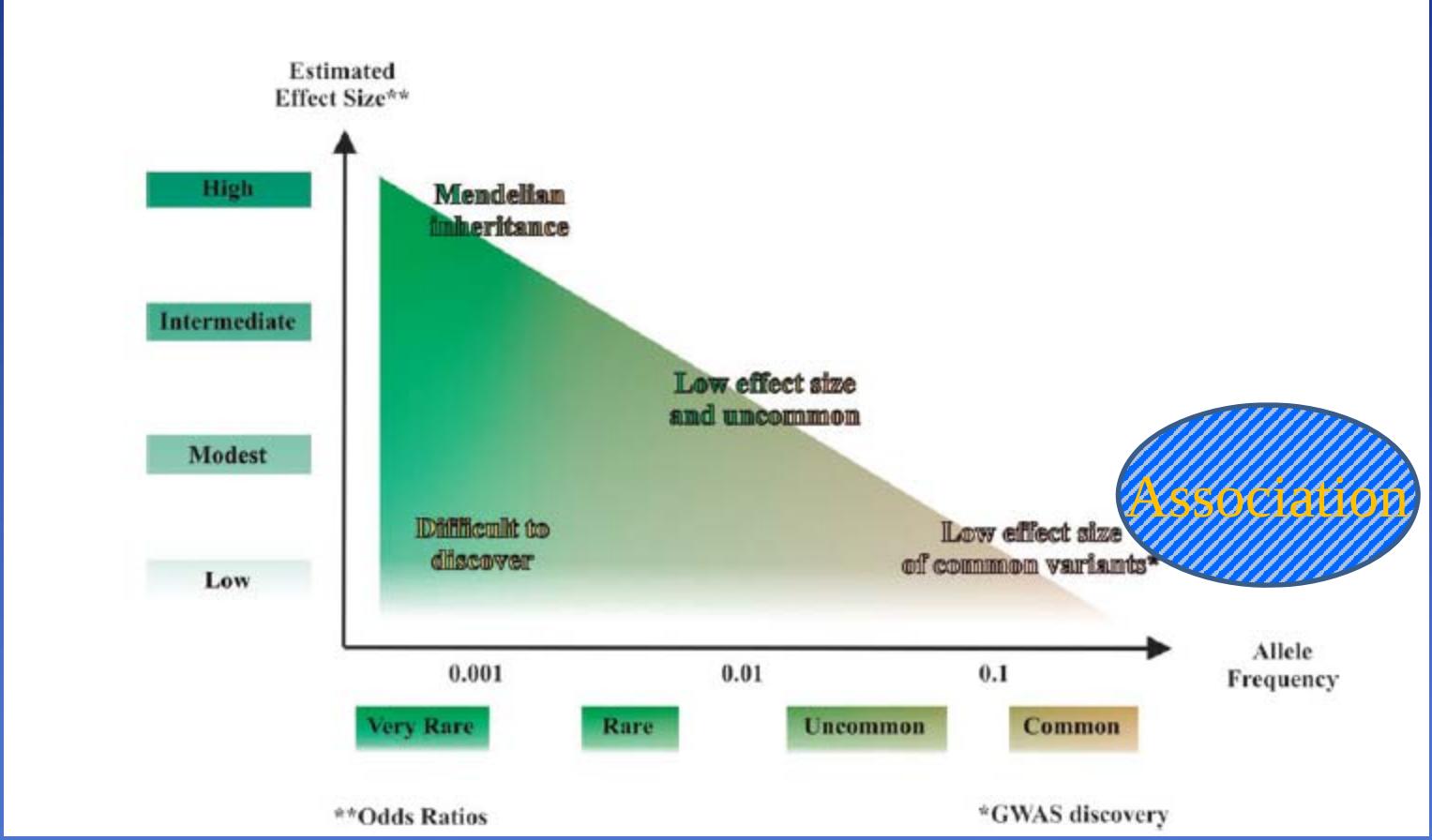
Understanding Genetic Architecture of Lung and Common Cancers

Christopher Amos
Ivan Gorlov
Changlu Liu

Dartmouth-Hitchcock
NORRIS COTTON
CANCER CENTER

The continuum of variation

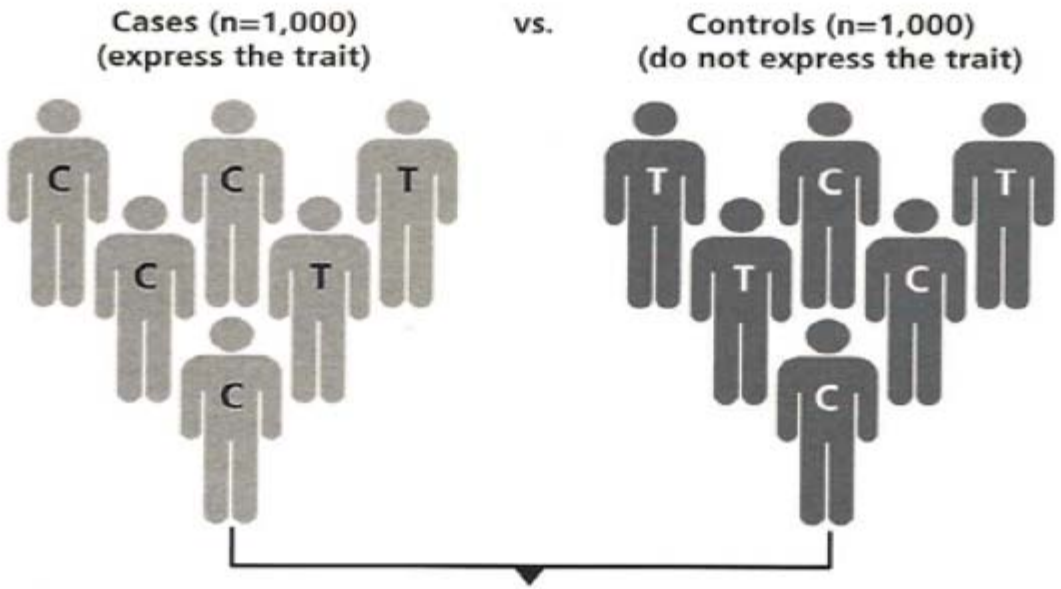


Chung et al.Carcinogenesis 31:111-120, 2010

SNPs are the most common type of polymorphism in the human genome

- SNPs occur ~ every 400 bases
- Minor allele frequency describes prevalence of rarer variant (0,0.5]
- Very Few (<0.01%) SNPs are associated with diseases
- SNPs are the bread of Genome Wide Association Studies (GWAS)
- Success requires inference about effects from Causal Variants

Conduct of Association Studies



	C	T
Cases	62%	38%
Controls	49%	51%

$$\chi^2 = 34.2,\ p\text{-}value = 4.9\text{x}10^{-9}$$

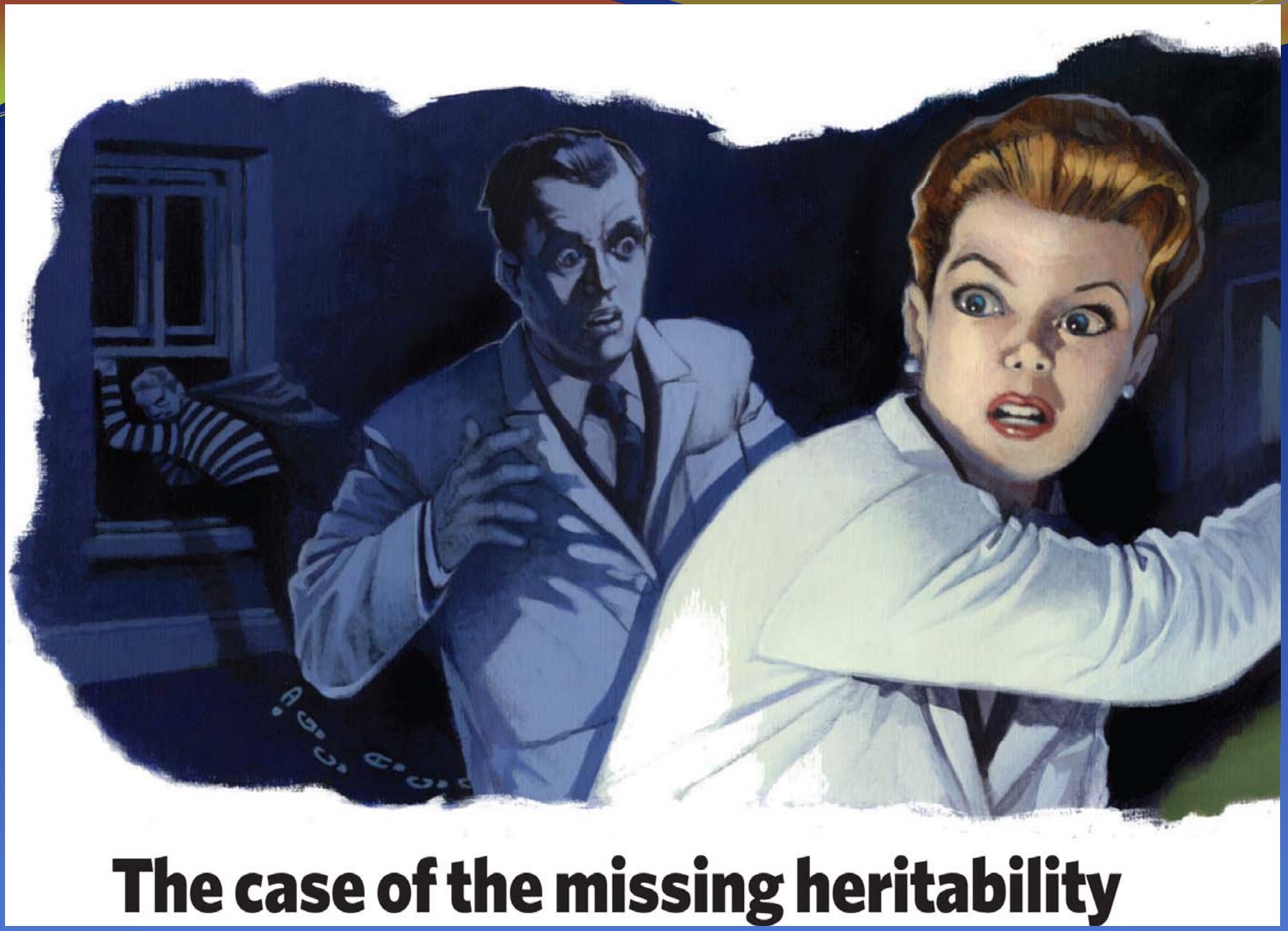
$$OR = \frac{P(case|C)/P(control|C)}{P(case|T)/P(control|T)}$$

Table 2. Array heritability explained by all autosomes (h^2_g) compared with the total heritability estimates from the largest twin studies

Cancer types	Total heritability from twins studies		Array heritability	
	Lichtenstein (2000) %	Mucci (2013) %	h^2_g (95% CI) %	h^2_g (95% CI), after removing known loci %
Bladder	31 (0-45)	n.a.	1 (0-11)	0 (0-10)
Breast	**27 (4-41)**	**28 (12-52)**	13 (0-56)	5 (0-46)
Endometrial cancer Aus	0 (0-42)	**24 (14-87)**	**39 (2-76)**	**39 (2-76)**
Endometrial cancer UK			**23 (1-45)**	**23 (1-45)**
Esophageal adenocarcinoma	n.a.	n.a.	**24 (14-34)**	**24 (14-34)**
Esophageal squamous cell carcinoma	n.a.	n.a.	**19 (7-31)**	**19 (7-31)**
Gastric Cancer	n.a.	n.a.	11 (0-27)	8 (0-22)
Kidney	n.a.	**23 (11-42)**	**18 (4-32)**	**15 (1-31)**
Lung	26 (0-49)	**25 (12-44)**	10 (0-24)	8 (0-22)
Melanoma QLD	n.a.	**39 (8-81)**	**30 (10-50)**	**21 (1-41)**
Melanoma USA			**19 (1-37)**	8 (0-28)
Ovary	22 (0-41)	**28 (15-47)**	**30 (18-42)**	**29 (17-41)**
Pancreas	36 (0-53)	n.a.	**18 (6-30)**	**16 (4-28)**
Prostate	**42 (29-50)**	**58 (52-63)**	**81 (32-100)**	**59 (12-100)**

Ones in bold are significantly different from zero ($P<0.05$).

Lu et al., 2014



Nature 456, 18-21 (2008)

Theft of Heritability

- Inadequate Markers
  - GWAS markers are incomplete
  - Inadequate coverage of copy number change
- Common Disease-Common Variant Hypothesis is inadequate
- Epigenetic Changes
- Etiological heterogeneity

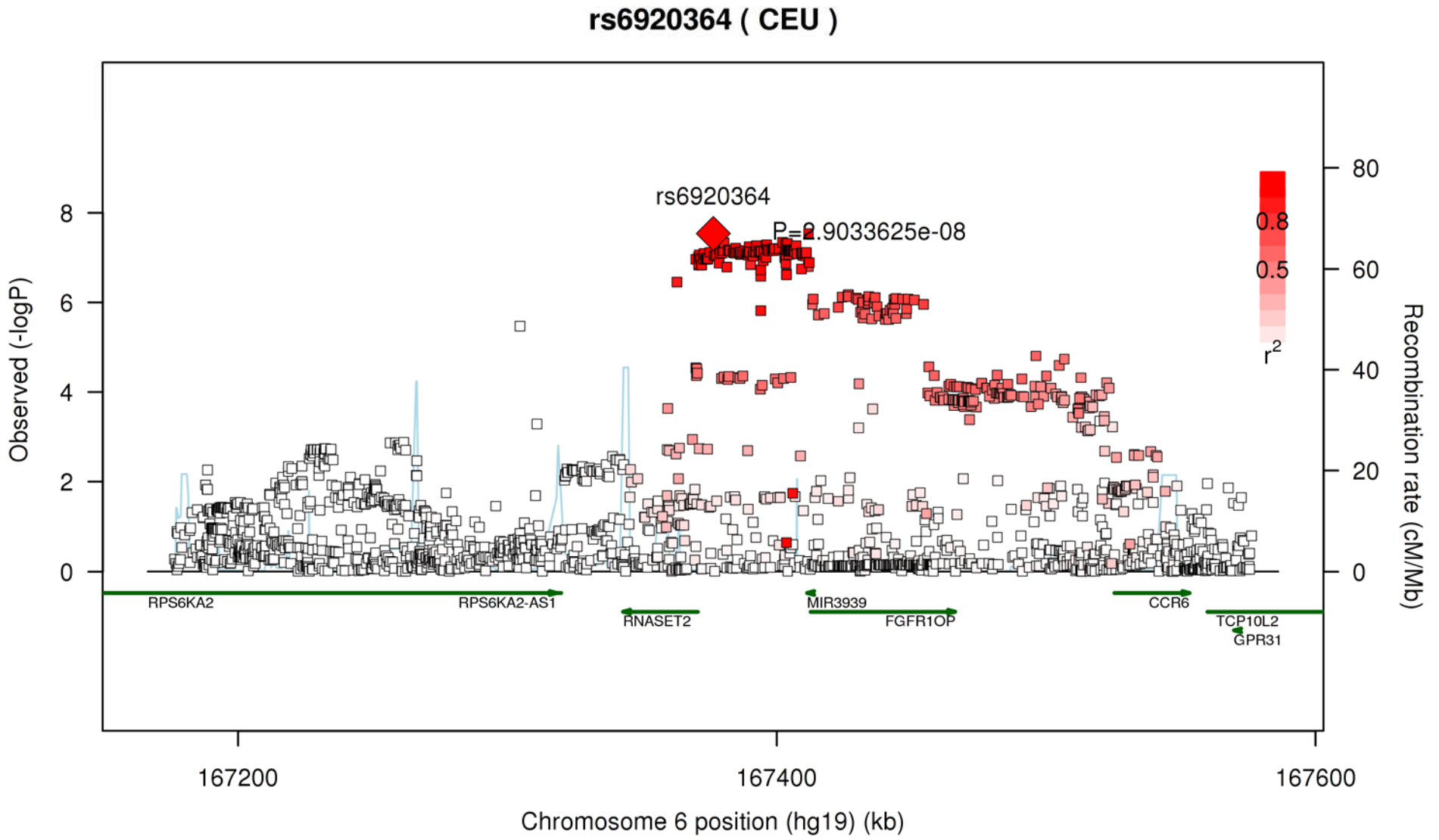
Hypothesize that the Common Disease Common Variant Hypothesis is incorrect

Slightly deleterious SNPs are mildly impair gene function and increase disease risk.

Effects of sdSNPs are not strong enough for selection to eliminate them from population

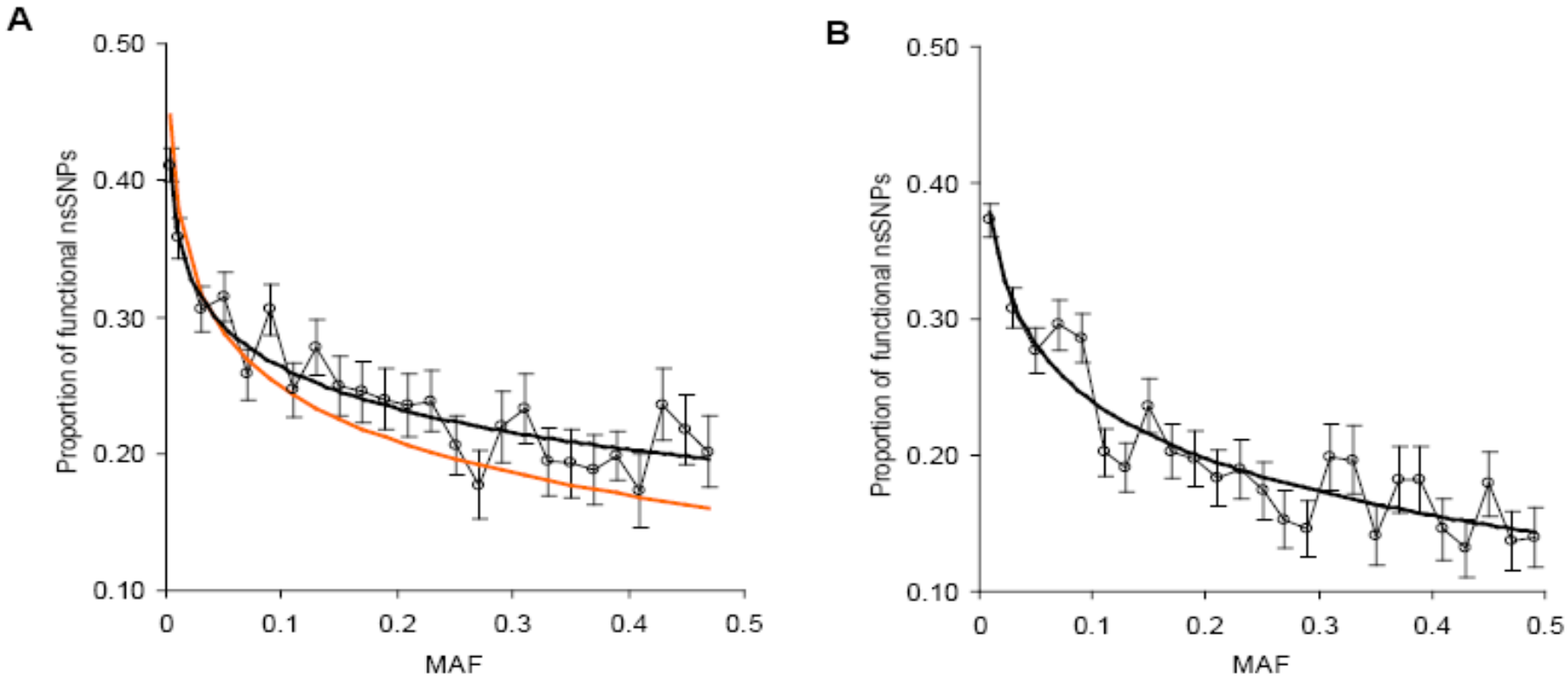
Gorlov, PLoS Genet. 2015 Jul 22;11(7)
Gorlov Hum Genet. 2014 Dec;133(12):1477-86.

Fine mapping to identify causal variants often difficult



Proportions of functional SNPs in different MAF categories

We estimated the proportion of nonsynonymous SNPs predicted to be functional by PolyPhen and SIFT in different MAF categories



Hypothesize that:

- Large fraction of the genetic susceptibility influenced by rare (<5%) variants with relatively strong effect size

- Targeting rarer variants for analysis may detect causal variants **when sample size is large.**

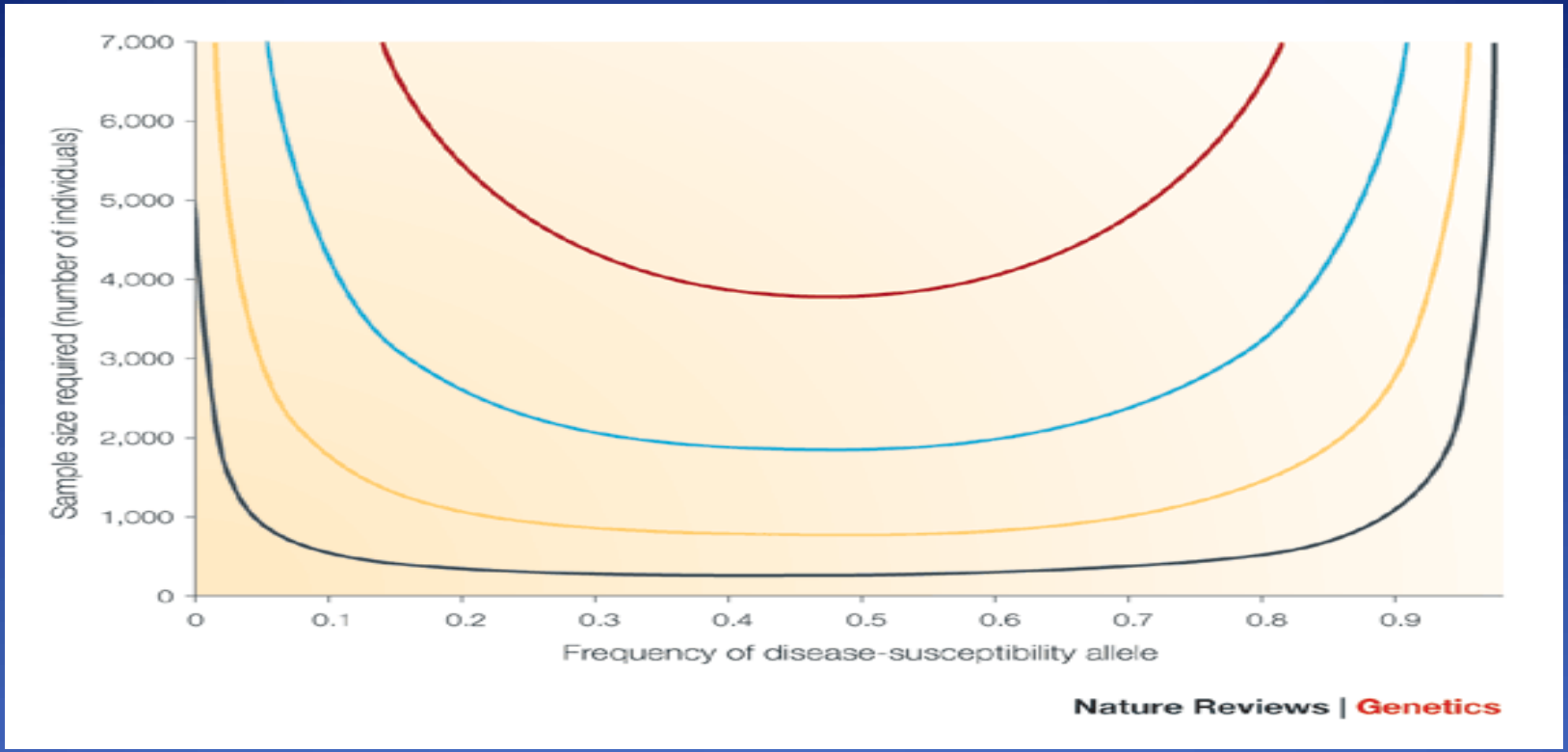
The majority of the significant SNPs detected by GWAS are SNPs tagging untyped causal variants – greatly reduces power – but indicates region for further study

Brute force analysis may not be sufficient to overcome power loss for multiple comparison – have to upweight analyses for most likely causal variants

Using GWAS to identify rarer SNPs associated with disease

- Newer SNP panels include rarer SNPs that can query rarer causal variants more effectively
- Using newer SNP platforms allow inference from 1000 genomes projects of unmeasured variants
- Sample sizes must be very large to derive sufficient power
- Assembling a very large study for cancer –OncoArray - 450K individuals for 530K variants, cost is $40/sample

MAF and a Required Sample Size



The numbers of cases and controls that are required in an association study to detect disease variants with allelic odds ratios of 1.2 (red), 1.3 (blue), 1.5 (yellow), and 2 (black). Numbers shown are for a statistical power of 80% at a significance level of P <10-6.

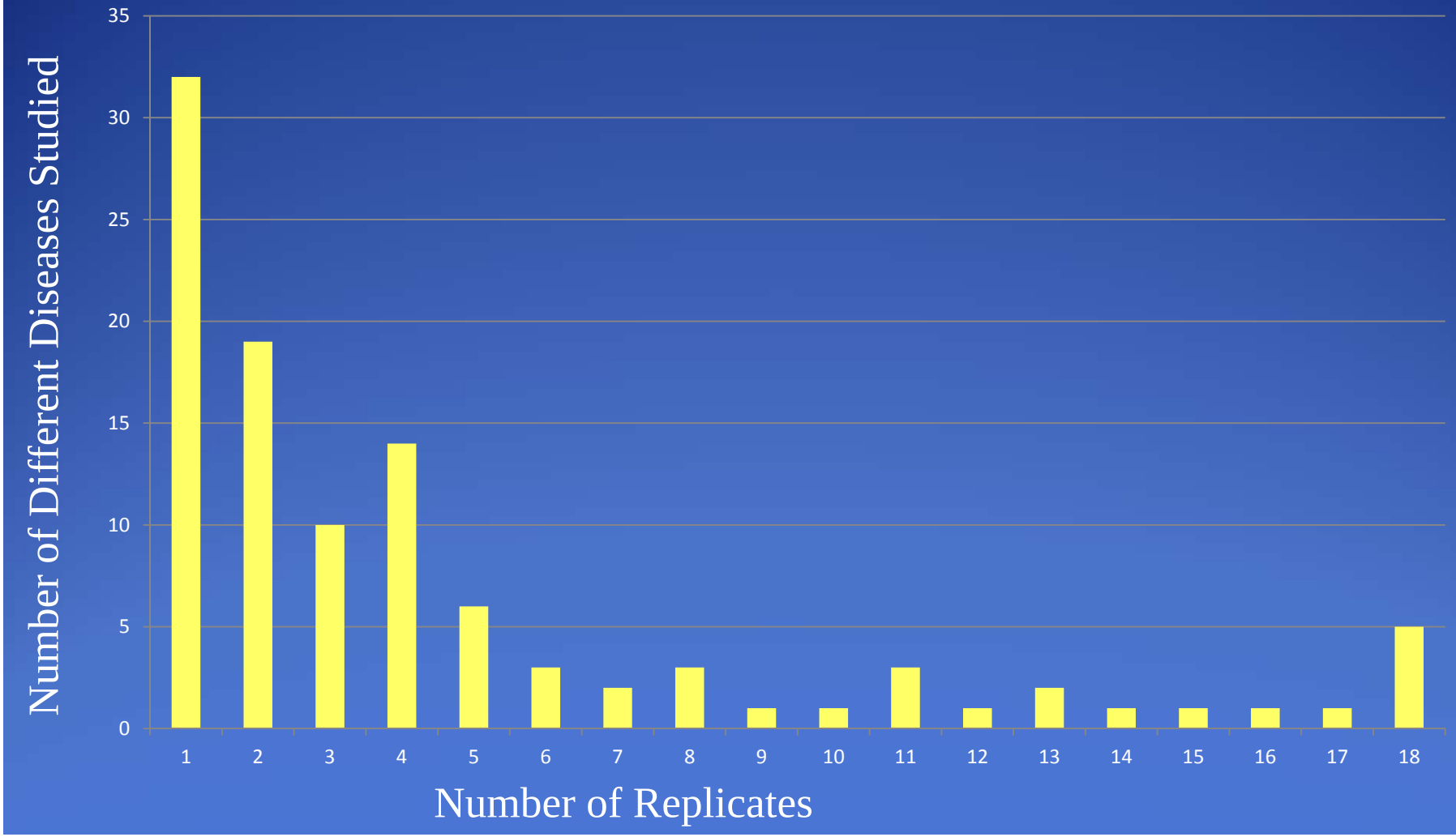
SNP Replication Rate

- GWAS replication rate at GWAS significance is low: about 5%
- This suggest that GWAS produce a considerable number of false discoveries
- Development of methods to predict which SNP will be replicated and which will not is important
- We studied results of published GWASs

Evaluating SNP replications

- Retrieved all data from the Catalog of Published GWA Studies (http://www.genome.gov/26525384/), Sep 13
- Restricted analysis to SNPs mapping to a single gene and associated with a disease rather than variation
- SNPs related to 106 diseases were studied, 2659 SNPs from 512 studies
- SNP findings sorted by date with the first report denoted as a discovery and subsequent studies may be replications
- Reproducibility score modeled as the ratio of successful replications over the total number of subsequent studies.

Number of GWA Studies Evaluated



Characteristics Studied

Name	Description	Source of the data
Conservation index	The level of evolutionary conservation of the protein based on the most distant homolog of the human gene	NCBI HomoloGene database: http://www.ncbi.nlm.nih.gov/homologene
eQTL	SNP reported as an eQTL for HapMap Data	eQTL SNPs identified in lymphoblastic cell lines from HapMap project [16]
Gene Size	Size of the gene region in nucleotides	NCBI RefSeq database: http://www.ncbi.nlm.nih.gov/refseq/
Growth factor	The protein encoded by the linked gene is a growth factor	Gene Ontology (GO) database http://geneontology.org/
Kinase	The protein encoded by the linked gene is a kinase	Gene Ontology (GO) database http://geneontology.org/
-Log(P)	Minus LOG(P) where P is the P-value reported in CPG	Catalog of Published GWAS (CPG): http://www.genome.gov/26525384
MAF	Minor allele frequency	Catalog of Published GWAS (CPG): http://www.genome.gov/26525384 . MAF reported in control group were used.
Nuclear Localization	The protein encoded by the linked gene is localized in the nucleus	Gene Ontology (GO) database http://geneontology.org/
OMIM	Was associated gene in OMIM	http://www.ncbi.nlm.nih.gov/omim
Plasma Membrane	The protein encoded by the linked gene is localized in plasma membrane	Gene Ontology (GO) database http://geneontology.org/
Receptor	The protein encoded by the linked gene is a receptor	Gene Ontology (GO) database http://geneontology.org/
SNP type	Type of the SNP. For the details see materials and methods	Catalog of the Published GWAS (CPG): http://www.genome.gov/26525384
Tissue specific	The expression of the linked gene is tissue specific	Tissue specific Gene Expression and Regulation (TiGER) database: http://bioinfo.wilmer.jhu.edu/tiger/
Transcription factor	The protein encoded by the linked gene is a transcription factor	Gene Ontology (GO) database http://geneontology.org/

3' UTR, 3' Downstream, 5' Upstream, 5' UTR, Coding nonsynonymous, Coding synonymous, Intergenic, Intronic, Non-coding, and Non-coding intronic.

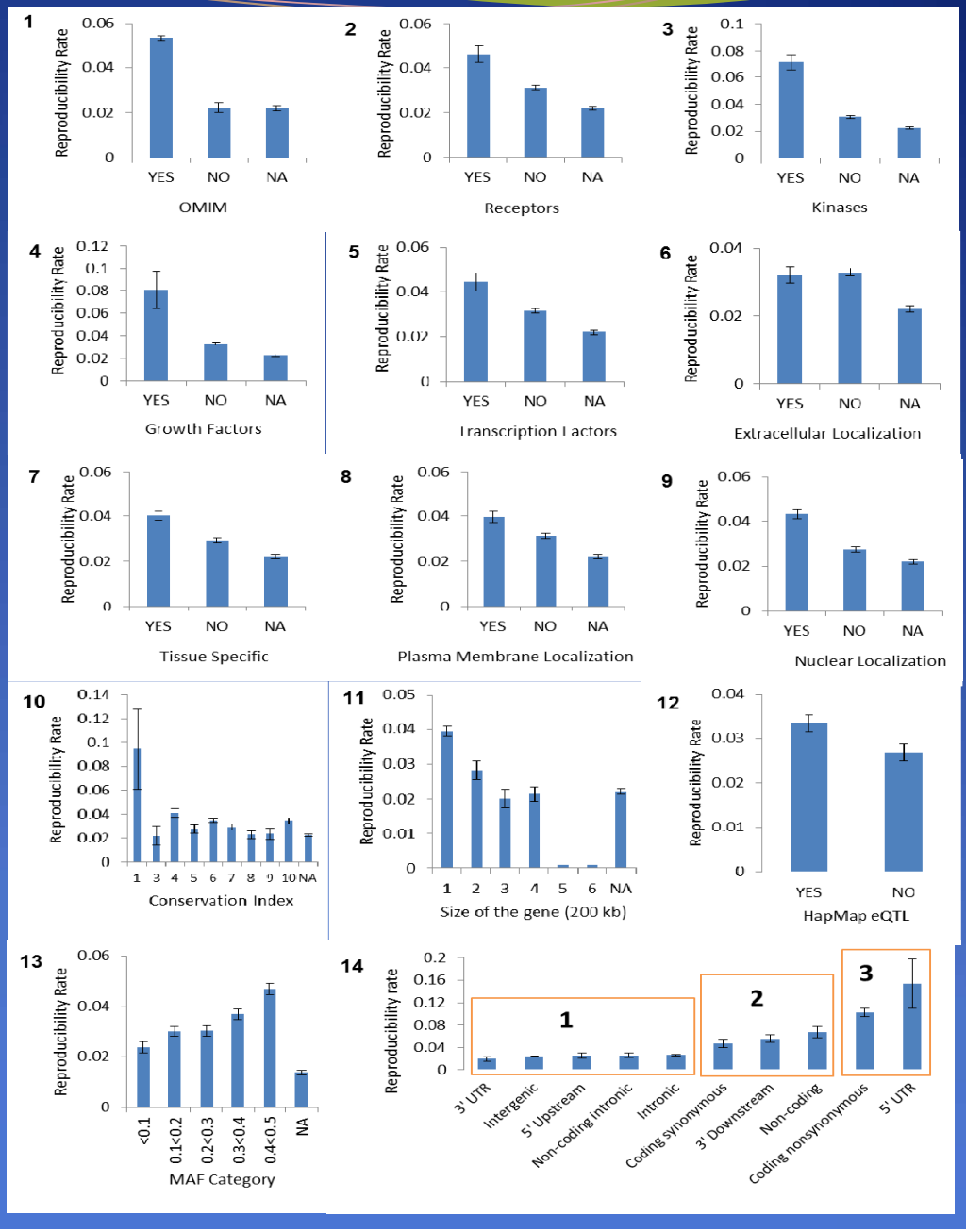
Univariate Test Results

Predictor	Test	Statistics	P-value
Conservation Index	Spearman correlation	Spearman R = 0.09	0.0007
eQTL	Mann-Whitney U test	(M-W U test) Z adjusted = -5.39	1.90E-07
Gene Size	Spearman correlation	Spearman R = -0.11	0.00001
Growth Factor	Mann-Whitney U test	(M-W U test) Z adjusted = -2.79	0.008
Kinase	Mann-Whitney U test	(M-W U test) Z adjusted = -7.84	2.40E-14
-Log(P)	Spearman correlation	Spearman R = 0.36	2.30E-08
MAF	Spearman correlation	Spearman R = 0.09	0.0007
Nuclear Localization	Mann-Whitney U test	(M-W U test) Z adjusted = -7.18	2.20E-12
OMIM	Mann-Whitney U test	(M-W U test) Z adjusted = 8.16	2.20E-15
Plasma Membrane	Mann-Whitney U test	(M-W U test) Z adjusted = -6.23	1.80E-09
Receptor	Mann-Whitney U test	(M-W U test) Z adjusted = -5.17	8.90E-07
SNP Type	Kruskal-Wallis (KW) test	(KW test) Chi-Square = 391. 8, df = 9	5.50E-79
Tissue Specific	Mann-Whitney U test	(M-W U test) Z adjusted = -3.97	0.0001
Transcription Factor	Mann-Whitney U test	(M-W U test) Z adjusted = 3.32	0.008

Further evaluating SNP Type

- (1) All pair-wise comparisons inside the group should be insignificant; and (2) All pairwise comparisons between the groups should be significant.
- SNP reproducibility was lowest in the group 1 (5' UTR, Intergenic, 5' Upstream, Non-coding intronic, Intronic), intermediate in the group 2 (Coding synonymous, 3' Downstream, Non-coding), and highest in the group 3 (Coding nonsynonymous, 5'UTR).
- 43 SNPs had >1 annotation

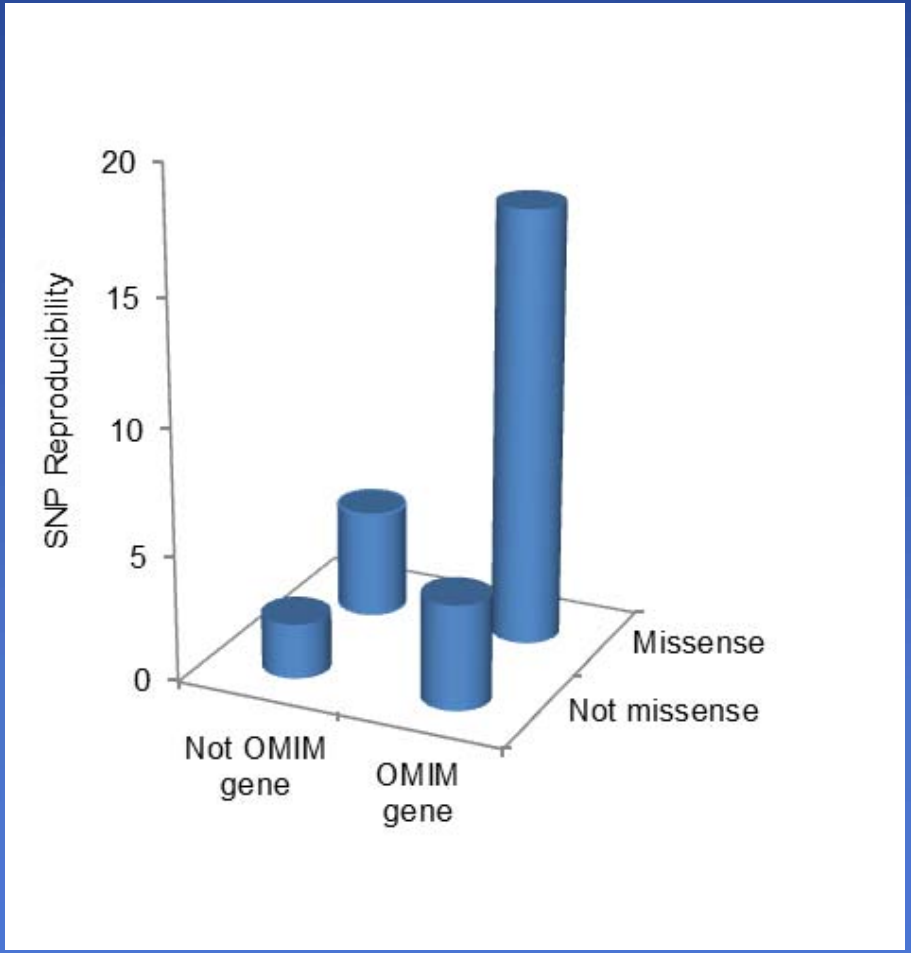
Univariate Predictors of Replication



Multivariable Analysis of Predictors

Characteristic	B	se(B)	p-value
SNP_TYPE	0.034687	0.006443	7.31E-08
Pvalue_mlog	0.000936	0.000187	5.34E-07
5_MAF_Groups	0.005509	0.001293	2.03E-05
OMIM	0.021424	0.007765	0.0058
Nuclear localization	0.025213	0.008232	0.002192
Kinases	-0.01213	0.014694	0.409105
Conservation index	0.052759	0.04749	0.266581
Growth factor	-0.01789	0.020355	0.379
Tissue_specific	-0.00375	0.007339	0.609778
eQTL HapMap	-0.00122	0.007148	0.864
Receptors	0.019558	0.012282	0.111304
Transcription factors	-0.02182	0.012491	0.080629
Gene Size	0.011653	0.007745	0.132435

Interactions among factors

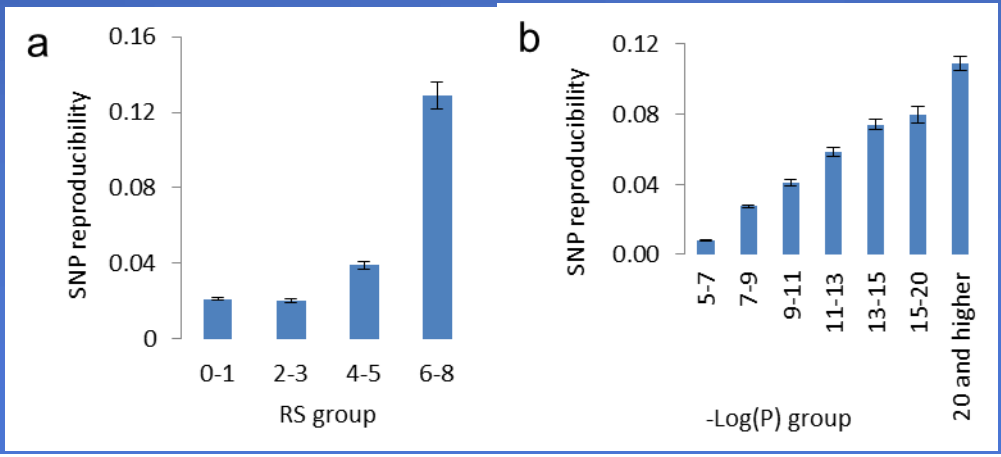


Identification of Causal Variants

- Showing a variants has a causal effect on a disease process is challenging
- Good candidate for the causal variant should be linked to relevant biological mechanism. For example it may have effect on gene expression or protein structure/folding.
- Functional analysis is needed to prove that (1) SNP change the important biological function, and that (2) alteration of function is associated with disease risk.

Conclusions

- Multivariable analysis showed that 11% of variability was explained by SNP predictors with 5% attributed to –log(P) value alone.
- SNP characteristics are second most prominent, followed by MAF (in wrong direction for weighting)
- Could also define profile measurement



OncoArray – A platform for Pan Cancer Discovery

Common Content – 40K

Fine-mapping of common cancer susceptibility loci (*TERT*, 8q24 (proximal and distal to *MYC*), *HNF1B, TET2, RAD51B*, 11q13, *MERIT40*, *MDM4)*

Ancestry Informative Markers

Cross-Site meta analysis

Pharmacogenetic components

eQTL (Height, Weight, BMI, WHR, Menarche, Menopause etc)

Other cancers published GWAS variants

Chromosome X and mitochondrial DNA variants

GWAS Backbone

260K

Illumina Core

OncoChip

600K

beadtypes

Cancer Specific Variants

Lung

Colon

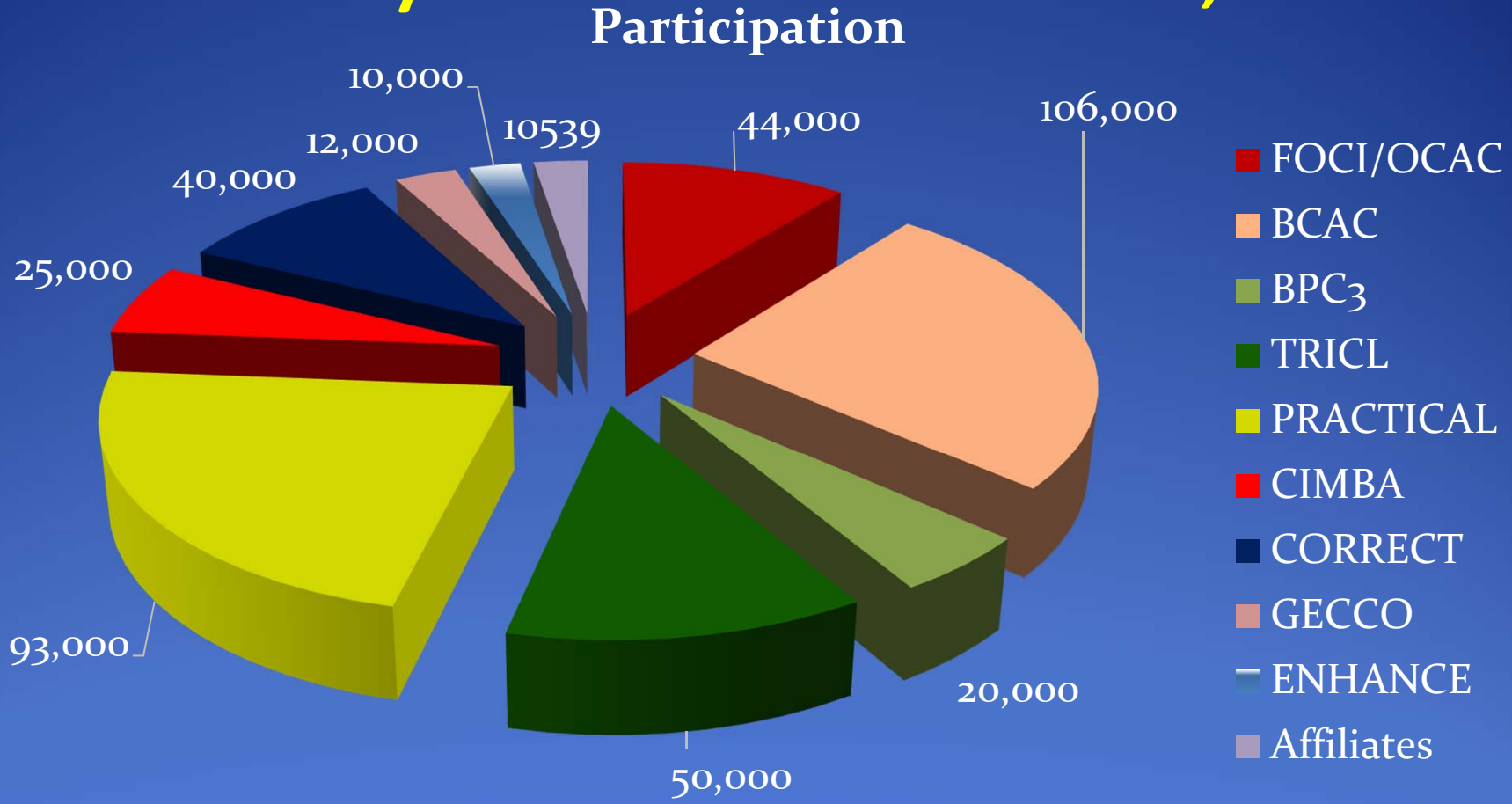
Breast

Prostate

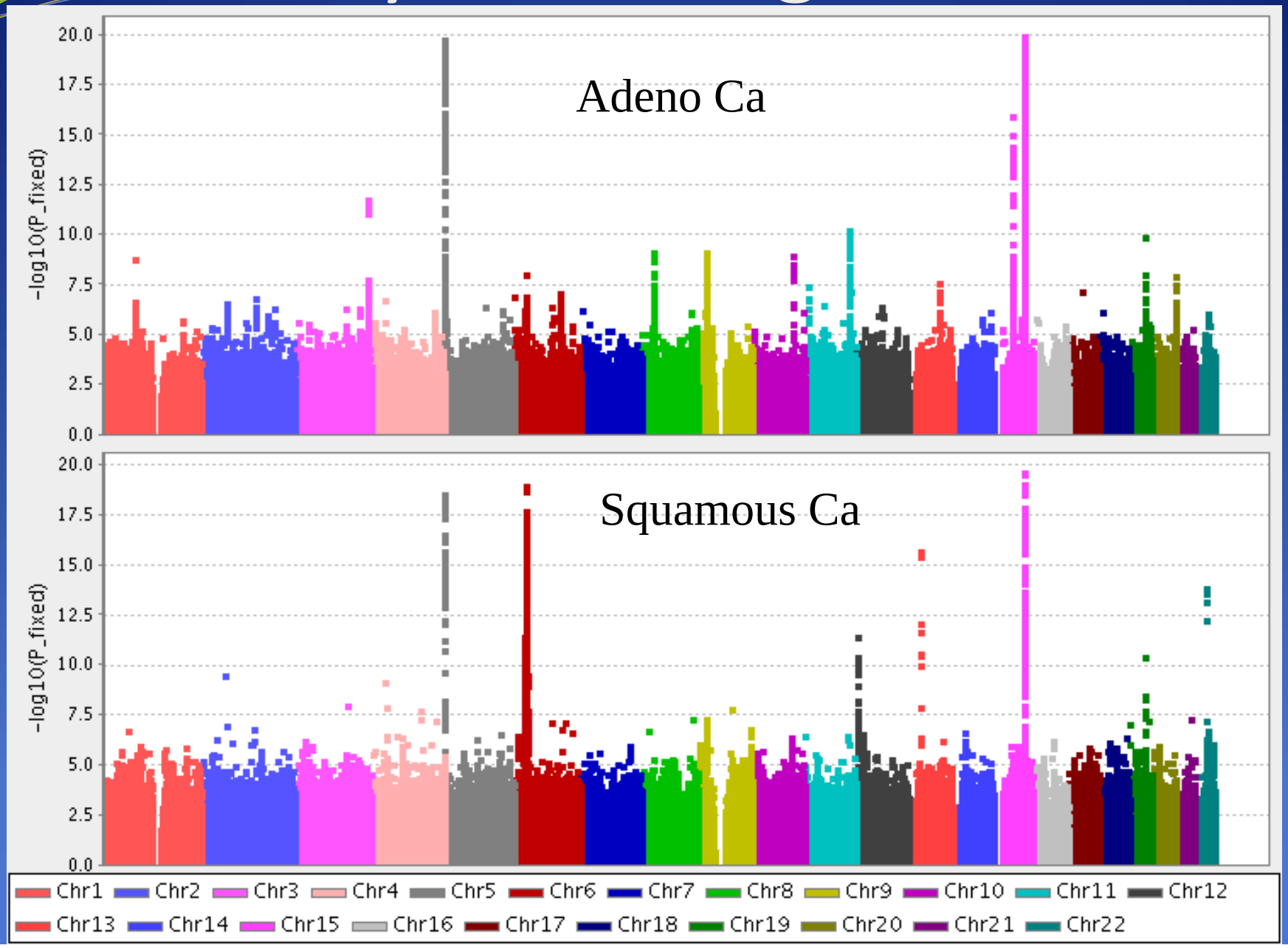
Ovarian

(proportional allocation)

Major Participants in GAME-ON Oncoarray Network n=410,539

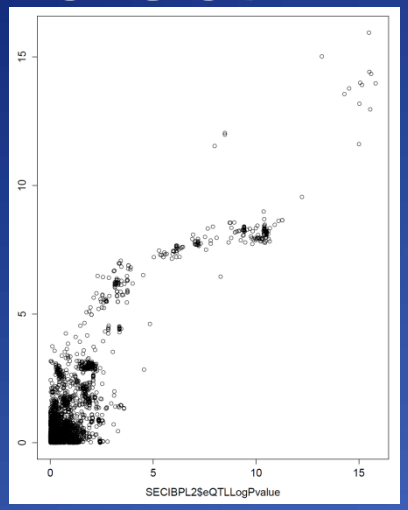


Oncoarray - Histologies

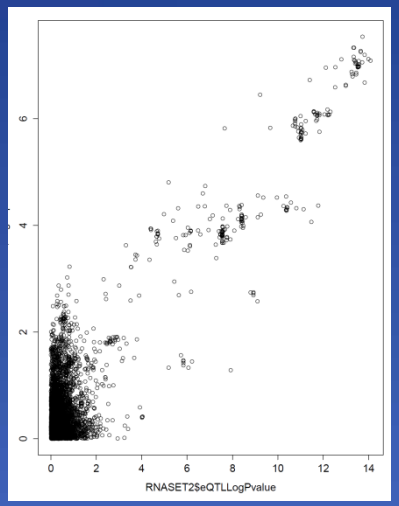


Understanding eQTL effects

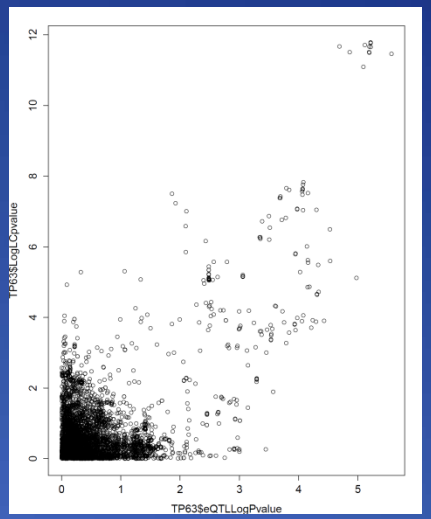
SECISBP2L



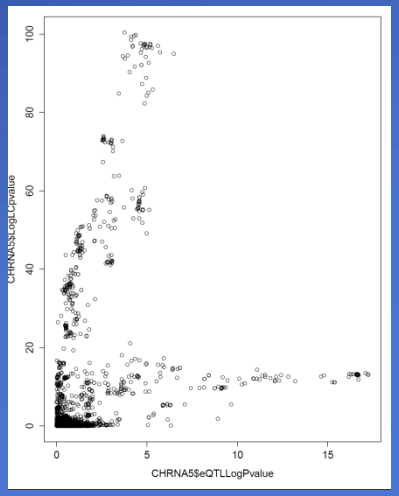
RNASET2



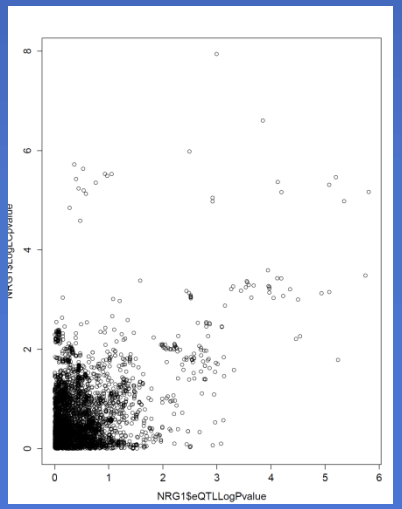
TP63



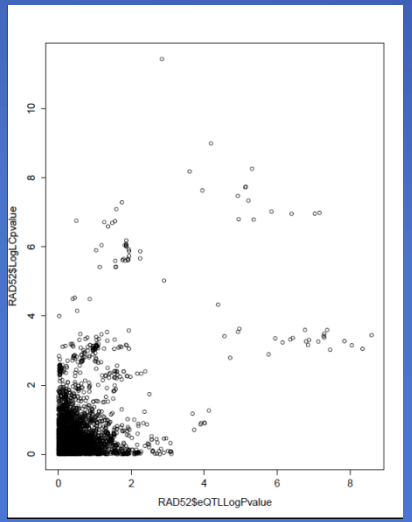
CHRNA5



NRG1



RAD52



Why are some people protected?

Identifying Interactions

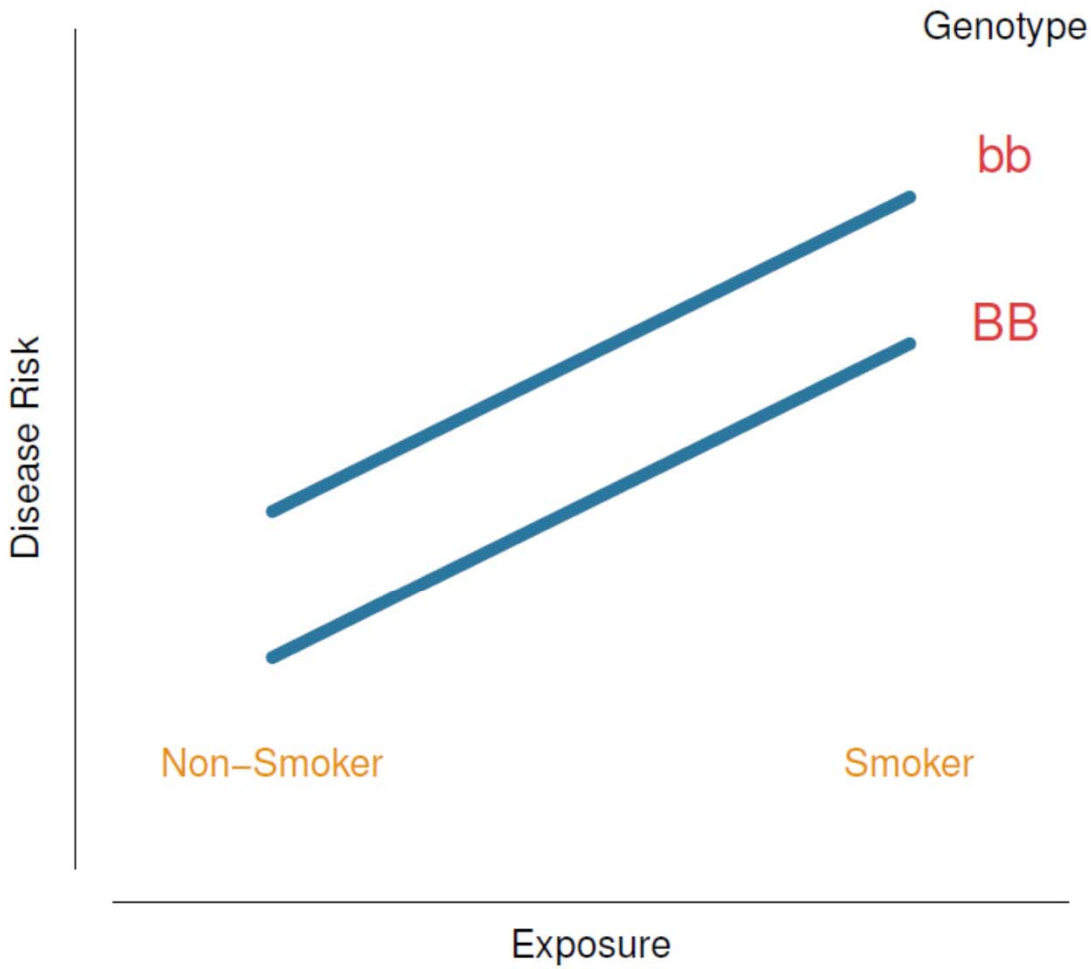
- Definition: Non-additivity of the effects of factors
- Gene-Gene and Gene-Environment Interactions: Different
- For linear models,

$$f(E(y_i)) = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2}$$

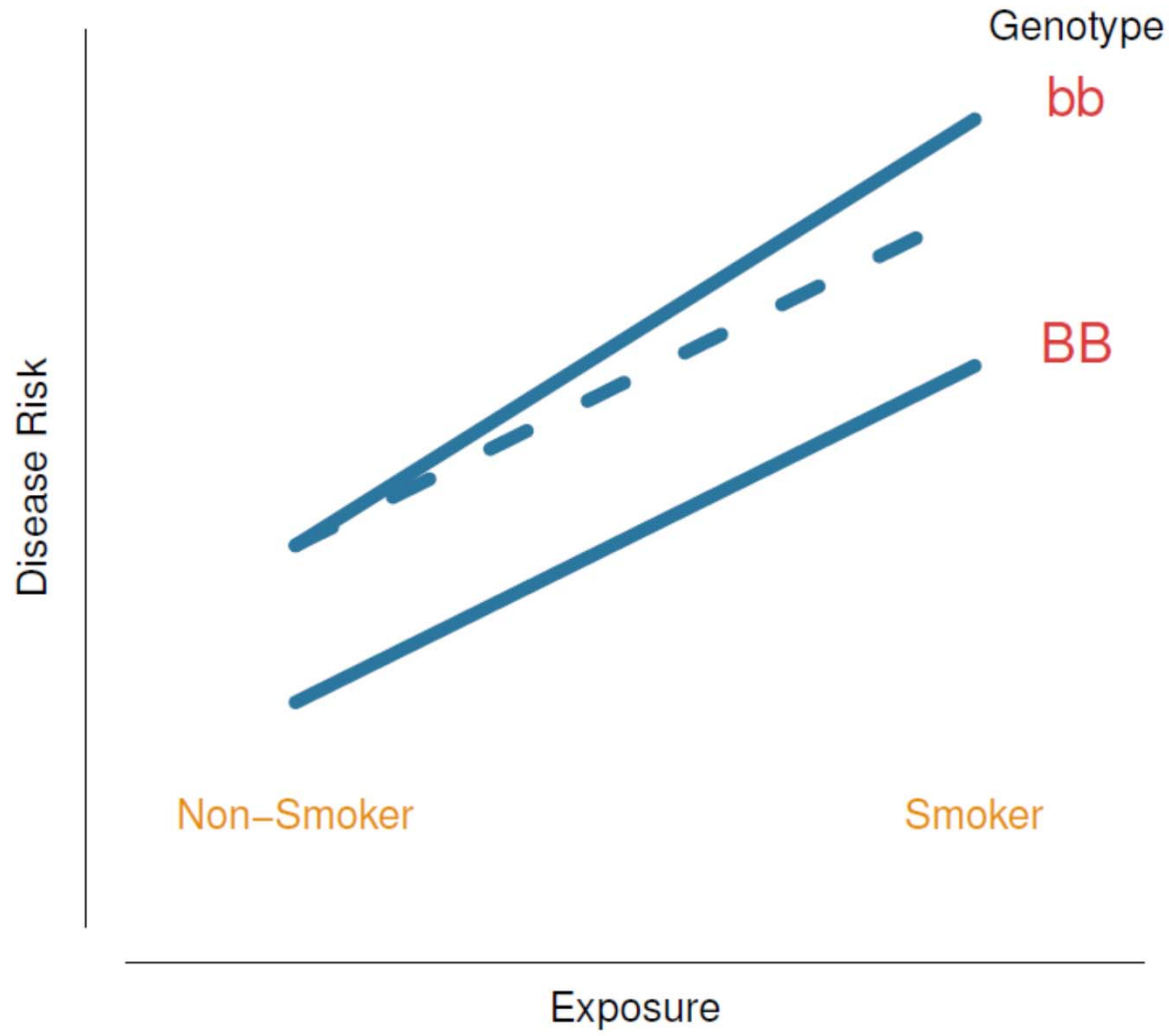
$$\beta_{12} \neq 0$$

Liu, et al. Hum Genet. 2015 Jan;134(1):23-36

GxE Interaction



GxE Interaction



Evaluating interaction effects using Bayesian Framework

- Inference: $f(\Theta|data) \propto f(\Theta) \times L(data|\Theta)$
- Computation: Markov chain Monte Carlo (WinBUGS)
- Related to Frequentist: e.g.,
  - Penalized Likelihood Method:

$$J(\Theta) = log(L(data|\Theta)) - \pi(\Theta)$$

  - Bayesian Model:

$$log(f(\Theta|data)) = log(L(data|\Theta)) + log(f(\Theta)) + C$$

1. Objective: Identify the genetic and environmental factors and interactions that are related to the disease status

$$y_i \overset{F}{\leftarrow} X_i,\ Z_i,\ X_i \otimes Z_i,\ X_i \otimes X_i$$

2. Develop a novel Bayesian variable selection model with the hierarchical constraints on the main effects and interactions

Stochastic Search Variable Selection for Interactions

$$E(f(y_i)) = \alpha + BX_i + \Gamma Z_i + \Theta(X_i \otimes Z_i) + K(X_i \otimes X_i)$$

where,

$$\beta_s \sim N(0, I_s\sigma_s^2 + (1 - I_s)\sigma_{s_\epsilon}^2)$$

$$\gamma_e \sim N(0, I_e\sigma_e^2 + (1 - I_e)\sigma_{e_\epsilon}^2)$$

$$\theta_{se} \sim N(0, I_{se}\sigma_{se}^2 + (1 - I_{se})\sigma_{se_\epsilon}^2)$$

$$k_{ss} \sim N(0, I_{ss}\sigma_{ss}^2 + (1 - I_{ss})\sigma_{ss_\epsilon}^2)$$

then,

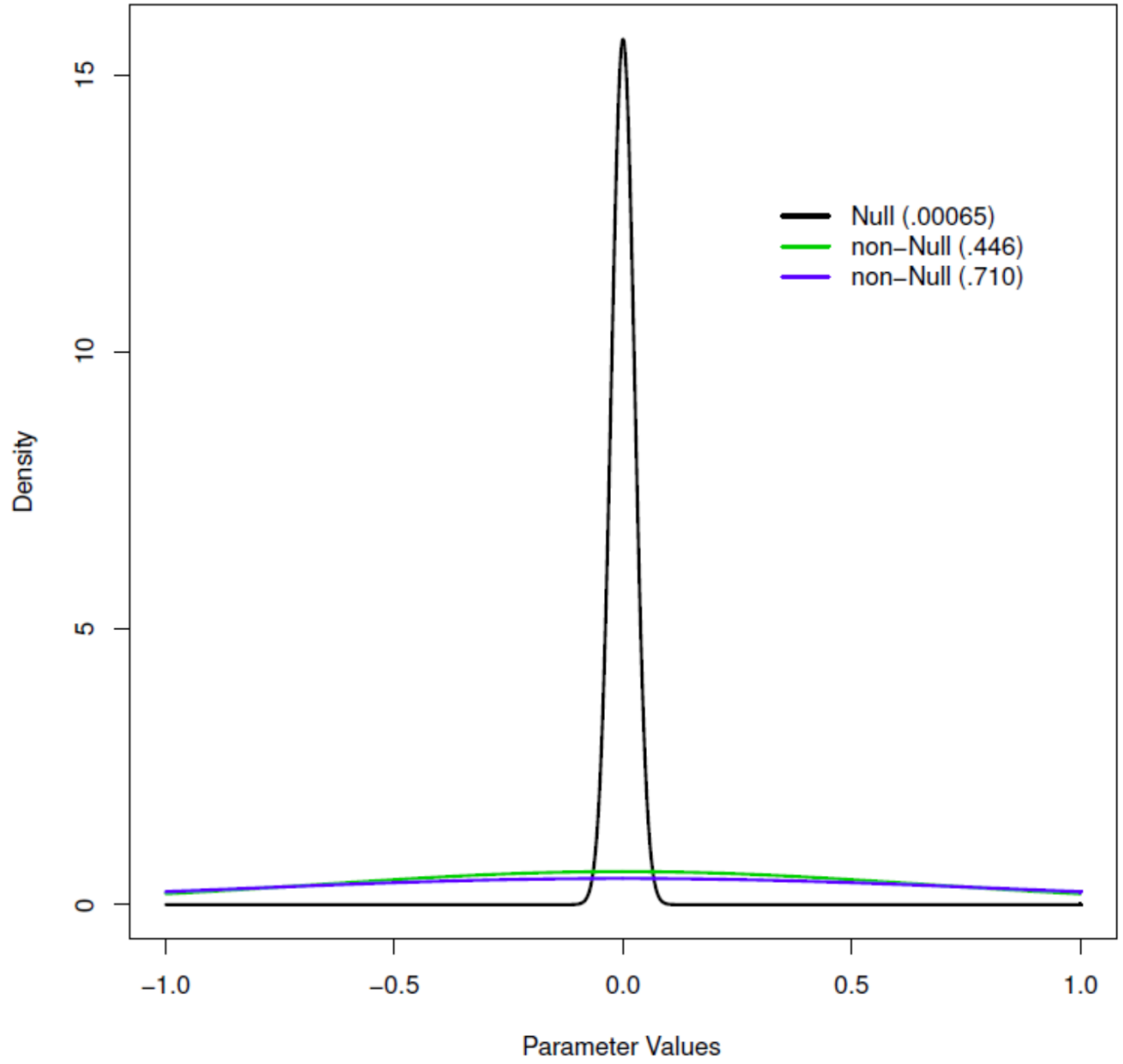
$$I_s \sim Bin(\pi_s)$$

$$I_e \sim Bin(\pi_e)$$

$$I_{es} \sim Bin(\pi_{es})$$

$$I_{ss} \sim Bin(\pi_{ss})$$

Spike and Slab Prior for SSVS



- Two constraint patterns:
  1. Strong Hierarchical: an interaction term can be included in model only if the corresponding main effects are also included in the model

$$\beta_{12} \neq 0 \Rightarrow \beta_1 \neq 0 \cap \beta_2 \neq 0$$

  2. Weak Hierarchical: an interaction term can be included in model only if at least one of the corresponding main effects are included in the model

$$\beta_{12} \neq 0 \Rightarrow \beta_1 \neq 0 \cup \beta_2 \neq 0$$

Lung Cancer

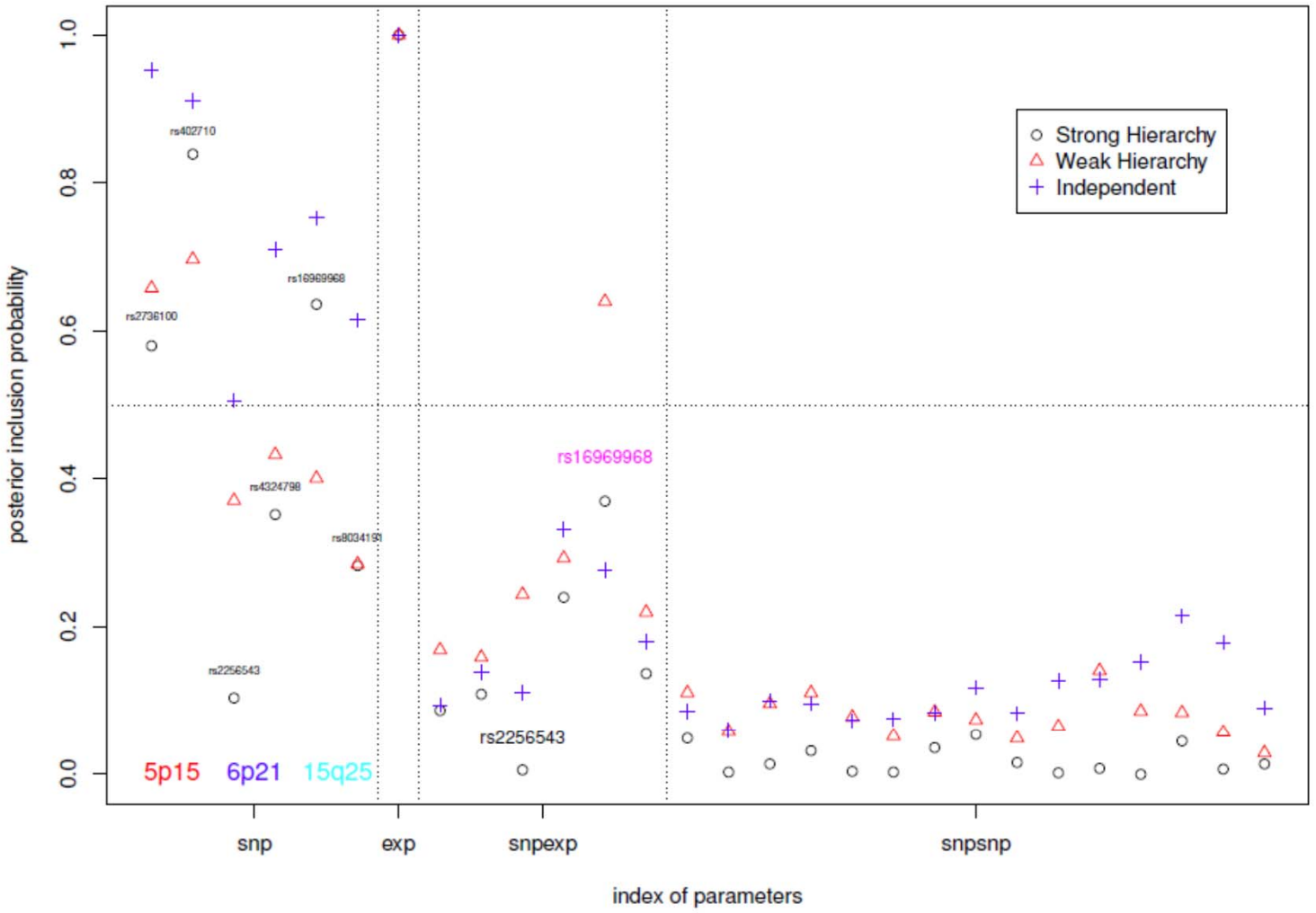


Figure : Real data result for the lung cancer study.

Summary of findings

- Genome wide association studies provide insights into regions but often do not identify specific causal variants
- Including biological information may improve identification of causal variants
- Interaction analysis may need to impose constraints given large number of possible interactions to improve accuracy of inferences