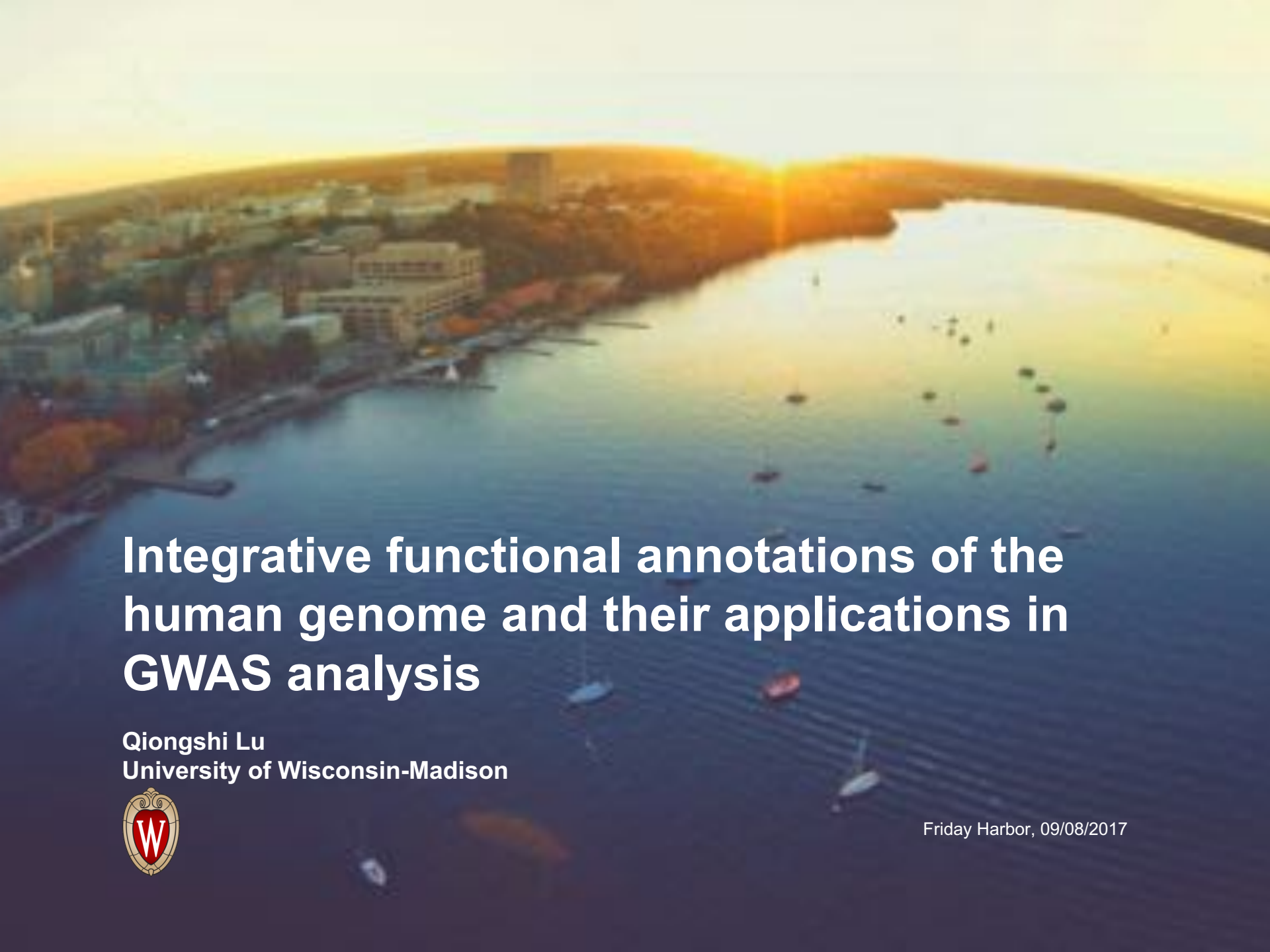


Integrative functional annotations of the human genome and their applications in GWAS analysis

Qiongshi Lu
University of Wisconsin-Madison

Friday Harbor, 09/08/2017

Outline

1. Background

2. Introduction to (narrow-sense) functional annotations

- Functional annotation in protein-coding genes
- Functional annotation in non-coding regions
- Other useful tools

3. Applications of functional annotations

- Functional SNP fine-mapping
- Partitioning heritability and genetic covariance
- Gene-level analysis
- Effect size estimation and risk prediction

Background

Despite the advancements, GWAS has its limitations

It is difficult to identify all associations

- Polygenicity and low effect size

It remains challenging to interpret the findings

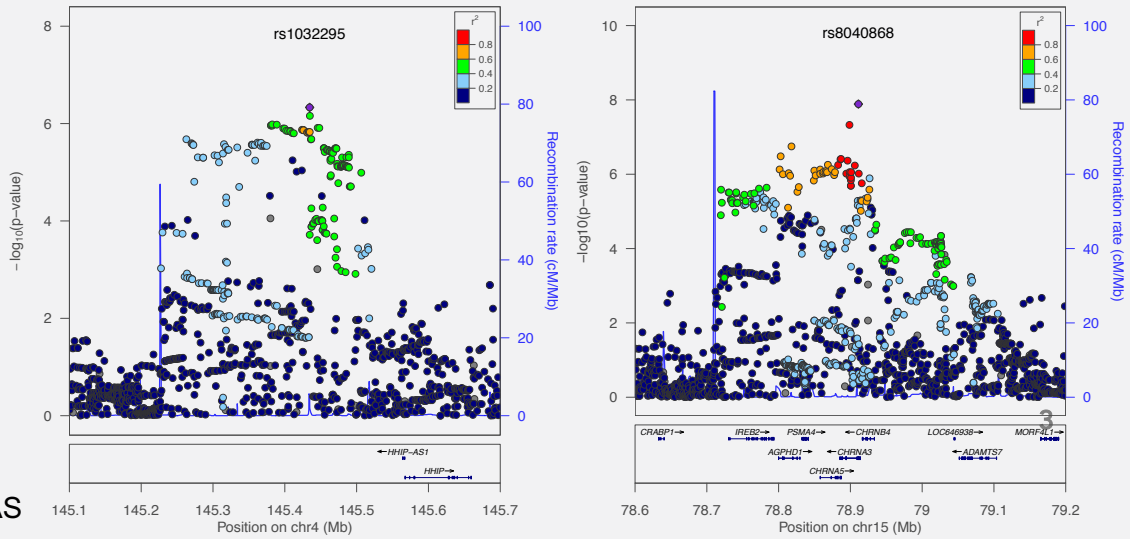
- 88% of significant associations are in the non-coding genome
- LD makes it challenging to identify biologically functional SNPs

Cell
Volume 169, Issue 7, 15 June 2017, Pages 1177-1186

Perspective

An Expanded View of Complex Traits: From Polygenic to Omnigenic

Evan A. Boyle [1], Yang I. Li [1], Jonathan K. Pritchard [1, 2, 3]

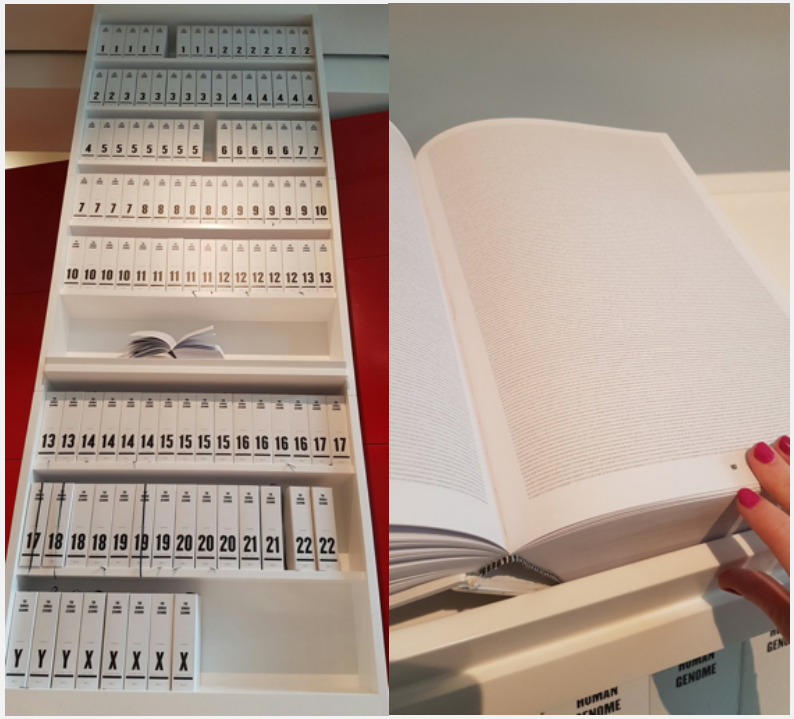


Two examples from COPD GWAS

Background

To solve these problems, we need better **functional annotations** for the (non-coding) human genome.

Only ~2% of the human genome encodes proteins. However, the rest 98% may be critically involved in a variety of regulatory machinery.



@susejohnston

Outline

1. Background

2. Introduction to (narrow-sense) functional annotations

- Functional annotation in protein-coding genes
- Functional annotation in non-coding regions
- Other useful tools

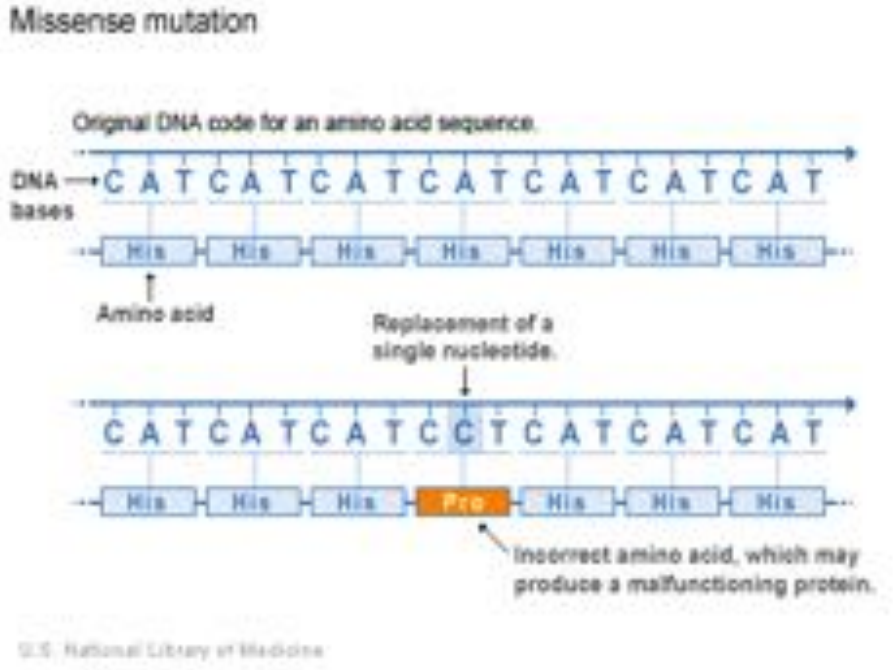
3. Applications of functional annotations

- Functional SNP fine-mapping
- Partitioning heritability and genetic covariance
- Gene-level analysis
- Effect size estimation and risk prediction

Functional annotation in protein-coding genes

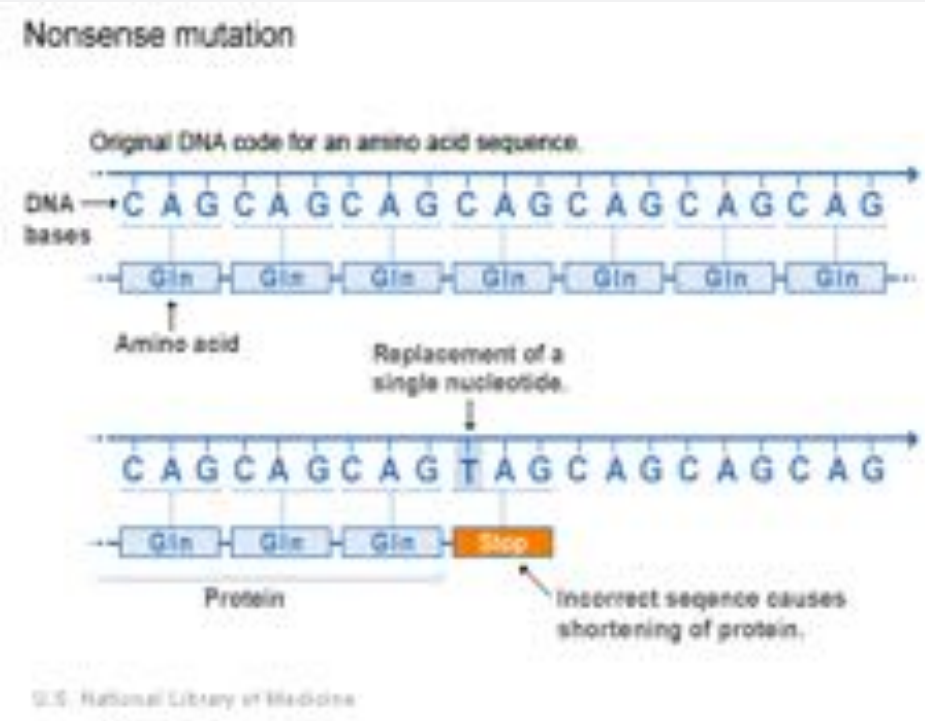
- Synonymous variant
- Missense variant

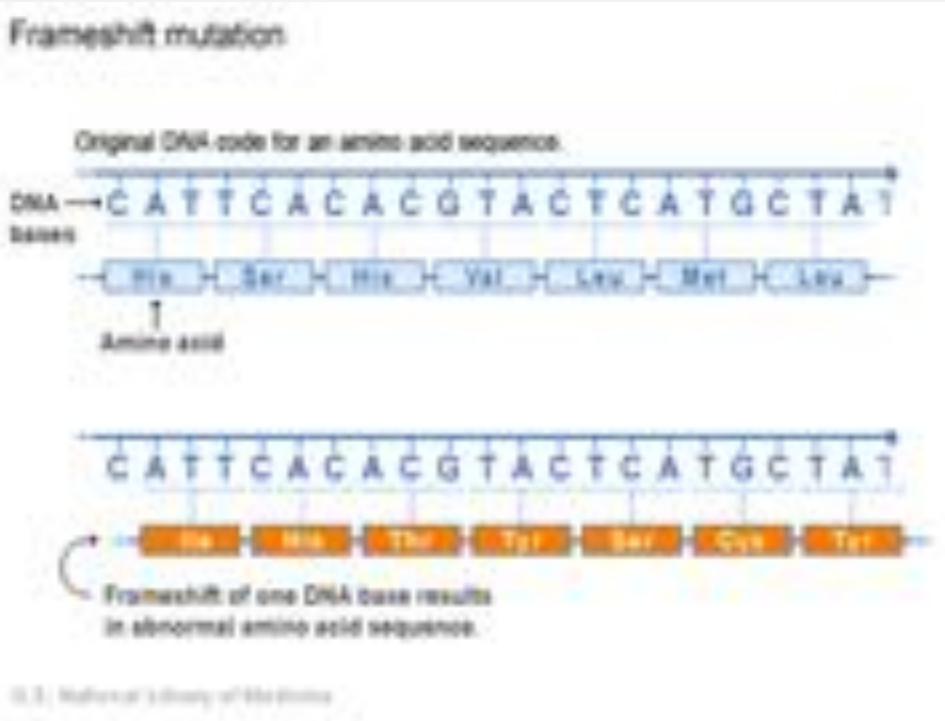
1st letter	Second Letter: U	Second Letter: C	Second Letter: A	Second Letter: G	3rd letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G



Functional annotation in protein-coding genes

- Loss-of-function variant



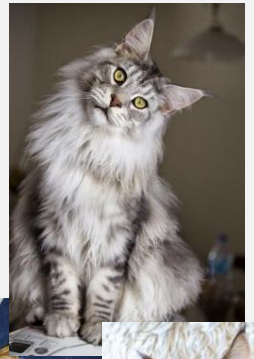


Functional annotation in protein-coding genes

Computation methods (supervised)

We understand the functional mechanism of genes. Training data are also available (OMIM, ClinVar)

- SIFT
- PolyPhen2
- MetaSVM



algorithm

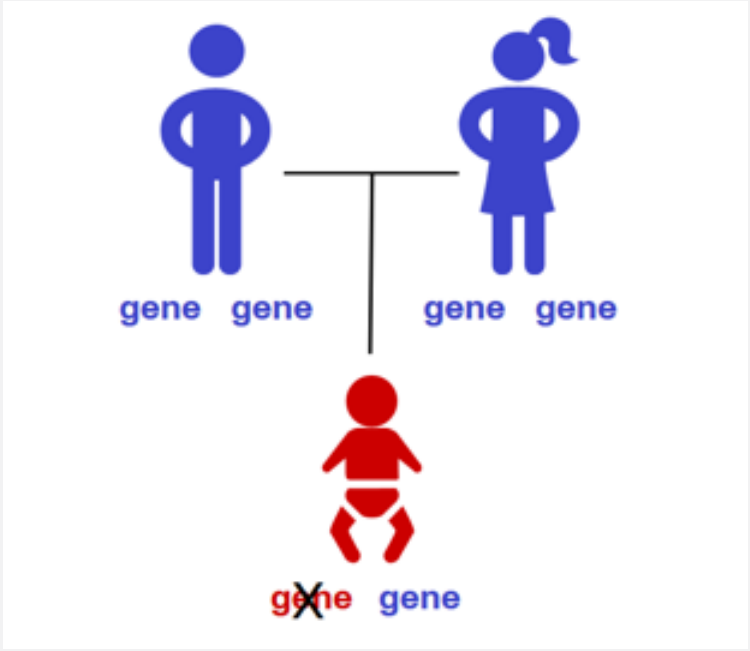
Maine Coon

?

Russian Blue

Functional annotation in protein-coding genes

Application – de novo mutation analysis



Functional annotation in protein-coding genes

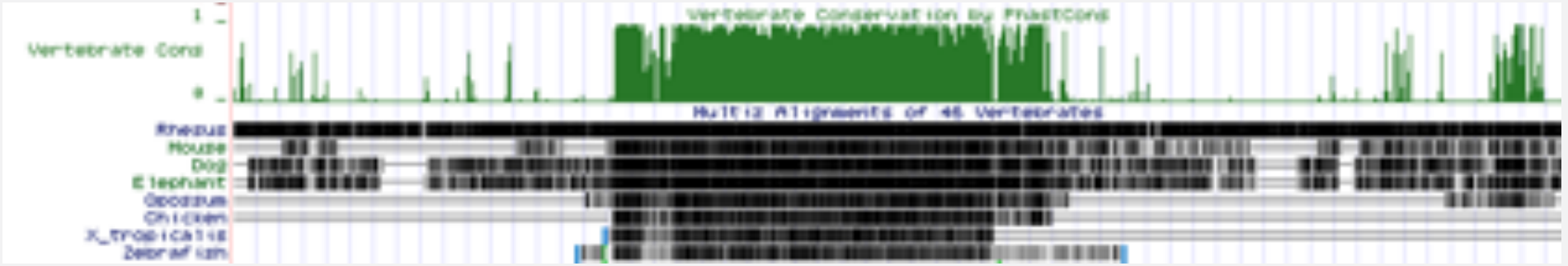
Application – de novo mutation analysis

	Cases, N = 1213						Controls, N = 900					
	Observed		Expected		Enrichment	P	Observed		Expected		Enrichment	P
	n	Rate	n	Rate			n	Rate	n	Rate		
All genes												
Total	1273	1.05	1312.7	1.08	1.0	0.87	925	1.03	979.7	1.09	0.9	0.96
Synonymous	277	0.23	371.4	0.31	0.7	1	229	0.25	277.4	0.31	0.8	1
Missense	846	0.70	824.9	0.68	1.0	0.24	614	0.68	615.6	0.68	1.0	0.53
D-Mis	212	0.17	133.1	0.11	1.6	$\mathbf{1.8 \times 10^{-10}}$	119	0.13	99.3	0.11	1.2	0.03
LoF	150	0.12	116.5	0.10	1.3	**0.0016**	82	0.09	86.7	0.10	0.9	0.71
Damaging	362	0.30	249.5	0.21	1.4	$\mathbf{1.5 \times 10^{-11}}$	201	0.22	186.0	0.21	1.1	0.14
HHE genes												
Total	448	0.37	372.4	0.31	1.2	$\mathbf{7.8 \times 10^{-05}}$	271	0.30	277.7	0.31	1.0	0.66
Synonymous	81	0.07	103.5	0.09	0.8	0.99	80	0.09	77.3	0.09	1.0	0.39
Missense	288	0.24	234.3	0.19	1.2	**0.00038**	163	0.18	174.7	0.19	0.9	0.82
D-Mis	99	0.08	40.6	0.03	**2.4**	$\mathbf{7.7 \times 10^{-15}}$	37	0.04	30.3	0.03	1.2	0.13
LoF	79	0.07	34.5	0.03	**2.3**	$\mathbf{6.2 \times 10^{-11}}$	28	0.03	25.7	0.03	1.1	0.35
Damaging	178	0.15	75.1	0.06	**2.4**	$\mathbf{5.1 \times 10^{-24}}$	65	0.07	55.9	0.06	1.2	0.13
LHE genes												
Total	825	0.68	940.3	0.78	0.9	1	654	0.73	702.1	0.78	0.9	0.97
Synonymous	196	0.16	267.8	0.22	0.7	1	149	0.17	200.1	0.22	0.7	1
Missense	558	0.46	590.5	0.49	0.9	0.91	451	0.50	440.9	0.49	1.0	0.32
D-Mis	113	0.09	92.4	0.08	1.2	0.021	82	0.09	69.0	0.08	1.2	0.069
LoF	71	0.06	82.0	0.07	0.9	0.9	54	0.06	61.1	0.07	0.9	0.83
Damaging	184	0.15	174.4	0.14	1.1	0.24	136	0.15	130.1	0.14	1.1	0.31

Functional annotation in non-coding regions

What about non-coding regions?

DNA conservation

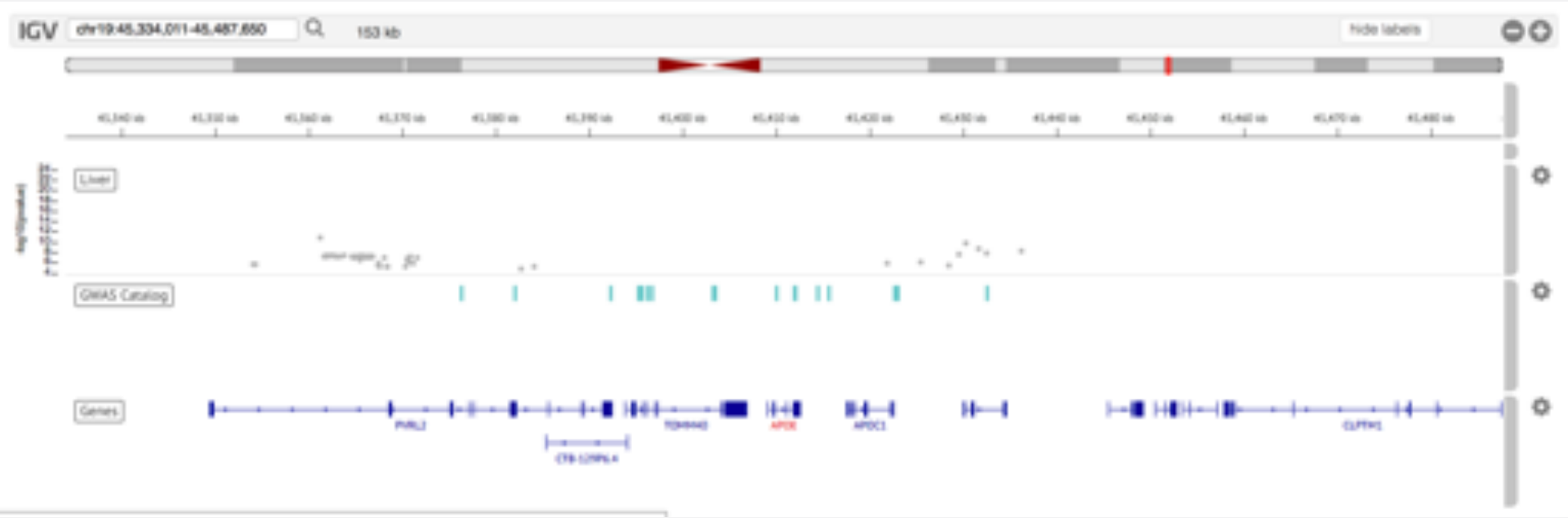


UCSC genome browser

Functional annotation in non-coding regions

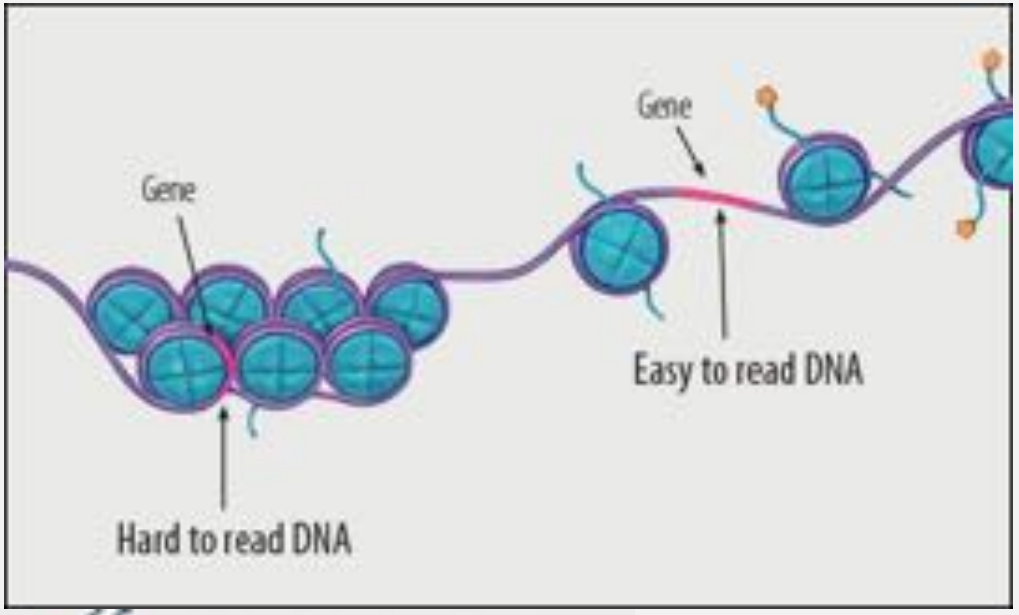
Transcriptomic information

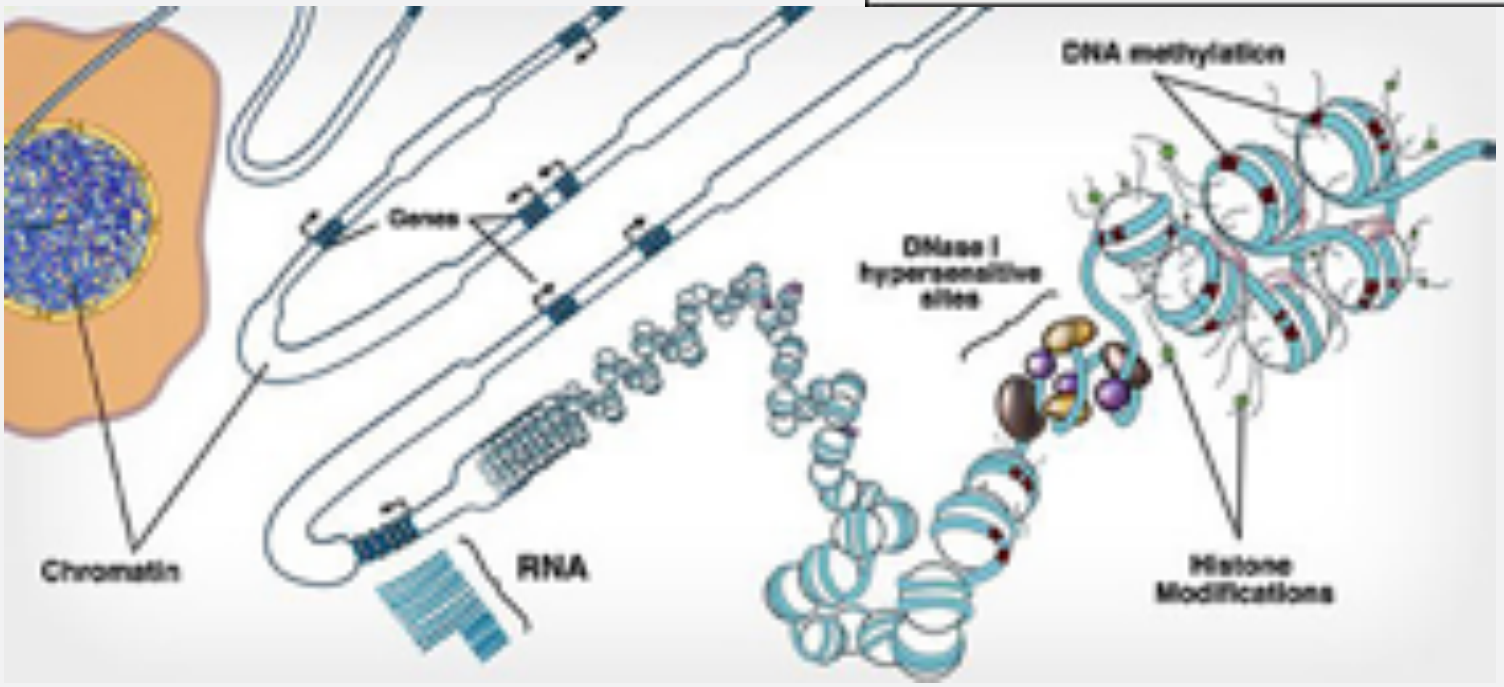
- ncRNA
- eQTL



Functional annotation in non-coding regions

Epigenetic information

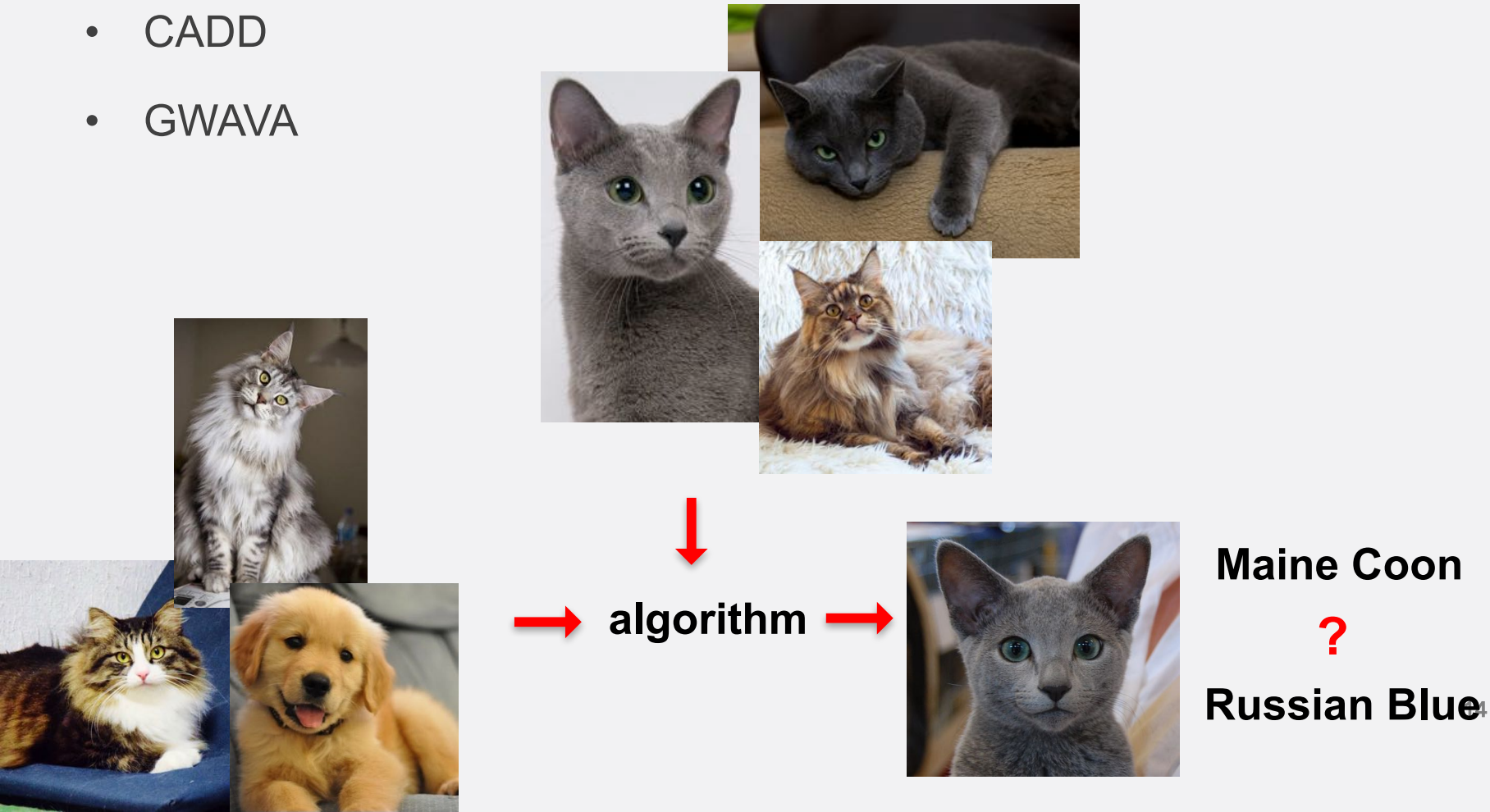




Functional annotation in non-coding regions

Computational methods based on supervised learning

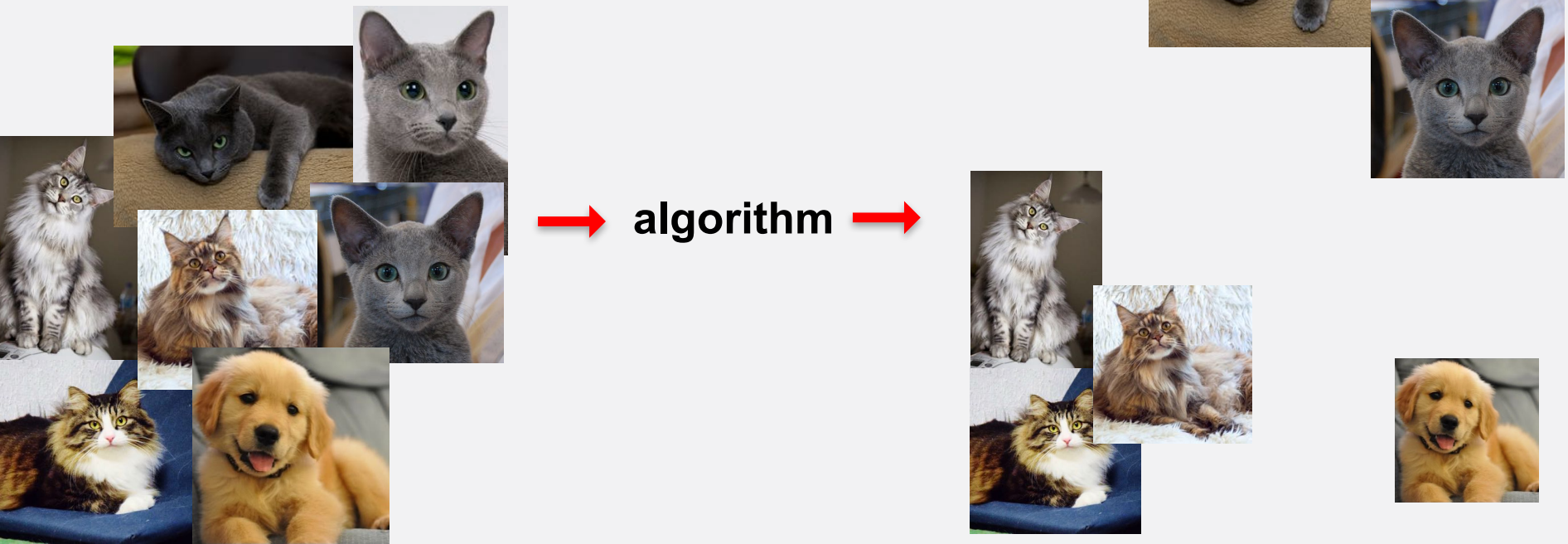
- **Labeled data** + Predictive features + Algorithm = Score
- CADD
- GWAVA



Functional annotation in non-coding regions

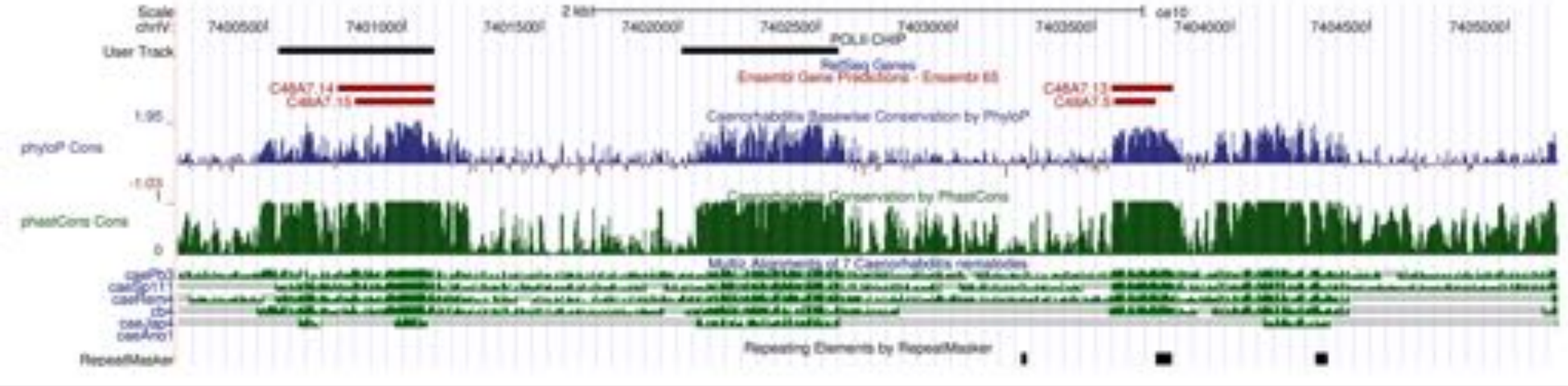
Computational methods based on unsupervised learning

- **Unlabeled data** + Predictive features + Algorithm = Score
- GenoCanyon, GenoSkyline, GenoSkyline-Plus
- EIGEN
- ChromHMM



Other useful tools

UCSC genome browser



Annovar – annotate your variant list using many functional annotations

Gene.refG	GeneDeta	ExonicFun	AAChange	pLi.refGer	pRec.refG	pNull.refG	Gene_full	Function_description.refG	Disease_descriptio	Tissue_specificity(Uni	Expression(ege	Expression(GNF/Atla	P(HI).refG	P(rec).refG	RVIS.refG
ISG15	NM_0051(	.	.	0.009848	0.608249	0.389903	ISG15 ubic	FUNCTION: Ubiquitin-like	DISEASE: Immunod	TISSUE SPECIFICITY: Ex	.	.	0.1	0.22633	-0.11561
ATAD3C	NM_0010(	.	.	4.91E-05	0.867306	0.132645	ATPase fa	.	.	.	.	.	0.16989	.	2.888598
NPHP4	NM_0015	.	.	1.29E-17	0.420065	0.579935	nephrono	FUNCTION: Involved in the	DISEASE: Note=Oli	TISSUE SPECIFICITY: Ex	.	.	0.12343	0.16808	0.569183
DDR2	.	.	.	0.990992	0.009008	3.79E-09	discoidin	FUNCTION: Tyrosine kinas	DISEASE: Spondylo	TISSUE SPECIFICITY: Ex	.	.	0.85011	0.1349	-0.77519
DNASE2B	.	.	.	3.79E-14	0.003092	0.996908	deoxyribo	FUNCTION: Hydrolyzes DN	.	TISSUE SPECIFICITY: Hi	.	.	0.20864	0.10705	0.88376
PRAMEF3	dist=1156	.	.	.	.	.	.	.	.	.	.	.	.	.	.
UBIAD1;D	dist=4396	.	.	.	.	.	.	.	.	.	.	.	.	.	.
LOC10012	dist=8725	.	.	.	.	.	.	.	.	.	.	.	.	.	.
IL23R	.	nonsynon	IL23R:NM_	0.006495	0.98949	0.004014	interleuki	FUNCTION: Associates wit	DISEASE: Inflamma	TISSUE SPECIFICITY: Ex	.	.	0.11254	0.33107	0.795569
ATG16L1	.	nonsynon	ATG16L1:?	0.999737	0.000263	1.67E-11	autophag	FUNCTION: Plays an essen	DISEASE: Inflamma	.	myocardium;sr	dorsal root ganglion;	0.3463	0.10686	0.15076

Outline

1. Background

2. Introduction to (narrow-sense) functional annotations

- Functional annotation in protein-coding genes
- Functional annotation in non-coding regions
- Other useful tools

3. Applications of functional annotations

- Functional SNP fine-mapping
- Partitioning heritability and genetic covariance
- Gene-level analysis
- Effect size estimation and risk prediction

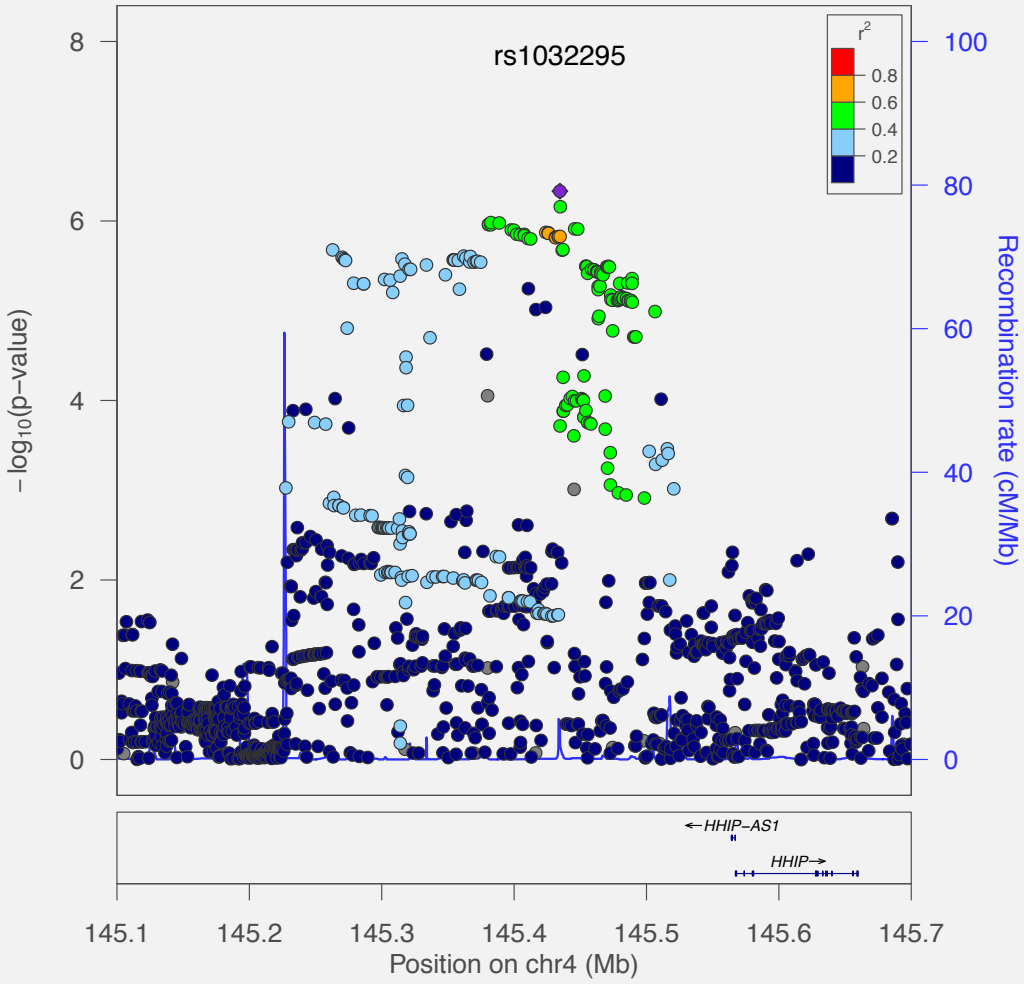
Functional SNP fine-mapping

How to identify functional SNP at a GWAS locus?

- PAINTOR
- FGWAS
- GenoWAP

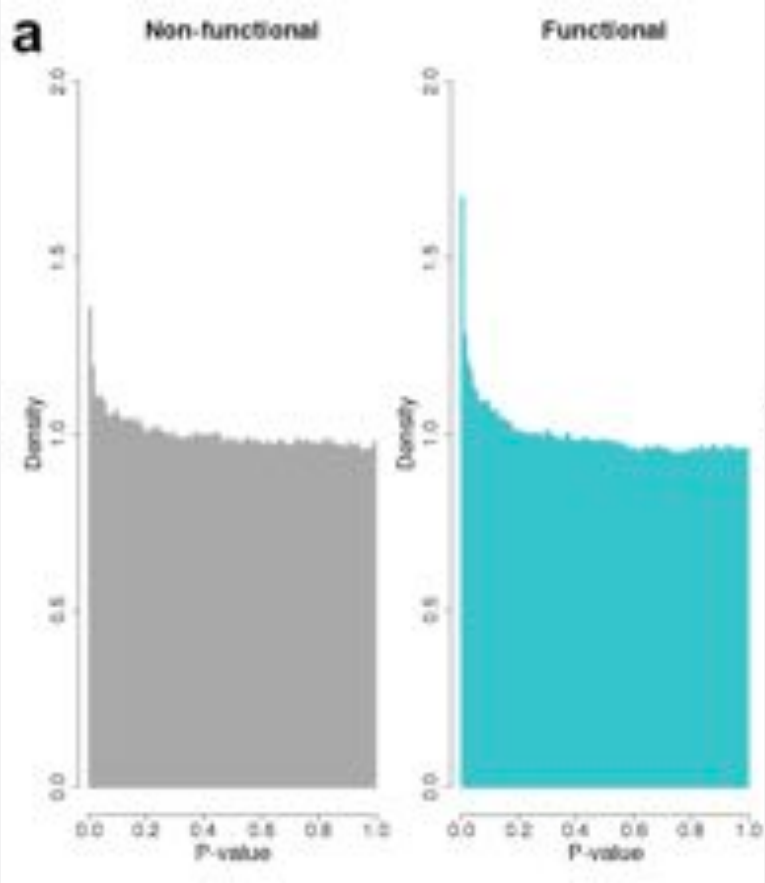
Goal:

$$\mathbb{P}(C \mid P, A)$$

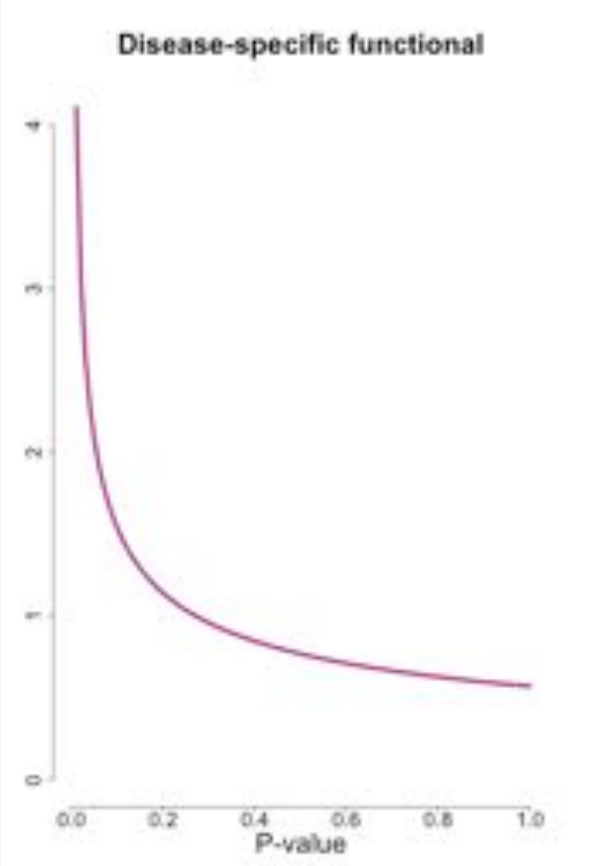


Functional SNP fine-mapping

- GenoWAP



EM algorithm

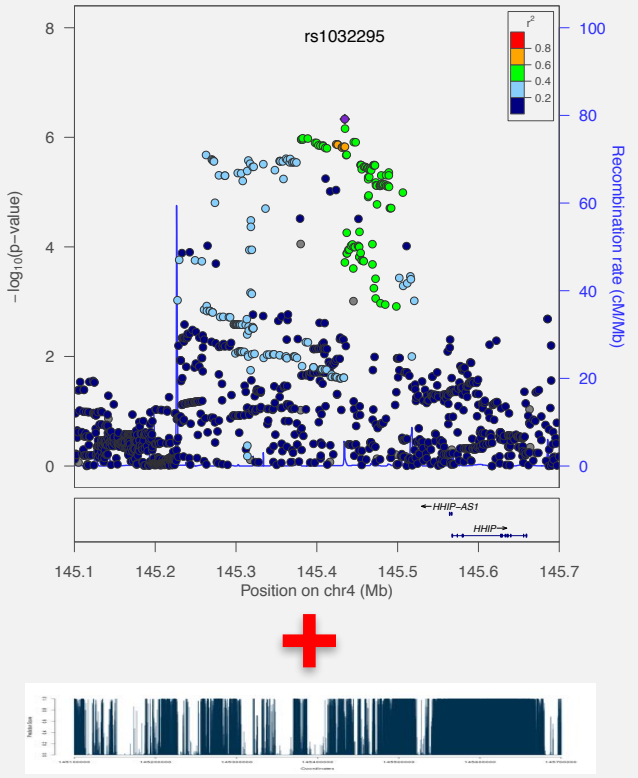


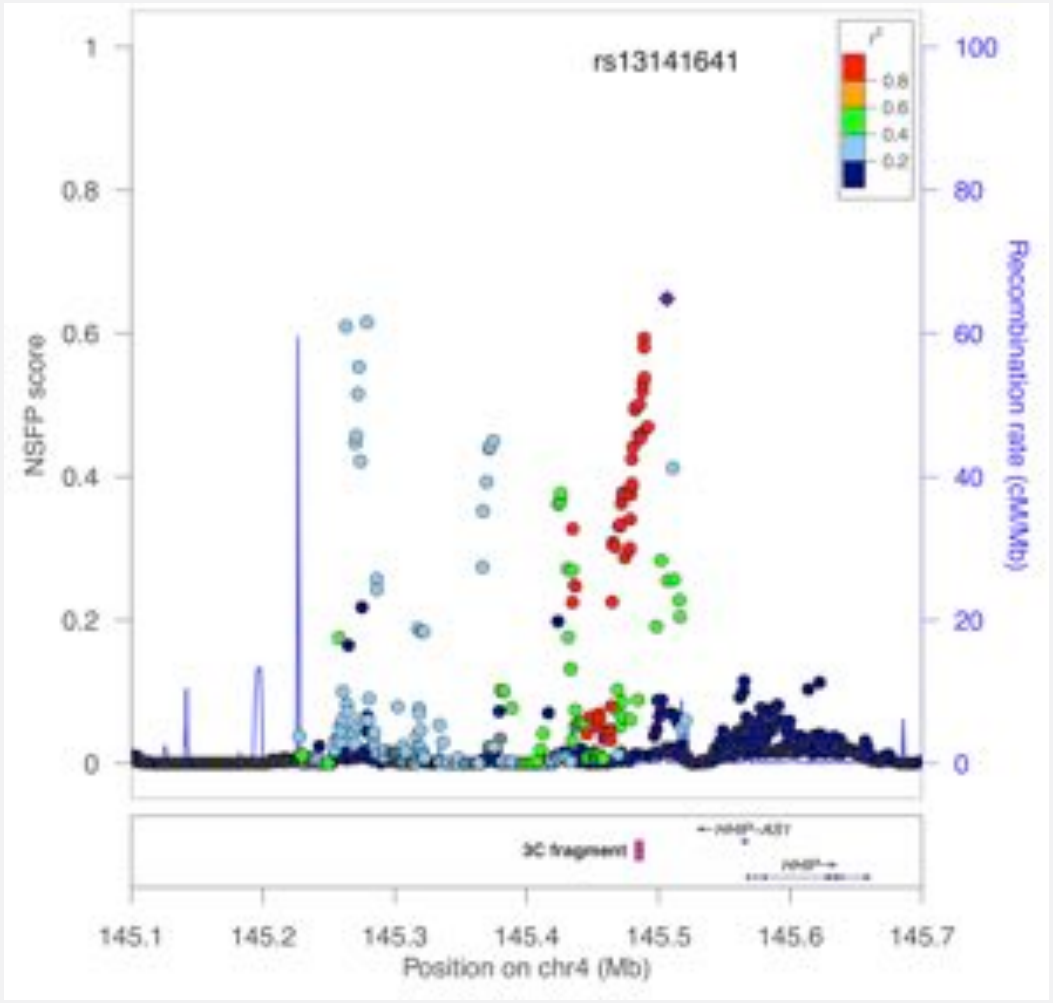
$$\mathbb{P}(C \mid P, A)$$

Lu et al. 2015

Functional SNP fine-mapping

- GenoWAP





Lu et al. 2016

Outline

1. Background

2. Introduction to (narrow-sense) functional annotations

- Functional annotation in protein-coding genes
- Functional annotation in non-coding regions
- Other useful tools

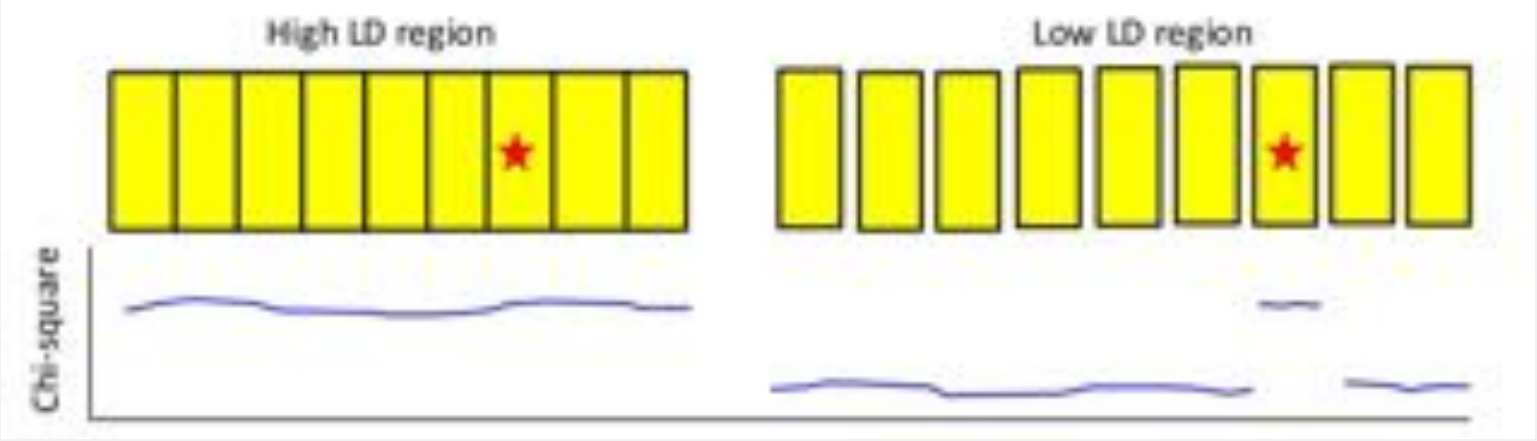
3. Applications of functional annotations

- Functional SNP fine-mapping
- Partitioning heritability and genetic covariance
- Gene-level analysis
- Effect size estimation and risk prediction

Partitioning heritability and genetic covariance

LD score regression

- Estimate (partition) heritability using GWAS summary statistics
- **We expect to see stronger associations in regions with high LD**



Finucane et al. 2015

- It can be shown that this relationship is linear!

sample size

heritability

$$\mathbb{E}(z_j^2) = \frac{Nh^2}{m} l_j + 1$$

GWAS associations

number of SNPs

LD score

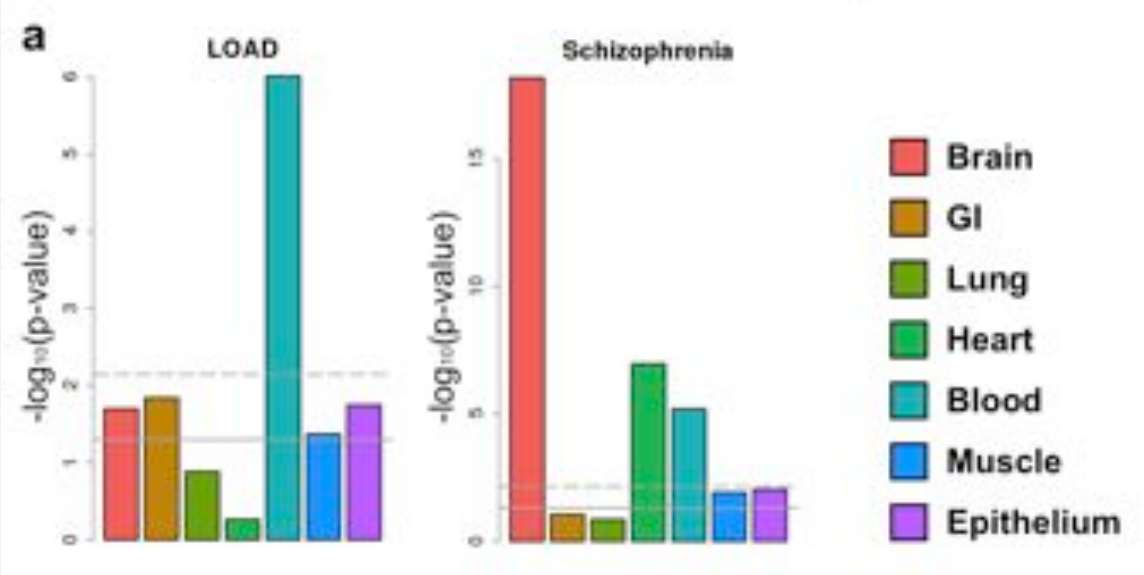
Bulik-Sullivan et al. 2014

Partitioning heritability and genetic covariance

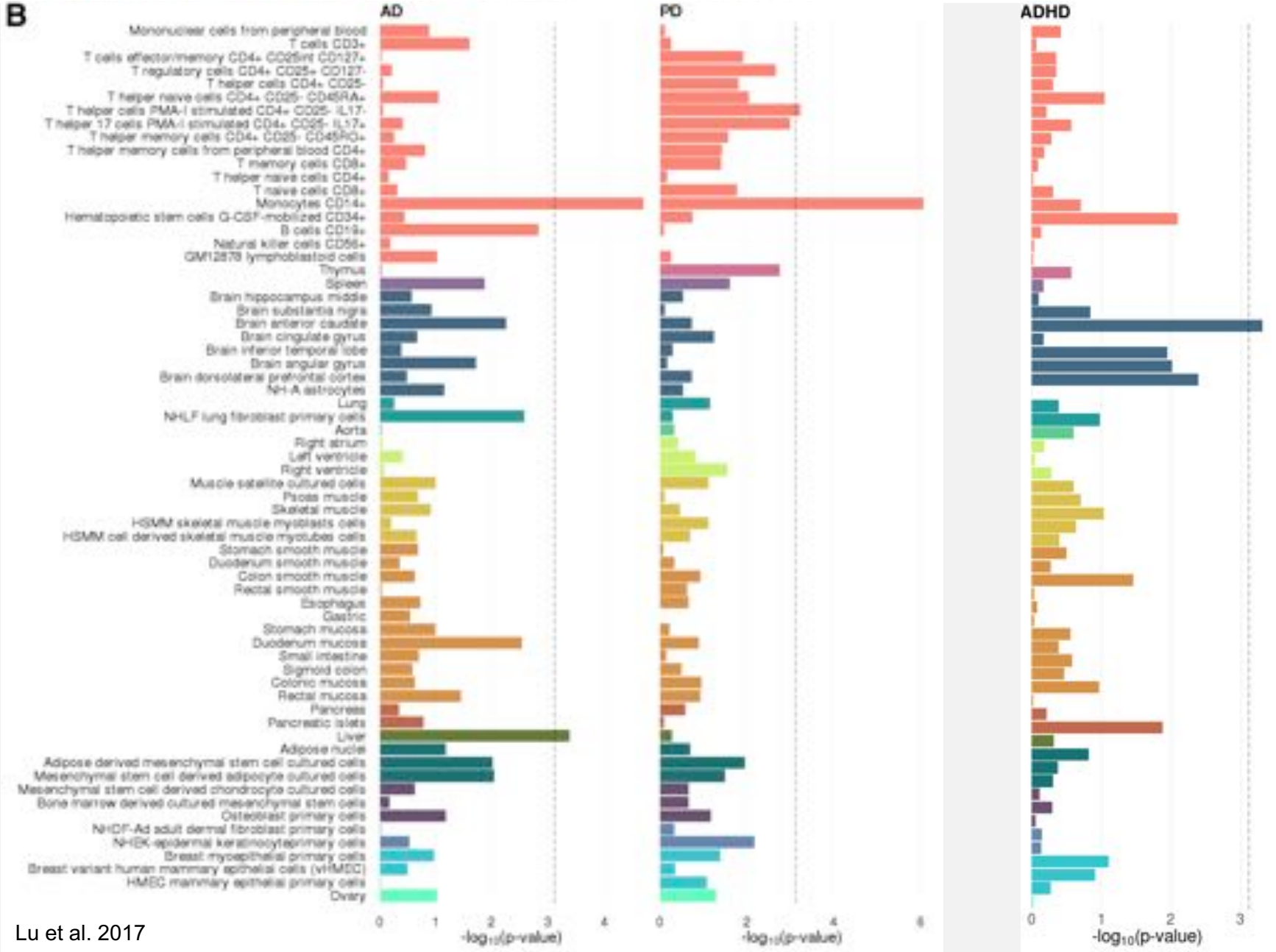
LD score regression

- The model can be extended to partition heritability by **functional annotation**
- This makes it possible to calculate enrichment

$$Enrichment = \frac{\%\ heritability\ explained}{\%\ genome\ covered}$$



Partitioning heritability and genetic covariance



Lu et al. 2017

Partitioning heritability and genetic covariance

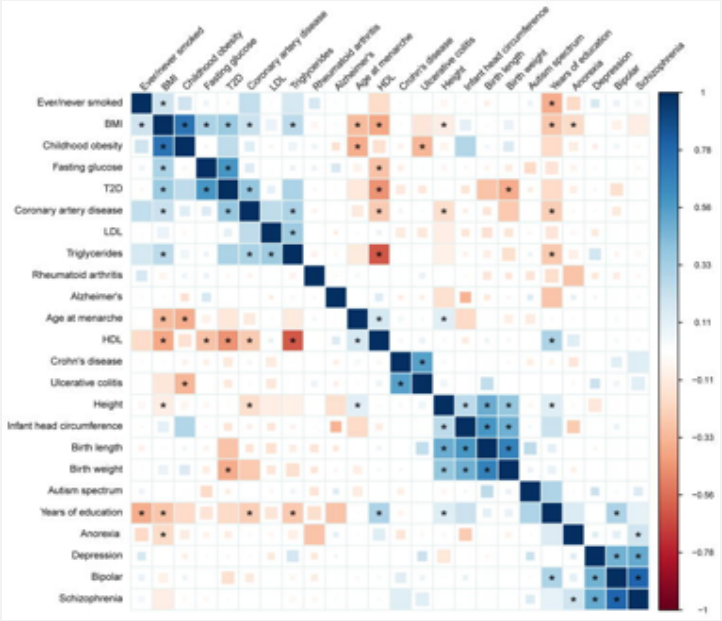
- **Genetic covariance** quantifies shared genetics among complex traits

$$y_1 = \sum_{i=1}^{K} X_i \beta_i + \epsilon$$

$$y_2 = \sum_{i=1}^{K} Z_i \gamma_i + \delta$$

genetic covariance

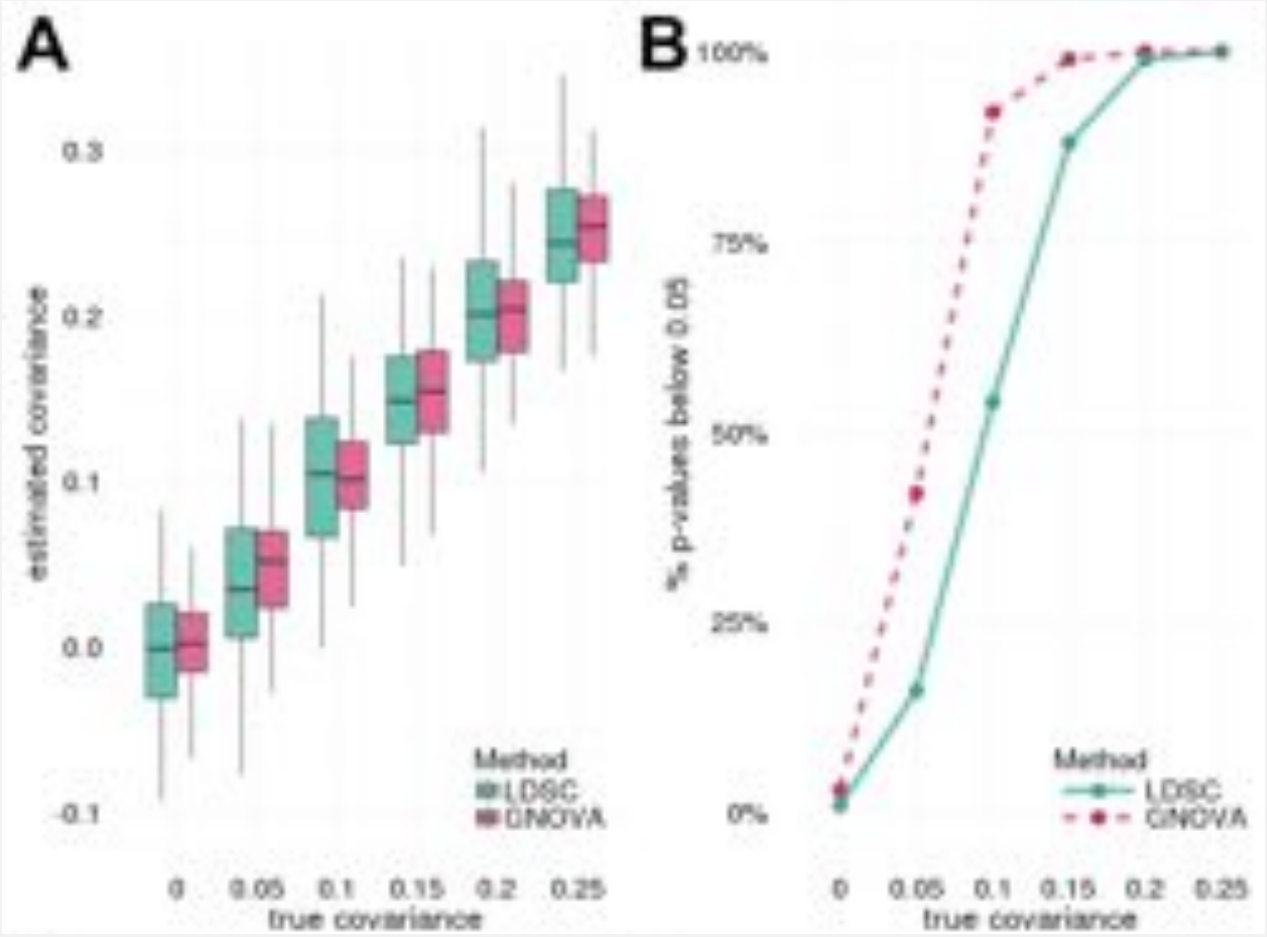
$$\mathbb{E}(\beta_i) = \mathbb{E}(\gamma_i) = 0 \ \ and \ \ \mathbb{E}\left(\gamma_i \beta_i^T\right) = \frac{\rho_i}{m_i} I, \quad i = 1, \ldots, K$$



Bulik-Sullivan et al. 2015

Partitioning heritability and genetic covariance

- **GNOVA**, a principled framework to perform annotation-stratified genetic covariance estimation



Lu et al. 2017

Partitioning heritability and genetic covariance

We dissected the genetic covariance between late-onset Alzheimer's disease (LOAD) and amyotrophic lateral sclerosis (ALS)

LOAD: IGAP phase-I (17,008 cases, 37,154 controls)

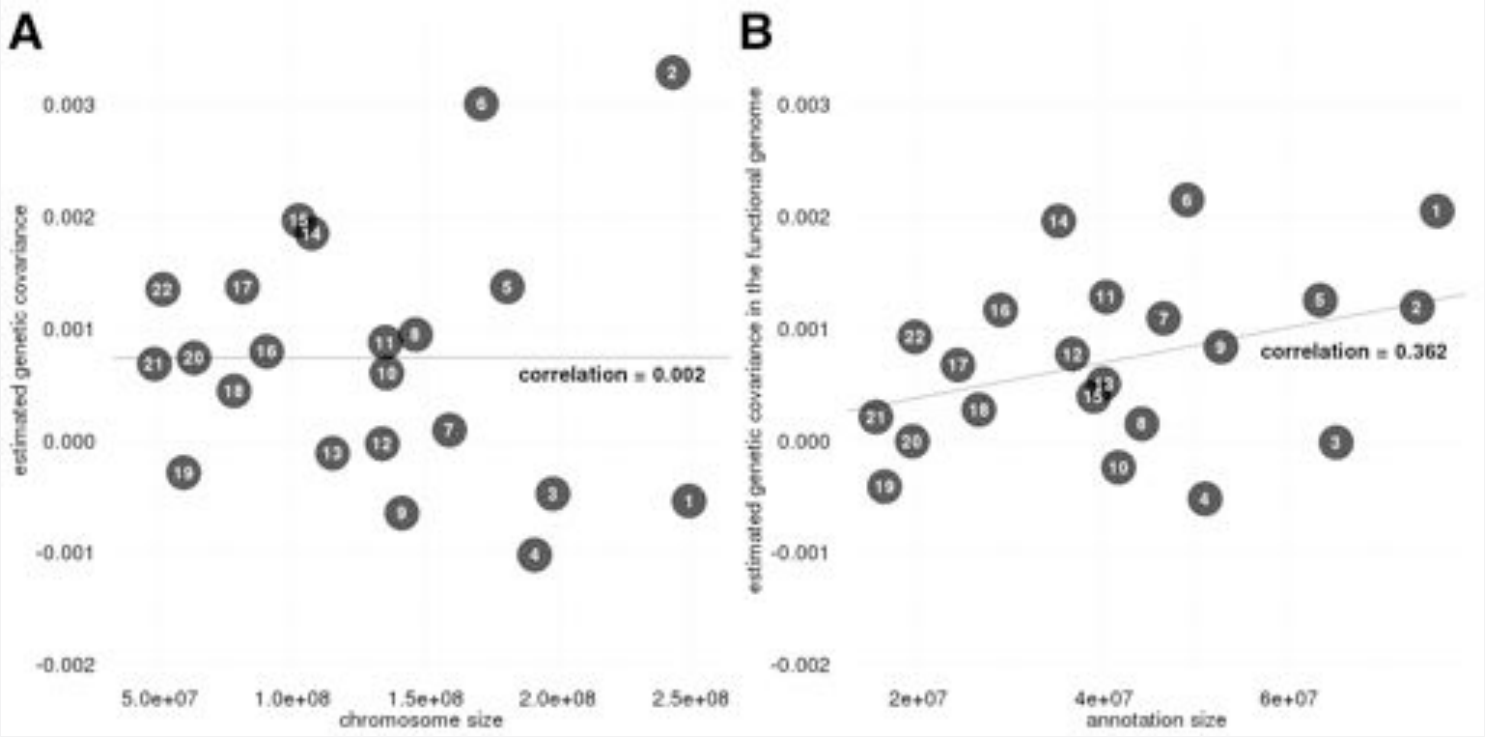
ALS: MinE project (12,577 cases, 23,475 controls)

Annotation	Category	Covariance	P-value
Non-stratified	GNOVA	0.016 (0.004)	**2.0×10^{-4}**
	LDSC	0.012 (0.007)	0.075
GenoCanyon	functional	0.016 (0.004)	**8.2×10^{-5}**
	non-functional	0.003 (0.004)	0.377
MAF	Q1	-0.001 (0.003)	0.842
	Q2	0.003 (0.004)	0.361
	Q3	0.004 (0.004)	0.327
	Q4	0.008 (0.003)	**0.005**

Lu et al. 2017

Partitioning heritability and genetic covariance

Genetic covariance between LOAD and ALS is proportional to the size of the functional genome on each chromosome. This suggests a **polygenic** genetic covariance structure.



Lu et al. 2017

Outline

1. Background

2. Introduction to (narrow-sense) functional annotations

- Functional annotation in protein-coding genes
- Functional annotation in non-coding regions
- Other useful tools

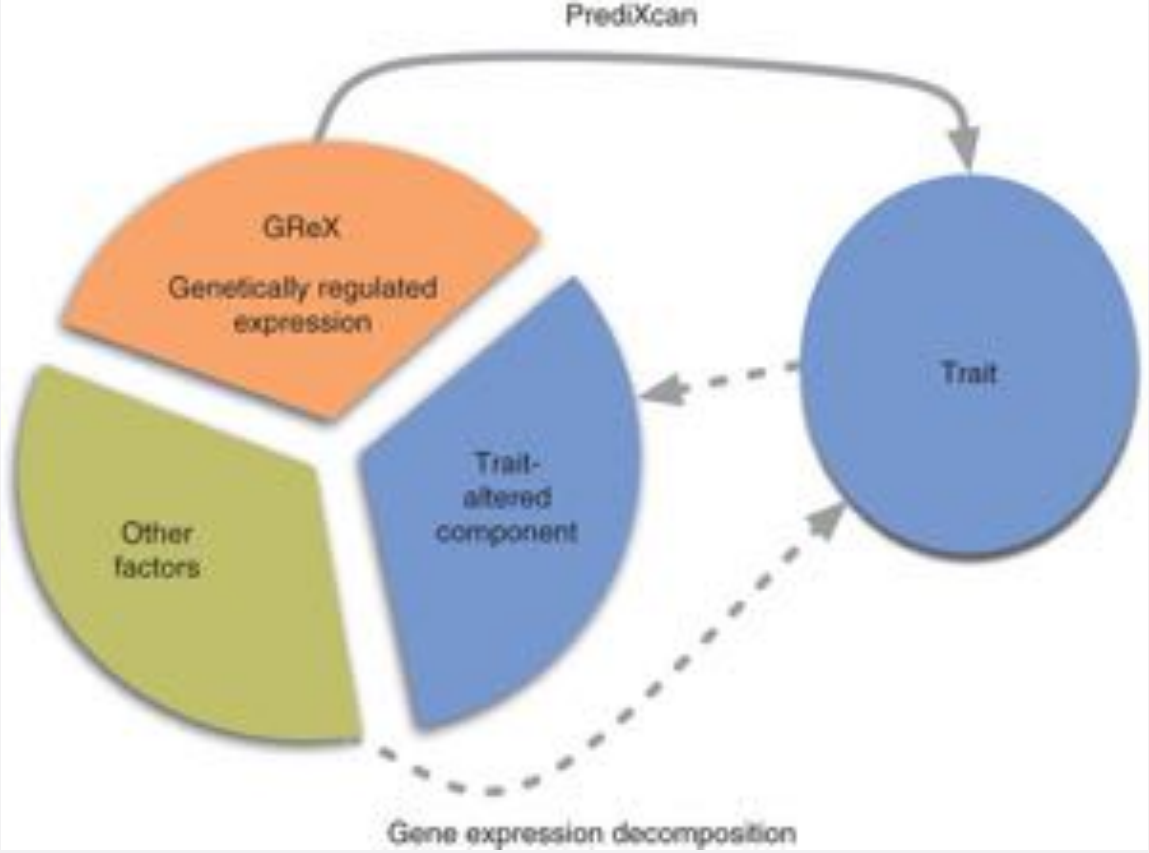
3. Applications of functional annotations

- Functional SNP fine-mapping
- Partitioning heritability and genetic covariance
- Gene-level analysis
- Effect size estimation and risk prediction

Gene-level analysis

PrediXcan and TWAS

- Impute gene expression using genotype information
- Perform association test using imputed expression



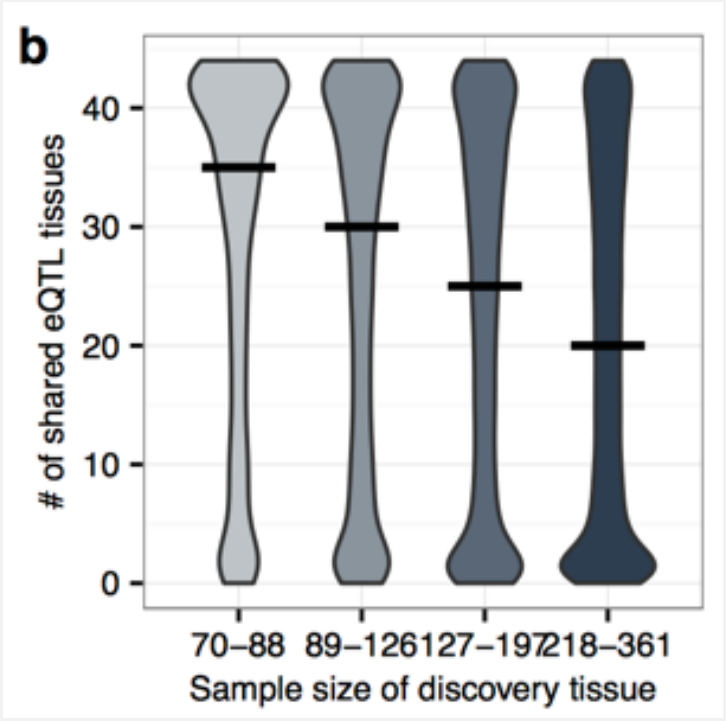
Gamezon et al. 2015

Gene-level analysis

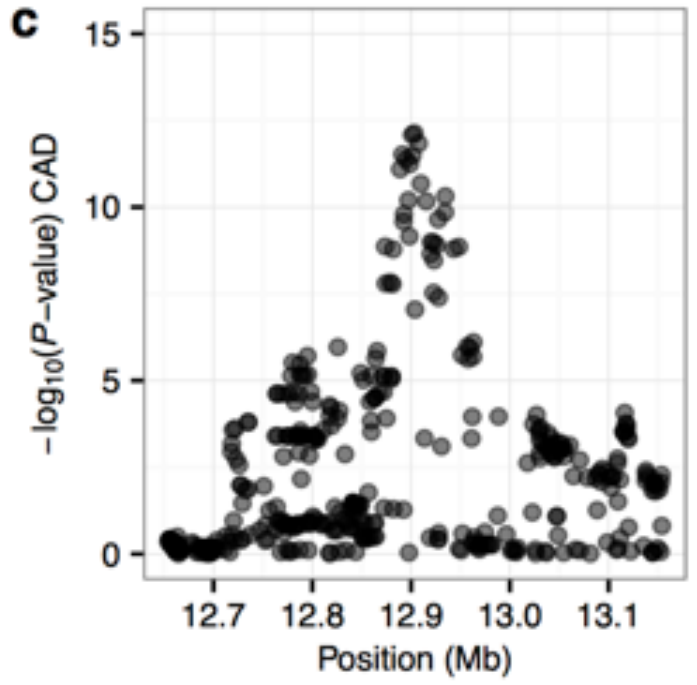
PrediXcan and TWAS

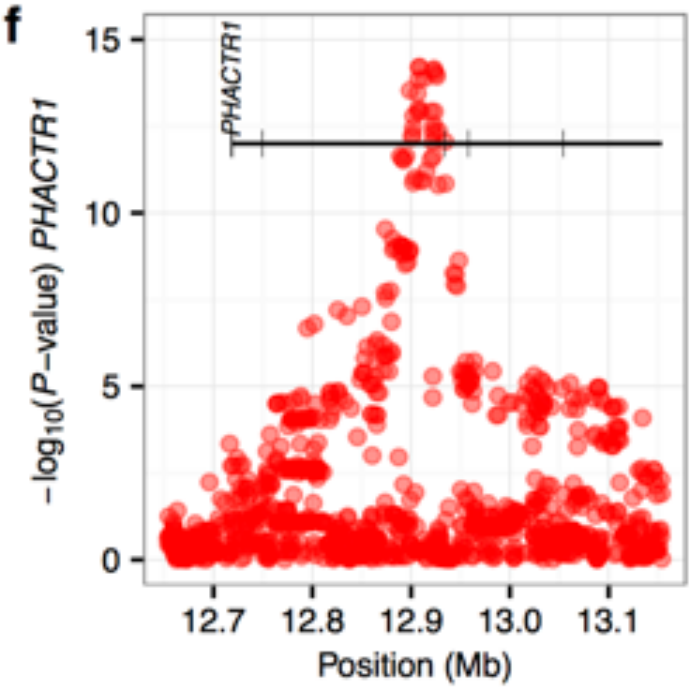
Idea is similar to colocalization

Often identify signals in irrelevant tissues



Aguet et al. 2016



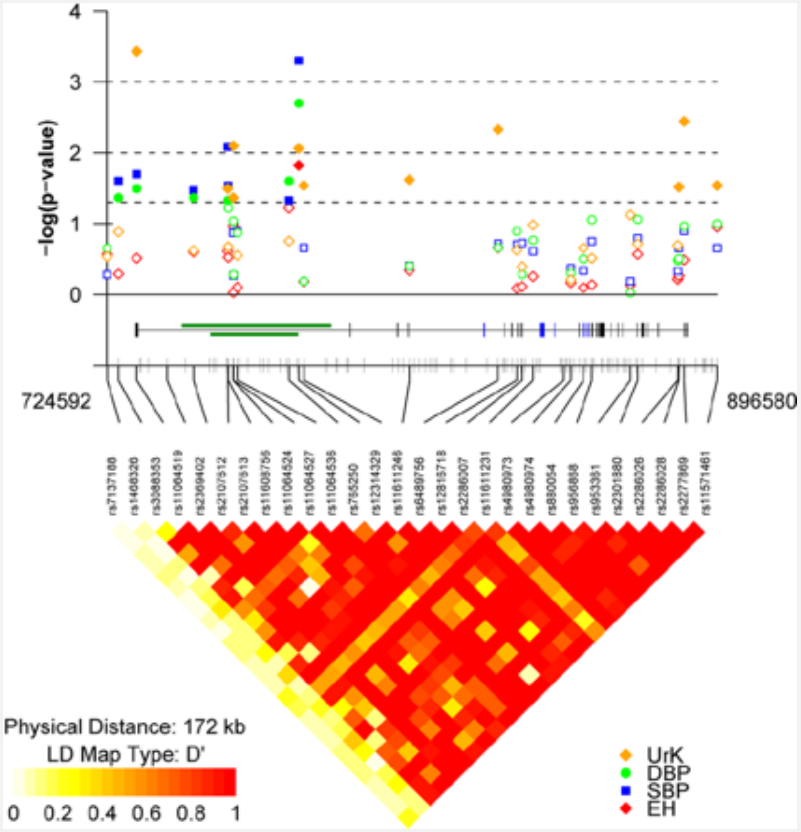


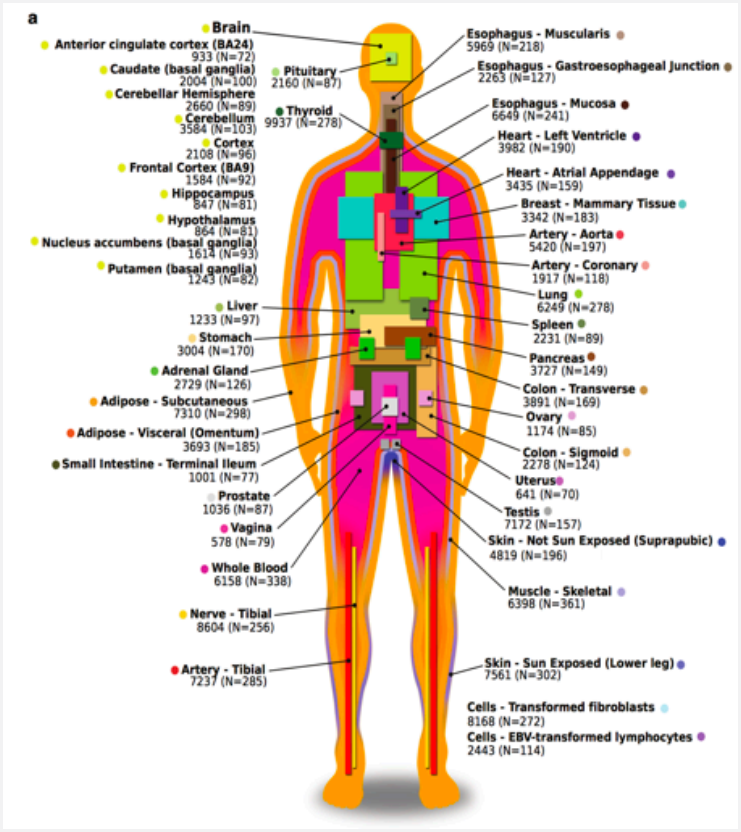
Gene-level analysis

PrediXcan and TWAS

Need a metric to summarize information across all tissues

This is very similar to "**burden test for common SNPs**"





Outline

1. Background

2. Introduction to (narrow-sense) functional annotations

- Functional annotation in protein-coding genes
- Functional annotation in non-coding regions
- Other useful tools

3. Applications of functional annotations

- Functional SNP fine-mapping
- Partitioning heritability and genetic covariance
- Gene-level analysis
- Effect size estimation and risk prediction

Effect size estimation and risk prediction

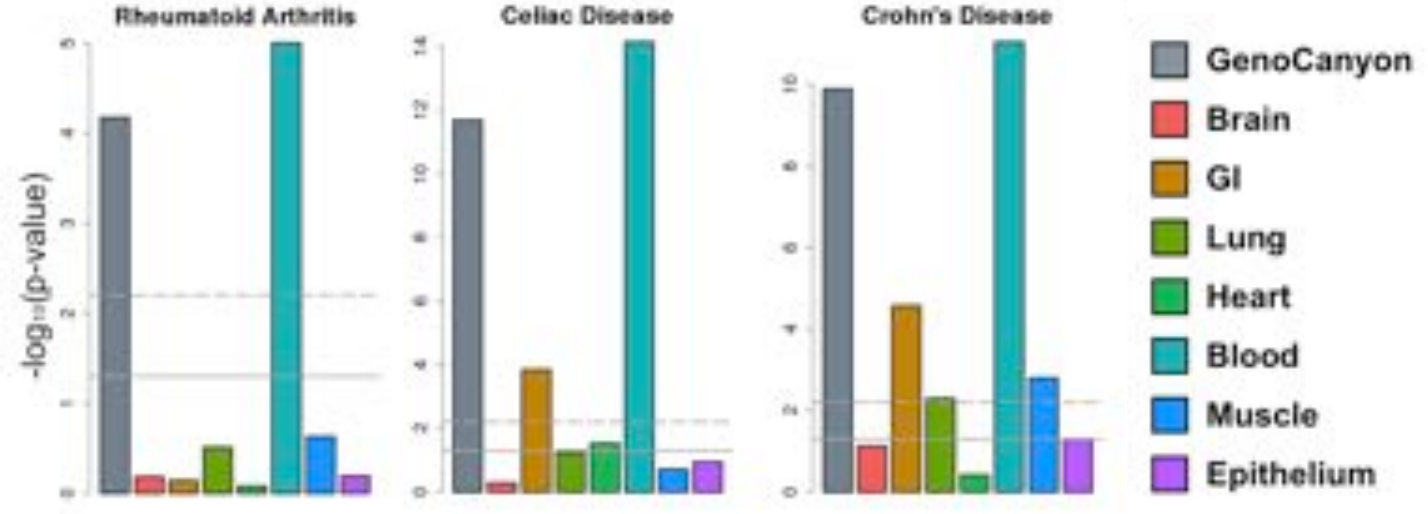
AnnoPred

Remember we can connect disease with tissues?

$$Enrichment = \frac{\%\ heritability\ explained}{\%\ genome\ covered}$$

Such connections can help estimate SNPs' effect sizes

$$\mathbb{E}\left(\beta \mid \hat{\beta}, \widehat{D}\right)$$

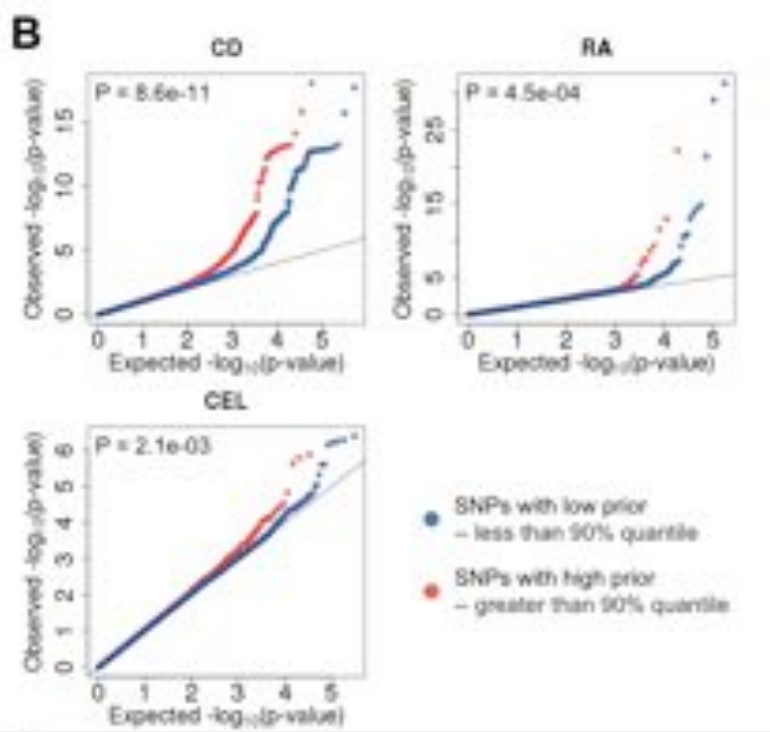


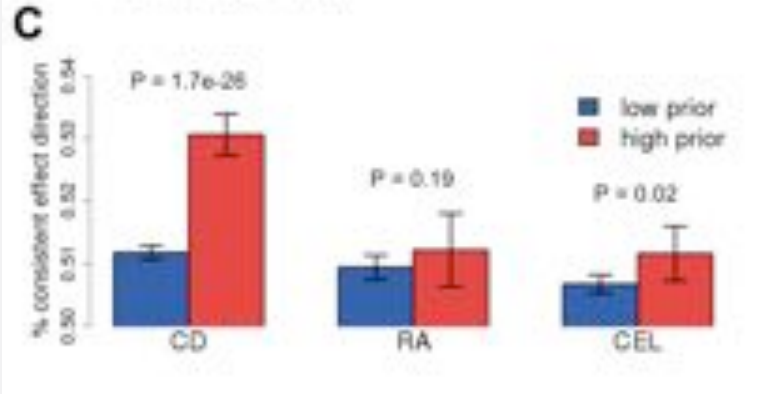
Hu et al. 2017

Effect size estimation and risk prediction

AnnoPred

SNPs with high prior show stronger associations and more consistent effect directions in validation cohorts





Hu et al. 2017

Effect size estimation and risk prediction

AnnoPred

We achieved higher risk prediction accuracy across five complex diseases

$$\hat{y} = X\mathbb{E}(\beta|\hat{\beta}, \hat{D})$$

Table 2. CORs of different methods. The highest CORs are highlighted in boldface.

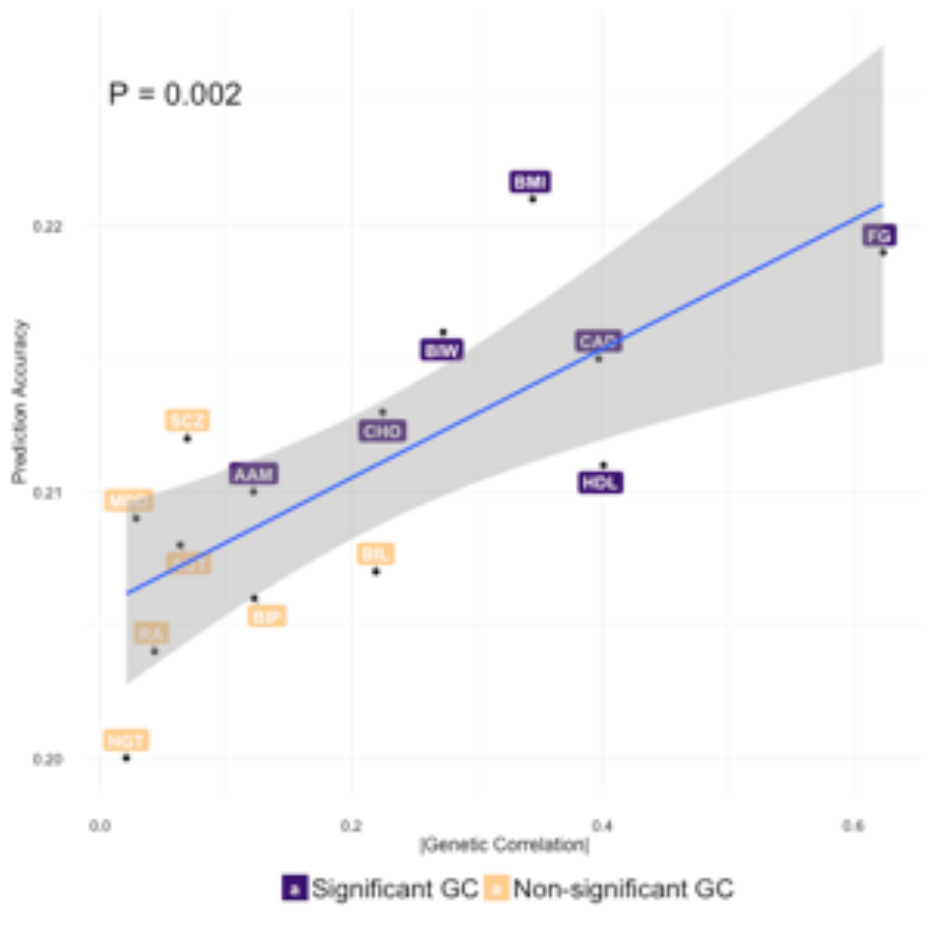
Disease/Trait	PRS_{sig}	PRS_{all}	PRS_{P+T}	LDpred	AnnoPred
Crohn's Disease	0.27	0.229	0.32	0.325	**0.343**
Breast Cancer	0.084	0.055	0.12	0.122	**0.137**
Rheumatoid Arthritis	0.204	0.114	0.248	0.282	**0.287**
Type-II Diabetes	0.165	0.156	0.204	0.202	**0.22**
Celiac Disease	0.11	0.136	0.18	0.197	**0.213**

Hu et al. 2017

Effect size estimation and risk prediction

PleioPred

We have further extended the model to incorporate multiple GWAS for genetically correlated diseases



Hu et al. 2017

Summary

Annotation = External Information

- Conservation
- Epigenetic data
- Transcriptomic data
- Functional prediction scores (supervised / unsupervised)
- Quantitative trait loci (eQTL, sQTL, pQTL)
- Allele frequency (ExAC, gnomad)
- LD (1000 Genomes)
- Additional GWAS
- Chromatin interaction

Summary

Applications

- Fine-mapping (GenoWAP)
- Partition heritability and infer relevant tissue (LDSC+GenoSkyline)
- Partition genetic covariance (GNOVA)
- Gene-level association test (new method coming soon)
- Risk prediction (AnnoPred, PleioPred)


THANK YOU