Annotating gene sequence variation

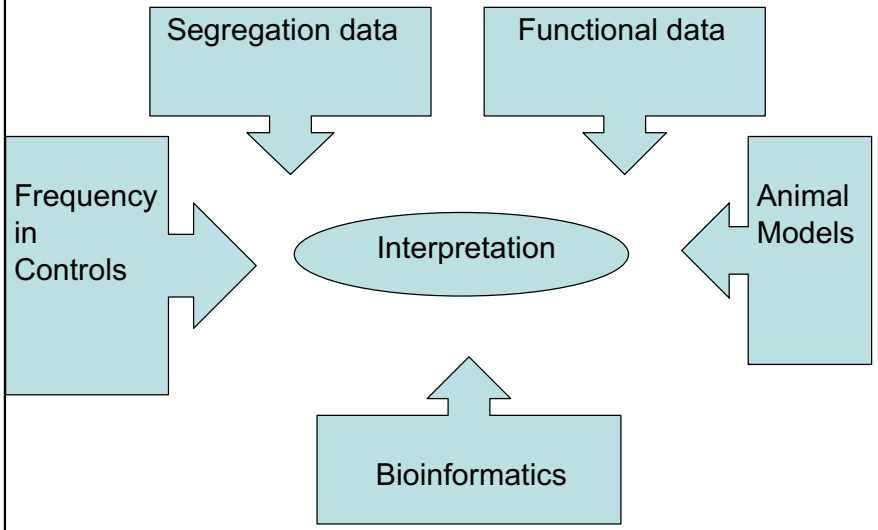
Shamil Sunyaev

Department of Biomedical Informatics
Harvard Medical School

Division of Genetics
Department of Medicine
Brigham and Women's Hospital / Harvard Medical School

Broad Institute of M.I.T. and Harvard

Identifying functionally significant causal variants in



Map variants on genomic annotation

Watch for multiple transcripts!

Watch for conflicting annotations!

Nonsense variants

One of most significant types of variants usually leading to the complete loss of function.

Nonsense variants are enriched in sequencing artifacts

Important considerations: i) location along the gene, ii) does the variant cause NMD? iii) is the variant in a commonly skipped exon?

Tool: LOFTEE

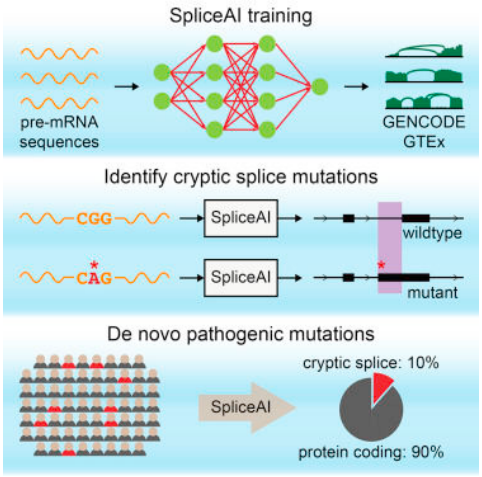
Variants involved in splicing

Variants in canonic splicing sites

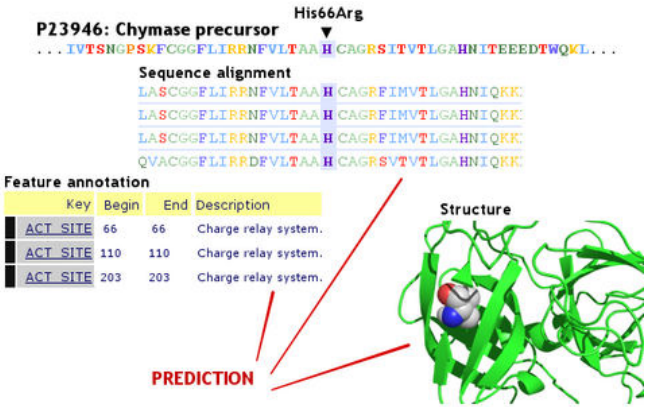
Variants in exonic or intronic splicing enhancers

Gain of splicing variants

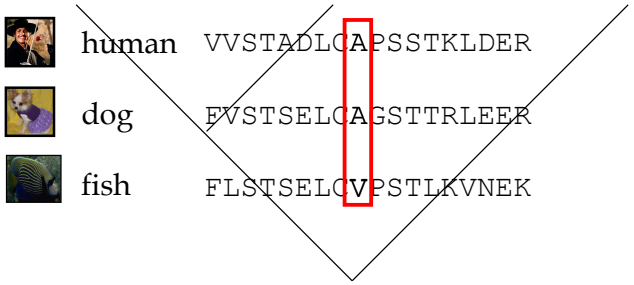
SpliceAI



Missense variants: computational predictions



Does the mutation fit the pattern of past evolution?



Statistical issues:
-sequences are related by phylogeny
-generally, we have too few sequences

Does the mutation fit the pattern of past evolution?

- We assume a constant fitness landscape: what is good for fish is good for human!
- We can estimate whether the mutation fits the pattern of amino acid changes.
- We can also estimate rate of evolution at the amino acid site

Continuous time Markov model

GLY → *VAL* → *ALA* → *GLY* → *ALA*

Continuous time Markov model

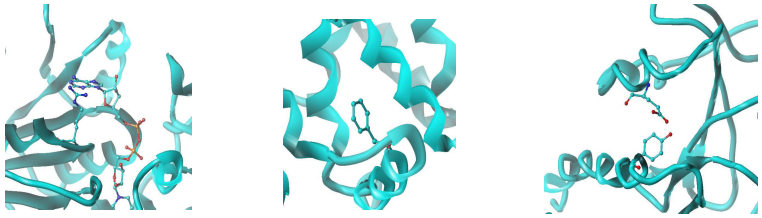
P – matrix of transition probabilities

$$P(t) = e^{Qt}$$

π – stationary distribution

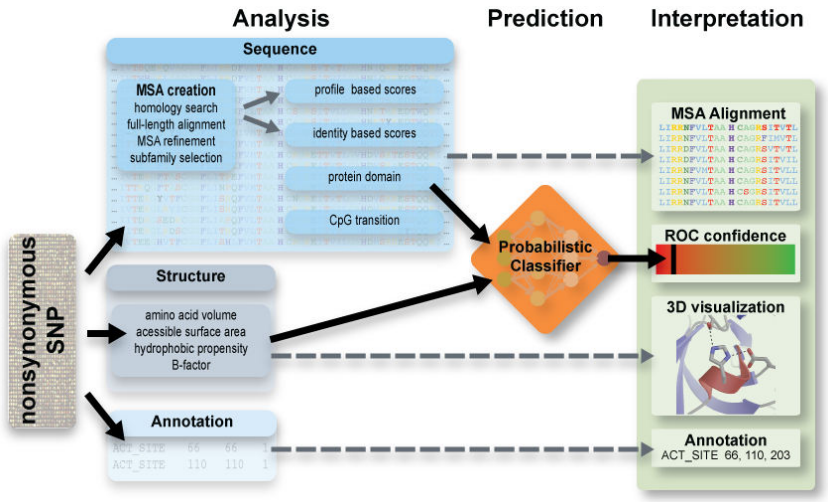
$$Q\bar{\pi} = 0$$

Protein structure view



- **Most of pathogenic mutations are important for stability (good news?).**
- ***ΔΔG* is difficult to estimate.**
- **Unfolded protein response pathway has to be taken into account.**
- **Heuristic structural parameters help but less than comparative genomics.**

PolyPhen2



www.genetics.bwh.harvard.edu/pph2 Adzhubei, et al. Nature Methods 2010

Weakly deleterious mutations

- Multiple independent lines of evidence suggest abundance of weakly deleterious alleles in humans
- Weakly deleterious variants may occur in highly conserved positions
- Weakly deleterious alleles probably contribute to complex phenotypes but not to simple Mendelian phenotypes

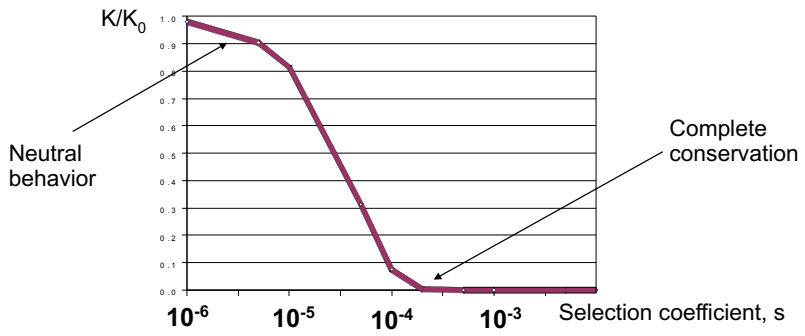
Conservation can be due to very weak selection

Every new mutation eventually will be either fixed or lost

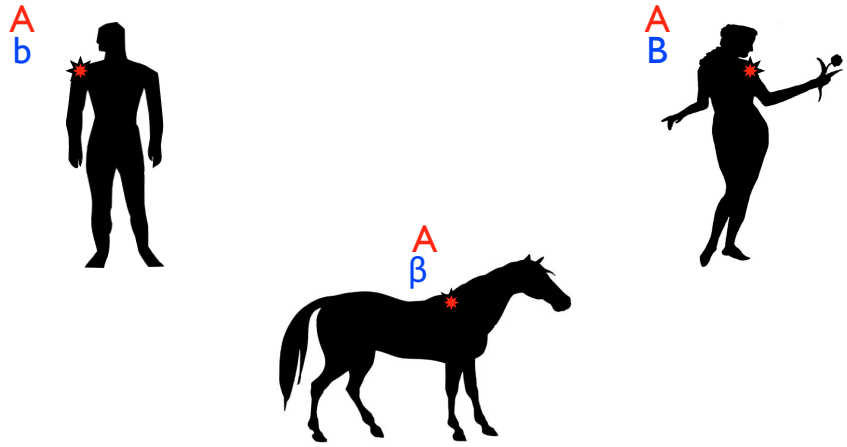
$$K = K_0 2 N_e \frac{(1-e^{-2s})}{(1-e^{-4N_e s})}$$

s – selection coefficient
N_e - effective population size

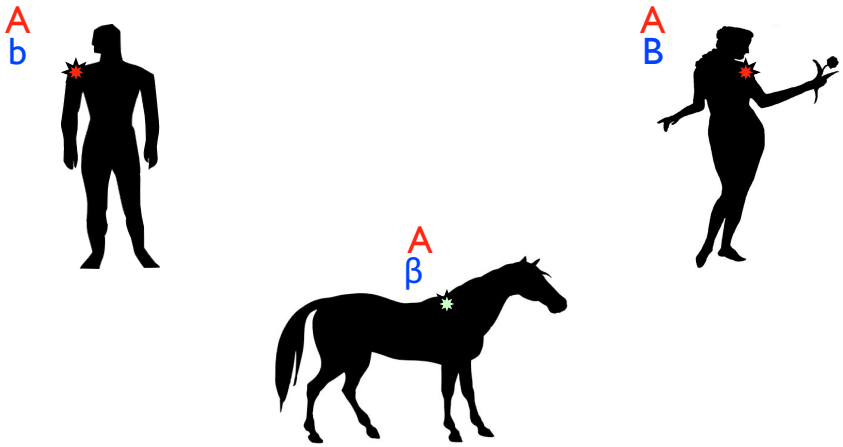
For humans estimated to be ~ 10 000



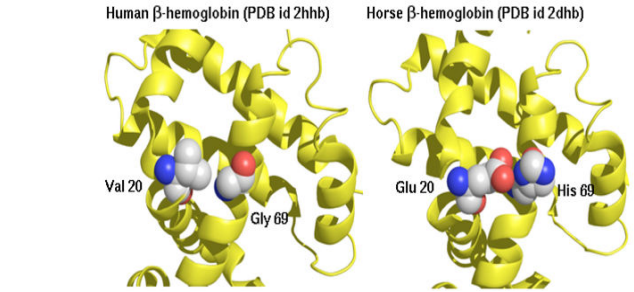
Constant evolutionary landscape



Epistatic interactions



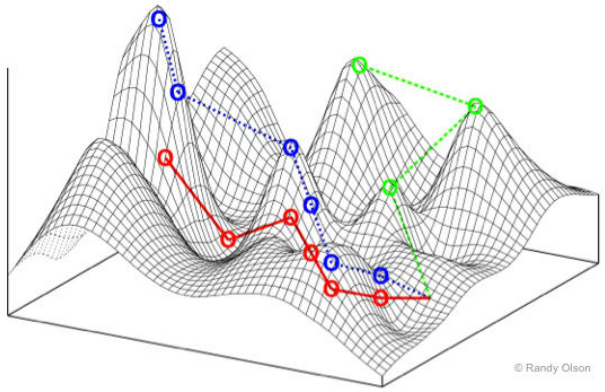
Compensatory mutations



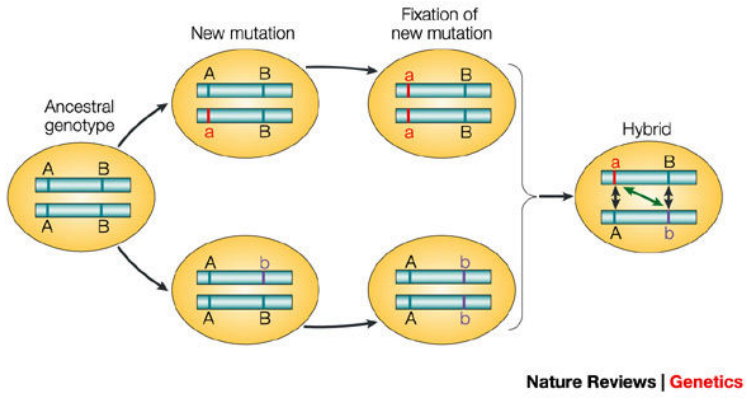
The phenomenon of compensatory mutations in different fields

- Biochemistry – protein stability, allosteric effects
- Genetics – incomplete penetrance
- Evolutionary biology – speciation, epistatic models of evolution

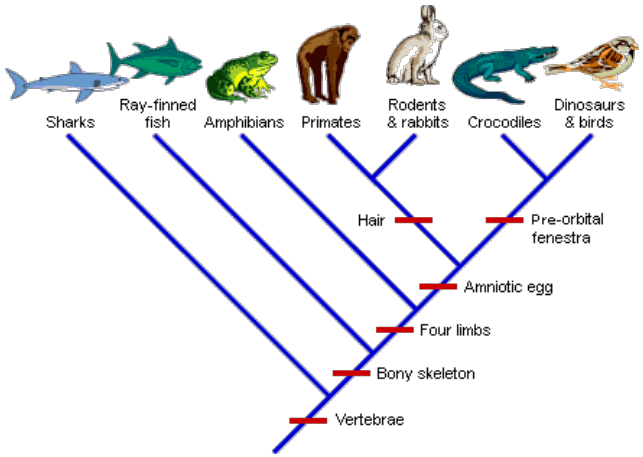
Ridges on the fitness landscape



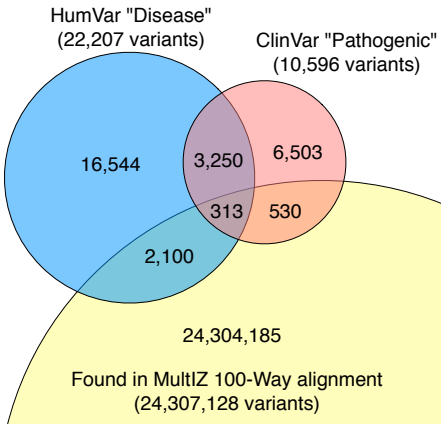
Dobzhansky-Muller incompatibility



Looking at vertebrate species



Many human disease mutations are found in vertebrates



5.5-6.5% of presumably pathogenic human mutations are detected in mammals

How complex genetic suppression can be?

LETTER

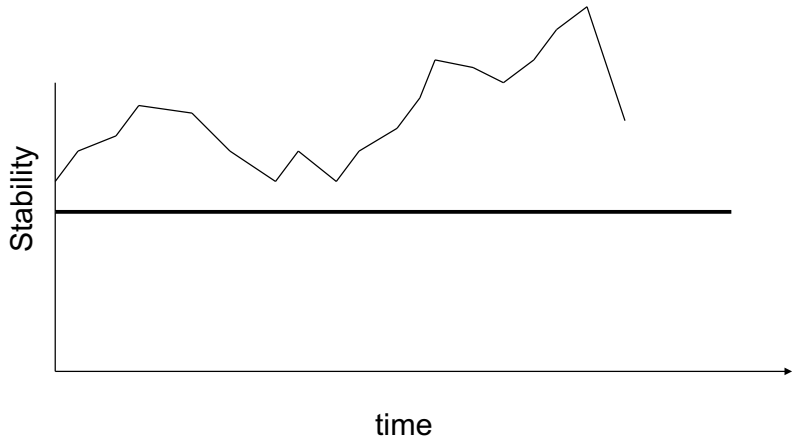
doi:10.1038/nature12678

Genetic incompatibilities are widespread within species

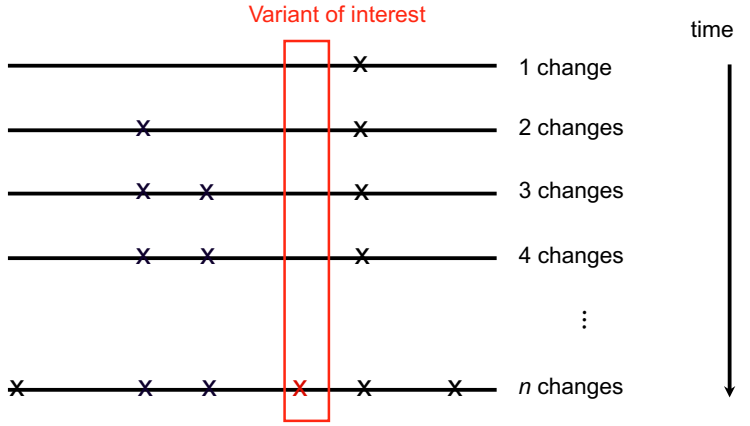
Russell B. Corbett-Detig[1], Jun Zhou[1], Andrew G. Clark[2,3], Daniel L. Hartl[1] & Julien F. Ayroles[1,2,4]

Numerous Dobzhansky-Muller incompatibilities in fly population

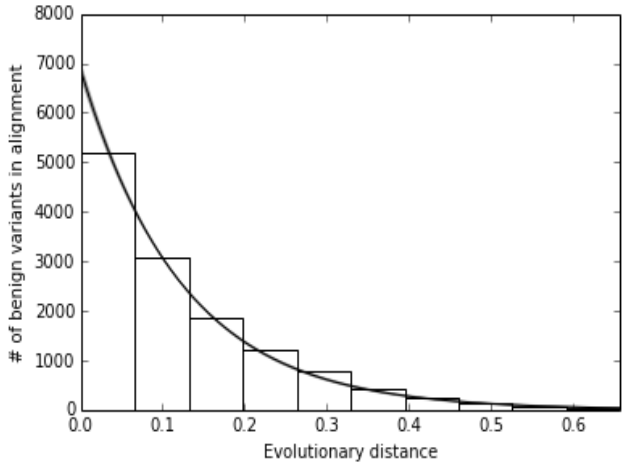
How complex genetic suppression can be?



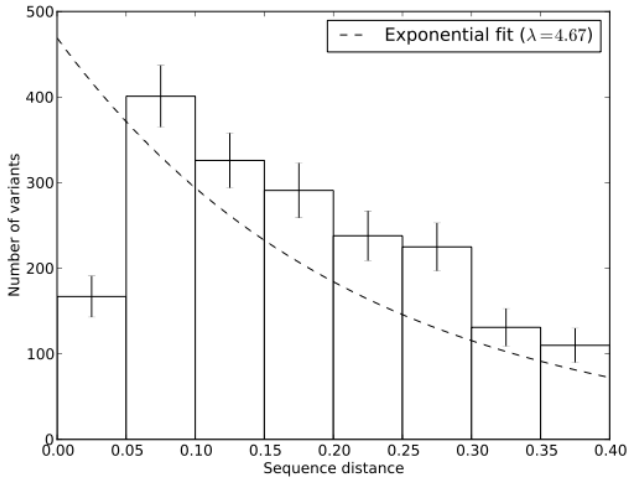
Model: accumulation of neutral variants



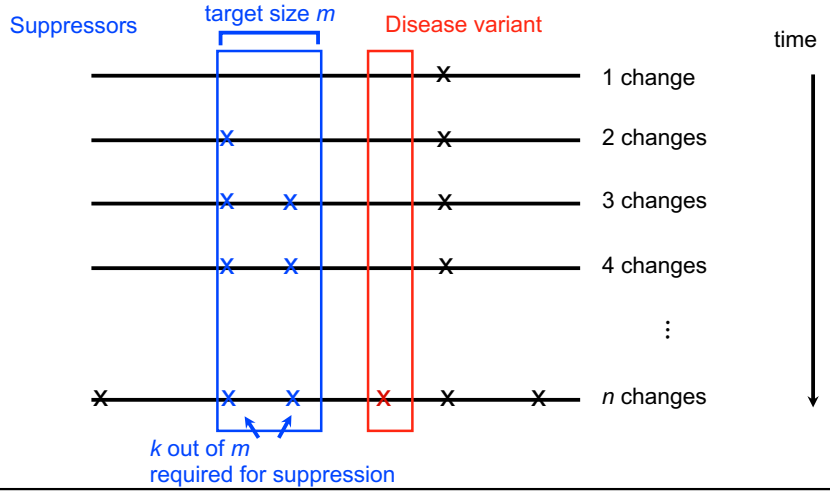
Exponential fit



Disease variants do not fit the Poisson expectation



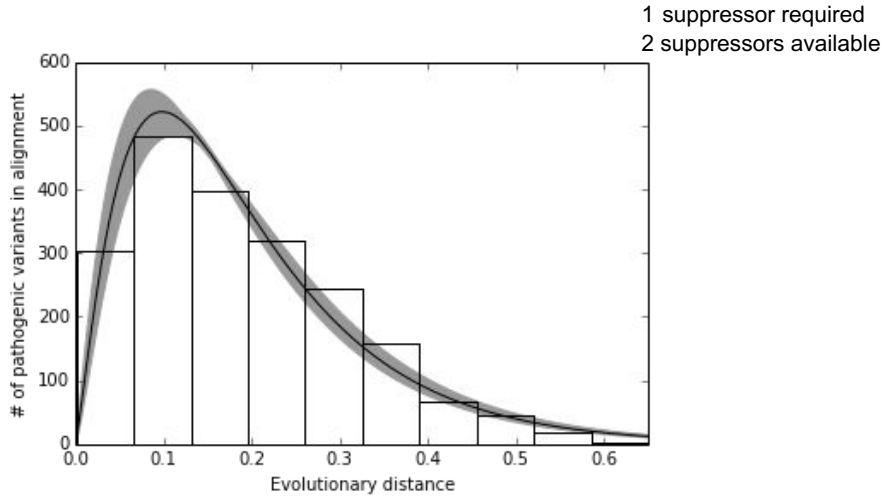
Model: accumulation of disease variants



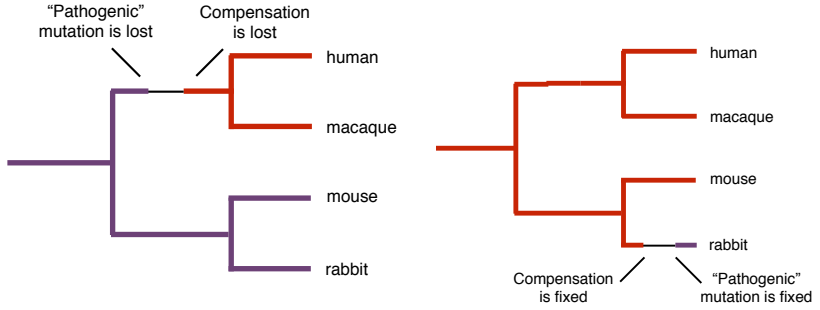
Model: Erlang distribution

$$L(k,t,\lambda) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{(k-1)!}$$

Model: fit for target size and number of compensatory changes



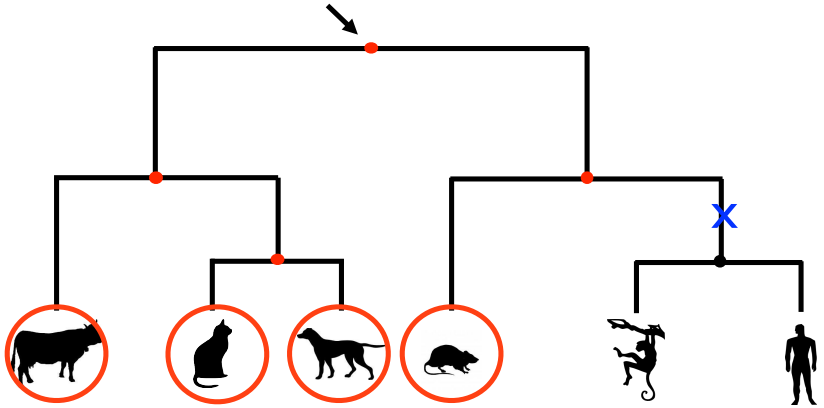
Evolutionary model of *cis*-complementation



Model Summary

- ~5-6% of human disease mutations have potential suppressors (i.e. are present in another mammalian species)
- In most cases, one large-effect suppressor is sufficient, out of only 1-2 available
- These values allow simple experiment to identify suppressors

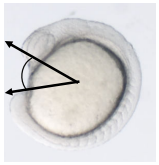
Predicting and testing suppressors



Zebrafish model

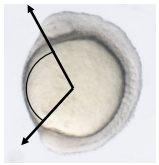
- Model of Bardet-Biedl Syndrome (obesity, renal failure, vision loss)
- Caused by defects in primary cilium
- Embryonic convergence / extension phenotype in zebrafish
- Easily scorable phenotype

Normal



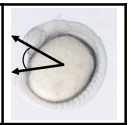
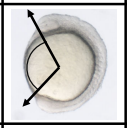
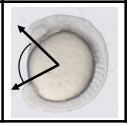
Class I

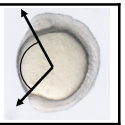
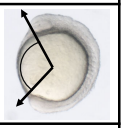
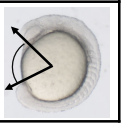
Class II



Images: Phoebe

Experiment interpretation

No injection	
Knockdown	
Rescue with human gene	

Human gene with disease mutant	
Double mutant (no suppression)	
Double mutant (full suppression)	

Images: Phoebe

Bardet-Biedl syndrome – BBS4 N165H

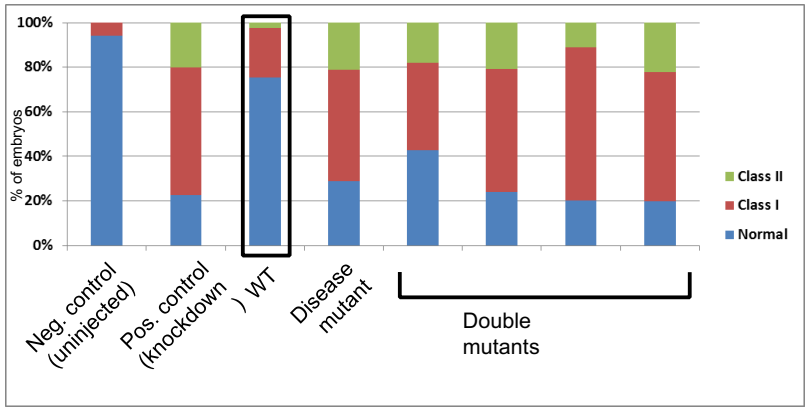
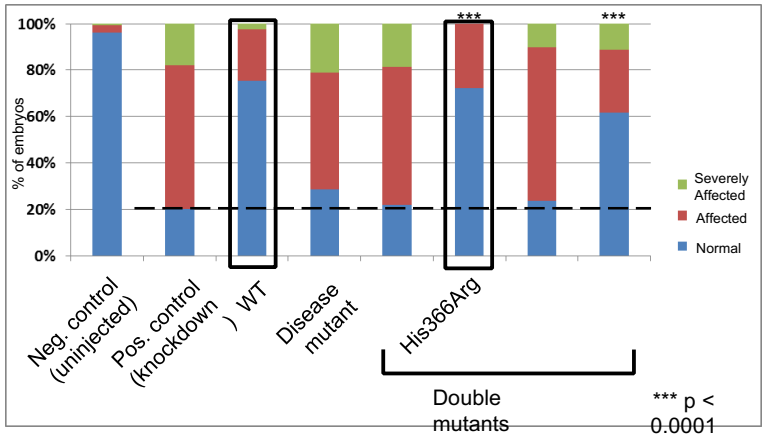


Figure: Stephan Frangakis

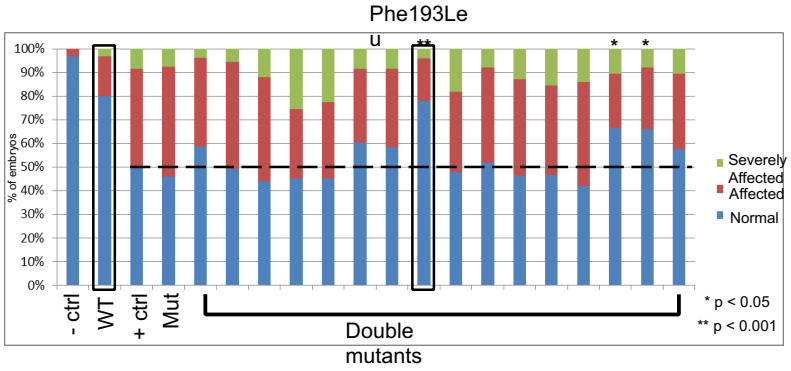
Bardet-Biedl syndrome – BBS4 N165H



Out of 9 candidates (7 shown here), 1 complete rescue

Figure: Stephan Frangakis

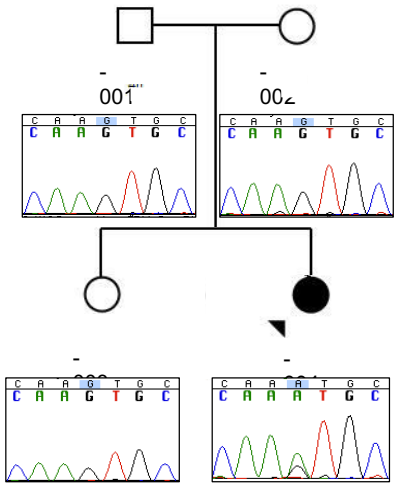
Bardet-Biedl syndrome – RPGRIP1L R937L



Out of 32 candidates, 1 complete rescue

Figure: Stephan Frangakis

A newly identified gene



Clinical features

Global developmental delay
microcephaly
feeding issues
failure to thrive
abnormal muscle tone
low immunoglobulins
frequent respiratory infections

Clinical testing

normal female microarray
metabolic testing – negative
extensive genetic testing – negative

BTG2 De novo

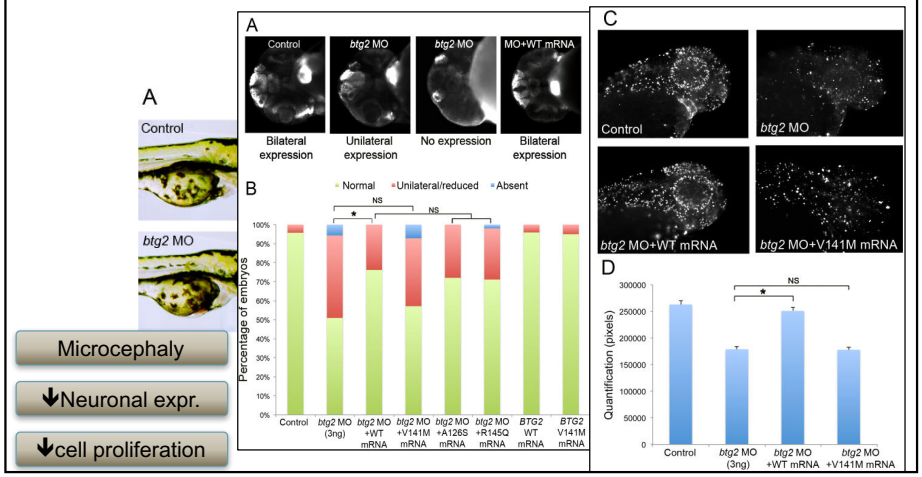
TTN Compound het

NOS2 De novo

LAMA1 Compound het

Stephan Frangakis

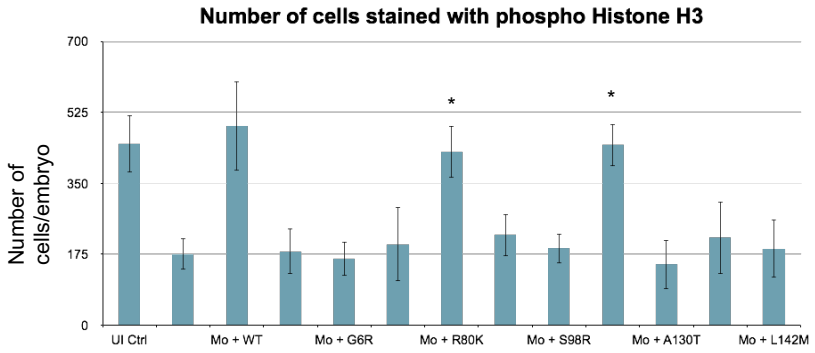
BTG2 is the disease driver



The mutation is a reversal to the mammalian ancestral state

BTG2	R80	L128	Q140	V141	L142
H. sapiens	R	L	Q	V	L
P. troglodytes	•	•	•	•	•
G. gorilla	•	•	•	•	•
M. musculus	K	V	•	M	M
R. norvegicus	K	V	•	M	M
H. glaber	•	V	•	M	M
S. domesticus	K	V	•	M	M
B. primigenius	K	V	•	M	M
E. ferus caballus	K	V	•	M	M
F. catus	K	V	•	M	M
C. lupus familiaris	K	V	•	M	M
D. novemcinctus	K	V	•	M	M
G. gallus	K	P	•	M	M

BTG2 has two compensatory mutations



*P<0.01 vs V141M rescue alone

Injection

Stephan Frangakis

Methods

PolyPhen2

SIFT

LRT

MutationTaster

FatHMM

SNPs3D

DeepSequence

Umbrella methods

Condel

REVEL

CADD

M-CAP

Incorporating regional constraint

CCR

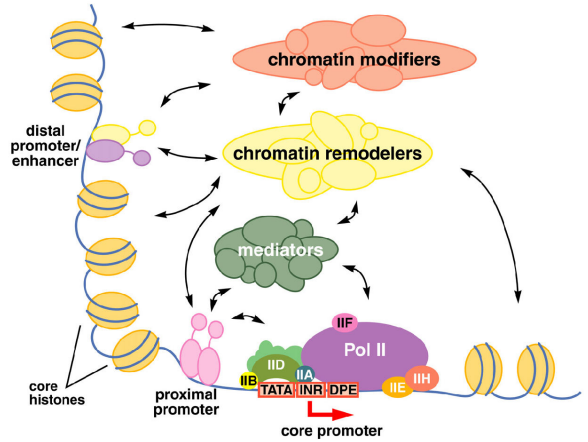
M-CAP

PrimateAI

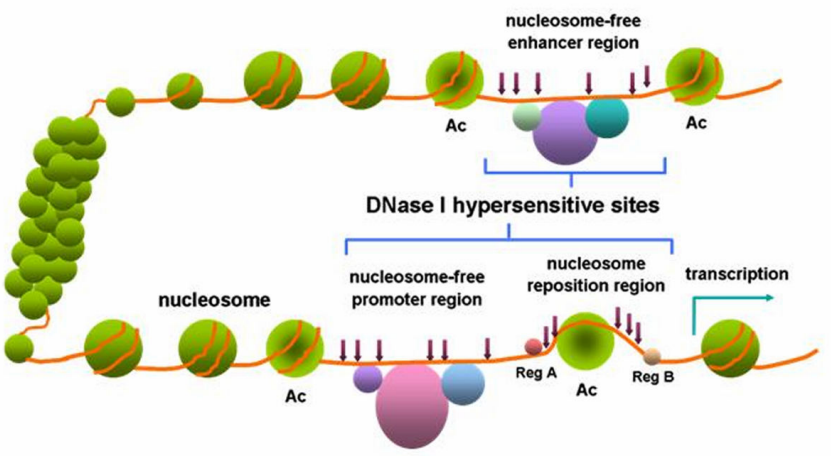
Non-coding variants

Regulatory variants

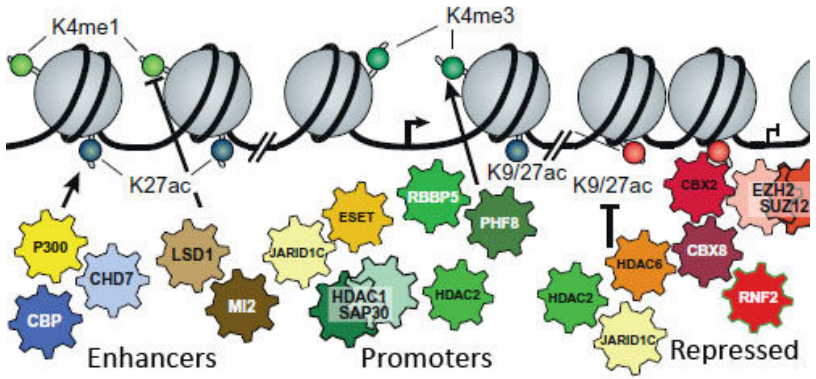
- Regulation: variants in promoters, enhancers, silencers, insulators



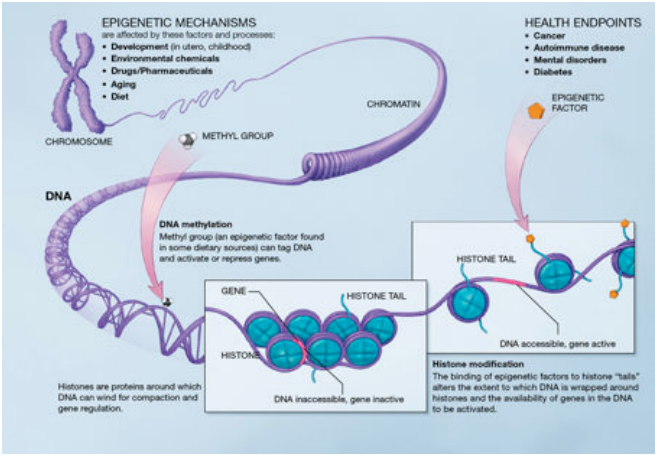
Chromatin accessibiliy



Chromatin modification



Epigenomics



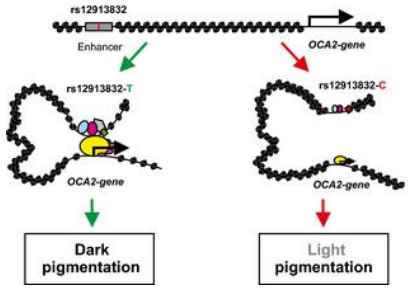
Why do we think that non-coding variation is of importance?

Regions and individual nucleotides conserved along phylogeny show signals of purifying selection in humans

Epigenetic studies report many well-localized regulatory marks

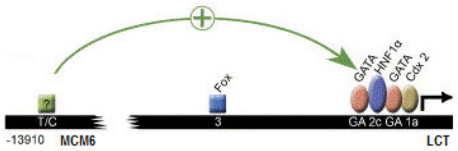
GWAS signals are predominantly located in non-coding regions

Non-disease alleles of large effect



Eye color

Lactase tolerance



Ultraconserved elements

OPEN ACCESS Freely available online PLOS BIOLOGY

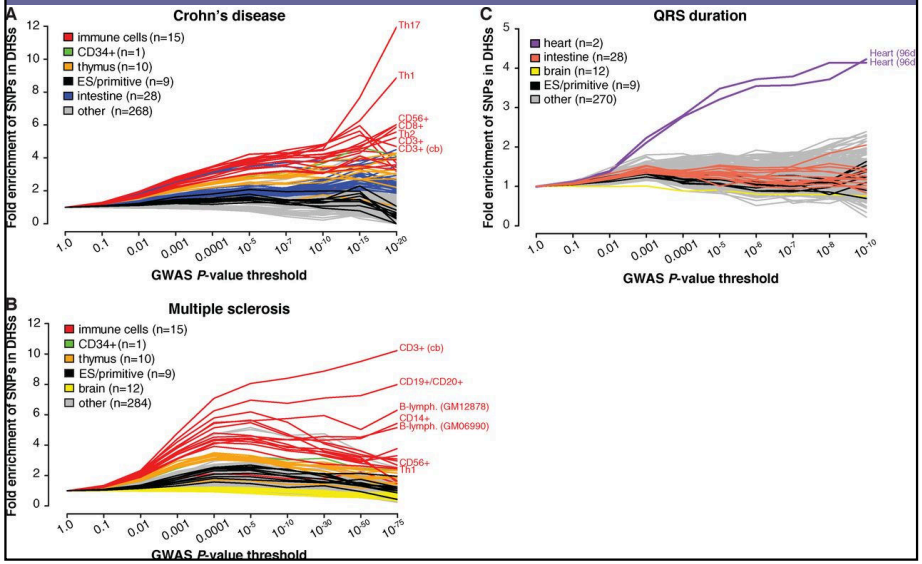
Deletion of Ultraconserved Elements Yields Viable Mice

Nadav Ahituv[1,2¤], Yiwen Zhu[1], Axel Visel[1], Amy Holt[1], Veena Afzal[1], Len A. Pennacchio[1,2], Edward M. Rubin[1,2*]

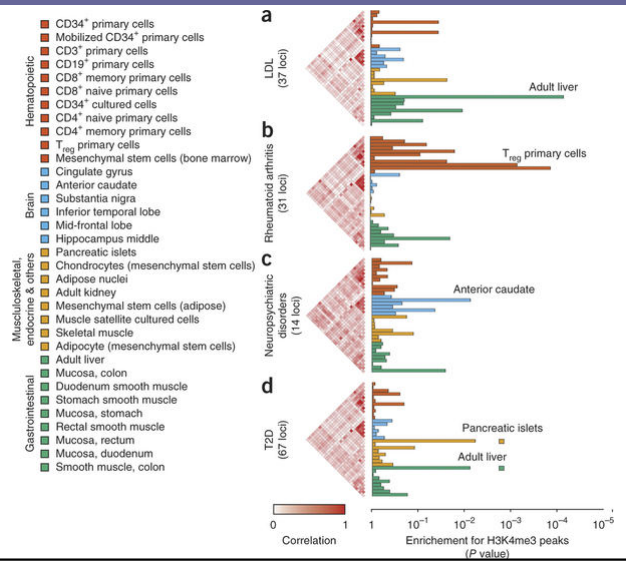
1 Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, 2 United States Department of Energy Joint Genome Institute, Walnut Creek, California, United States of America

Ultraconserved elements have been suggested to retain extended perfect sequence identity between the human, mouse, and rat genomes due to essential functional properties. To investigate the necessities of these elements in vivo, we removed four noncoding ultraconserved elements (ranging in length from 222 to 731 base pairs) from the mouse genome. To maximize the likelihood of observing a phenotype, we chose to delete elements that function as enhancers in a mouse transgenic assay and that are near genes that exhibit marked phenotypes both when completely inactivated in the mouse and when their expression is altered due to other genomic modifications. Remarkably, all four resulting lines of mice lacking these ultraconserved elements were viable and fertile, and failed to reveal any critical abnormalities when assayed for a variety of phenotypes including growth, longevity, pathology, and metabolism. In addition, more targeted screens, informed by the abnormalities observed in mice in which genes in proximity to the investigated elements had been altered, also failed to reveal notable abnormalities. These results, while not inclusive of all the possible phenotypic impact of the deleted sequences, indicate that extreme sequence constraint does not necessarily reflect crucial functions required for viability.

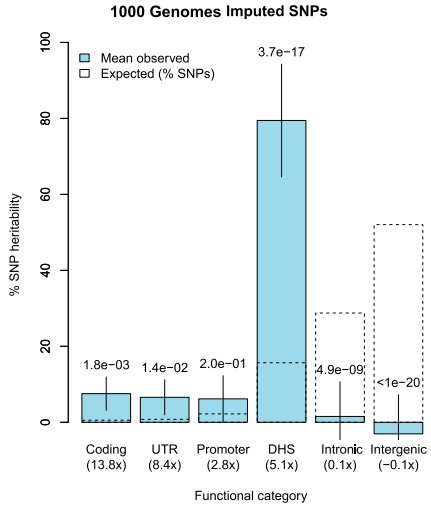
Enrichment of GWAS signals in putative regulatory elements



Enrichment of GWAS signals in putative regulatory elements



Partitioning heritability



GWAS SNPs co-localize with eQTLs

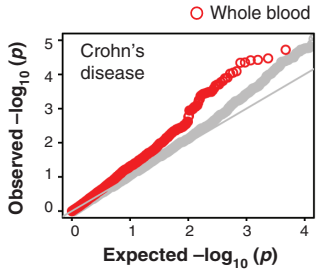
OPEN ACCESS Freely available online PLoS GENETICS

Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS

Dan L. Nicolae[1,2,3], Eric Gamazon[1], Wei Zhang[1], Shiwei Duan[1¤], M. Eileen Dolan[1,2], Nancy J. Cox[1,2*]

The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans

The GTEx Consortium*†



Translating GWAS findings into mechanistic models

GWAS peak

↓

Controlled model system

↓

Biochemistry

Human Genetics All the Way

GWAS peak

Endophenotype **Endophenotype**

Gene expression (eQTL)

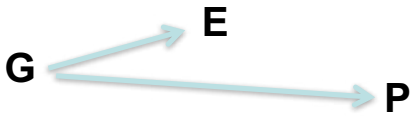
↓

Molecular phenotype (molecular_QTL)

Causality

Mediation

Independent effects

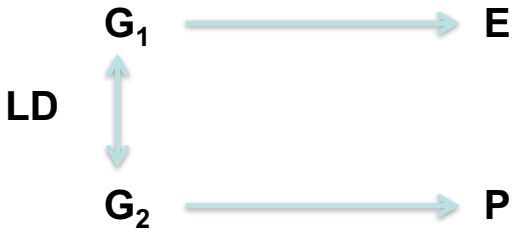


Reverse causation

Co-localization

Same causal variant

Distinct variants



Methods

Coloc

eCAVIAR

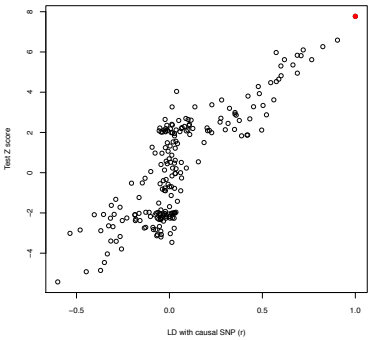
JLIM

PAINTOR's fine mapping model

- If a SNP is causal, then r^2 should predict association of other SNPs in the area:

  $$Z_2 \sim N(r_{12}\lambda_1, 1)$$

- Correlation between test statistics *Z* are approximated by MVN given local pairwise LD structure.

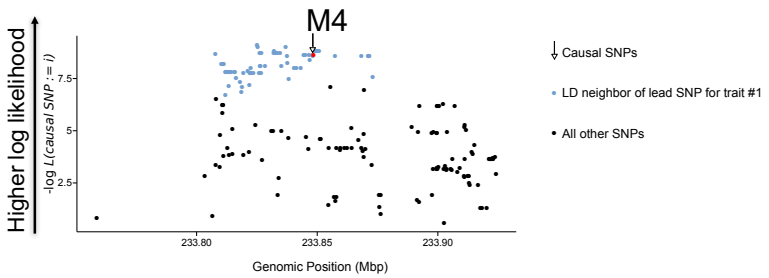


$$L(\boldsymbol{Z}|C = \{m\}; \lambda_m) = N(\boldsymbol{Z}; \Sigma(\lambda \circ C), \Sigma)$$
$$\propto e^{-\frac{1}{2}(\boldsymbol{Z}^T \Sigma^{-1} \boldsymbol{Z} - z_m^2)}$$

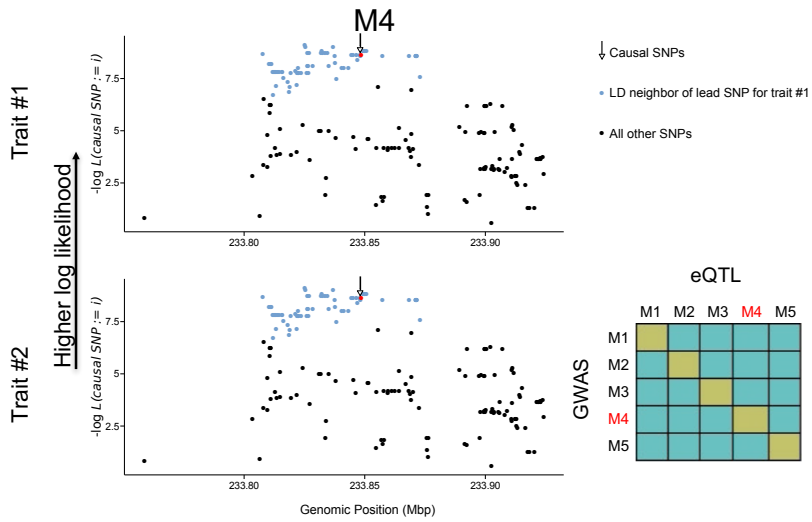
Parameters
λ: standardized effect size
Z: association statistic
C: indicator of causality
m: SNP considered

Kichaev *et al*. PLoS Genet. 2014; Chen *et al.* Genetics. 2015

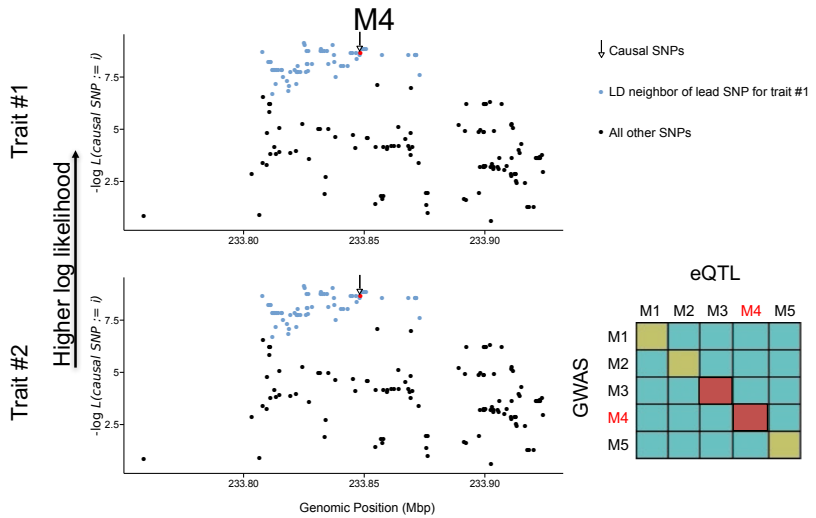
Likelihood of causal SNP



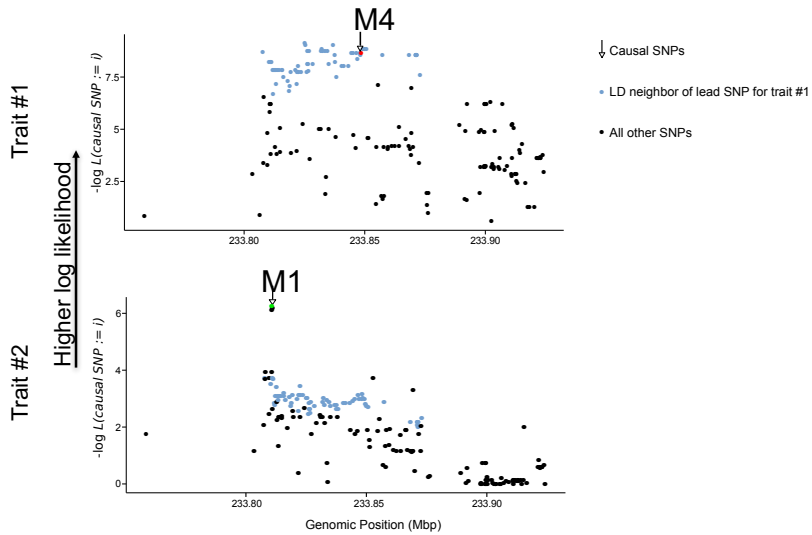
Joint likelihood: same causal variant



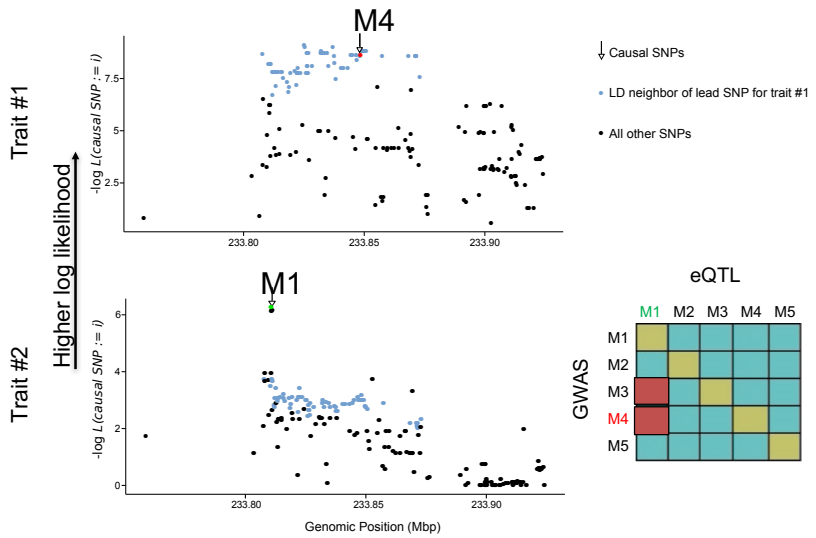
Joint likelihood: same causal variant



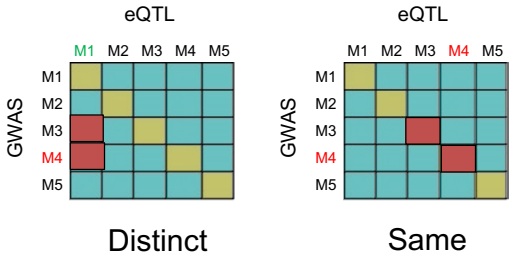
Joint likelihood: distinct variants



Joint likelihood: distinct variants



Joint likelihood test



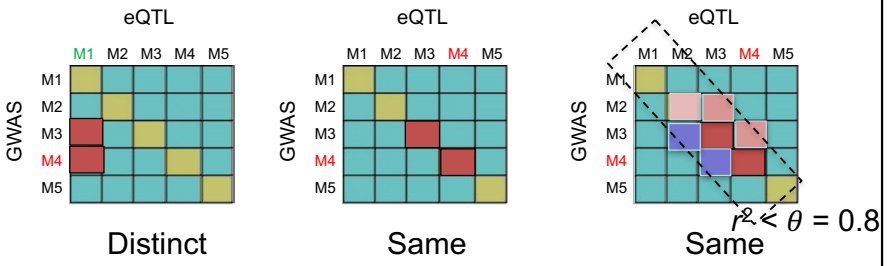
$$\Lambda = \sum_{r^2_{i,m^*>\theta}} L_1(i) \cdot \log \frac{L_1(\boldsymbol{i})L_2(\boldsymbol{i})}{\max_{r^2_{i,j}<\theta} L_1(\boldsymbol{i})L_2(\boldsymbol{j})}$$

$$= \sum_{r^2_{i,m^*>\theta}} e^{-\frac{1}{2}(z_i^2 - z_{m^*}^2)} \cdot (w_i^2 - \max_{r^2_{i,j}<\theta} w_j^2)$$

Parameters
z: disease association statistic
w: eQTL association statistic
m^*: lead disease-associated SN
θ: r^2 resolution limit

P-values can be estimated by permuting eQTLs

Joint likelihood test



$$\Lambda = \sum_{r^2_{i,m^*>\theta}} L_1(i) \cdot \log \frac{L_1(\boldsymbol{i})L_2(\boldsymbol{i})}{\max_{r^2_{i,j}<\theta} L_1(\boldsymbol{i})L_2(\boldsymbol{j})}$$

$$= \sum_{r^2_{i,m^*>\theta}} e^{-\frac{1}{2}(z_i^2 - z_{m^*}^2)} \cdot (w_i^2 - \max_{r^2_{i,j}<\theta} w_j^2)$$

Parameters
z: disease association statistic
w: eQTL association statistic
m^*: lead disease-associated SN
θ: r^2 resolution limit

P-values can be estimated by permuting eQTLs

Real data: autoimmune/inflammatory diseases

- Highly successful GWAS
- ImmunoChip: custom fine-mapping array
- Free availability of summary statistics on ImmunoBase
- Accessibility of disease relevant cell types and eQTLs data

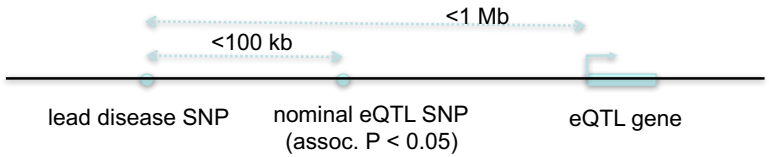
Disease	Densely genotyped [a]
MS	59
IBD	69
Crohn	19
UC	10
T1D	47
RA	34
CEL	34
Overall	272

* Excluding conditional hits
** Defined by ImmunoChip's densely genotyped fine-mapping intervals. Excluding MHC

Disease	Densely genotyped [a]	Number of loci: eQTL present [b] CD4+	CD14+	LCL	Total
MS	59				
IBD	69				
Crohn	19				
UC	10				
T1D	47				
RA	34				
CEL	34				
Overall	272				

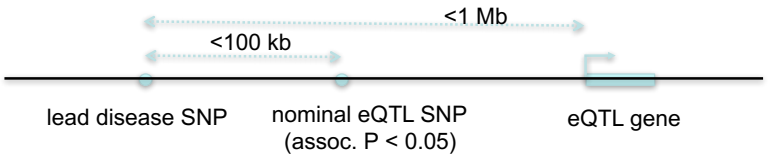
**** CD4/CD14+ (n=211/213) from Raj et al. Science 2014;
LCL (n=278) from Lappalainen et al. Nature 2013

Disease	Densely genotyped [a]	Number of loci: eQTL present [b] CD4+	CD14+	LCL	Total
MS	59				
IBD	69				
Crohn	19				
UC	10				
T1D	47				
RA	34				
CEL	34				
Overall	272				



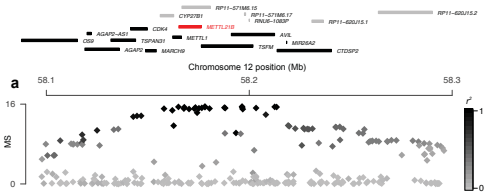
**** CD4/CD14+ (n=211/213) from Raj et al. Science 2014;
LCL (n=278) from Lappalainen et al. Nature 2013

Disease	Densely genotyped [a]	Number of loci: eQTL present [b] CD4+	CD14+	LCL	Total
MS	59	54	55	55	56
IBD	69	69	69	68	69
Crohn	19	18	18	18	18
UC	10	10	9	10	10
T1D	47	39	40	36	40
RA	34	34	34	34	34
CEL	34	34	34	34	34
Overall	272	258	259	255	261

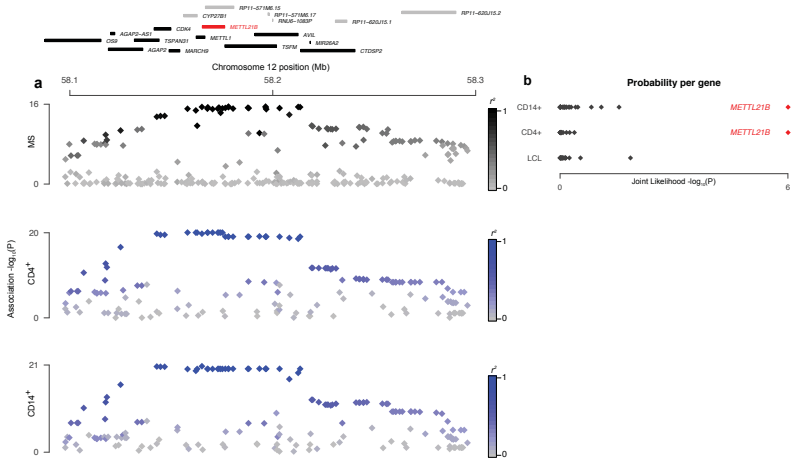


**** CD4/CD14+ (n=211/213) from Raj et al. Science 2014;
LCL (n=278) from Lappalainen et al. Nature 2013

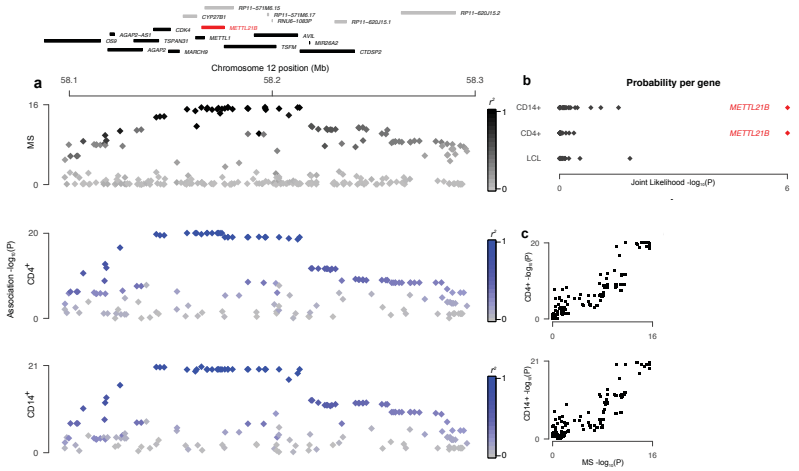
METTL21B eQTLs in CD4⁺ and CD14⁺ are consistent with association to MS



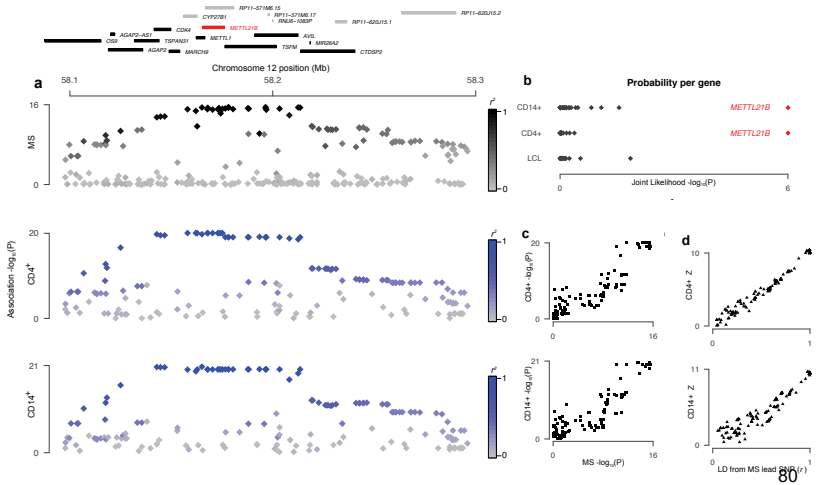
METTL21B eQTLs in CD4⁺ and CD14⁺ are consistent with association to MS



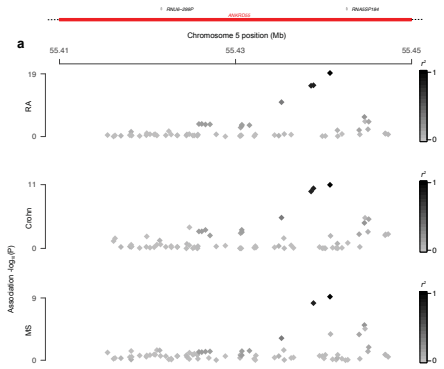
METTL21B eQTLs in CD4⁺ and CD14⁺ are consistent with association to MS



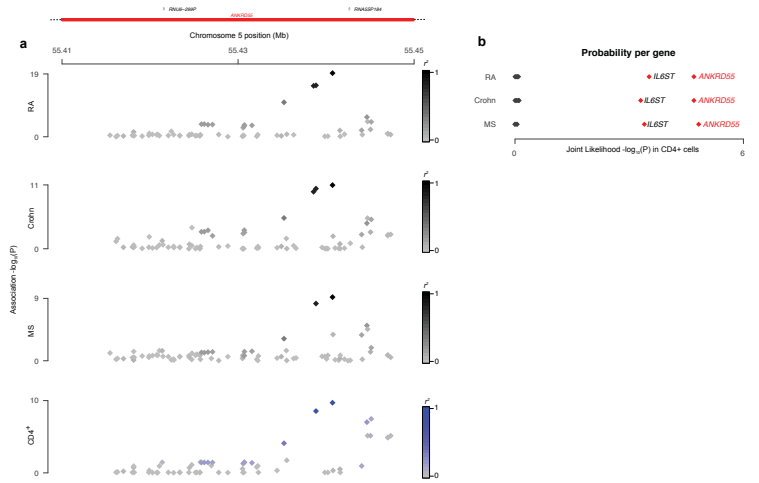
METTL21B eQTLs in CD4⁺ and CD14⁺ are consistent with association to MS



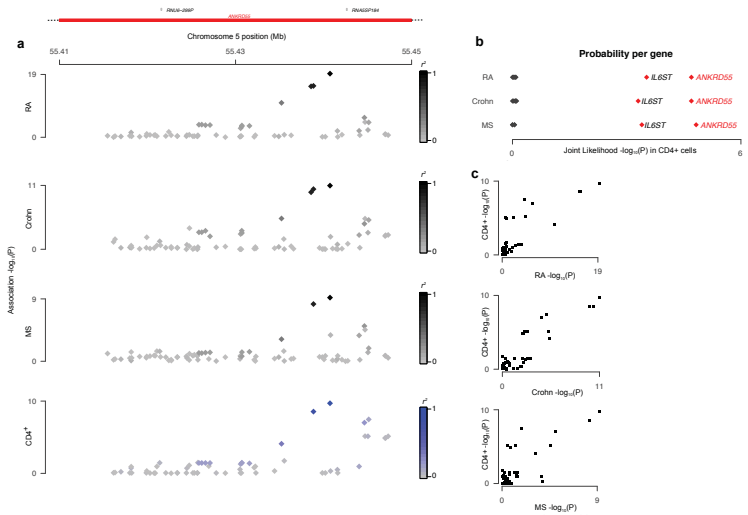
A GWAS peak shared across RA, Crohn, and MS is consistent with *ANKRD55* eQTL in T cells



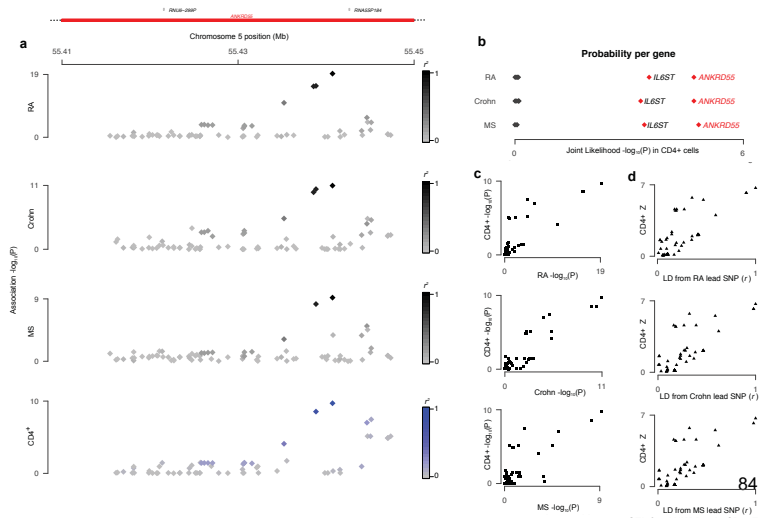
ANKRD55 eQTL in CD4+ is consistent with association with MS, Crohn, and RA



ANKRD55 eQTL in CD4+ is consistent with association with MS, Crohn, and RA



ANKRD55 eQTL in CD4+ is consistent with association with MS, Crohn, and RA

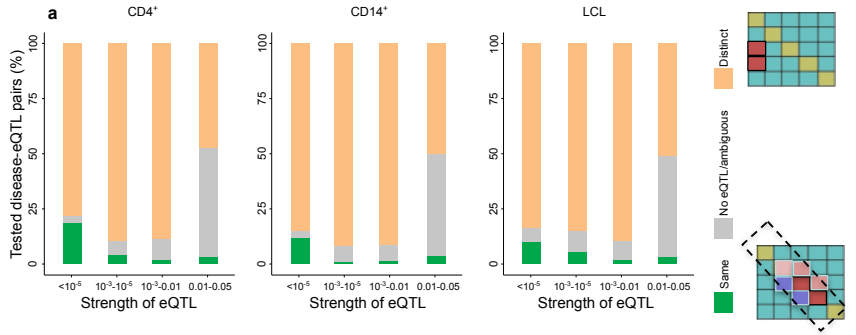


15% of disease loci have eQTL driven by the same variant (FDR 5%)

Disease	Densely genotyped [a]	Number of loci: eQTL present [b] CD4+	CD14+	LCL	Total	Driven by same effect [c] CD4+	CD14+	LCL	Total
MS	59	54	55	55	56	8	3	6	12
IBD	69	69	69	68	69	6	9	1	12
Crohn	19	18	18	18	18	2	1	0	3
UC	10	10	9	10	10	2	1	3	4
T1D	47	39	40	36	40	2	0	0	2
RA	34	34	34	34	34	2	0	1	3
CEL	34	34	34	34	34	3	2	0	5
Overall	272	258	259	255	261	25	16	11	41

* 75% of hits pass Bonferroni threshold as well.

~75% of tests disease eQTL pairs are driven by distinct variants



n=3,256 n=3,360 n=2,653

Summary on eQTLs

- ~15% of GWAS loci were mapped to eQTL genes.
- ~25% of GWAS loci are driven by the eQTLs of same effect.
- JLIM software

Methods

GWAVA (supervised)

CADD (predicts loss of genetic variation)

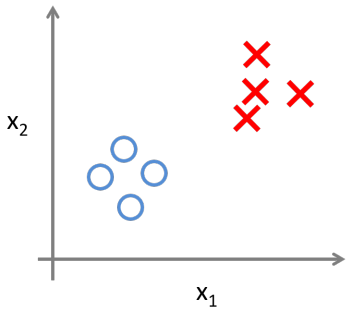
INSIGHT / LINSIGHT (population genetics)

Eigen (eigenvector in the annotation space)

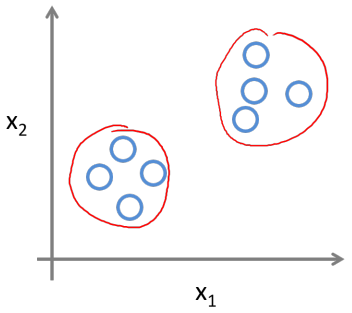
PINES (phenotype-specific)

Prediction Methods

Supervised Learning

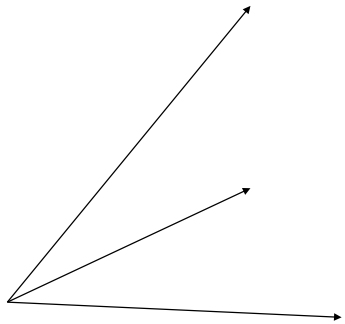


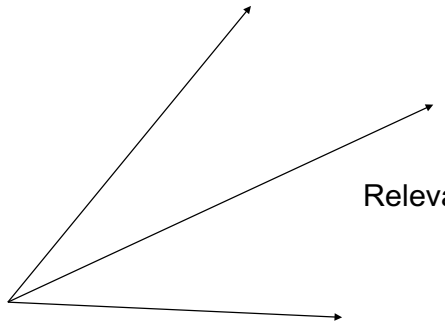
Unsupervised Learning

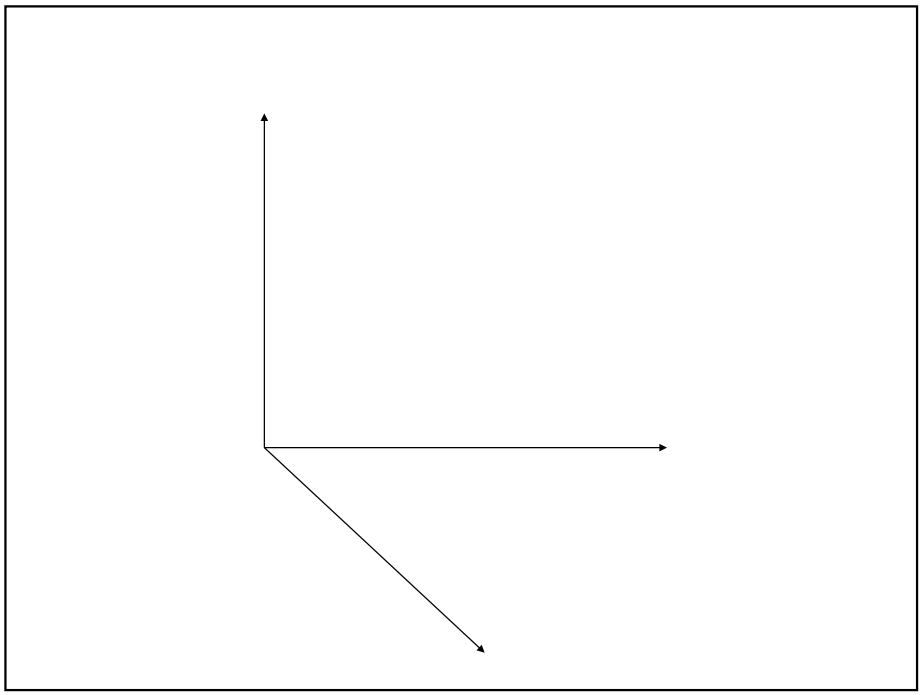


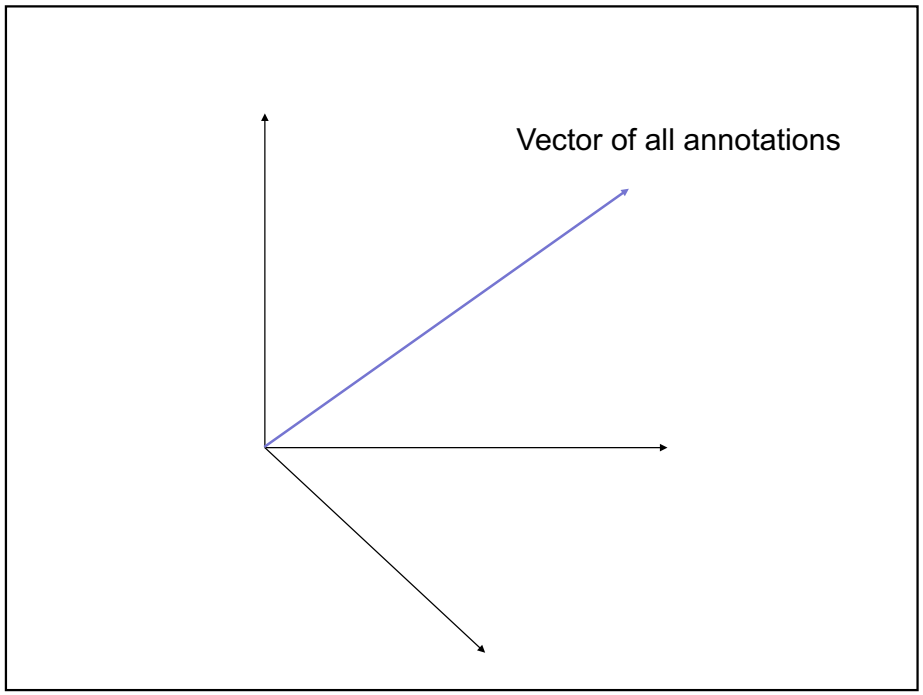
PINES

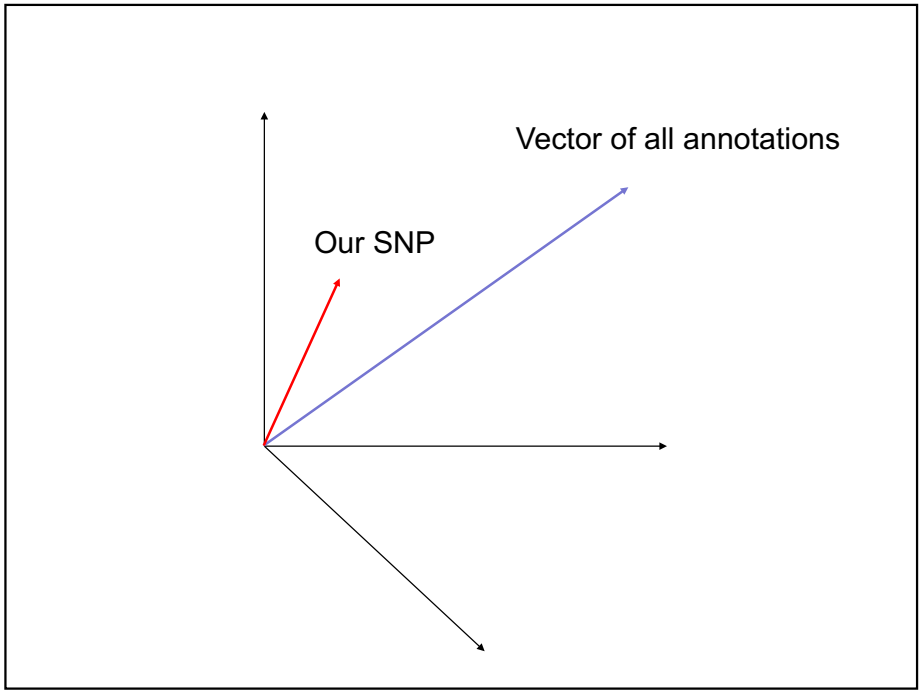
- Take all annotations and create an uncorrelated space
- Upweight axes corresponding to relevant cell/tissue types
- The score is based on the angle with the direction of the maximal possible annotation

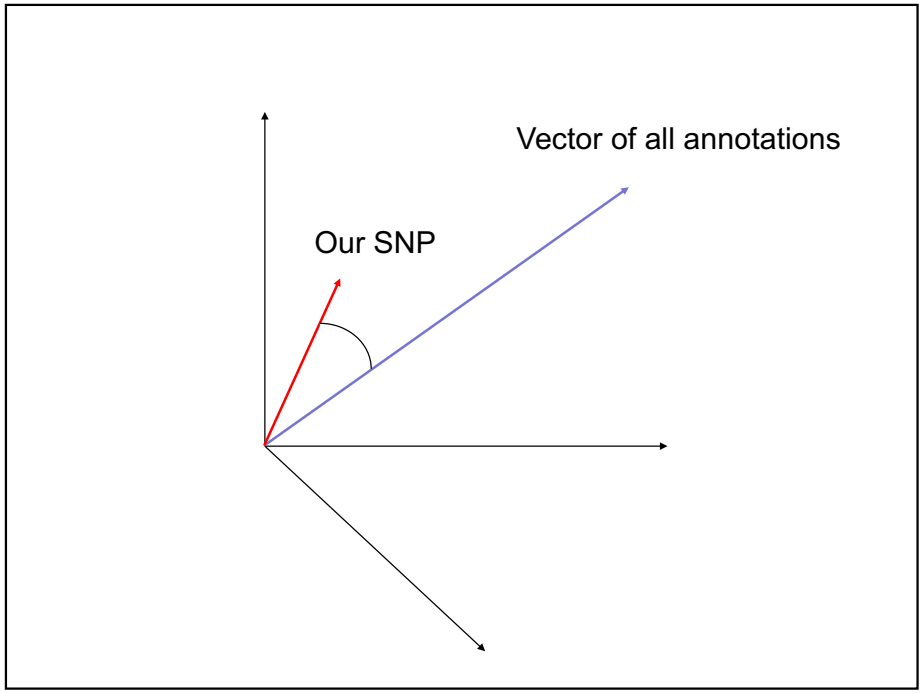




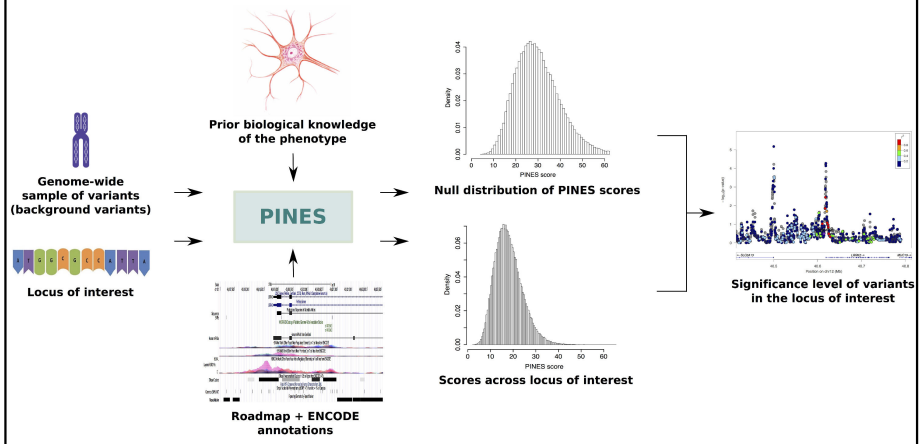


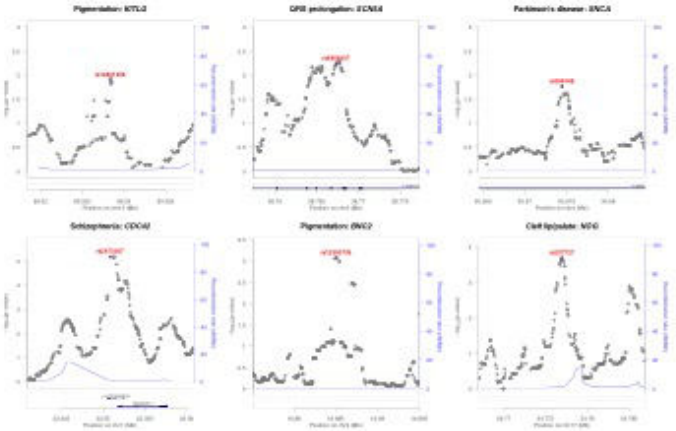

Vector of all annotations


Vector of all annotations
Our SNP


Vector of all annotations
Our SNP

PINES





PINES: http://genetics.bwh.harvard.edu/pines/