Variant association and prioritization

Edinburgh Genomics

Edinburgh, UK

23rd October 2015

Marta Bleda Latorre

mb2033@cam.ac.uk

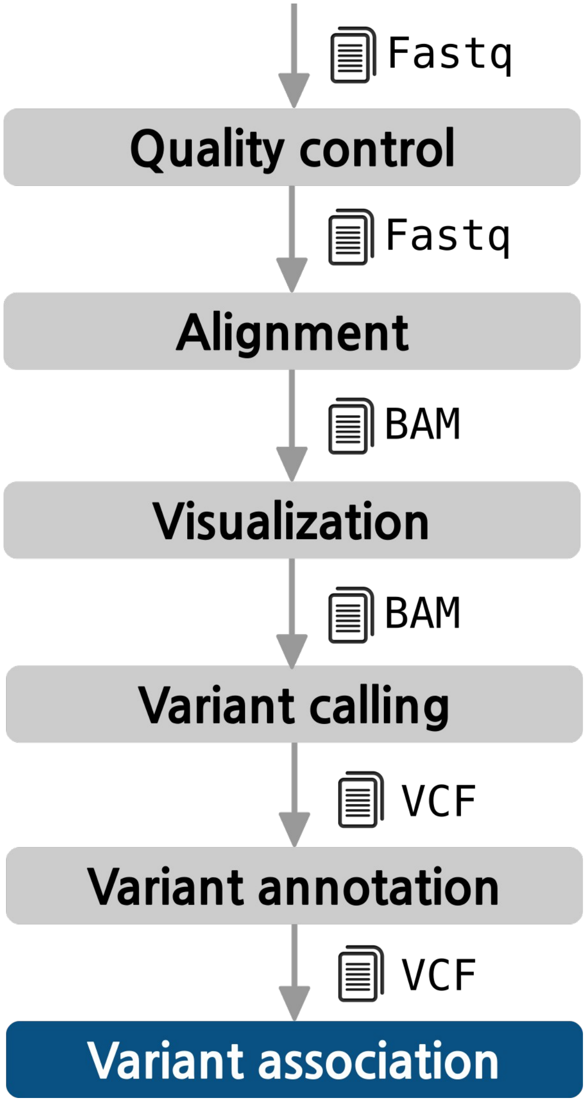
Research Assistant at the Department of Medicine

University of Cambridge

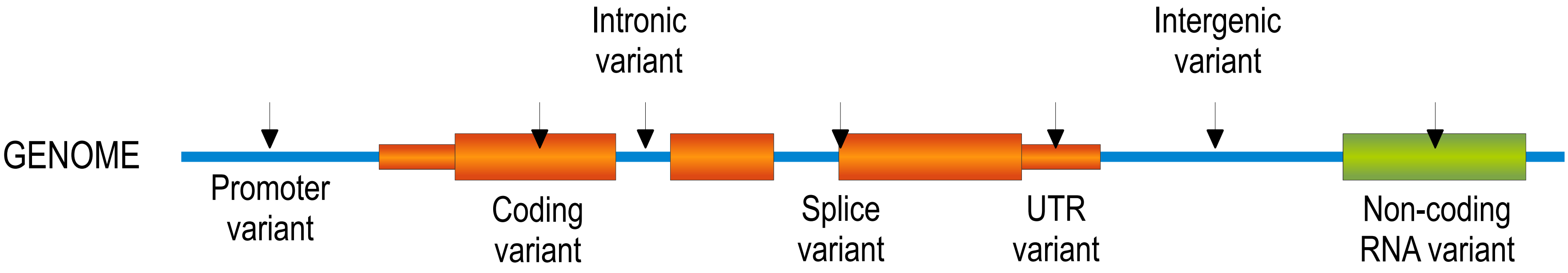
Cambridge, UK

UNIVERSITY OF CAMBRIDGE

The pipeline



The challenge



Still a challenge

- Each individual **exome** carries between 25,000 and 50,000 variants
- A **whole genome** can carry 3.5 million variants on average
- After annotating there will be **hundreds** of **deleterious** variants

CAUTION!

On average, each *normal* person is found to carry:

~11,000 synonymous variants

~11,000 non-synonymous variants

250 to 300 los-of-function variants in annotated genes

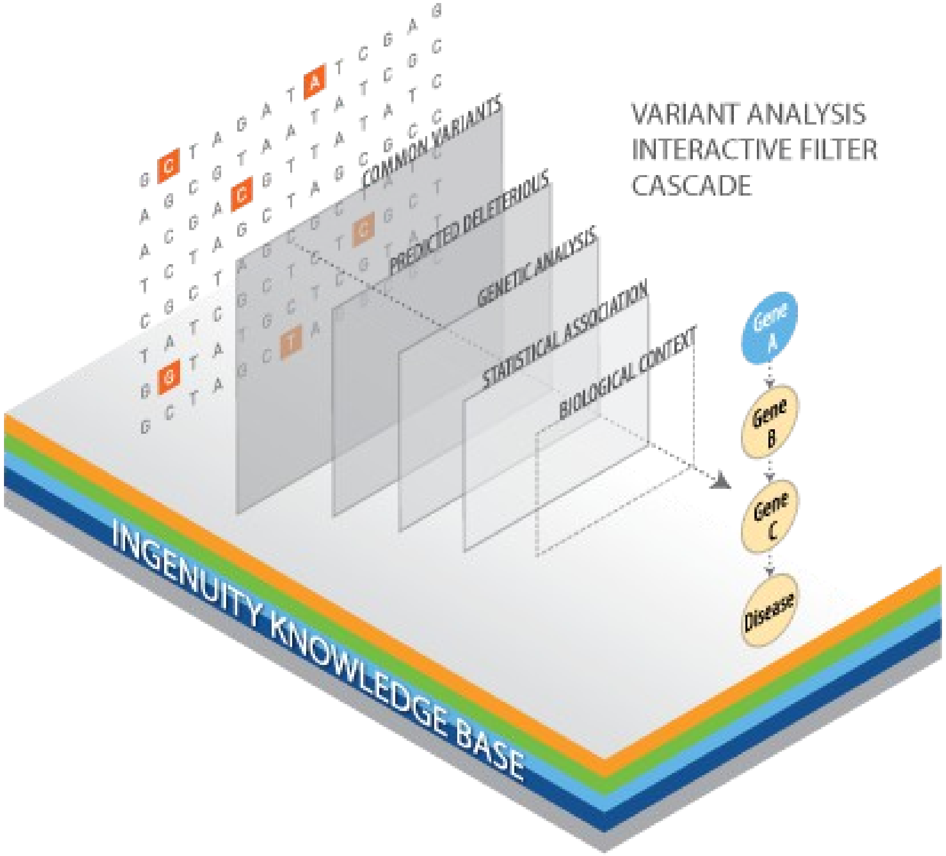
50 to 100 variants previously implicated in **inherited disorders**

1000 Genomes Project Consortium. *A map of human genome variation from population-scale sequencing.* **Nature**. 2010 Oct 28;467(7319):1061-73. PubMed PMID: 20981092

The challenge

The challenge

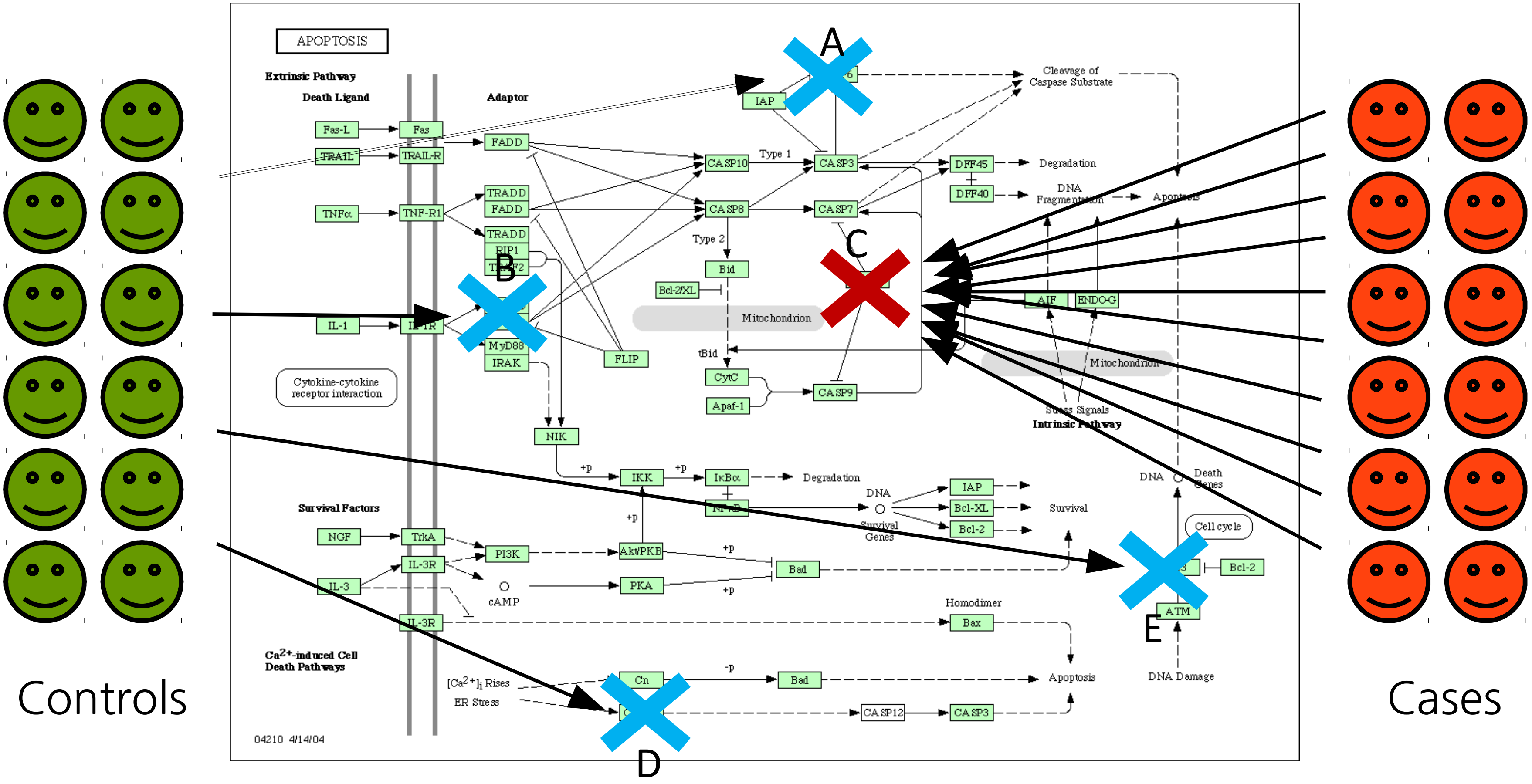
The objective



And now what?

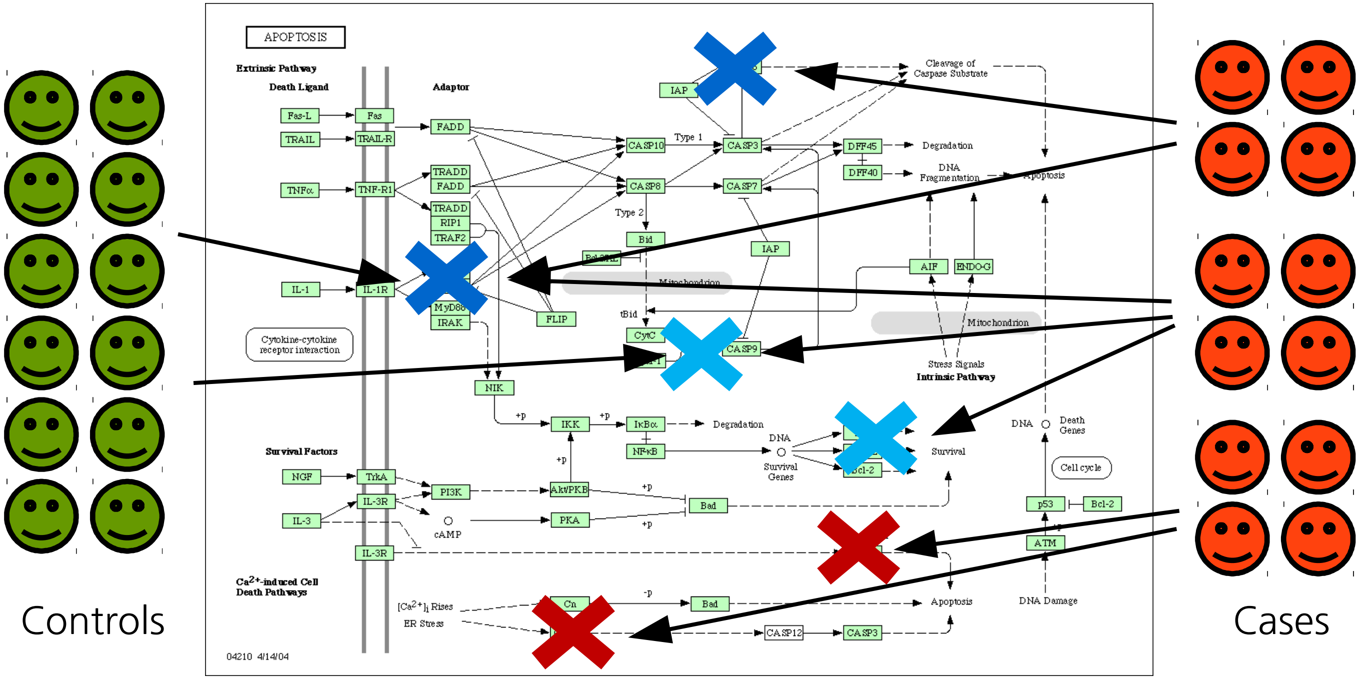
Finding the mutations causative of diseases

The **simplest case**: monogenic disease due to a single gene



And now what?

Finding the mutations causative of diseases



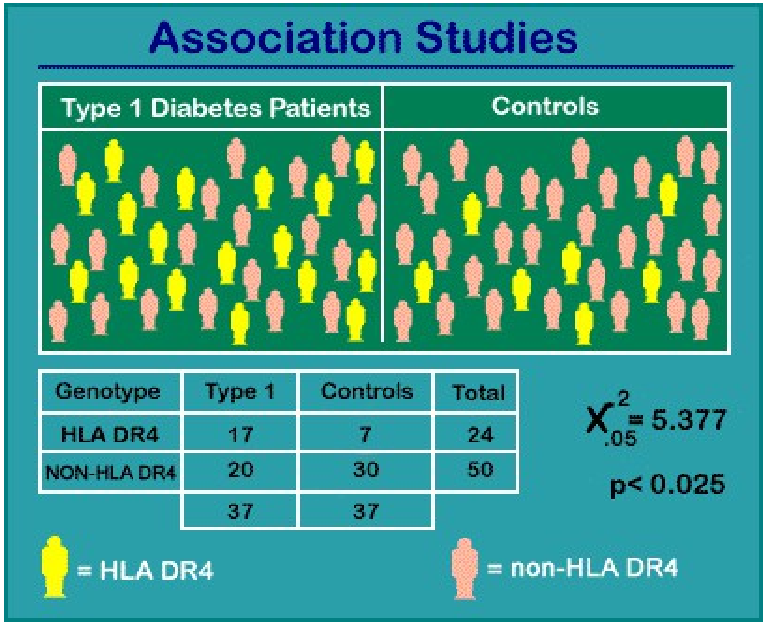
Clear individual **gene associations are difficult to find** in some diseases

Same phenotype can be due to **different mutations and different genes** (or combinations)

Many cases have to be used to obtain significant associations to many markers

The only common element is the **pathway** (yet unknown) affected

Genome-Wide Association Studies (GWAS)



Association Studies

Type 1 Diabetes Patients	Controls

Genotype	Type 1	Controls	Total
HLA DR4	17	7	24
NON-HLA DR4	20	30	50
	37	37	

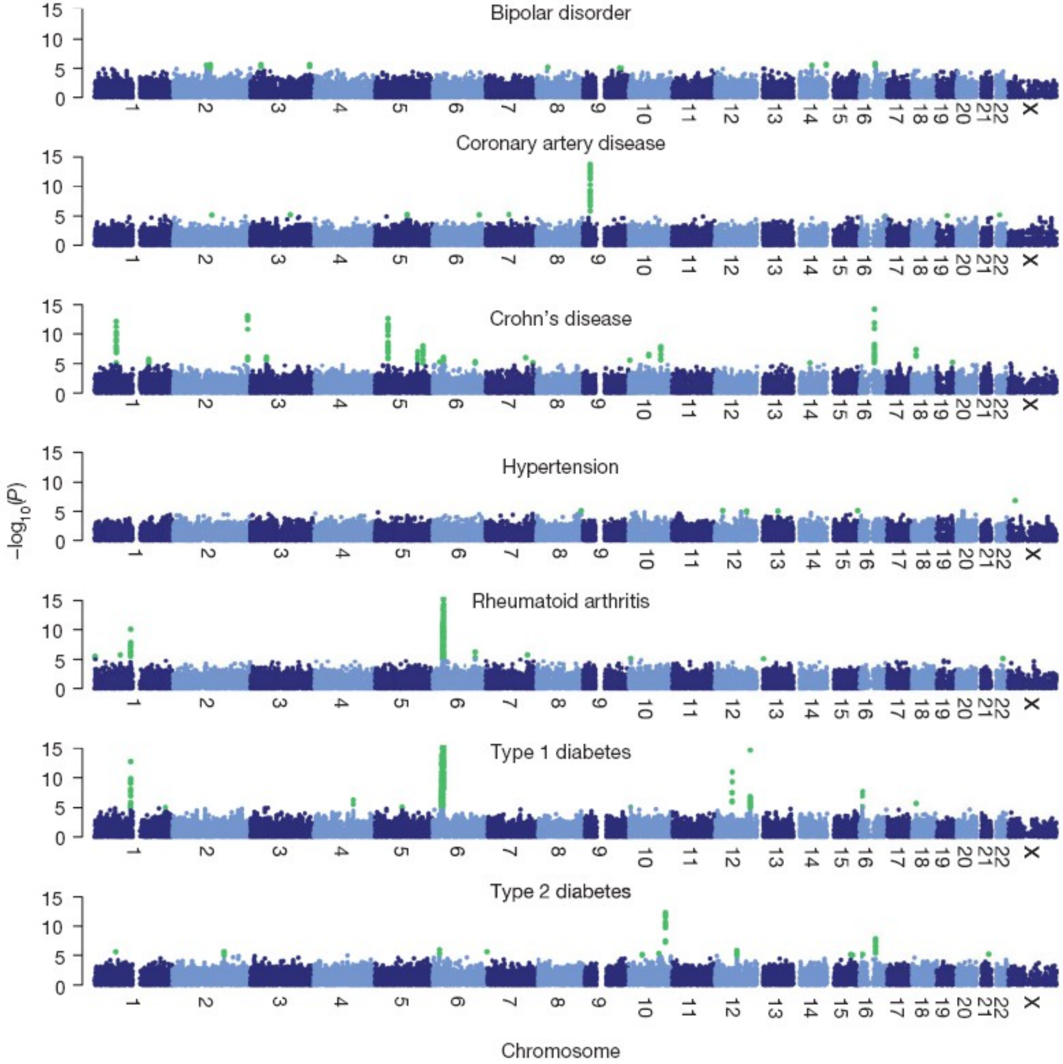
$X^2_{.05} = 5.377$

$p < 0.025$

= HLA DR4

= non-HLA DR4

Odds Ratio: 3.6
95% CI = 1.3 to 10.4



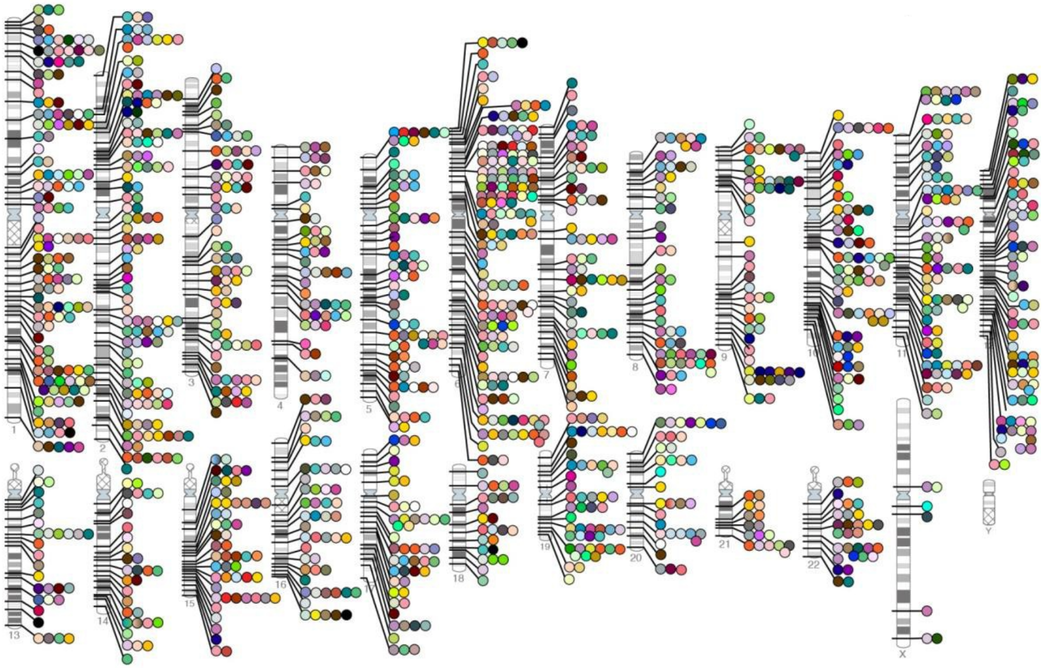
The Wellcome Trust Case Control Consortium

Genome-Wide Association Studies (GWAS)

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 **15,396** SNPs significantly associated to more than **1,500** traits.



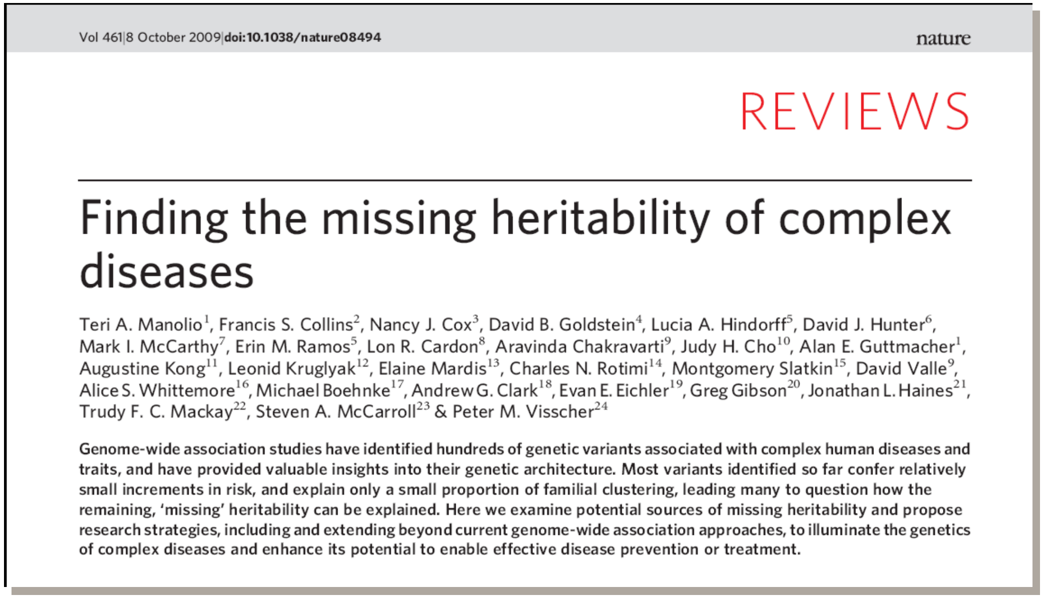
NHGRI GWA Catalog
www.genome.gov/GWAStudies

The missing heritability problem

NEWS FEATURE PERSONAL GENOMES NATURE|Vol 456|6 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. 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 this problem?

Rare Variants, rare CNVs, epigenetics or epistatic effects?

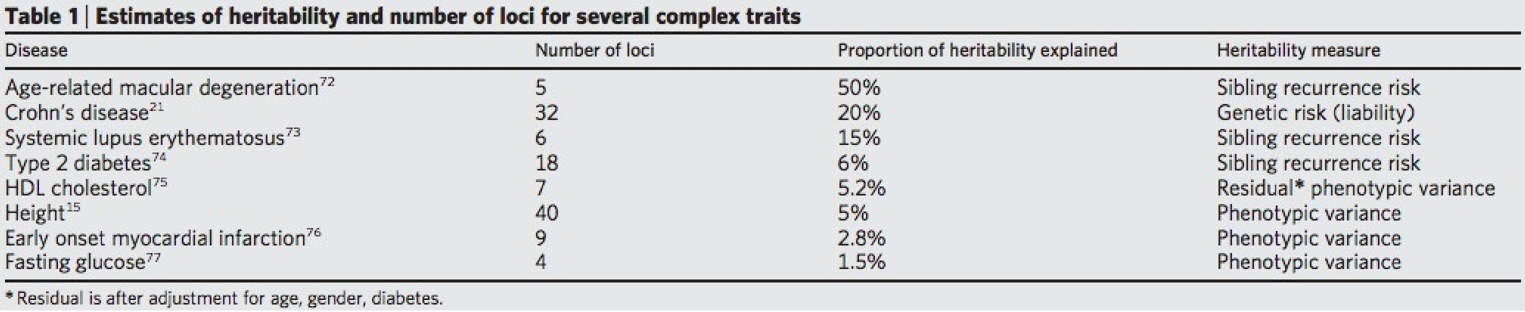
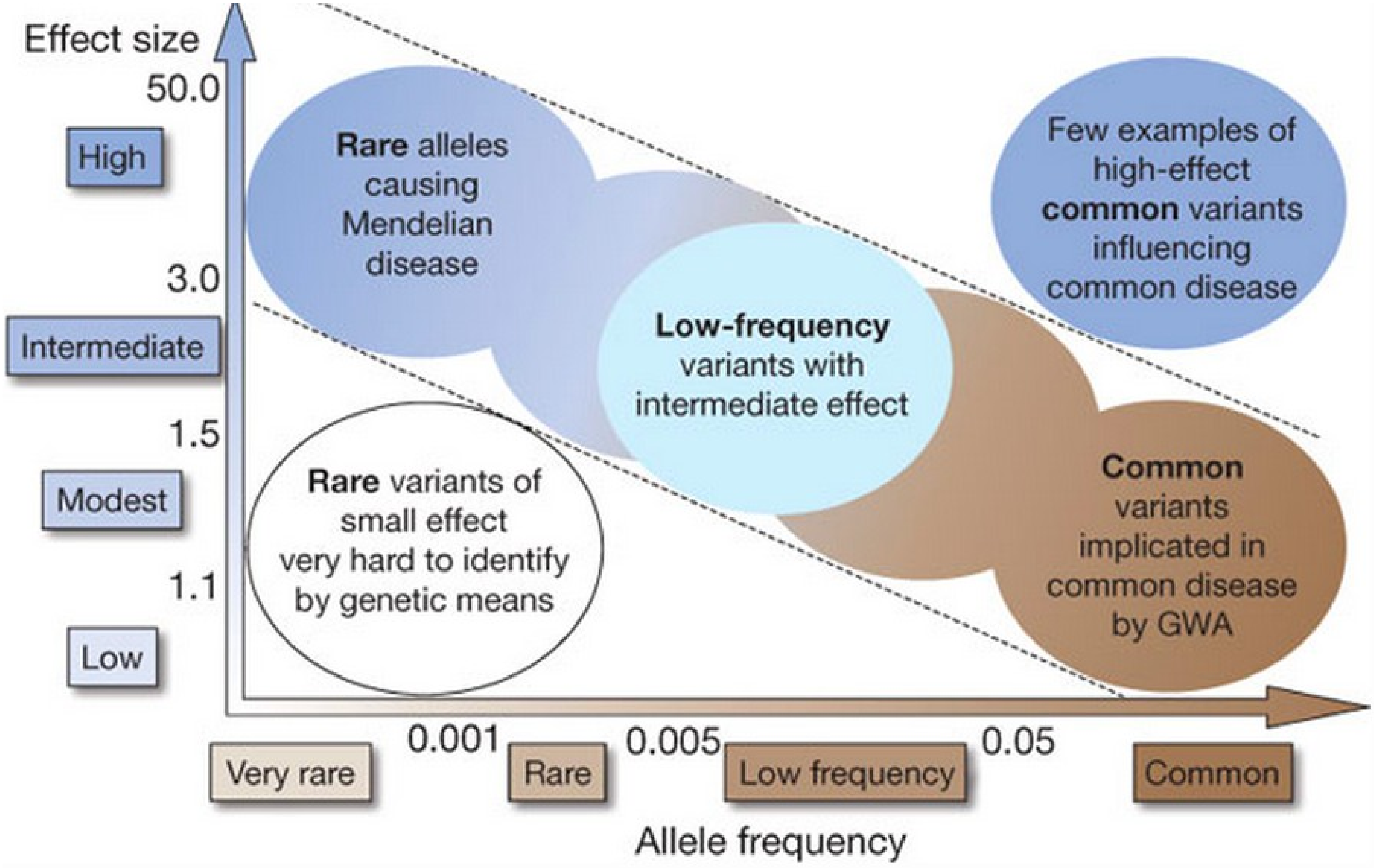


Table 1 | Estimates of heritability and number of loci for several complex traits

Disease	Number of loci	Proportion of heritability explained	Heritability measure
Age-related macular degeneration[72]	5	50%	Sibling recurrence risk
Crohn's disease[21]	32	20%	Genetic risk (liability)
Systemic lupus erythematosus[73]	6	15%	Sibling recurrence risk
Type 2 diabetes[74]	18	6%	Sibling recurrence risk
HDL cholesterol[75]	7	5.2%	Residual* phenotypic variance
Height[15]	40	5%	Phenotypic variance
Early onset myocardial infarction[76]	9	2.8%	Phenotypic variance
Fasting glucose[77]	4	1.5%	Phenotypic variance

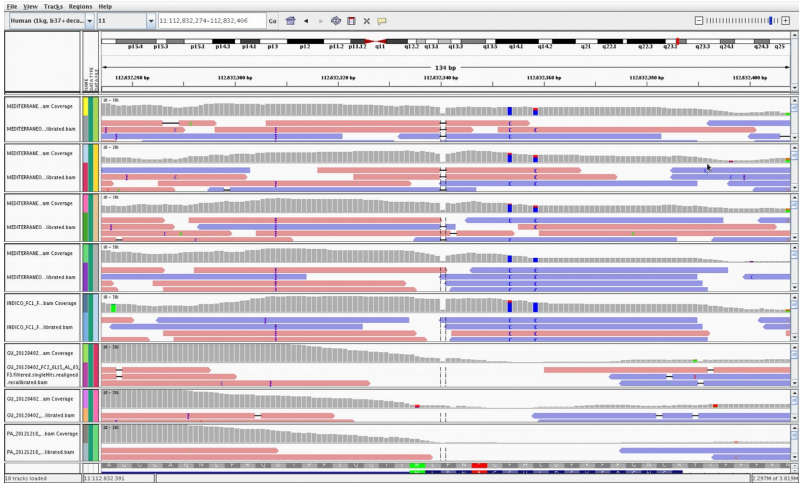
*Residual is after adjustment for age, gender, diabetes.

Distribution of genetic variation


Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele frequency

Teri A. Manolio, et al. **Finding the missing heritability of complex diseases.** Nature 461, 747-753(8 October 2009)

Strategies

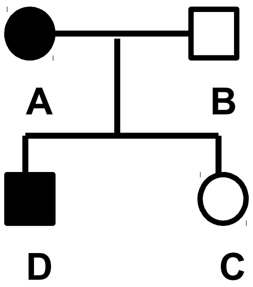


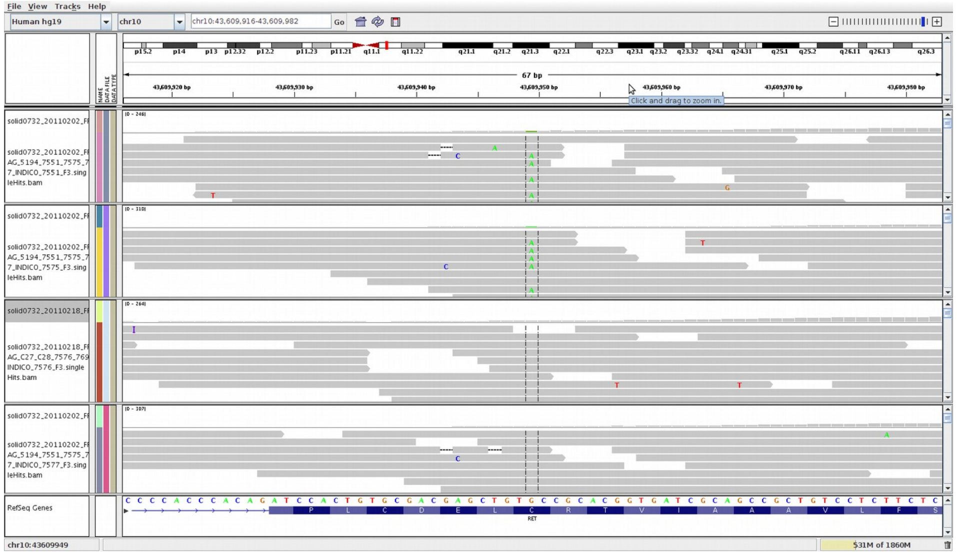
- Case - control
- Filtering using **family information**
- Rare variant **association**
  - Single variant tests
  - Gene or region-based aggregation tests
- **Network** (Systems biology) approaches
  - PPIs
  - Gene regulatory elements (miRNAs, Tfs)
  - GO terms

Using family information

- Families containing **control and disease** individuals can help us to **reduce** the number of variants obtained
- Individuals from the same family → **less variability**
- **Filter** variants present in healthy relatives

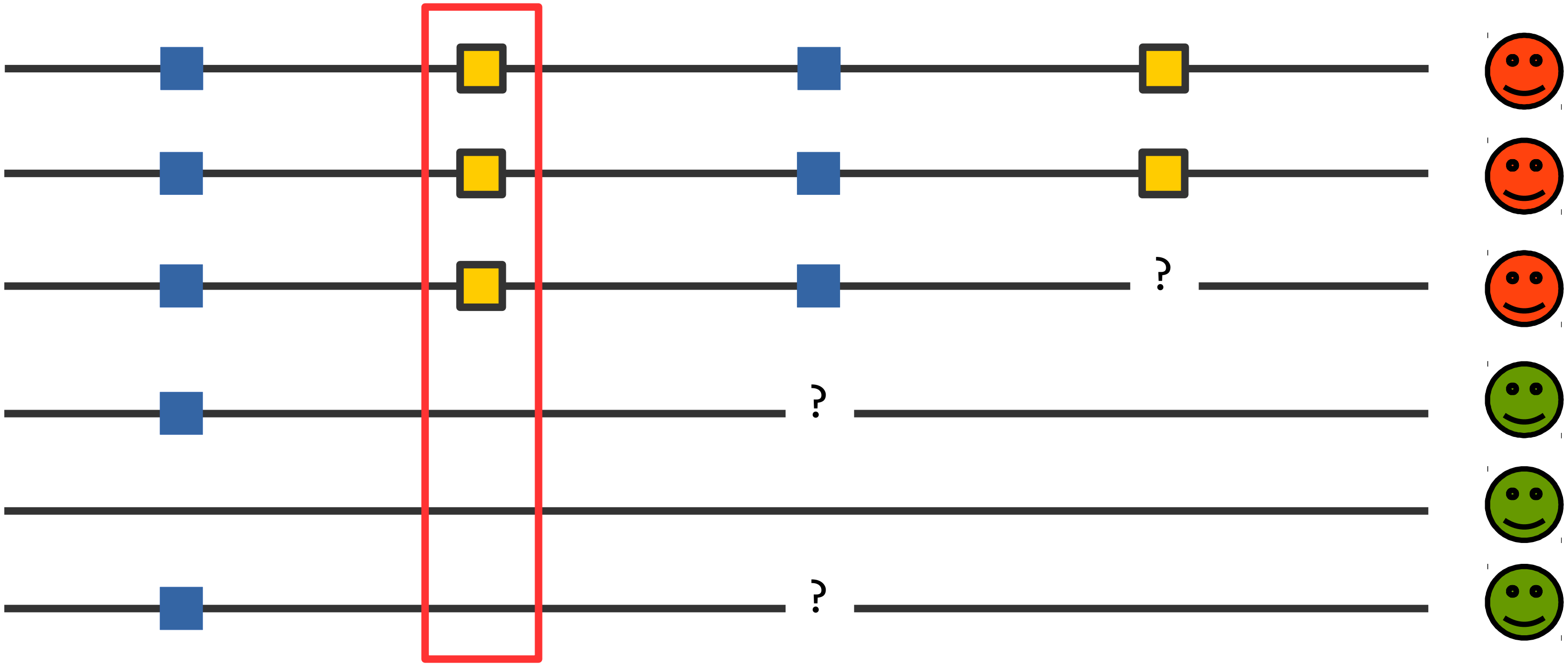
Segregation within a pedigree




File View Tracks Help
Human hg19
chr10
chr10:43,609,916-43,609,982
Go
p15.2 p14 p13 p12.32 p12.2 p11.23 p11.21 q11.1 q11.22 q21.1 q21.2
67 bp
43,609,920 bp
43,609,930 bp
43,609,940 bp
43,609,950 bp
43,609,960 bp
43,609,970 bp
43,609,980 bp
Click and drag to zoom in.
NAME
DATA FILE
DATA TYPE
solid0732_20110202_FR
solid0732_20110202_FRAG_5194_7551_7575_77_INDICO_7551_F3.singleHits.bam
[0 - 246]
solid0732_20110202_FR
solid0732_20110202_FRAG_5194_7551_7575_77_INDICO_7575_F3.singleHits.bam
[0 - 310]
solid0732_20110218_FR
solid0732_20110218_FRAG_C27_C28_7576_769INDICO_7576_F3.singleHits.bam
[0 - 264]
solid0732_20110202_FR
solid0732_20110202_FRAG_5194_7551_7575_77_INDICO_7577_F3.singleHits.bam
[0 - 307]
RefSeq Genes
RET
chr10:43609949
531M of 1860M

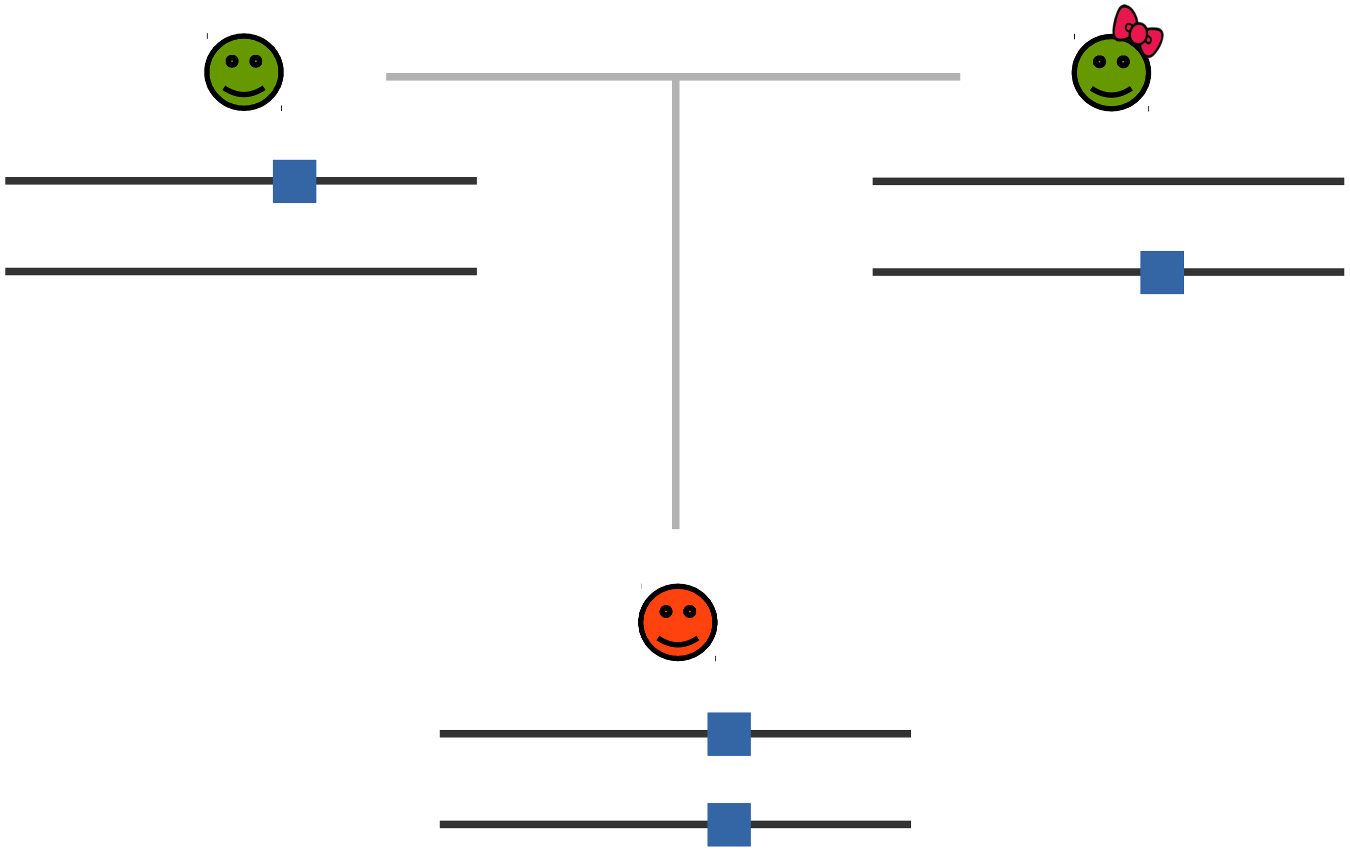
Using family information

Dominant inheritance



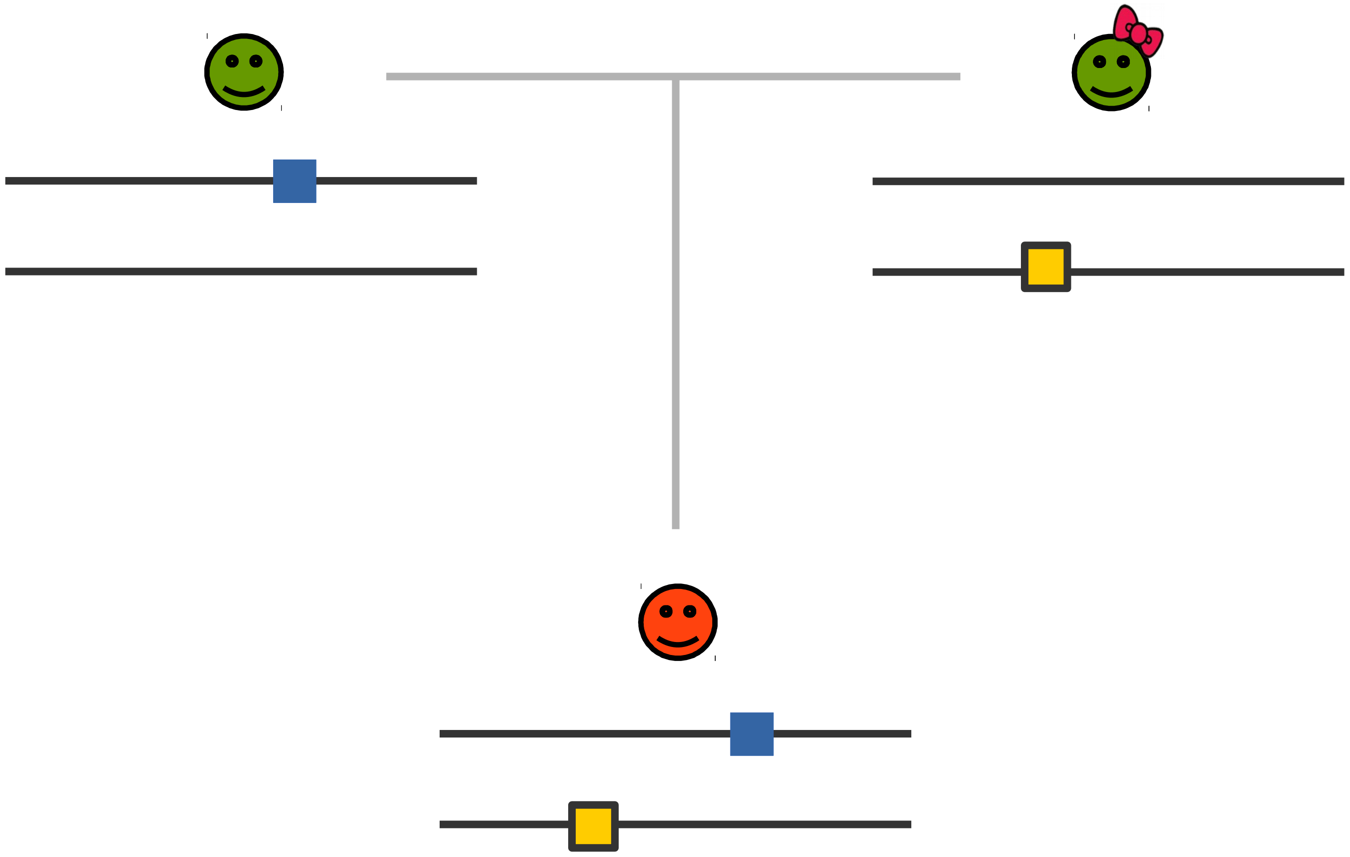
Using family information

Recessive homozygous



Using family information

Recessive - Compound heterozygosity



Rare variant association

- Genome-wide association studies (**GWAS**) have been widely used with microarray data to evaluate **common genetic variants** (MAF $> 5\%$)

- Despite many discoveries, much of the genetic contribution is still unexplained → **missing heritability**

- **Low frequency** ($0.5\% \leq$ MAF $\leq 5\%$) and **rare variants** ($< 0.5\%$) could explain additional disease risk

Methods for rare variant association

- **Single variant tests**

  Evaluate each variant for association with a trait individually

  Less powerful for rare variants than for common variants with same sample size

- **Gene or region-based aggregation tests of multiple variants**

  Evaluate cumulative effects of multiple genetic variants in a gene or region, increasing power when multiple variants in the group are associated with a given disease or trait.

Methods for rare variant association

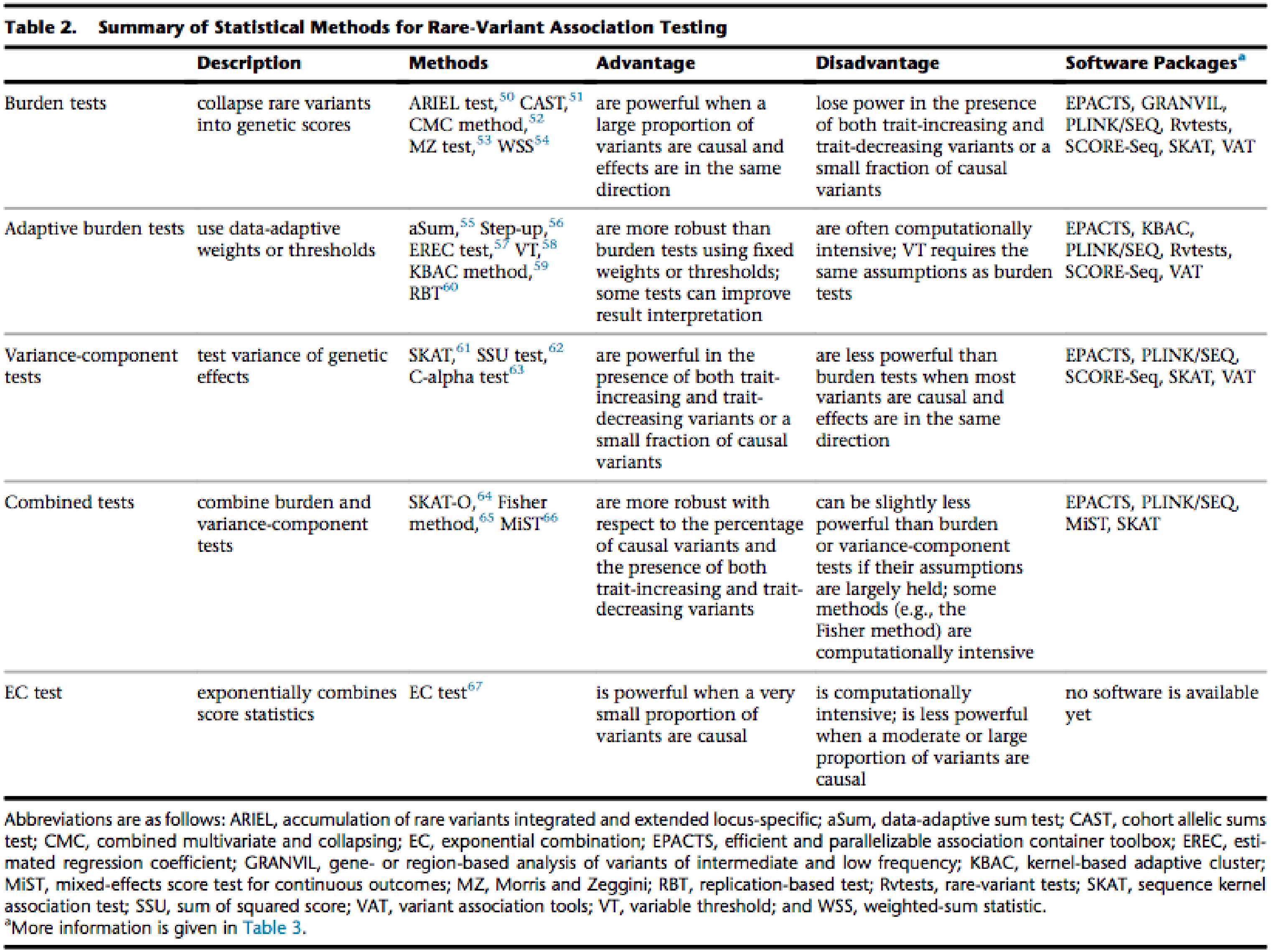


Table 2. Summary of Statistical Methods for Rare-Variant Association Testing

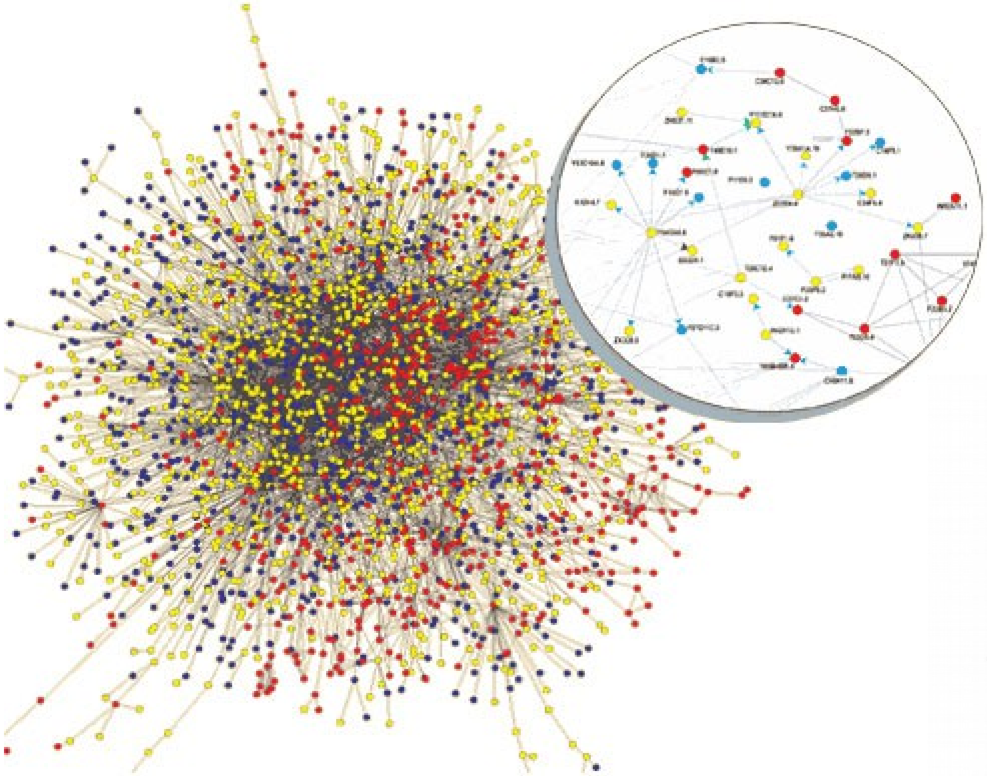
	Description	Methods	Advantage	Disadvantage	Software Packages[a]
Burden tests	collapse rare variants into genetic scores	ARIEL test,[50] CAST,[51] CMC method,[52] MZ test,[53] WSS[54]	are powerful when a large proportion of variants are causal and effects are in the same direction	lose power in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	EPACTS, GRANVIL, PLINK/SEQ, Rvtests, SCORE-Seq, SKAT, VAT
Adaptive burden tests	use data-adaptive weights or thresholds	aSum,[55] Step-up,[56] EREC test,[57] VT,[58] KBAC method,[59] RBT[60]	are more robust than burden tests using fixed weights or thresholds; some tests can improve result interpretation	are often computationally intensive; VT requires the same assumptions as burden tests	EPACTS, KBAC, PLINK/SEQ, Rvtests, SCORE-Seq, VAT
Variance-component tests	test variance of genetic effects	SKAT,[61] SSU test,[62] C-alpha test[63]	are powerful in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	are less powerful than burden tests when most variants are causal and effects are in the same direction	EPACTS, PLINK/SEQ, SCORE-Seq, SKAT, VAT
Combined tests	combine burden and variance-component tests	SKAT-O,[64] Fisher method,[65] MiST[66]	are more robust with respect to the percentage of causal variants and the presence of both trait-increasing and trait-decreasing variants	can be slightly less powerful than burden or variance-component tests if their assumptions are largely held; some methods (e.g., the Fisher method) are computationally intensive	EPACTS, PLINK/SEQ, MiST, SKAT
EC test	exponentially combines score statistics	EC test[67]	is powerful when a very small proportion of variants are causal	is computationally intensive; is less powerful when a moderate or large proportion of variants are causal	no software is available yet

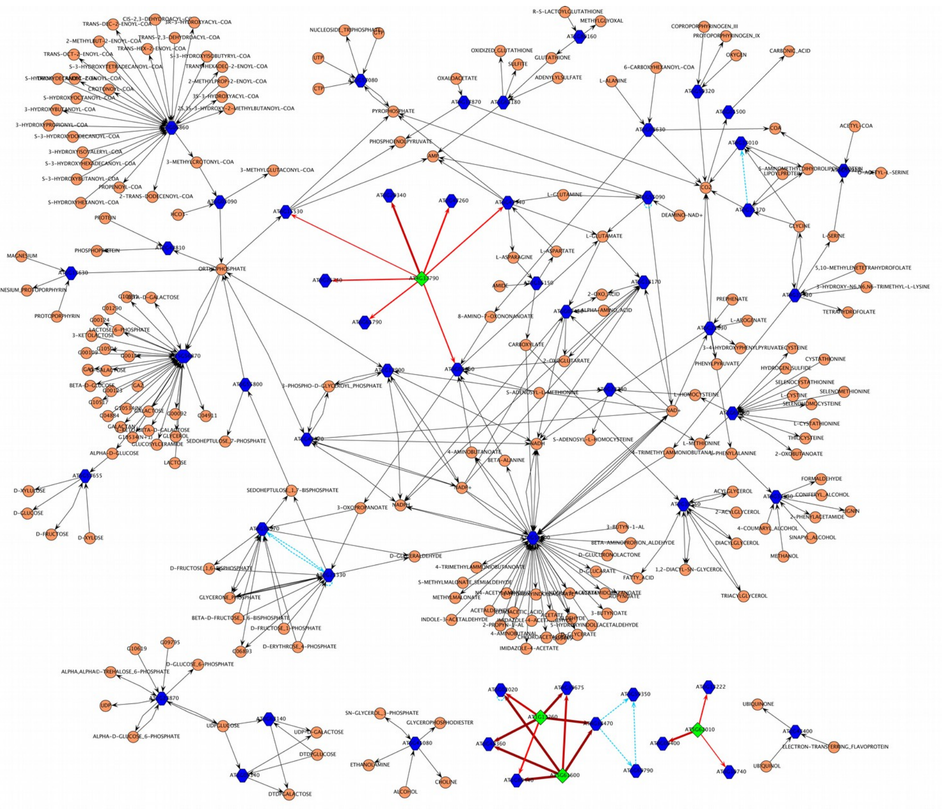
Abbreviations are as follows: ARIEL, accumulation of rare variants integrated and extended locus-specific; aSum, data-adaptive sum test; CAST, cohort allelic sums test; CMC, combined multivariate and collapsing; EC, exponential combination; EPACTS, efficient and parallelizable association container toolbox; EREC, estimated regression coefficient; GRANVIL, gene- or region-based analysis of variants of intermediate and low frequency; KBAC, kernel-based adaptive cluster; MiST, mixed-effects score test for continuous outcomes; MZ, Morris and Zeggini; RBT, replication-based test; Rvtests, rare-variant tests; SKAT, sequence kernel association test; SSU, sum of squared score; VAT, variant association tools; VT, variable threshold; and WSS, weighted-sum statistic.

[a] More information is given in Table 3.

Lee, Seunggeung, et al. "**Rare-variant association analysis: study designs and statistical tests.**" *The American Journal of Human Genetics* 95.1 (2014): 5-23.

Using network information



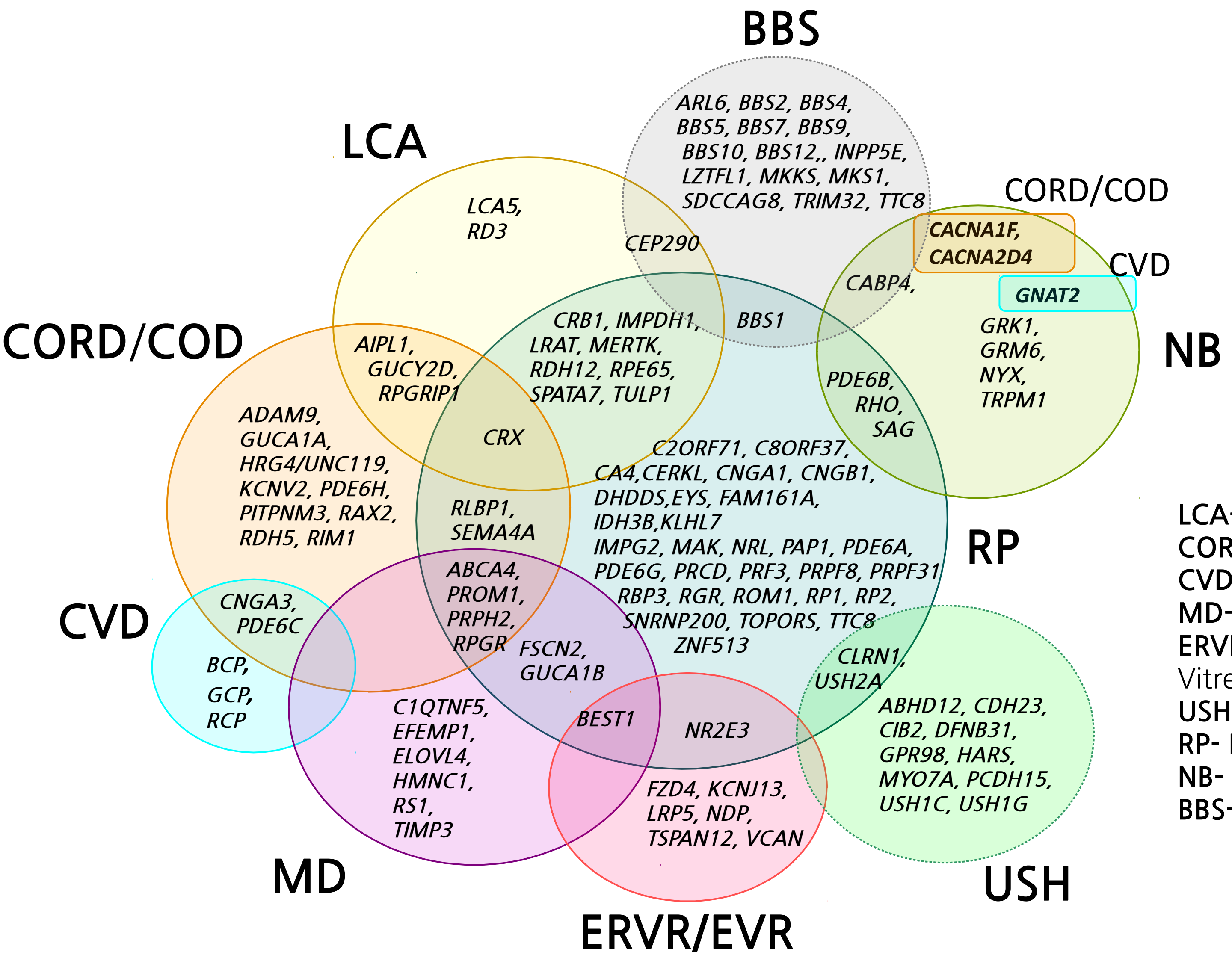


Example with Inherited Retinal Dystrophies (IRD)

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

Is genetic overlapping among IRDs related to protein interaction?

Example with Inherited Retinal Dystrophies (IRD)

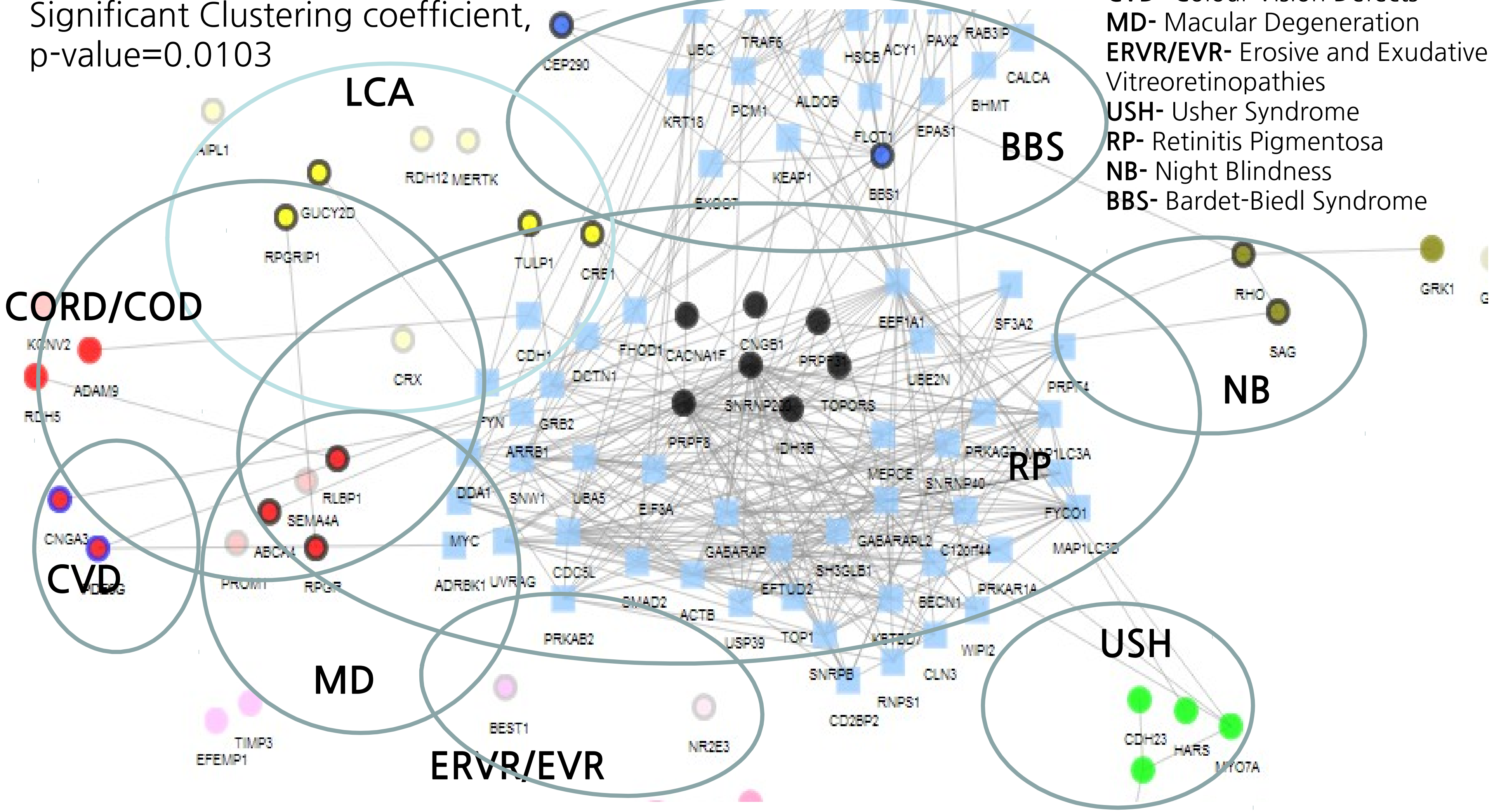


LCA-Leber Congenital Amaurosis
CORD/COD- Cone and cone-rod dystro.
CVD- Colour Vision Defects
MD- Macular Degeneration
ERVR/EVR- Erosive and Exudative Vitreoretinopathies
USH- Usher Syndrome
RP- Retinitis Pigmentosa
NB- Night Blindness
BBS- Bardet-Biedl Syndrome

Example with Inherited Retinal Dystrophies (IRD)

Significant Clustering coefficient,
p-value=0.0103

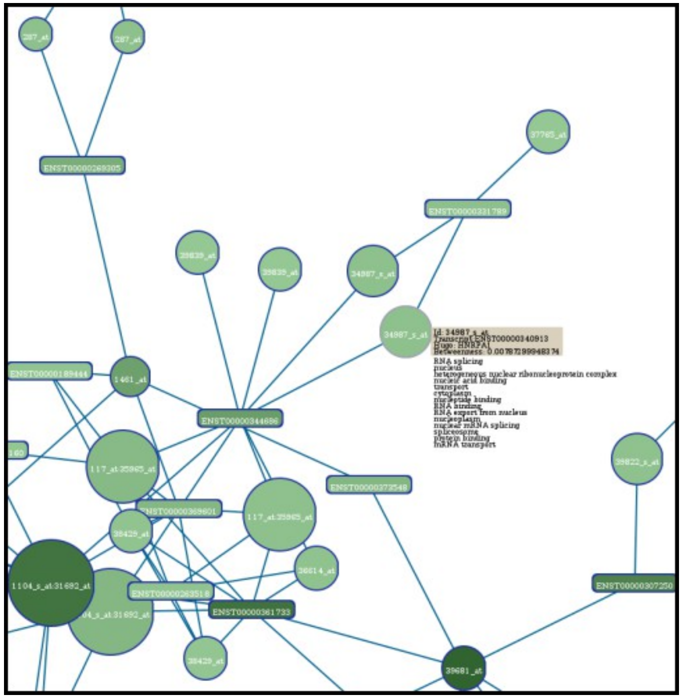
LCA-Leber Congenital Amaurosis
CORD/COD- Cone and cone-rod dystro
CVD- Colour Vision Defects
MD- Macular Degeneration
ERVR/EVR- Erosive and Exudative Vitreoretinopathies
USH- Usher Syndrome
RP- Retinitis Pigmentosa
NB- Night Blindness
BBS- Bardet-Biedl Syndrome



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

SNOW

- The SNOW tool introduces **protein-protein interaction data** into the functional profiling of genomic data
  - Evaluates **role of the list within the interactome**: identifies hubs in the list of proteins/genes (nodes) and evaluates the topological parameters of the within the interactome
  - Evaluates the list's cooperative behavior as a **functional module**



http://babelomics.bioinfo.cipf.es/functional.html

NetworkMiner

Prioritizing disease candidate genes

Scenario

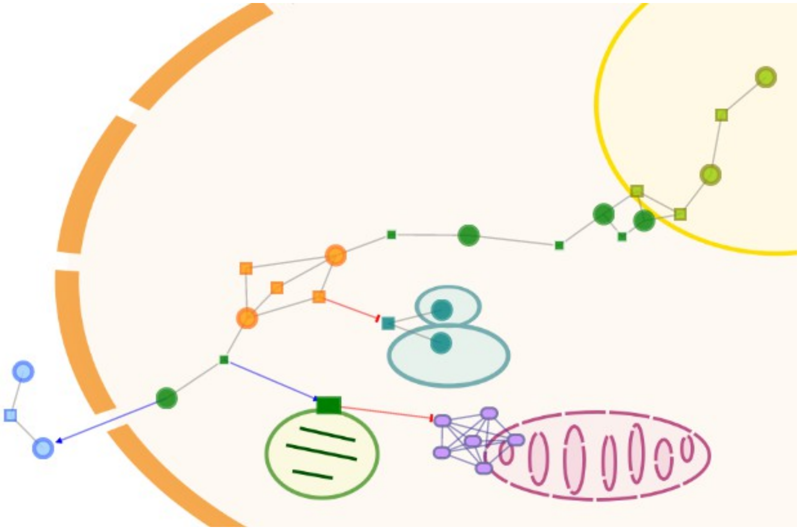
http://babelomics.bioinfo.cipf.es/functional.html

You have:

1. a list of **disease candidates** (ranked by their population frequency)
2. a list of **genes** that are known to be **associated to the disease**

You want to see:

which of your candidates are functionally related or interacting with the known disease genes



NetworkMiner Study

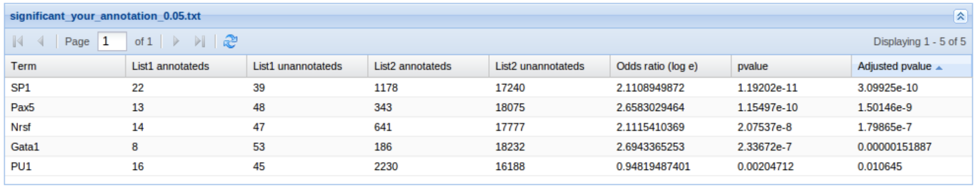
Tests whether any of the candidates is significantly located in the neighborhood of the known disease genes

RENATO (REgulatory Network Analsis TOol)

Identifying common regulatory elements

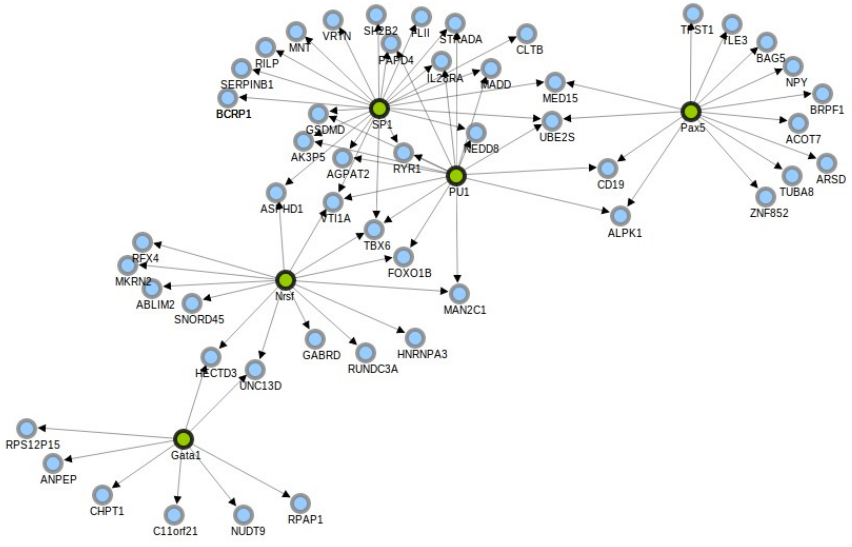
- Sometimes, the problem is not in the gene but in its regulators
- Tool for the **interpretation and visualization** of transcriptional (TFs) and post-transcriptional (miRNAs) **regulatory information**
- Designed to identify **common regulatory elements** in a list of genes
- RENATO maps these genes to the regulatory network, extracts the corresponding regulatory connections and evaluate each regulator for **significant over-representation** in the list.

http://renato.bioinfo.cipf.es



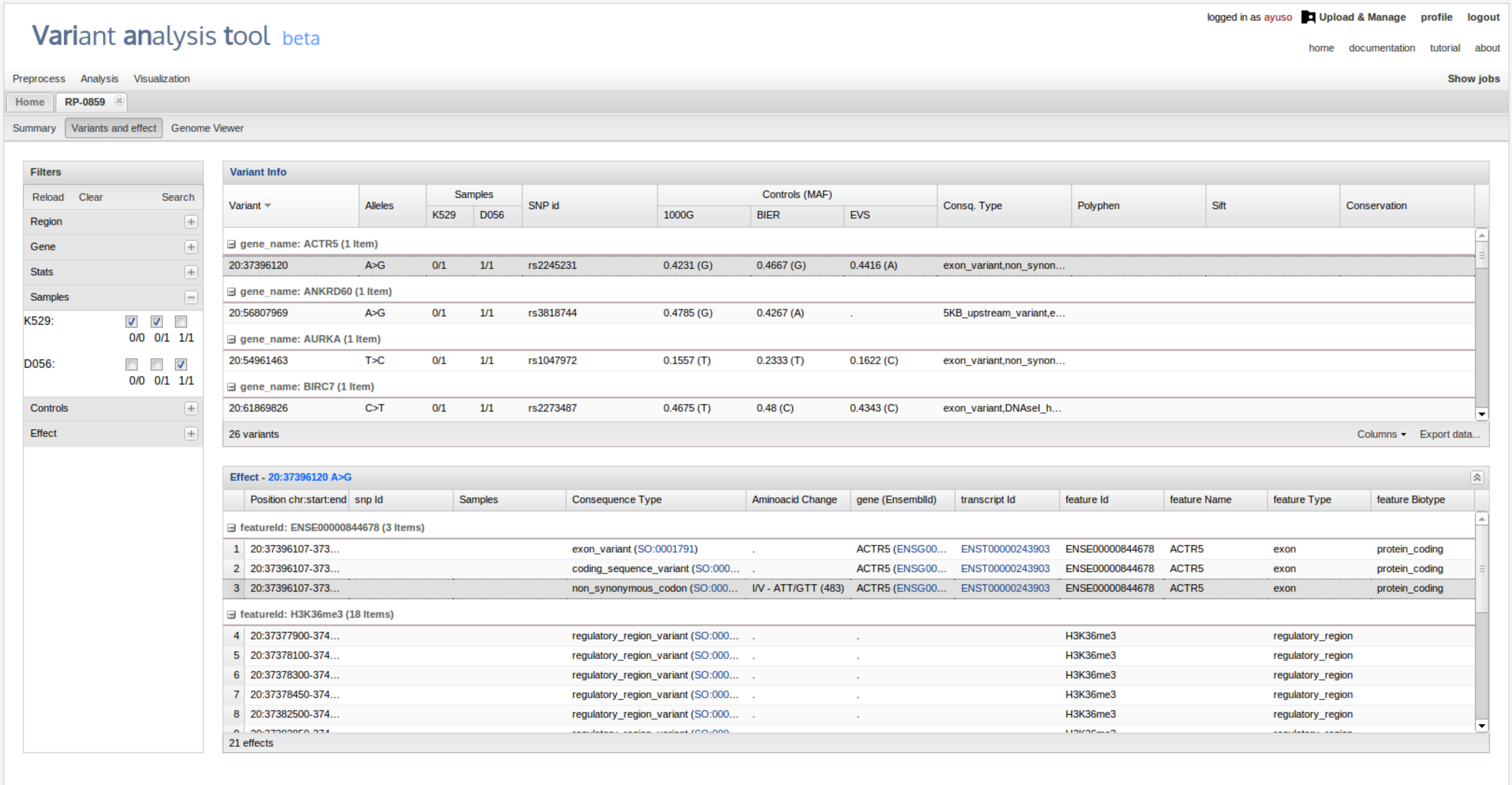
significant_your_annotation_0.05.txt

Term	List1 annotateds	List1 unannotateds	List2 annotateds	List2 unannotateds	Odds ratio (log e)	pvalue	Adjusted pvalue ▲
SP1	22	39	1178	17240	2.1108949872	1.19202e-11	3.09925e-10
Pax5	13	48	343	18075	2.6583029464	1.15497e-10	1.50146e-9
Nrsf	14	47	641	17777	2.1115410369	2.07537e-8	1.79865e-7
Gata1	8	53	186	18232	2.6943365253	2.33672e-7	0.00000151887
PU1	16	45	2230	16188	0.94819487401	0.00204712	0.010645



BierApp

Bierapp.babelomics.org



Variant Info

Variant	Alleles	Samples		SNP id	Controls (MAF)			Consq. Type	Polyphen	Sift	Conservation
		K529	D056		1000G	BIER	EVS				
gene_name: ACTR5 (1 Item)											
20:37396120	A>G	0/1	1/1	rs2245231	0.4231 (G)	0.4667 (G)	0.4416 (A)	exon_variant,non_synon...			
gene_name: ANKRD60 (1 Item)											
20:56807969	A>G	0/1	1/1	rs3818744	0.4785 (G)	0.4267 (A)	.	5KB_upstream_variant,e...			
gene_name: AURKA (1 Item)											
20:54961463	T>C	0/1	1/1	rs1047972	0.1557 (T)	0.2333 (T)	0.1622 (C)	exon_variant,non_synon...			
gene_name: BIRC7 (1 Item)											
20:61869826	C>T	0/1	1/1	rs2273487	0.4675 (T)	0.48 (C)	0.4343 (C)	exon_variant,DNAsel_h...			

26 variants

Effect - 20:37396120 A>G

	Position chr:start:end	snp Id	Samples	Consequence Type	Aminoacid Change	gene (EnsemblId)	transcript Id	feature Id	feature Name	feature Type	feature Biotype
featureId: ENSE00000844678 (3 Items)											
1	20:37396107-373...			exon_variant (SO:0001791)	.	ACTR5 (ENSG00...	ENST00000243903	ENSE00000844678	ACTR5	exon	protein_coding
2	20:37396107-373...			coding_sequence_variant (SO:000...	.	ACTR5 (ENSG00...	ENST00000243903	ENSE00000844678	ACTR5	exon	protein_coding
3	20:37396107-373...			non_synonymous_codon (SO:000...	I/V - ATT/GTT (483)	ACTR5 (ENSG00...	ENST00000243903	ENSE00000844678	ACTR5	exon	protein_coding
featureId: H3K36me3 (18 Items)											
4	20:37377900-374...			regulatory_region_variant (SO:000...	.	.		H3K36me3		regulatory_region	
5	20:37378100-374...			regulatory_region_variant (SO:000...	.	.		H3K36me3		regulatory_region	
6	20:37378300-374...			regulatory_region_variant (SO:000...	.	.		H3K36me3		regulatory_region	
7	20:37378450-374...			regulatory_region_variant (SO:000...	.	.		H3K36me3		regulatory_region	
8	20:37382500-374...			regulatory_region_variant (SO:000...	.	.		H3K36me3		regulatory_region	

21 effects

THANK YOU.