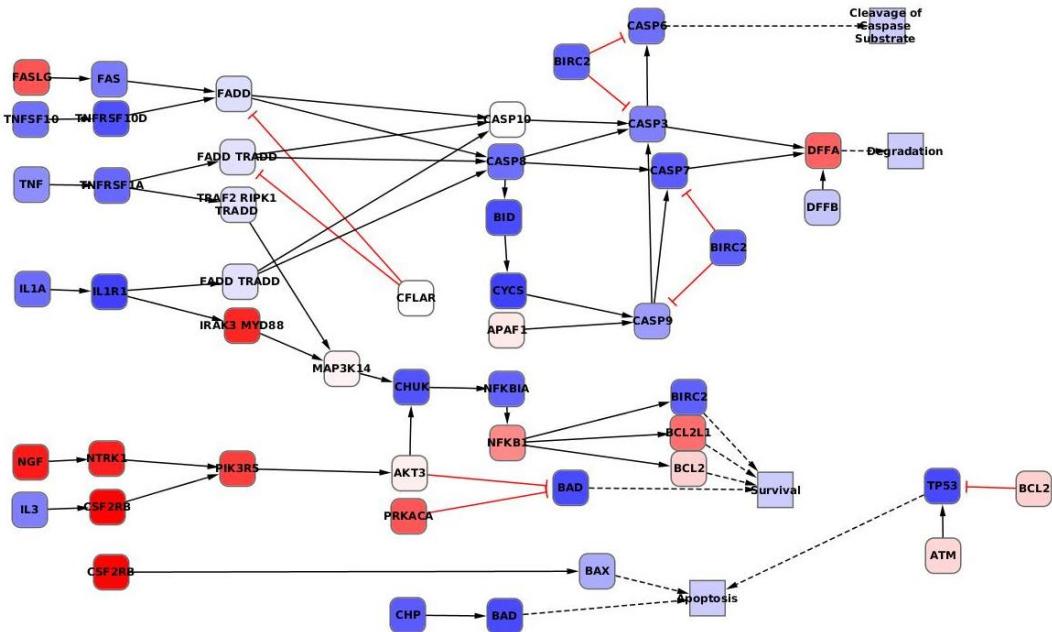
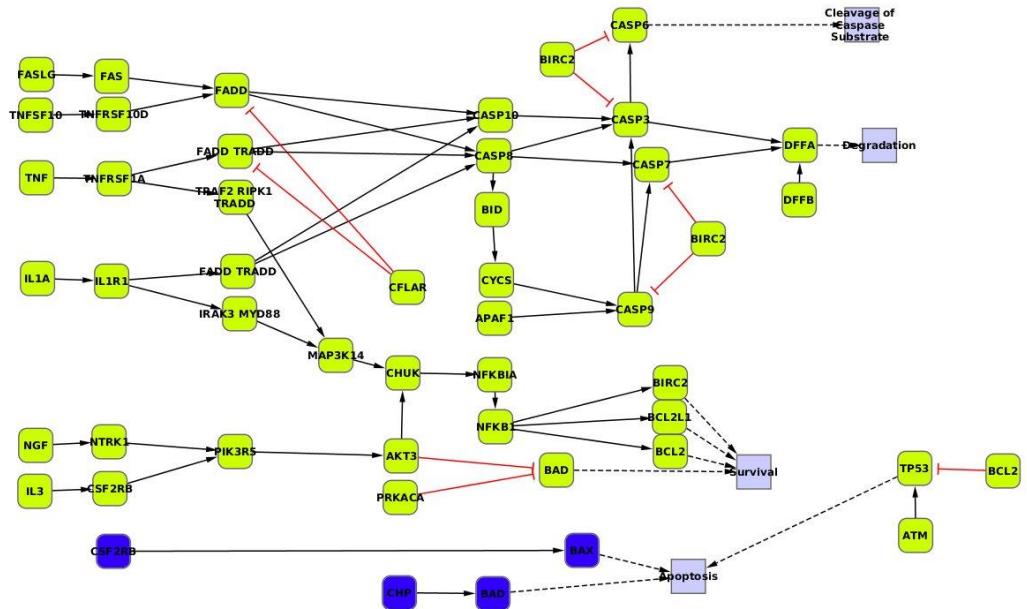
What would you predict about the consequences of gene activity changes in the apoptosis pathway in a case control experiment of colorectal cancer?

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

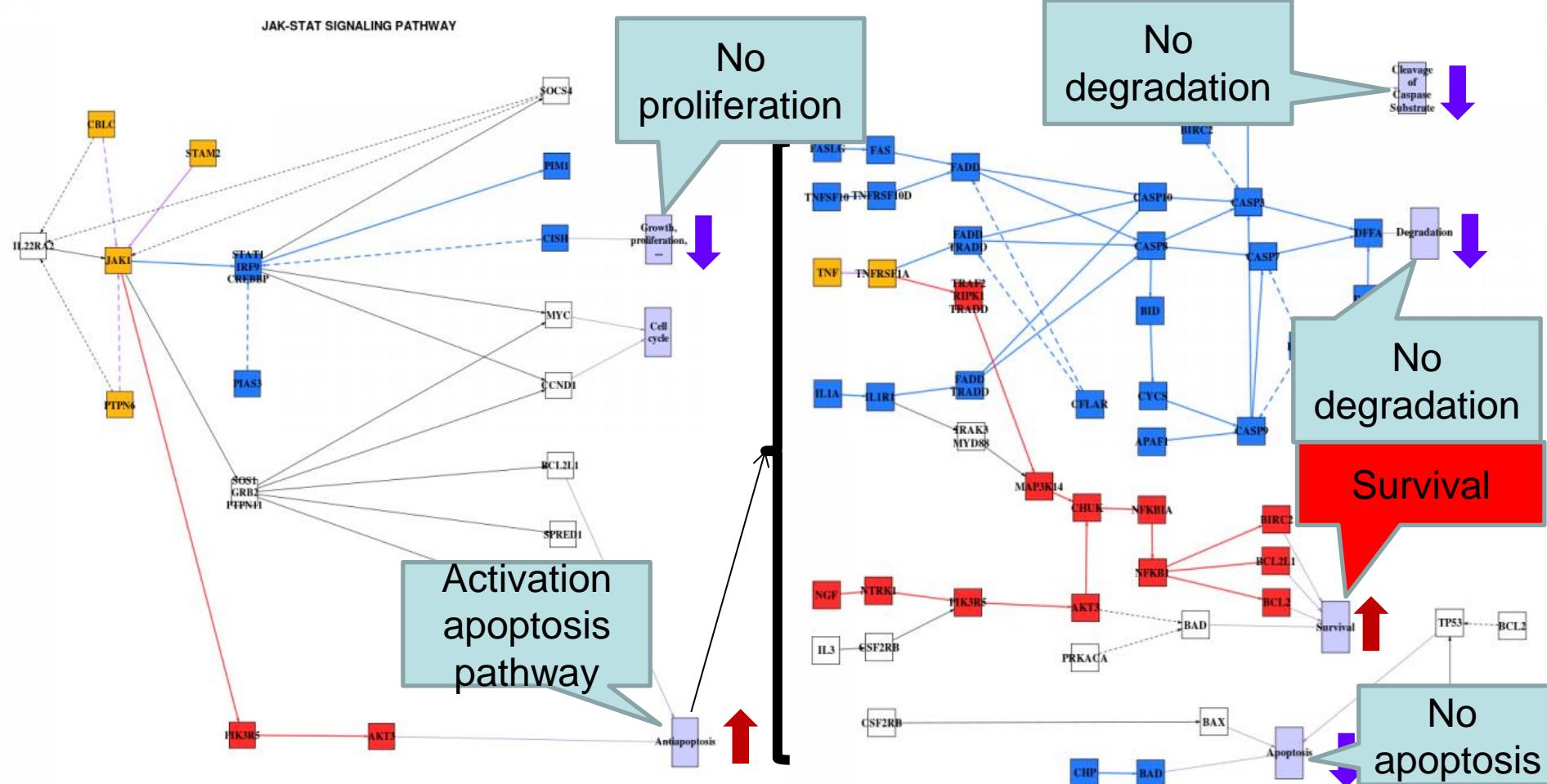


Apoptosis inhibition is not obvious from gene expression

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.

Tools to study the impact of gene deregulation or mutations over signaling pathways

Pathiways: study the impact of differences in gene expression levels on signaling

PathiPred: Builds a predictor based on signaling circuits using gene expression levels as data.

Available at:

<http://pathiways.babelomics.org/>

PathiVar: Study the impact of the variants found in a VCF file over the signaling circuits of the pathways.

Available at:

<http://pathivar.babelomics.org/>

The screenshot shows the PATHiWAYS web application interface. At the top, there are tabs for 'PATHiWAYS' and 'PATHiPRED'. Below the tabs, there are sections for 'Platform' (CEL compressed file or Normalized matrix selected), 'Input file' (Browse... button, No file selected), 'Experimental design' (Experimental design data: Browse... No file selected, Condition 1 and Condition 2 fields), 'Other parameters' (Summ: 90th percentile dropdown), 'Comparison tests' (Test: Multilevel test selected, Wilcoxon test option), and a 'Pathways' section with an 'All' button.

The screenshot shows the PATHiVar web application interface. It has a header with the URL 'pathivar.babelomics.org'. Below the header, it says 'PATHiVar input parameters' and includes a 'Run example' button. There are sections for 'VCF file' (File browser field), 'Sample name from VCF' (dropdown), 'Inheritance pattern' (Recessive dropdown), 'Should compound heterozygotes be included?' (Yes dropdown), and 'Consequence types' (checkboxes for splice acceptor variant, splice donor variant, stop gained, stop lost, and non synonymous codon, with 'Select all' and 'Deselect all' buttons). The background of the interface is dark blue.

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

Future prospects

Hospital Universitario La Paz

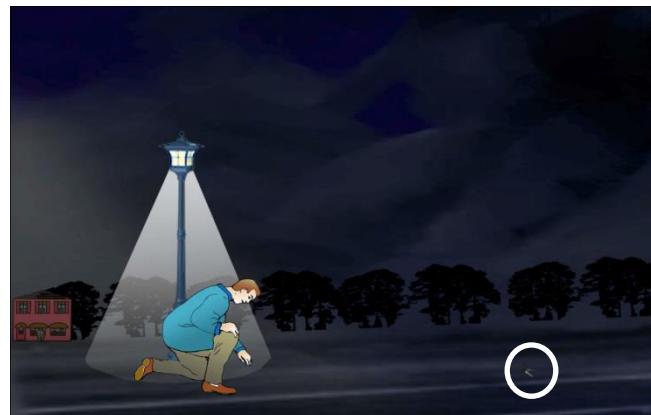
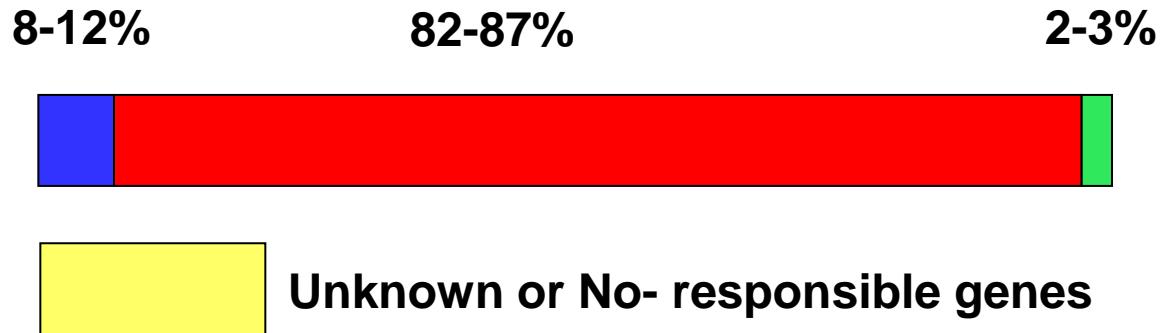
Known causes of Human Genetic Diseases



Pablo
Lapunzina,
Personal
communication

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

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



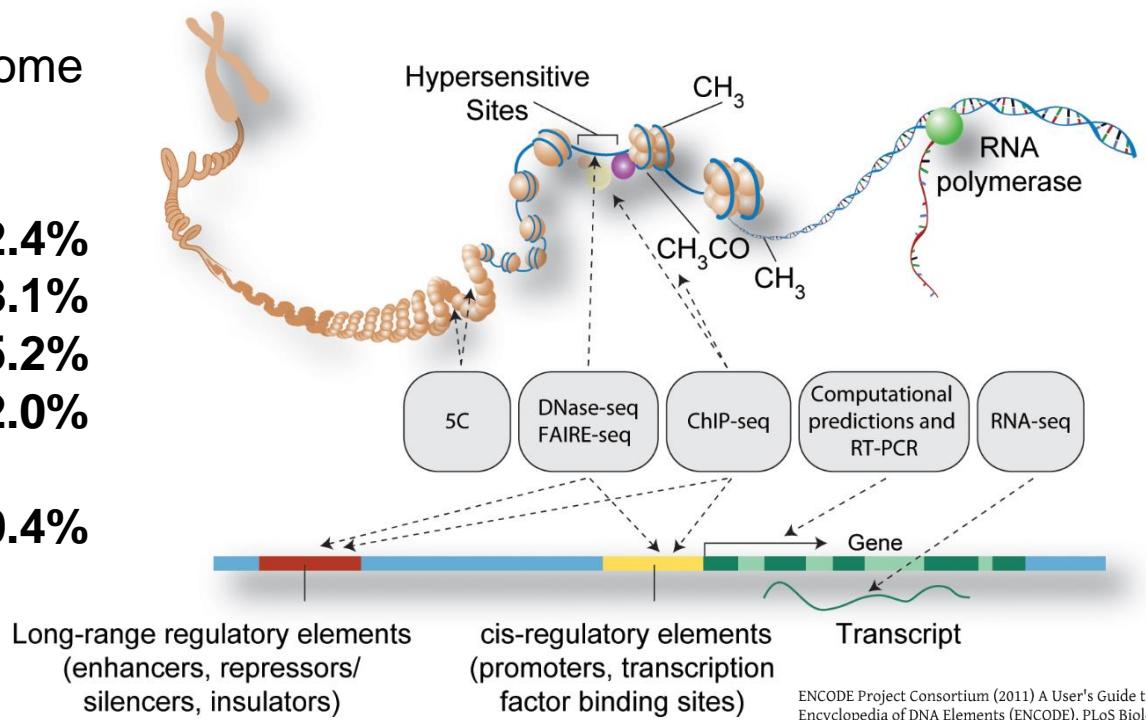
Fact: exons represent a comparatively small part of the complete genome
Other fact: there is still a lot of missing heritability

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: e1-24

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

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.

