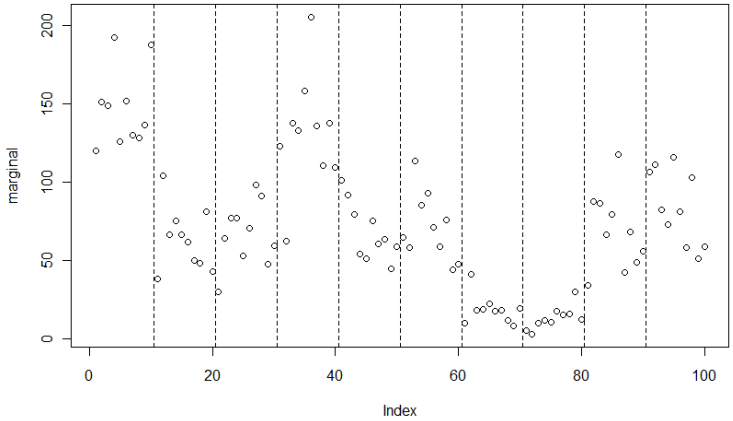
Statistical fine-mapping of GWAS signals

Bogdan Pasaniuc
Associate Professor
Computational Medicine
UCLA
@bpasaniuc
pasaniuc@ucla.edu

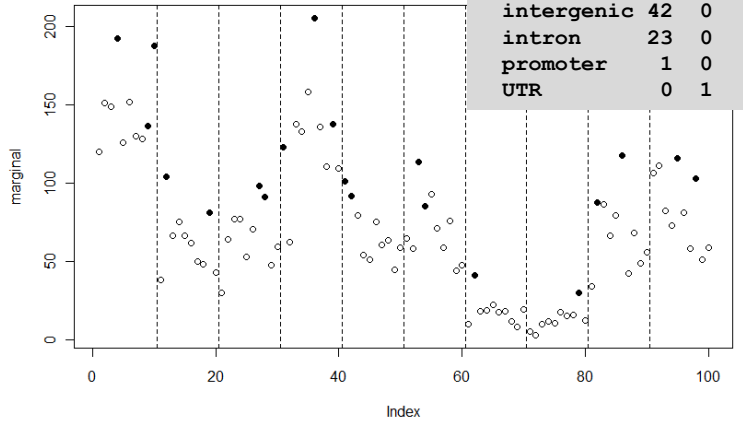
Advanced gene mapping course, January 2020

Motivating example:

- You've conducted a GWAS using in ~100,000 individuals.
- 10 regions are GWAS significant.
- You have resources to conduct follow-up studies of 10 SNPs.
- How do you chose the SNPs for functional follow-up?



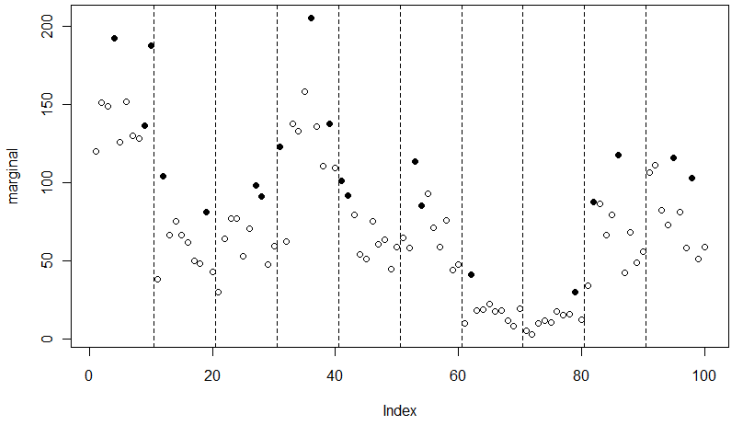
How would you choose the ten SNPs to follow up? How would you use the genetic association data? What other information could you use? How would you combine the association results with this additional information?



	truth	
	0	1
dhs	14	19
intergenic	42	0
intron	23	0
promoter	1	0
UTR	0	1

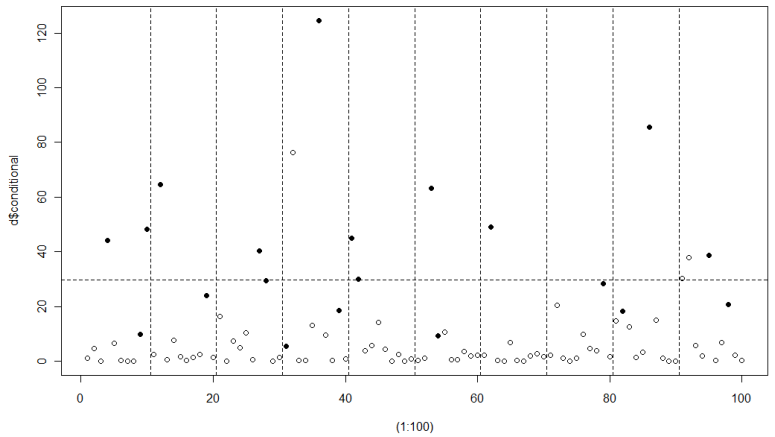
Filled in points = true causal SNPs

Global ranking emphasizes regions with high effect/LD



Filled in points = true causal SNPs

Conditional



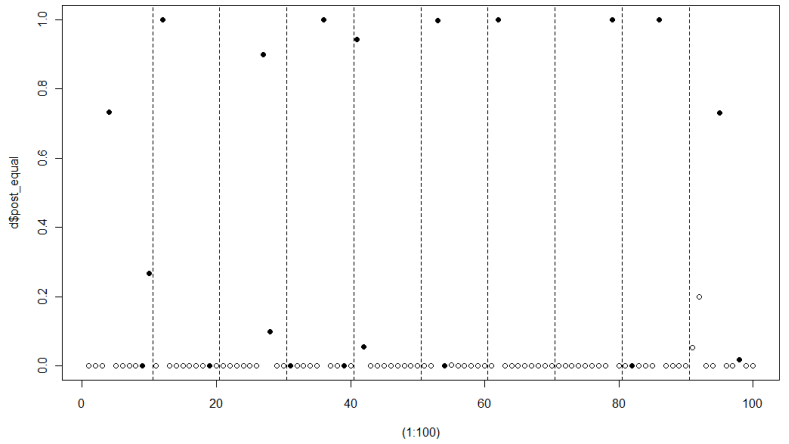
Stepwise conditional analysis

Does this analysis match the goal?

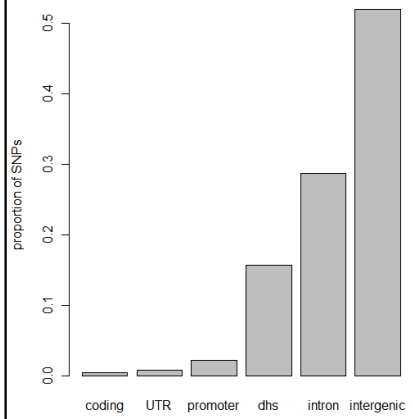
Maybe not: the more “LD friends” a causal variant has, the more likely the stepwise approach will pick one of its “friends”

This is particularly concerning when power is modest (unlike in our well-powered example)

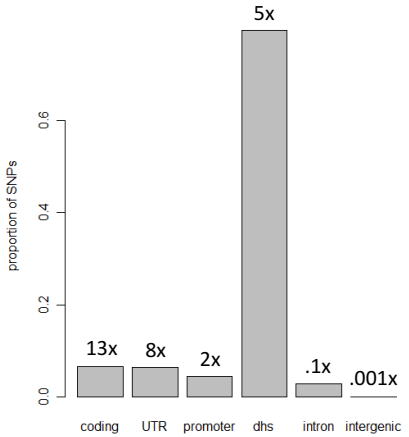
Posterior: 1 causal



Proportion of 1,000 Genomes SNPs in different categories



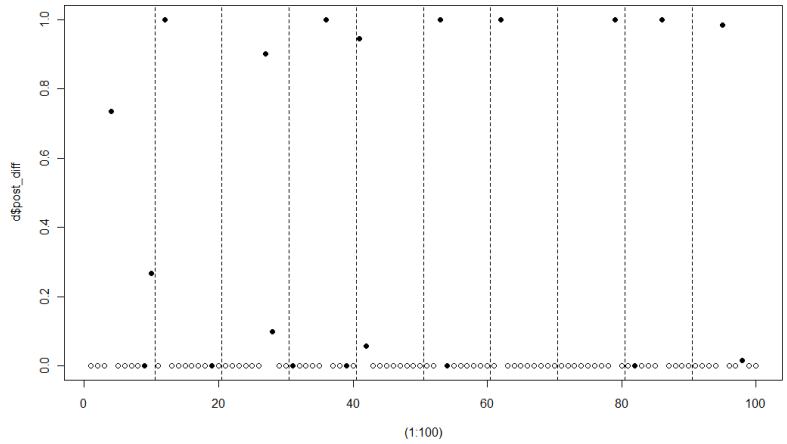
Proportion of 1,000 Genomes SNPs that are causally associated with complex traits



Gusev et al. Regulatory variants explain much more heritability than coding variants across 11 common diseases. AJHG 2014

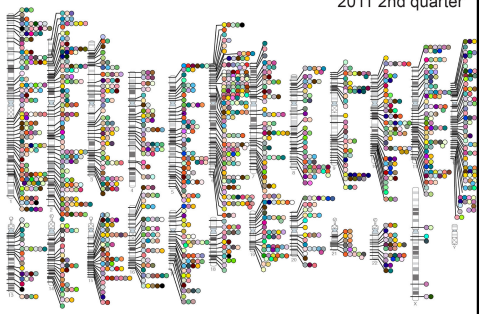
Multiple lines of evidence and several different types of analysis suggest that certain functional categories are enriched for SNPs associated with complex traits. We might infer that causal SNPs are similarly enriched.

Posterior: 1 causal with different weights



Thousands of loci found by GWAS

- GWAS ➔ >1,000 variants associated to various diseases
- Variants at any locus are highly correlated (LD)
- GWAS-Variants are **not causal**

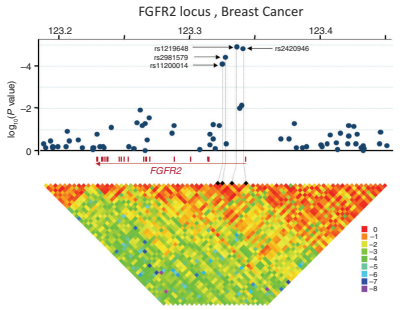


NHGRI GWA Catalog, www.genome.gov/GWAStudies

How do we identify causal variants at GWAS loci?

Fine mapping studies

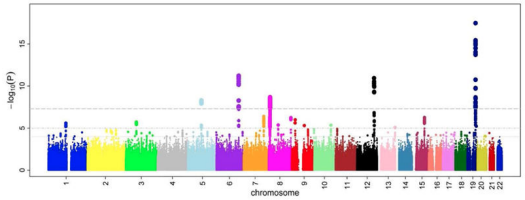
- **Goal:** find causal variants at GWAS associated loci
- Sample size required for differentiating SNPs in LD[Udler et al GenEpi 2010]:
  - $r^2(0.3)$ -> N=3,600 (maf=0.3, R=1.2)
  - $r^2(0.8)$ -> N=12,500 (maf=0.3, R=1.2)



[Hunter et al NG 2007]

Large sample sizes required for fine mapping!

An optimization approach to fine-mapping



- Many GWAS associated loci
  - Height (>180), BC(~30), PC(~30)…
- 2-step approach:
  1. Select set of SNPs for functional validation from considered loci
  2. Test set of variants in functional assays
- **Goal:** find as many causals within fixed budget

Methods for fine-mapping

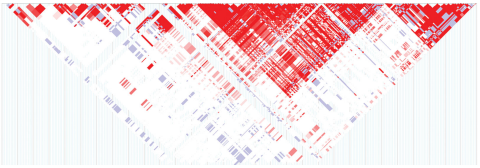
- GOAL: prioritize variants for functional assays
- Prioritization approaches:
  - Marginal association statistics
    - Ignores LD ➔ suboptimal performance
  - Conditioning approaches
    - No clear strategy for selecting variants to condition on
    - Stopping condition?
    - Does not correctly model LD for prioritizing causal variants
  - Probabilistic approach
    - Estimates probability of each variant to be causal
    - Can integrate ENCODE genomic features as priors

Statistical model ➔ causal SNP probabilities

(statistical fine-mapping, see Schaid et al NRG 2018)

Pervasive Linkage Disequilibrium (LD) in the human genome

- Neighboring SNPs are inherited together on haplotypes
- Block-like correlation structure across the human genome



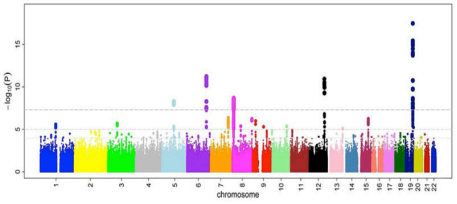
Barret et al. *Bioinformatics* 2005

Only need to measure a subset of the markers for effective GWAS!

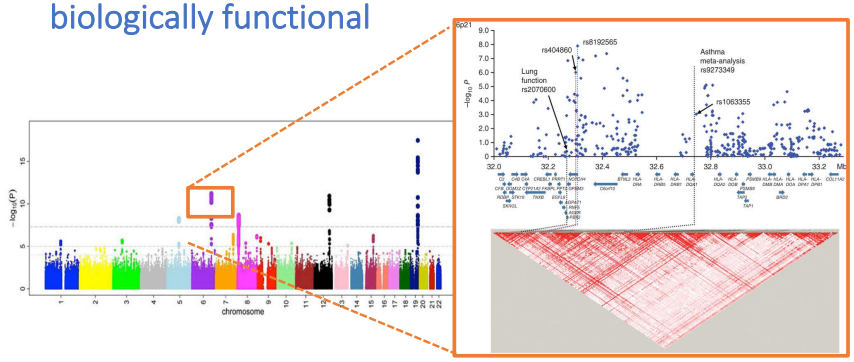
- Neighboring SNPs are inherited together on haplotypes
- Block-like correlation structure across the human genome
- Fill in the rest through imputation

Marchini et al. *Nat Rev Genet* 2010

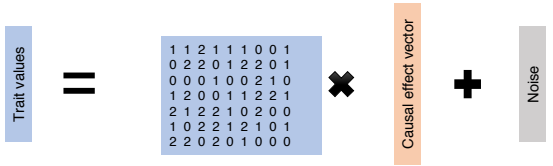
Groups of neighboring variants will display significant associations due to LD



Fine mapping aims to figure which variants are responsible for the observed association → biologically functional



Basic linear model for trait



$$\underset{[n,1]}{y} = \underset{[n,m]}{X} * \underset{[m,1]}{\beta} + \epsilon$$

[n,1]

n: individuals ***m***: SNPs ***X***: standardized genotype matrix var(***y***)=1

GWAS marginal effect is biased due to LD

GWAS effect size at SNP i:

$$\widehat{\boldsymbol{\beta}_{\boldsymbol{GWAS,i}}} = \frac{1}{n}X_i y = \frac{1}{n}X_i([X_1 \cdots X_p]\beta + \epsilon)$$
$$= \left[\frac{1}{n}X_i X_1 \cdots \frac{1}{n}X_i X_p\right]\beta + \frac{1}{n}X_i\epsilon$$
$$= \boldsymbol{\Sigma_{j=1}^{p}\, r_{ij}\beta_j + \frac{1}{n}X_i\epsilon}$$

Or in matrix notation:

$$\widehat{\boldsymbol{\beta}_{\boldsymbol{GWAS}}} = \boldsymbol{MVN}(\boldsymbol{V\beta}, \frac{\boldsymbol{V(1-h_g^2)}}{\boldsymbol{n}})$$

Am. J. Hum. Genet. 69:1–14, 2001

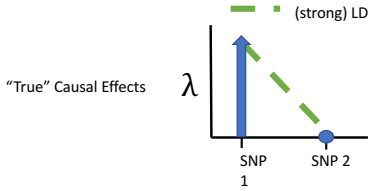
REVIEW ARTICLE
Linkage Disequilibrium in Humans: Models and Data

Jonathan K. Pritchard[1] and Molly Przeworski[2]
Department of Statistics, University of Oxford, Oxford

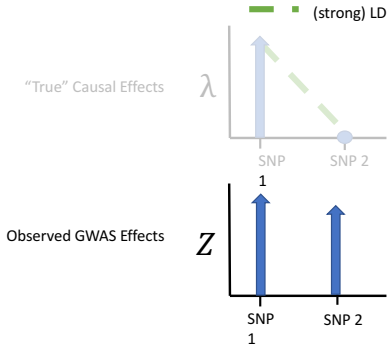
In this review, we describe recent empirical and theoretical work on the extent of linkage disequilibrium (LD) in the human genome, comparing the predictions of simple population-genetic models to available data. Several studies report significant LD over distances longer than those predicted by standard models, whereas some data from short, intergenic regions show less LD than would be expected. The apparent discrepancies between theory and data present a challenge—both to modelers and to human geneticists—to identify which important features are missing from our understanding of the biological processes that give rise to LD. Salient features may include demographic complications such as recent admixture, as well as genetic factors such as local variation in recombination rates, gene conversion, and the potential segregation of inversions. We also outline some implications that the emerging patterns of LD have for association-mapping strategies. In particular, we discuss what marker densities might be necessary for genomewide association scans.

Pritchard&Przeworski AJHG 2001, Shi et al, AJHG 2016, ...

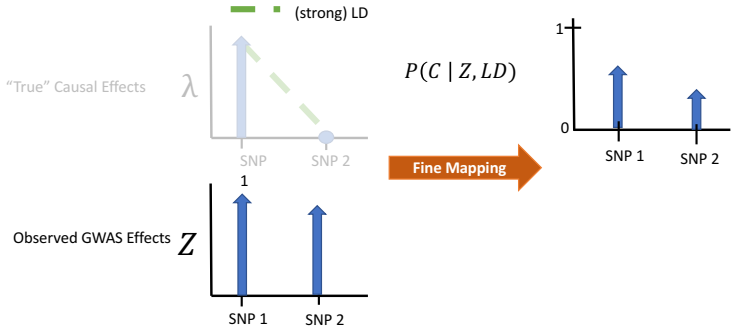
LD induces correlations between causal and non-causal SNPs



LD induces correlations between causal and non-causal SNPs



Given the correlation structure and association strength, quantify the probability that SNPs explain the signal



CAVIAR, CAVIARBF, FINEMAP, PAINTOR, Susie, etc…
(nicely reviewed in Schaid et al NRG 2018)

Approximate estimation of posterior probabilities (1-causal assumption)

- Assumptions
  - one causal SNP per locus
  - causal variant is typed

$$P(c \mid \bar{s}) = \frac{P(\bar{s} \mid c)P(c)}{P(c)}$$

- Marginal association statistics are sufficient to estimate posterior [Maller et al Nat Gen 2012]
  - Can be extended to frequentist approach
  - Statistics at nearby SNPs are independent of phenotype conditional on causal variant
- "Robust to misspecifications"
- Can estimate confidence sets

Allowing for multiple causals improves accuracy

- Causal status vector ➔

$$C = \{c_1, c_2, \cdots c_m\}$$

- Use Bayes to compute probability of each vector

$$P(C|S) = \frac{P(S|C)P(C)}{P(S)}$$

- Model for association statistics

Causal status (Unknown)

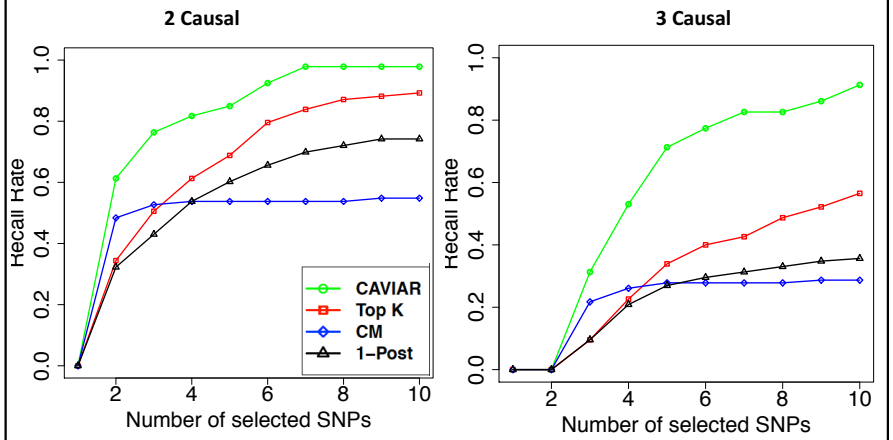
$$S \sim MVN(\lambda_c \Sigma C, \Sigma)$$

Z-score (Known)

NCP of causal SNP (Known)

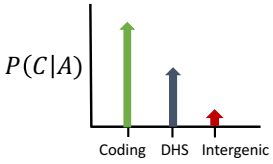
LD pattern (Known)

Allowing for multiple causals improves accuracy

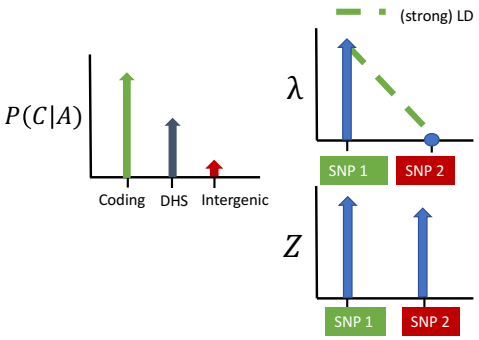


[Hormozdiari et al. Genetics 2014]

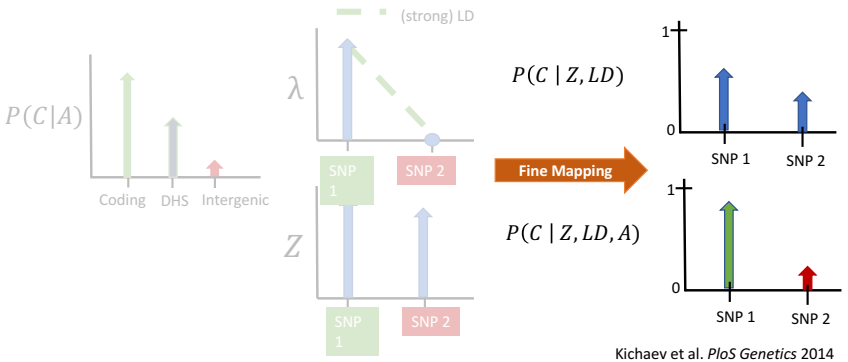
Probabilistic models provide a principled way to incorporate prior biological knowledge



Probabilistic models provide a principled way to incorporate prior biological knowledge



Probabilistic models provide a principled way to incorporate prior biological knowledge

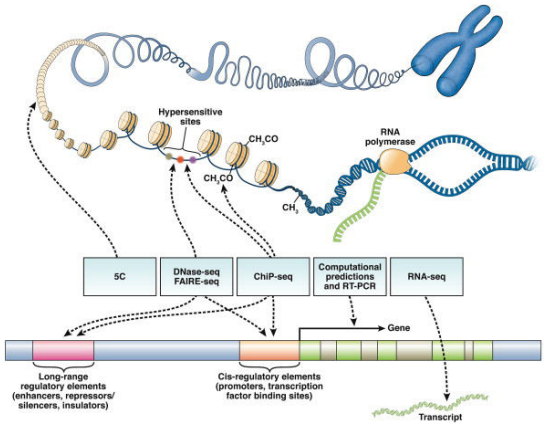


Kichaev et al. *PloS Genetics* 2014

(1) Where to get external functional information?
(e.g., classes of SNPs more likely to have function)

(2) How to quantify which classes of SNPs more useful for our trait of interest?
(e.g., regulatory in tissue X vs tissue Y)

(1) ENCODE/ROADMAP provides a functional map of the human genome (or use your own data)



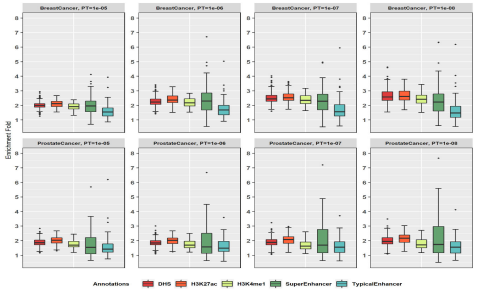
"Here, we assign biochemical functions for 80% of the genome"
—ENCODE Consortium 2012 Nature

(2) Functional enrichment

- Question:
  - Is GWAS signal concentrated in particular ”functional” areas of the genome?
- Functional enrichment = GWAS Signal / proportion of genome covered by functional annotation

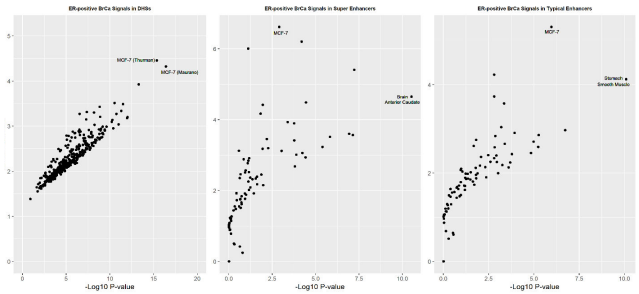
(2) Functional enrichment quantification
Option 1: count of biofeatures with GWAS signal

- Question:
  - Is GWAS signal concentrated in particular ”functional” areas of the genome?
- Functional enrichment = GWAS Signal / proportion of genome covered by functional annotation
- Example from Breast/Prostate Cancer (Chen..Linstroem Hum Genet 2019)

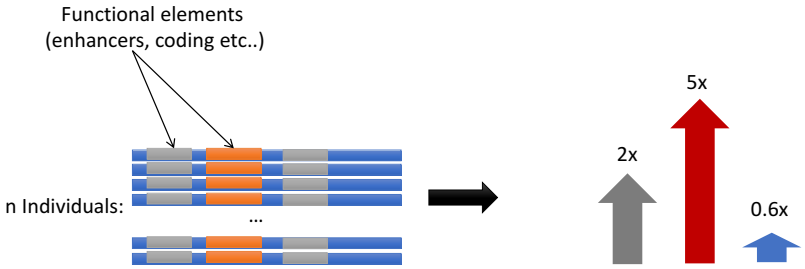


(2) Functional enrichment quantification
Option 1: count of biofeatures with GWAS signal

- Question:
  - Is GWAS signal concentrated in particular ”functional” areas of the genome?
- Functional enrichment = GWAS Signal / proportion of genome covered by functional annotation
- Example from Breast/Prostate Cancer (Chen..Linstroem Hum Genet 2019)

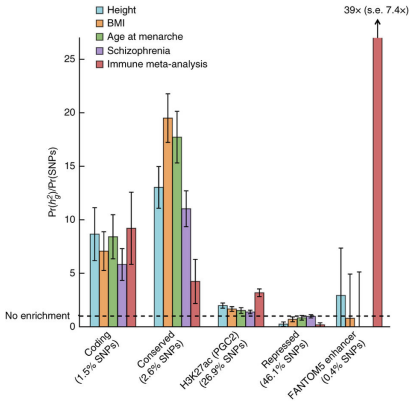


(2) Functional enrichment quantification
Option 2: SNP-heritability enrichment



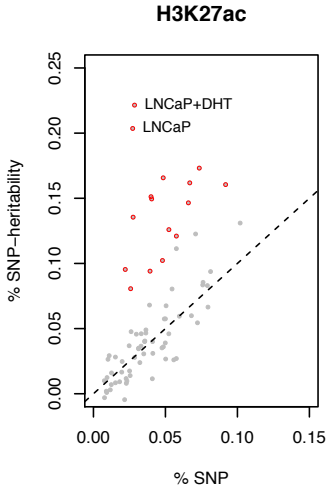
Can be estimated directly from summary GWAS (e.g., Finucane et al Nat Gen 2014)

Example of functional enrichment



- Finucane et al Nat Genet 2016

Top H3k27ac marks enriched in Prostate Cancer



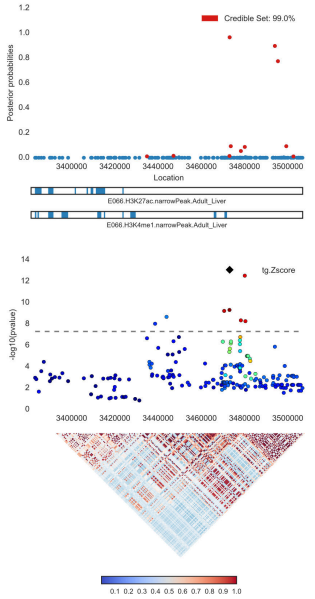
[Gusev et al NatComm 2016]

How do we put all together

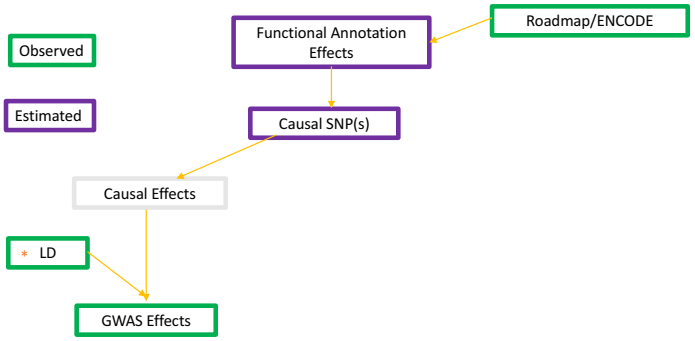
STATISTICAL FINE-MAPING

OUTPUT
- Probability of each SNP to be causal
- (optional: functional enrichment)

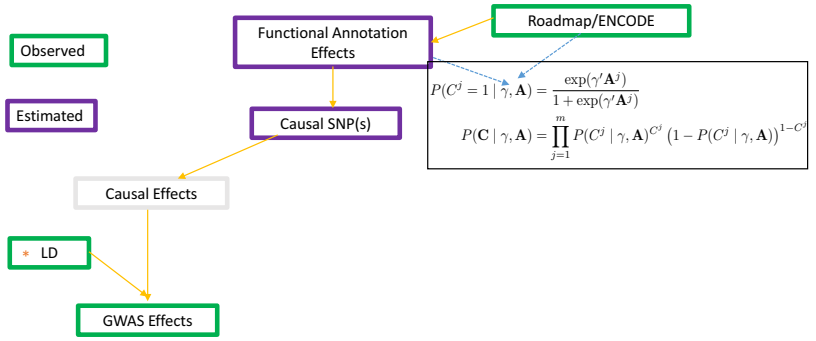
INPUT:
- Functional classes of SNPs (e.g., RoaMap/ENCODE)
- GWAS output (p-values for all SNPs)
- LD patterns (correlation structure among SNPs)



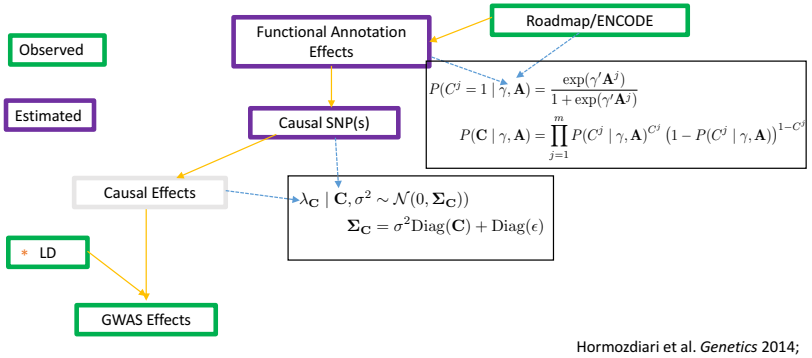
Combining this into a Bayesian Hierarchical Model: Big Picture Schematic



Combining this into a Bayesian Hierarchical Model: Big Picture Schematic

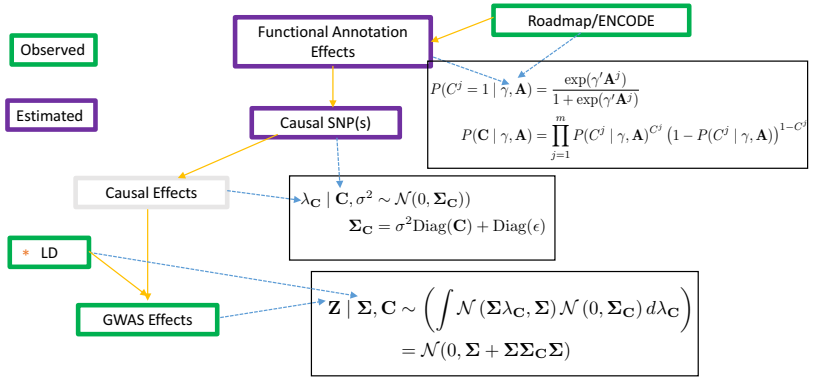


Combining this into a Bayesian Hierarchical Model: Big Picture Schematic

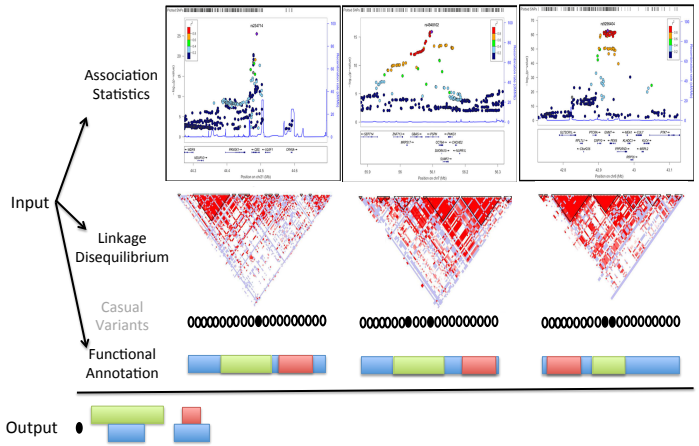


Hormozdiari et al. *Genetics* 2014;
Chen et al. *Genetics* 2015

Combining this into a Bayesian Hierarchical Model: Big Picture Schematic

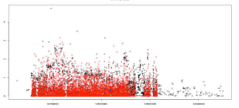


Model: visual representation



PAINTOR; RIVIERA; CAVIARBF;FINEMAP etc…

Association statistics (z-scores)

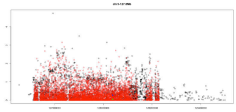


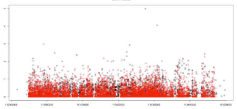
LD structure

Functional annotation

Locus 1

Association statistics (z-scores)

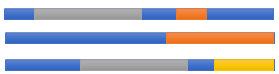


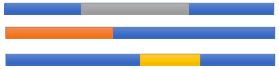


LD structure

Causal/non-causal SNPs

Functional annotation

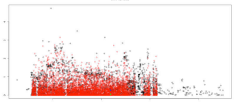




Locus 1 Locus 2

Statistical Model (N-SNPs, M functional classes)

Z: N-size vector of observed Association statistics



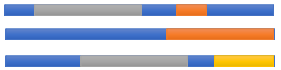
$$P(Z_j|C_j;\lambda_j) = \mathcal{N}(Z_j;\Sigma_j(\lambda_j \circ C_j),\Sigma_j) \quad \text{(p.d.f of multivariant normal)}$$

C: N-size vector 0/1 indicating causal status

$$P(C_j;\gamma) = \prod_i P(c_{ij};\gamma)$$

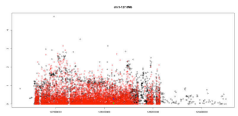
$$P(c_{ij};\gamma) = \left(\frac{1}{1+\exp(\gamma^T A_{ij})}\right)^{c_{ij}} \left(\frac{1}{1+\exp(-\gamma^T A_{ij})}\right)^{1-c_{ij}}$$

A: NxM matrix of 0/1 indicating annotation membership for every SNP



Statistical Model (N-SNPs, M functional classes)

Z: N-size vector of observed Association statistics



λ_j: effect size of SNP j

$$P(Z_j|C_j;\lambda_j) = \mathcal{N}(Z_j;\Sigma_j(\lambda_j \circ C_j),\Sigma_j) \quad \text{(p.d.f of multivariant normal)}$$

Σ_i: LD at locus i

C: N-size vector 0/1 indicating causal status

● ○ ○○ ● ○○○ ○

$$P(C_j;\gamma) = \prod_i P(c_{ij};\gamma)$$

$$P(c_{ij};\gamma) = \left(\frac{1}{1+\exp(\gamma^T A_{ij})}\right)^{c_{ij}} \left(\frac{1}{1+\exp(-\gamma^T A_{ij})}\right)^{1-c_{ij}}$$

A: NxM matrix of 0/1 indicating annotation membership for every SNP

γ: effect of annotation on probability of SNP to be causal

Getting output from statistical model: Bayes

$$P(\mathbf{C} \mid \mathbf{Z}, \mathbf{\Sigma}, \mathbf{A}, \gamma) = \frac{\mathcal{N}(0, \mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{\Sigma_C}\mathbf{\Sigma}) P(\mathbf{C} \mid \mathbf{A}, \gamma)}{\sum_{C \in \mathcal{C}} \mathcal{N}(0, \mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{\Sigma_C}\mathbf{\Sigma}) P(\mathbf{C} \mid \mathbf{A}, \gamma)}$$

- Causal vector formulation gives the flexibility to model **multiple causal variants** at any risk locus
- Iterative procedure to estimate annotation effects across all fine mapping loci using Maximum Likelihood and EM.

Practical considerations

Fine-mapping credible sets are miscalibrated when assuming a single causal variant.

Causal Assump.	Method	Functional Annotations	Causals Identified	90% Credible Set Size
Single	Maller et al.	None	64.2%	265.0
Multiple	CAVIAR	None	91.9%	510.3

90% credible set = set of SNPs that consume 90% of the total posterior probability mass

Maller et al. *Nat Gen* 2012;
Hormozdiari et al. *Genetics* 2014;

Leveraging functional enrichment improves fine-mapping resolution

Causal Assump.	Method	Functional Annotations	Causals Identified	90% Credible Set Size
Single	Maller et al.	None	64.2%	265.0
Multiple	CAVIAR	None	91.9%	510.3
Multiple	**PAINTOR**	**Included**	**91.2%**	**393.7**

up to ~30% improvement

90% credible set = set of SNPs that consume 90% of the total posterior probability mass

Maller et al. *Nat Gen* 2012; Hormozdiari et al. *Genetics* 2014; Kichaev et al. *Plos Genetics* 2014

Statistical fine-mapping identifies bona-fide causal variants

LETTERS

A thrifty variant in *CREBRF* strongly influences body mass index in Samoans

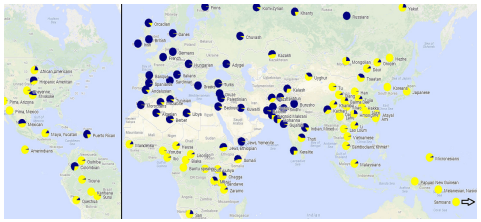
Ryan L Minster[1,13], Nicola L Hawley[2,13], Chi-Ting Su[1,12,13], Guangyun Sun[3,13], Erin E Kershaw[4], Hong Cheng[3], Olive D Buhule[5,12], Jerome Lin[3], Muagututi'a Sefuiva Reupena[6], Satupa'itea Viali[7], John Tuitele[8], Take Naseri[9], Zsolt Urban[1,14], Ranjan Deka[3,14], Daniel E Weeks[1,5,14] & Stephen T McGarvey[10,11,14]

*"Bayesian fine-mapping with PAINTOR strongly supported following up the missense variant. The two variants in the region with the highest posterior probability (PP) of being causal were **rs373863828** (PP = 0.80) and rs150207780 (PP = 0.22); when Encyclopedia of DNA Elements (ENCODE) functional annotation was included, these probabilities increased to 0.92 and 0.34, respectively"*

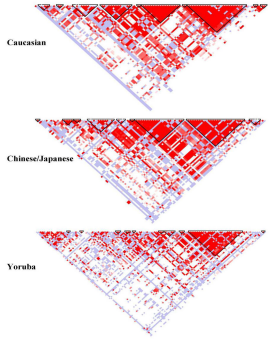
What if we have multiple ancestries? Can we still perform statistical fine-mapping?

Divergent population histories give rise to unique genetic backgrounds

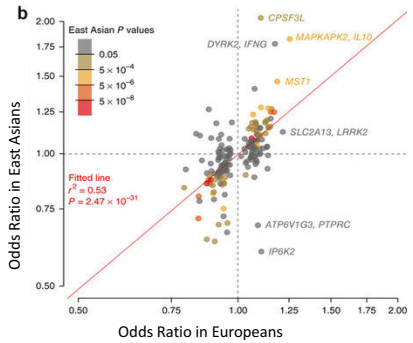
Allele Frequencies



LD Patterns



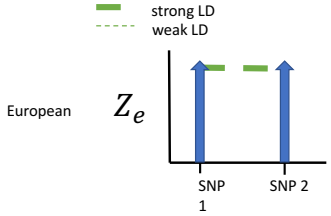
GWAS loci tend to replicate in non-European populations → shared causal variants



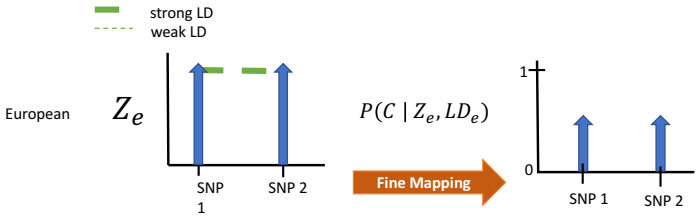
Liu et al. *Nat Genet* 2015

see also: Marigorta et al. *Plos Genet*. 2013
Zaitlen et al. *Am J Hum Genet.* 2011, PAGE consortium Nature 2019

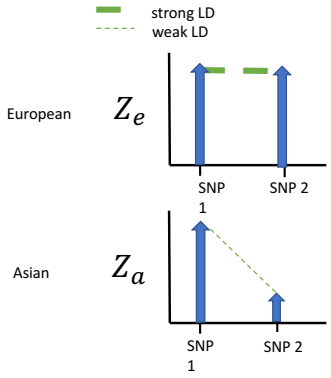
Leveraging genetic diversity to improve fine-mapping



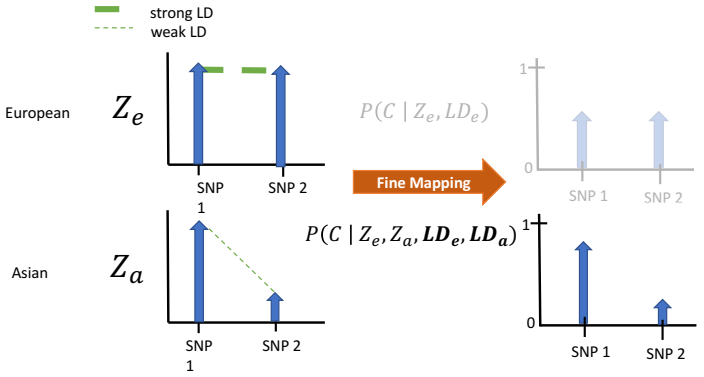
Leveraging genetic diversity to improve fine-mapping



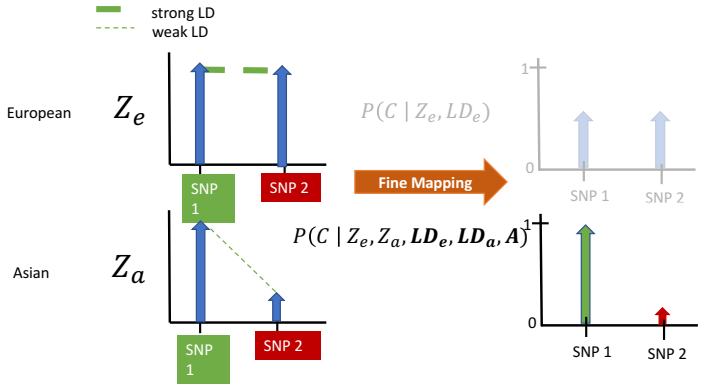
Leveraging genetic diversity to improve fine-mapping



Leveraging genetic diversity to improve fine-mapping



Leveraging genetic diversity and functional annotations to improve fine-mapping



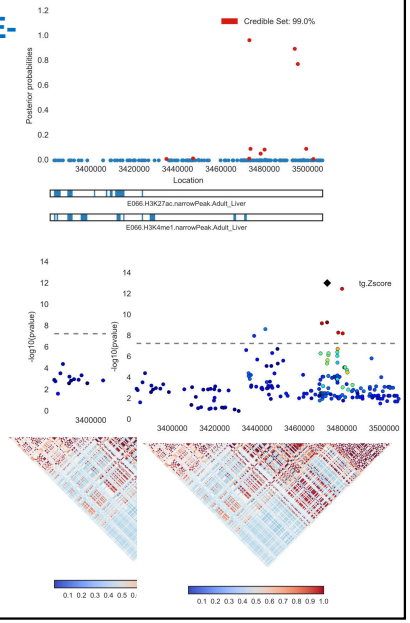
MULTI-ETHNIC STATISTICAL FINE-MAPING

OUTPUT

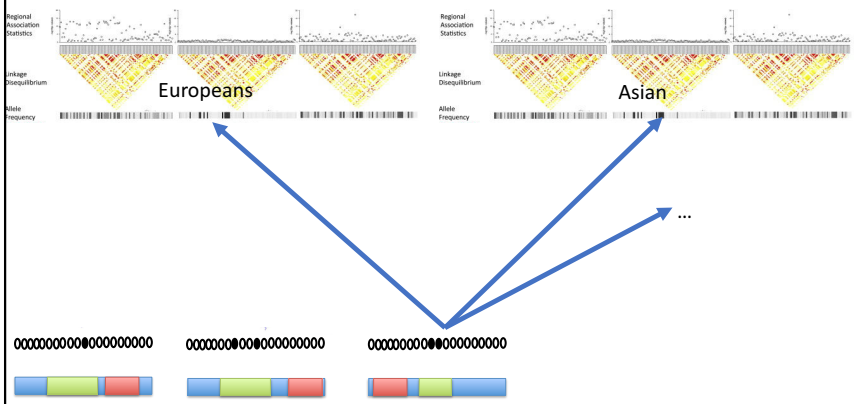
- Probability of each SNP to be causal
- (optional: functional enrichment)

INPUT:

- Functional classes of SNPs (e.g., RoaMap/ENCODE)
- GWAS output for each ancestry (p-values for all SNPs)
- LD patterns for each ancestry (correlation structure among SNPs)



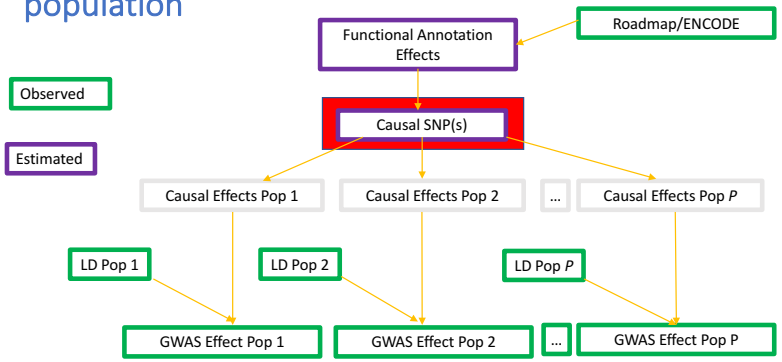
Integrating functional annotation data in trans-ethnic fine-mapping



Main assumption: same causal variants with causal probability from functional data

[Kichaev et al AJHG 2015; Morris Gen Epi 2011; Liu et al AJHG 2016; etc…]

Generalizing integrative fine-mapping from a single population



$$\mathbf{Z}_1 \cdots \mathbf{Z}_P \mid \mathbf{\Sigma}_1 \cdots \mathbf{\Sigma}_P, \mathbf{C}, \mathbf{A} \sim \left(\prod_{p=1}^{P} \mathcal{N}(0, \mathbf{\Sigma}_p + \mathbf{\Sigma}_p \mathbf{\Sigma}_\mathbf{C} \mathbf{\Sigma}_p) \right) P(\mathbf{C} \mid \mathbf{A})$$

Trans-ethnic fine-mapping reduces the size of the credible causal sets

LETTER

doi:10.1038/nature12373

Genetics of rheumatoid arthritis contributes to biology and drug discovery

A list of authors and their affiliations appears at the end of the paper

N Europeans ≈ 68K
N Asians ≈ 36K

Final Annotations selected by model:

DHS (Skin Keratinocytes, Th2, and B-lymphocytes), Immune Enhancers, Exonic regions

Average Size of 90% credible set

	Data	No Annotations	With Annotations
Single Population	Asian	35.2	31.9
	European	32.0	28.7
Multi Population	Meta analysis	28.5	25.0
	Asian, European	24.0	21.7

+32%

Example of multi-ethnic fine-mapping Europeans+East Asians

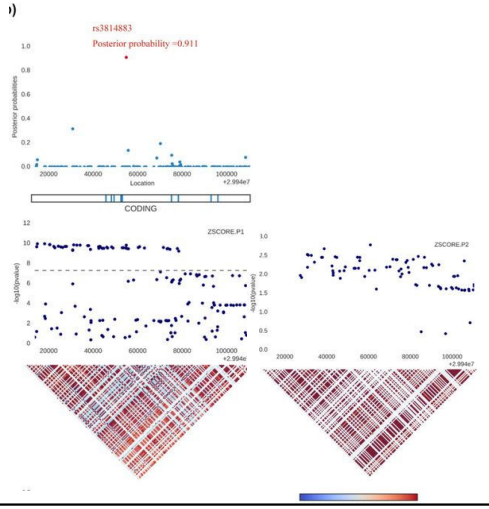
nature genetics

Article | Published: 09 October 2017

Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia

Zhiqiang Li, Jianhua Chen [...] Yongyong Shi

Nature Genetics 49, 1576–1583 (2017) | Download Citation



Example of multi-ethnic fine-mapping

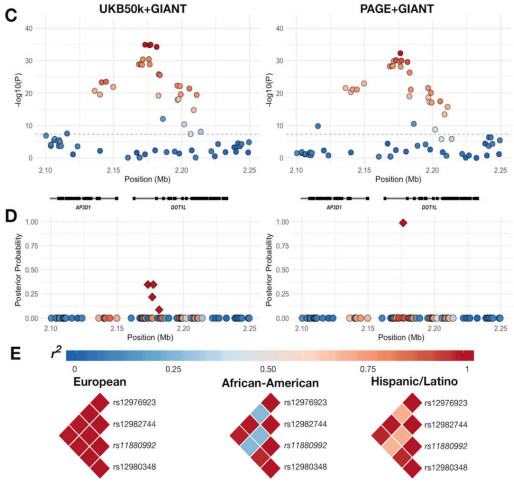
nature

Letter | Published: 19 June 2019

Genetic analyses of diverse populations improves discovery for complex traits

Genevieve L. Wojcik, Mariaelisa Graff, [...] Christopher S. Carlson

Nature 570, 514–518 (2019) | Download Citation



Wojcik et al Nature 2019

Big picture

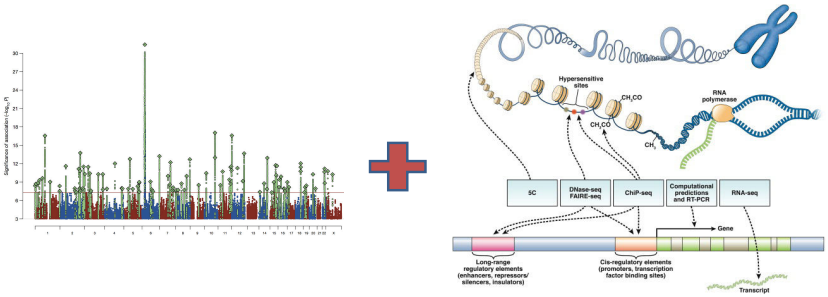
GWAS Marginal Summary Statistics + Linkage Disequilibrium Reference Panel + Functional Annotation

LD Score Regression (or many others) → (1) Find functional annotations of interest

PAINTOR/Susie (or many others) → (2) Estimate SNP Probabilities of Causality

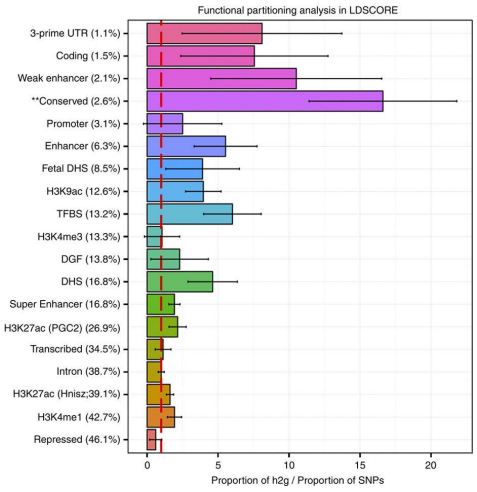
Questions

RECAP



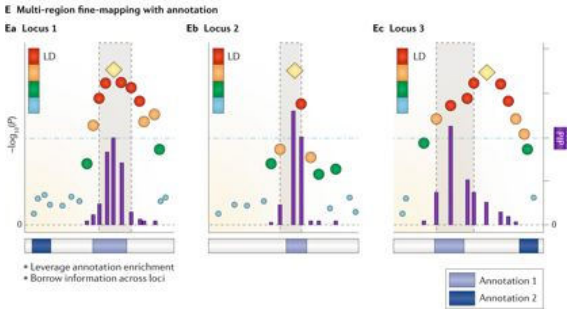
- Functional enrichment (LDSCORE regression, Finucane et al NG 2016)

Identify functional elements of interest



(hippocampal volume) Hibar et al Nat Comm 2017

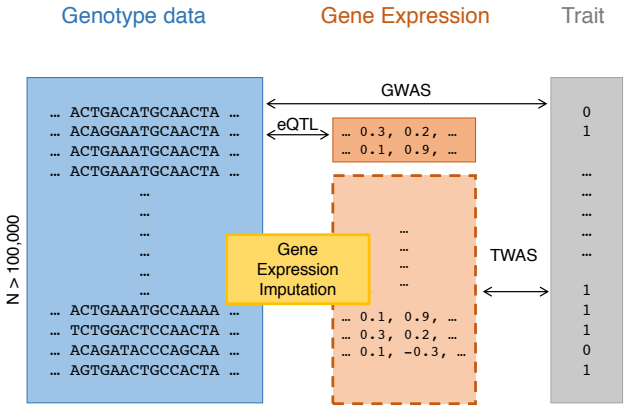
Prioritize genetic variants likely to be biologically causal



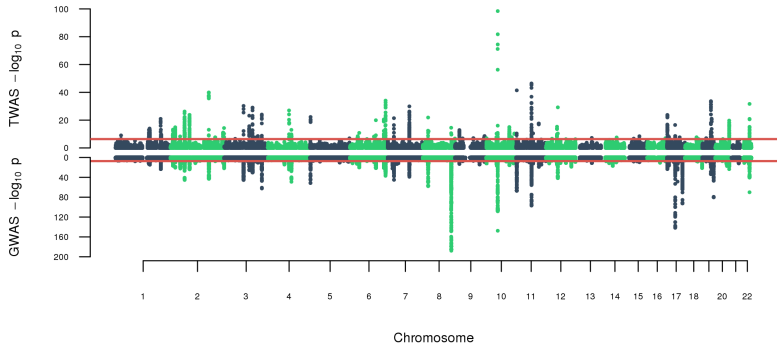
Schaid et al Nat Rev Genet 2018

Questions

Statistical fine-mapping at gene-level in TWAS/PrediXscan/etc…



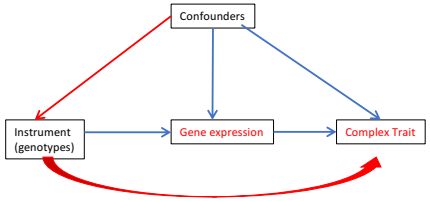
TWAS example



Mancuso et al. Nat Comm 2018

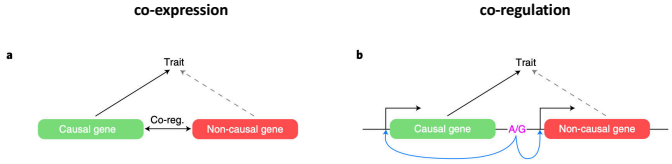
Are all TWAS significant genes causal?

TWAS = Mendelian Randomization under strong assumptions



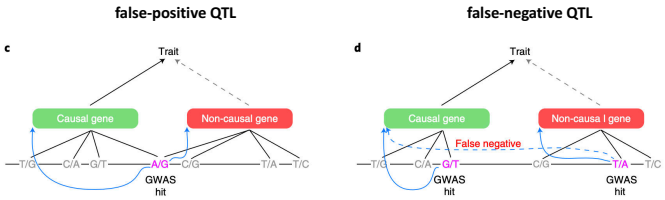
- Instrument must be associated with the exposure
- **Instrument must not have an association with outcome, except through the exposure**
- Instrument is not related to measured or unmeasured confounders

non-causal TWAS associations due to co-regulation



Wainberg et al. 2019 *Nat Genet*

non-causal TWAS associations due to tagging



Wainberg et al. 2019 *Nat Genet*

But we really want causal genes…

TWAS interpretation

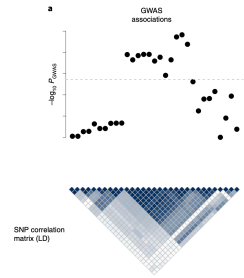
- Two sample Mendelian Randomization test
  - Estimate **mediating** effect of gene expression **under very strong assumptions!**
  - Zhu et al. Nat Gen 2016; Barfield et al Gen Epi 2018, …
- Test of association (genetic covariance) expression and trait
  - **Similarity** between trait / GE at local genetics
  - Gusev et al. Nat Gen 2016; Mancuso et al. AJHG 2017, Wainberg et al Nat Gen 2019; Mancuso et al Nat Genet 2019

Opportunities and challenges for transcriptome-wide association studies

Michael Wainberg[1], Nasa Sinnott-Armstrong[2], Nicholas Mancuso[3], Alvaro N. Barbeira[4], David A. Knowles[5,6], David Golan[2], Raili Ermel[7], Arno Ruusalepp[7,8], Thomas Quertermous[9], Ke Hao[10], Johan L. M. Björkegren[8,10,11,12], Hae Kyung Im[4], Bogdan Pasaniuc[3,13,14], Manuel A. Rivas[15] and Anshul Kundaje[1,2]

Probabilistic fine-mapping of TWAS

GWAS: Fine-mapping multiple associated SNPs in LD at a locus

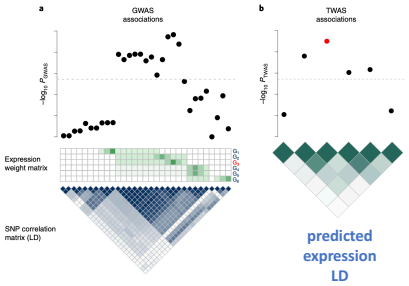


Mancuso et al. 2019 *Nat Genet*

Probabilistic fine-mapping of TWAS

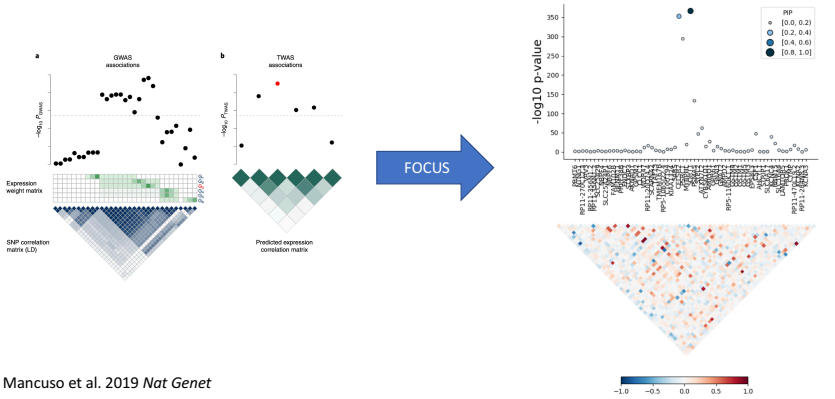
GWAS: Fine-mapping multiple associated SNPs in LD at a locus

TWAS: Fine-mapping multiple gene models ➔ posterior probability for each gene ➔ **credible sets** (FOCUS)



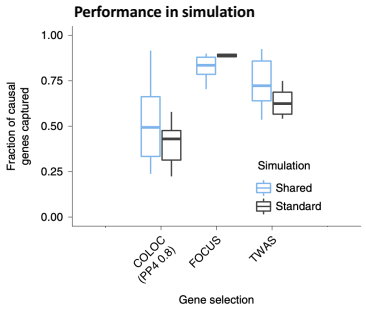
Mancuso et al. 2019 *Nat Genet*

Example of TWAS fine-mapping



Mancuso et al. 2019 *Nat Genet*

FOCUS accurately identifies credible genes



Performance in real data

Table 1 | Summary of gene-based fine-mapping in lipid GWAS risk regions

Lipid trait	GWAS risk regions	GWAS risk regions with TWAS-significant genes	TWAS genes at risk regions	Genes in 90%-credible sets
HDL	43	18	64	30
LDL	36	20	56	40
Total cholesterol	51	24	73	53
Triglycerides	30	13	33	25
Overall	160	75	226	148
Unique	89	46	146	100

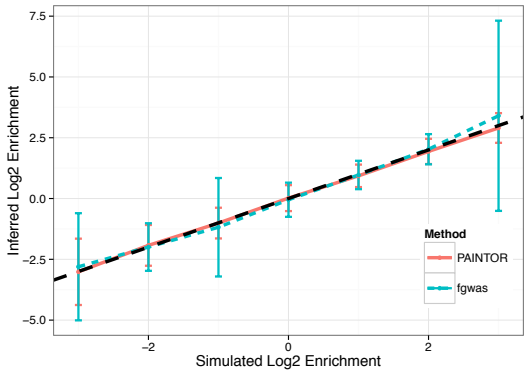
A GWAS risk region is defined to be an LD block defined by LDetect[41] harboring at least one genome-wide-significant SNP ($P < 5 \times 10^{-8}$) reported in ref.[16]. A TWAS gene is a gene whose predicted expression reaches transcriptome-wide significance of $P < 0.05/15,277$. Overall results are presented as total counts across traits with unique results discarding repeated elements.

1.5 genes on average in the 95% credible set

Mancuso et al. 2019 *Nat Genet*

Questions

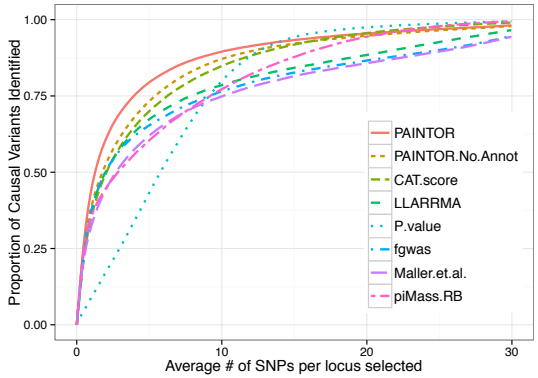
Unbiased estimation of enrichment of causal variants in simulations



Hapgen
simulations
100 loci 10Kb
Europeans 100G
h2=0.25
N=5,000

[Kichaev et al. Plos Genetics 2014]

Functional data improves fine-mapping accuracy

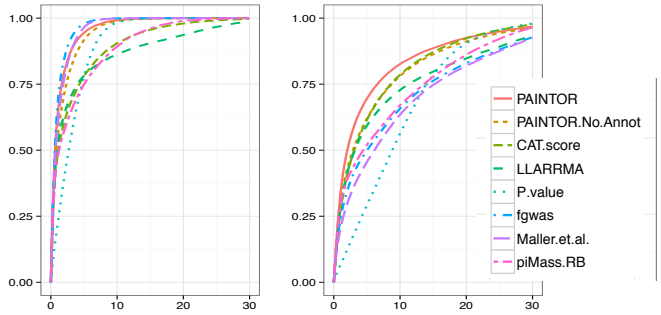


[Kichaev et al. Plos Genetics 2014]

Functional data improves fine-mapping accuracy

Loci with 1 causal

Loci with >1 causal



[Kichaev et al. Plos Genetics 2014]

How many variants to functionally test to find causal variants?

Method (assumptions)	50% of all causals	90% of all causals
p-value	5.7 (369.5)	12.3 (796.9)

Simulations:
100 loci 10Kb
Europeans 100G
h2=0.25
N=10,000

How many variants to functionally test to find causal variants?

Method (assumptions)	50% of all causals	90% of all causals
p-value	5.7 (369.5)	12.3 (796.9)
Probabilities (single causal per locus) (Maller et al NG'12)	2.7 (172.4)	25.0 (1616.1)

Simulations:
100 loci 10Kb
Europeans 100G
h2=0.25
N=10,000

How many variants to functionally test to find causal variants?

Method (assumptions)	50% of all causals	90% of all causals
p-value	5.7 (369.5)	12.3 (796.9)
Probabilities (single causal per locus) (Maller et al NG'12)	2.7 (172.4)	25.0 (1616.1)
PAINTOR (multiple causals)	1.7 (108.9)	11.4 (734.4)

Simulations:
100 loci 10Kb
Europeans 100G
h2=0.25
N=10,000

How many variants to functionally test to find causal variants?

Method (assumptions)	50% of all causals	90% of all causals
p-value	5.7 (369.5)	12.3 (796.9)
Probabilities (single causal per locus) (Maller et al NG'12)	2.7 (172.4)	25.0 (1616.1)
PAINTOR (multiple causals)	1.7 (108.9)	11.4 (734.4)
PAINTOR (multiple causals; with ENCODE)	1.2 (78.7)	9.7 (625.6)

Simulations:
100 loci 10Kb
Europeans 100G
h2=0.25
N=10,000

How many variants to functionally test to find causal variants?

Method (assumptions)	50% of all causals	90% of all causals
p-value	5.7 (369.5)	12.3 (796.9)
Probabilities (single causal per locus) (Maller et al NG'12)	2.7 (172.4)	25.0 (1616.1)
PAINTOR (multiple causals)	1.7 (108.9)	11.4 (734.4)
PAINTOR (multiple causals; with ENCODE)	1.2 (78.7)	9.7 (625.6)
PAINTOR (multiple causals; true ENCODE prior)*	*1.2 (77.9)*	*9.4 (610.6)*

Simulations:
100 loci 10Kb
Europeans 100G
h2=0.25
N=10,000

Assuming 1 causal per locus yields miss-calibrated causal sets

Causal Set	Method	Annotations	# Causals	# SNPs
90%	Maller et al.	-	63.2	325.5
95%	Maller et al.	-	68.8	409.8
99%	Maller et al.	-	76.7	579.4

[Kichaev et al. Plos Genetics 2014]

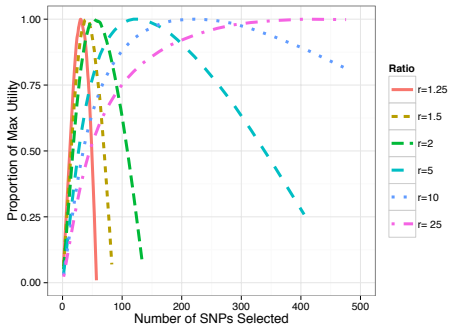
Assuming 1 causal per locus yields miss-calibrated causal sets

Causal Set	Method	Annotations	# Causals	# SNPs
90%	Maller et al.	-	63.2	325.5
	PAINTOR	-	90.1	569.4
	PAINTOR	+	89.9	461.0
95%	Maller et al.	-	68.8	409.8
	PAINTOR	-	95.8	741.4
	PAINTOR	+	95.6	628.8
99%	Maller et al.	-	76.7	579.4
	PAINTOR	-	101.5	1118.9
	PAINTOR	+	101.6	1001.8

[Kichaev et al. Plos Genetics 2014]

How many SNPs to follow-up?

- Depends on ratio of benefit of finding a causal variant to cost of testing a variant (r=B/C)
  - r=10 ➔the benefit of finding a causal outweighs 10 times the cost of testing 1 SNP

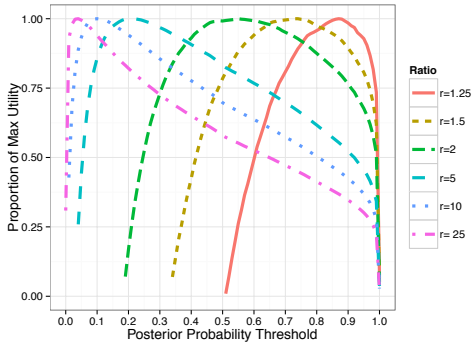


r=10, ~3.8 SNPs per locus to identify ~72.6% of all causals
r=100, ~13.1 SNPs per locus to identify ~96.2% of all causals

[Kichaev et al. Plos Genetics 2014]

How many SNPs to follow-up?

- Thresholding on posterior probability of causality gives a principled way of maximizing utility.



[Kichaev et al. Plos Genetics 2014]

Other considerations

- Locus size (function of LD):
  - Increasing locus size increases performance of PAINTOR vs existing methods
    - 10Kb: 27.4 vs 11.4 variants to follow-up to find 90% of causals
    - 50Kb: 110.7 vs 24.1 variants to follow-up to find 90% of causals
- Causal not in the data:
  - Median distance in Kb increases by ~6%
    - 21.6 vs 22.0 median Kb to the top 10 SNPs
    - 1.6 vs 4.0 minimum Kb to the top 10 SNPs
- Sample size
  - 19.0, 12.5, 10.8 variants per locus to find 90% of all causals for 2.5k, 5k and 10k samples