

Statistical and practical issues for replicating GWAS signals in meta-analysis

Dajiang Liu

Reproducibility Boot Camp

06/07/2019

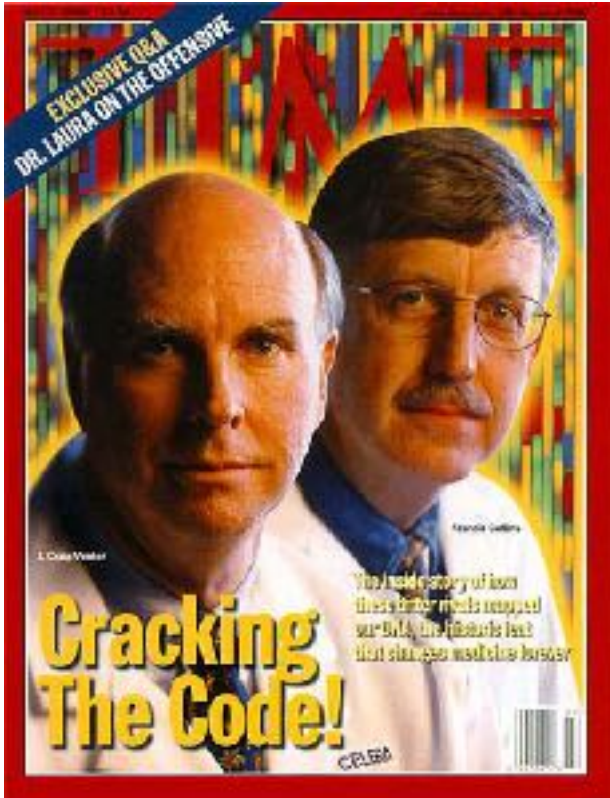
Outline

- Forewords
- Overview of GWAS Meta-analysis and follow-up studies
- Motivation of assessing reproducibility of GWAS signals
- MAMBA method
 - Joint work with Dan McGuire and Qunhua Li
- Conclusions

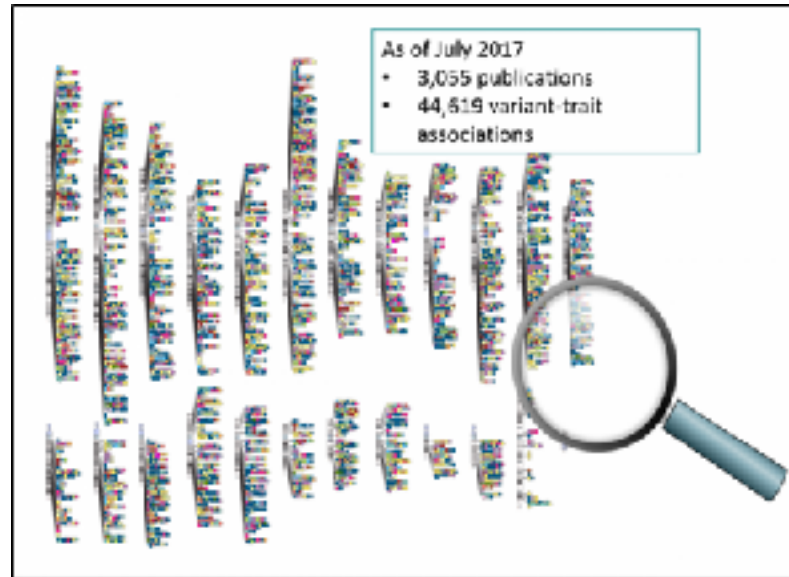
Forewords

- Why reproducibility?
 - People all make mistakes
 - People are inclined to rush to conclusions and publish
- Reproducibility does not make research error-free
 - But allows us to “reproduce errors” and seek their root causes
- Reproducibility can be achieved by
 - The right attitude
 - Careful documentation
 - Statistical methods

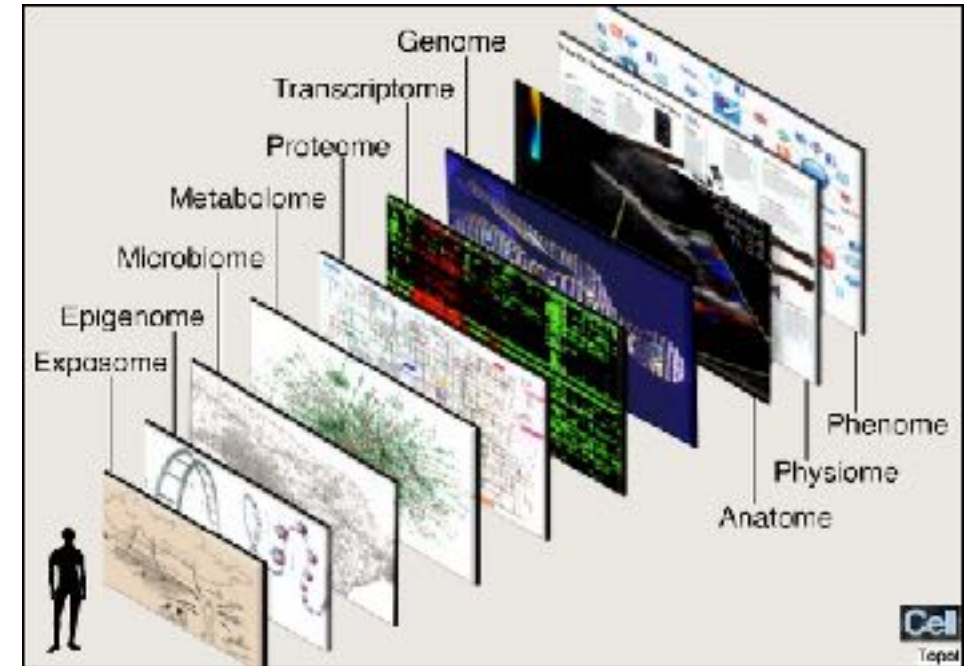
A Quick Look Back



Initial Optimism:
Human genome project, a historic feat that changes medicine forever



Booming of GWAS:
Lots of discoveries made; many hypotheses generated;
Even more unknowns to be explored



Mechanistic studies:
Dissect GWAS loci;
Learn new biology;
Personalized medicine;

Big Picture Questions

Mechanism

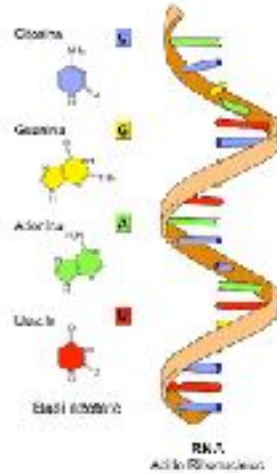
DNA



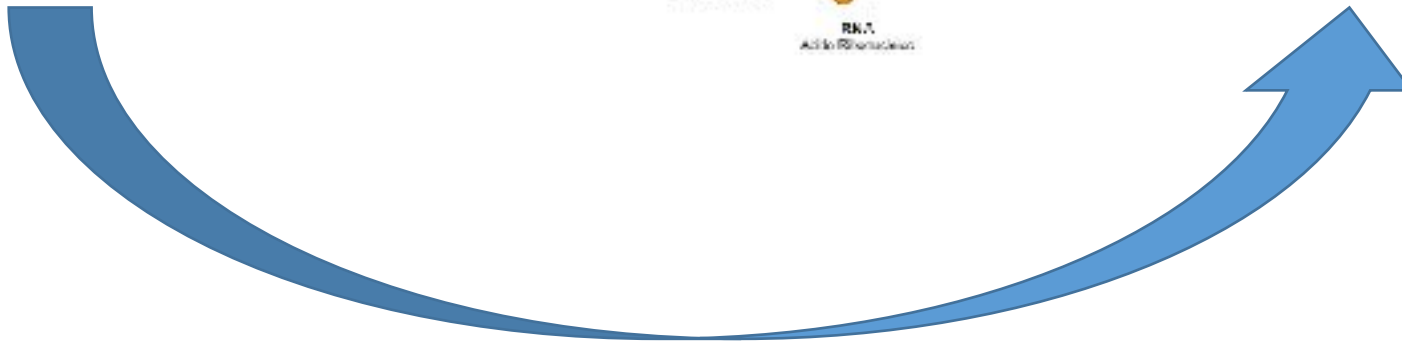
Molecular
Phenotypes



Complex
Phenotypes

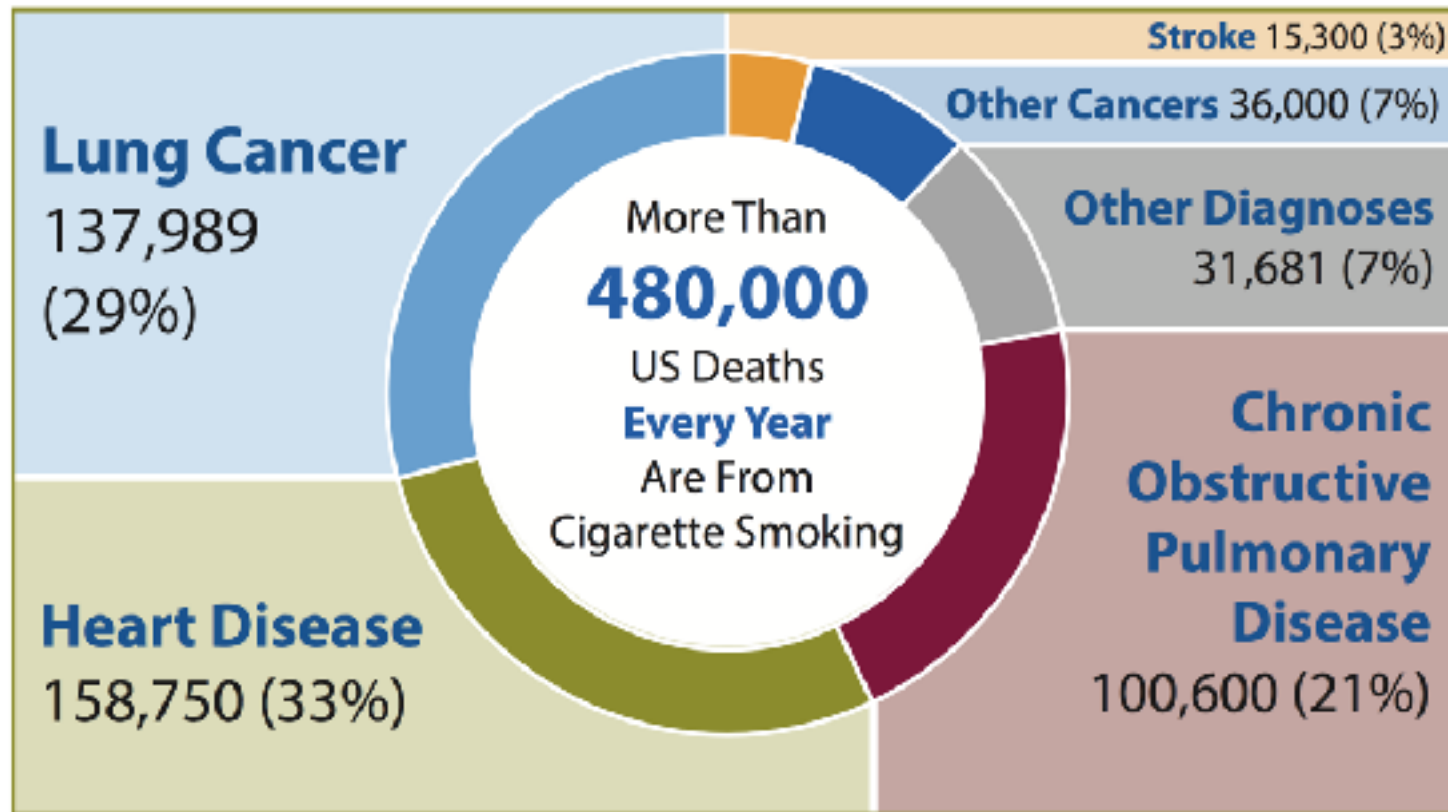


Clinical Translation



Smoking and Drinking – Biggest Modifiable Risk Factor for Death

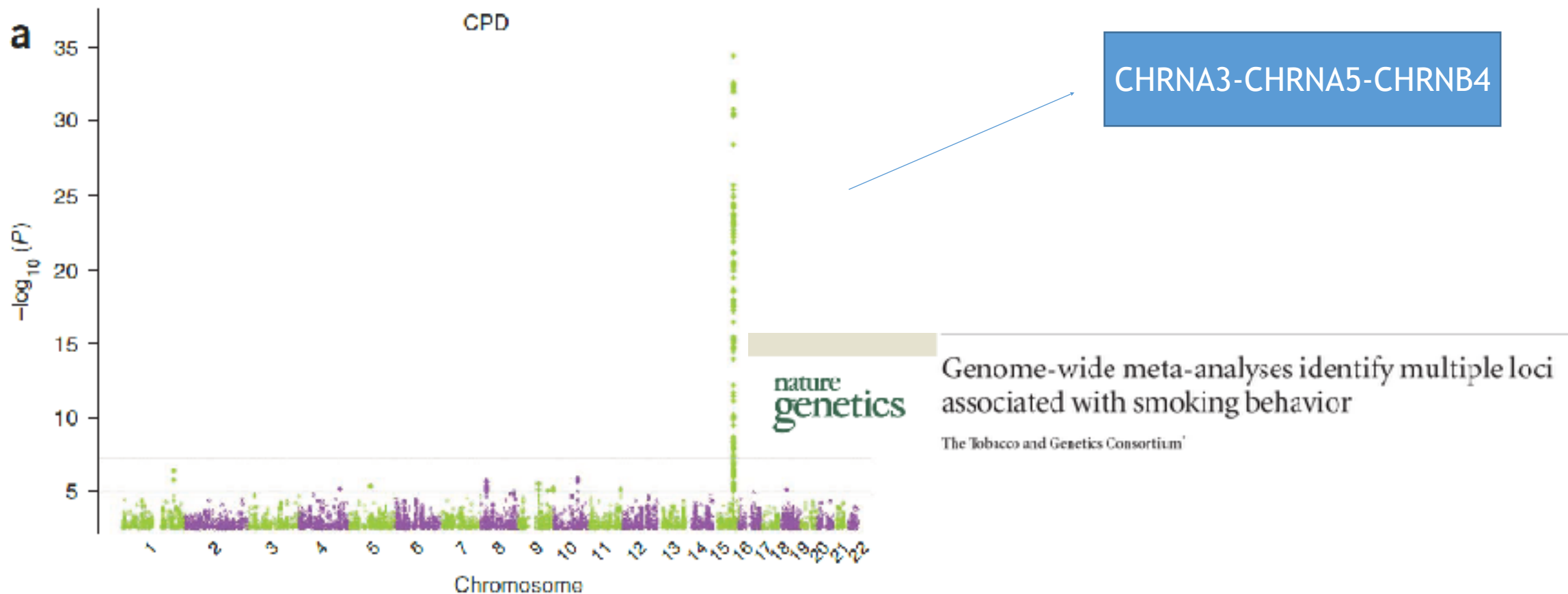
Annual Deaths from Smoking, United States



Note: Average annual number of deaths for adults aged 35 or older, 2005-2009.

Source: [2014 Surgeon General's Report, Table 12.4, page 660.](#)

Despite Being Heritable, Little was Found for Cigarette-Per-Day with N=75,000 (in 2010)



GSCAN Study Design

- GWAS and Sequencing Consortia of Alcohol and Nicotine Use (GSCAN)
 - Scott Vrieze, Univ. Minnesota
- Each study contribute summary association statistics
 - Genetic effect estimates and their standard deviations
- Imputation to Haplotype Reference Consortium panel
 - 15.9 million variant post filtering

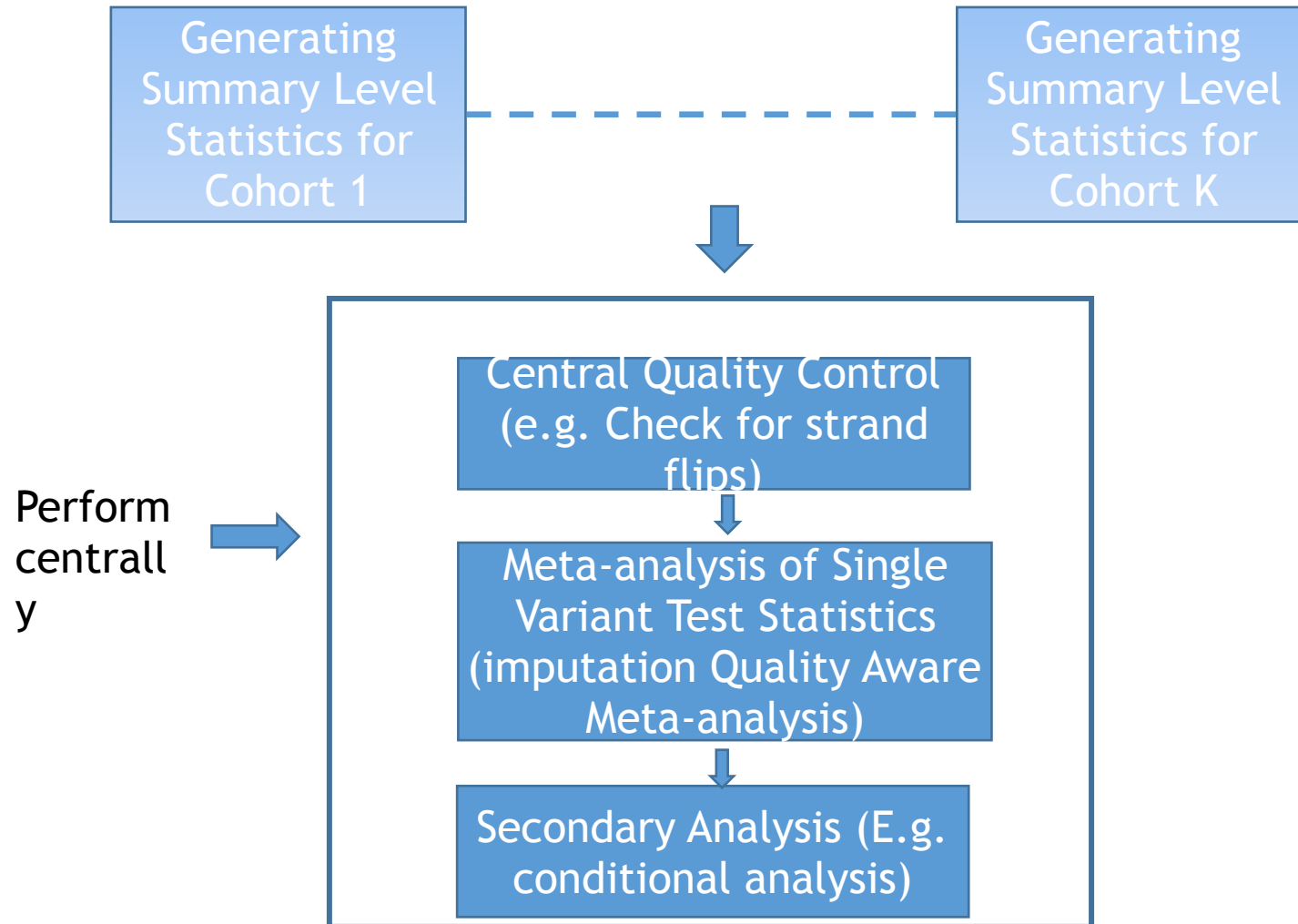


**Association studies of up to 1.2 million individuals
yield new insights into the genetic etiology of
tobacco and alcohol use**

GSCAN Study

Phenotypes	Abbreviations	Sample Size	Type
Age of Initiation of Smoking	AgeInit	341,427	Continuous
Smoking Initiation (Ever Smoker vs. Never Smoker)	SmkInit	1,232,091	Binary
Cigarettes Per Day	CigPerDay	337,334	Continuous
Smoking Cessation (Current vs. Former Smoker)	SmkCes	547,219	Binary
Drinks Per Week	DrnkWk	941,280	Continuous

Workflow for RAREMETAL

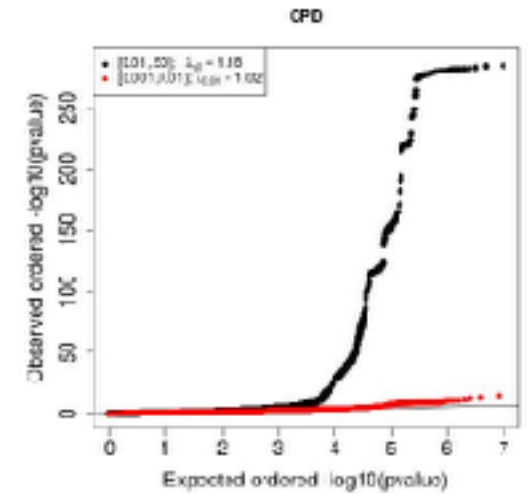
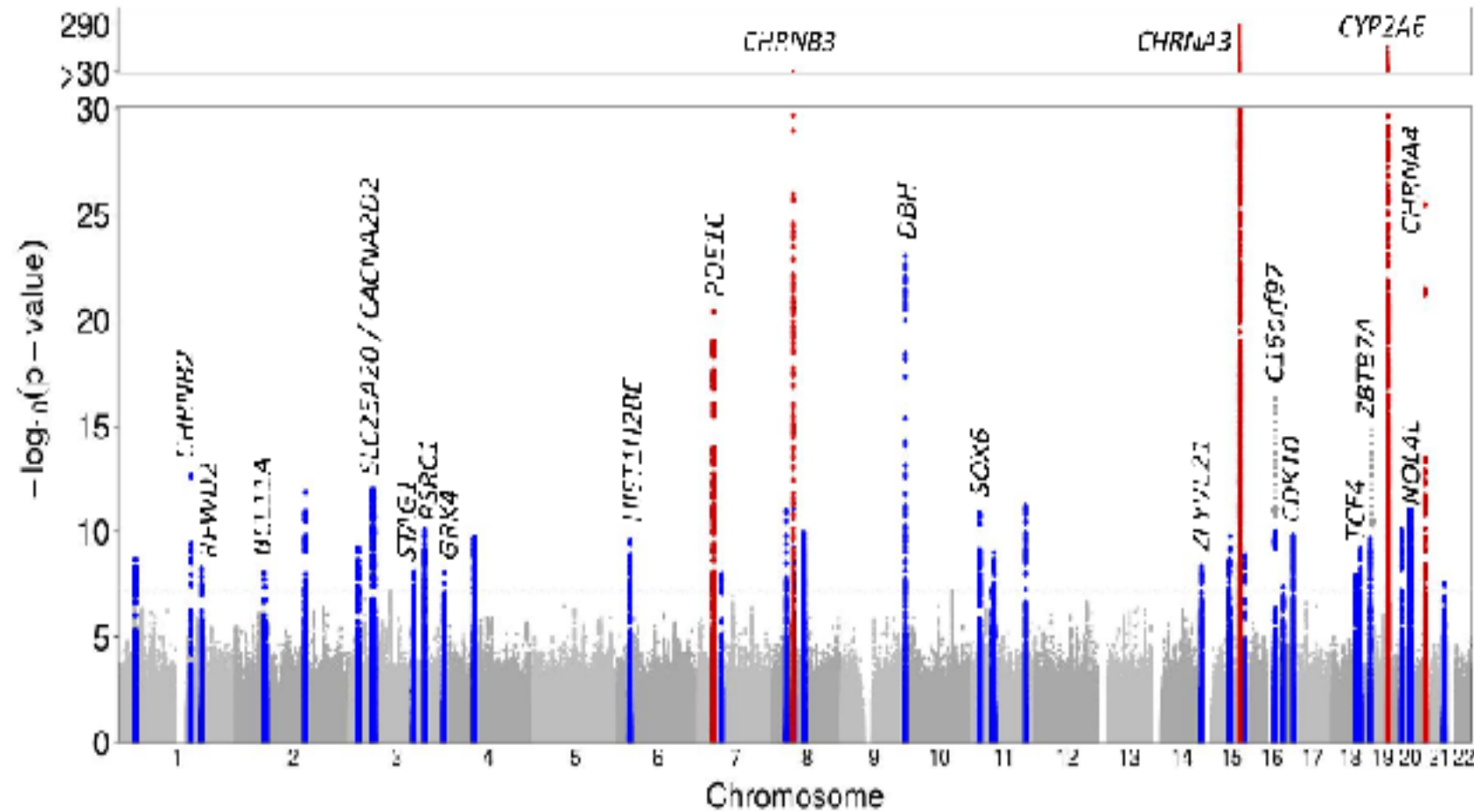


Zhan X, Hu Y, Li B, Abecasis G, Liu DJ
RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data, **Bioinformatics** 2016

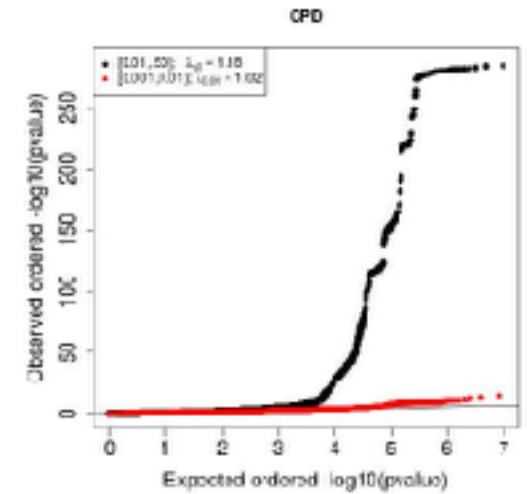
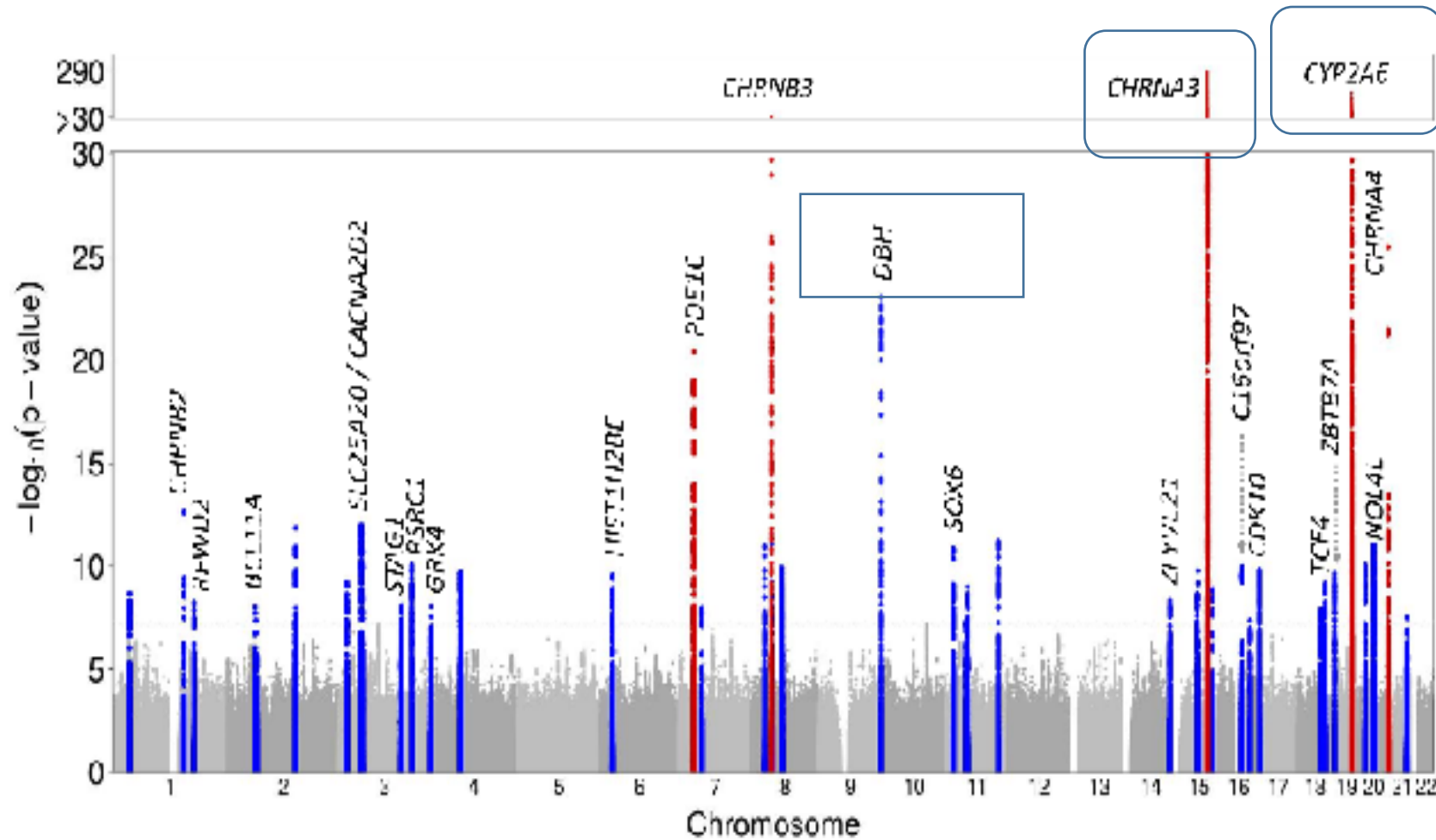
Liu DJ et al *Meta-analysis of Gene-level Tests of Rare Variant Association*, **Nature Genetics** 2014

Jiang Y et al, *Proper conditional analysis in the presence of missing data: Application to large scale meta-analysis of tobacco use phenotypes*, **PLOS Genetics**, 2018

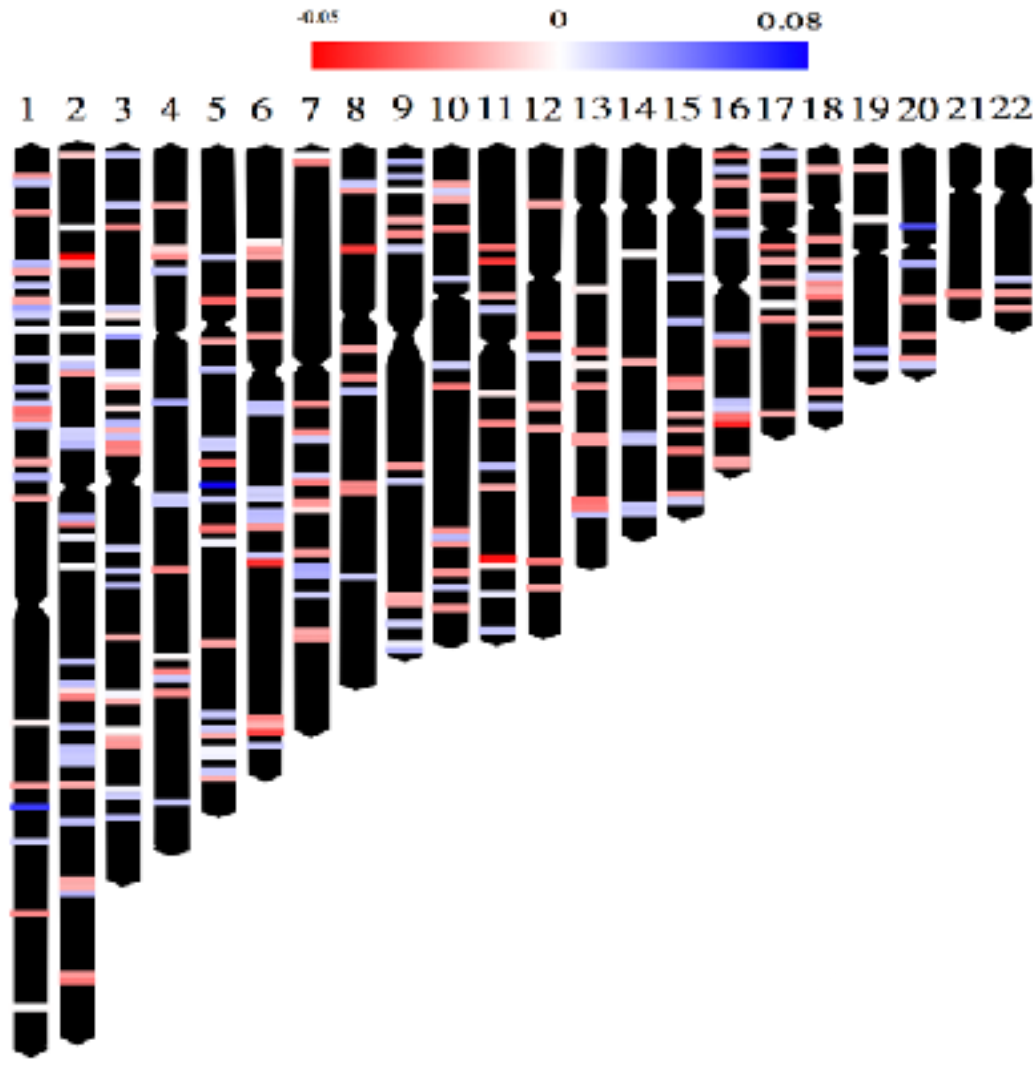
Results for CigDay N=1,000,000



Results for CigDay N=1,000,000



GSCAN GWAS results (All Traits)



- GSCAN GWAS phase 1, **completed!**
 - **406 loci uncovered**
- Next challenge: Interpret the **functional consequence** of these loci

How Many Loci Are “Real”?

- Population based studies can be subject to artefacts
 - Population structure;
 - cryptic relatedness;
 - Genotyping error
 - Etc
- Conventional approach to ensure the validity of the discovery is **replication**
- But
 - Having a well-powered replication study is almost impossible in modern-day genetic studies
 - Ad-hoc procedures can be applied, but hard to be broadly applied and generalize

The Field Recognized This Problem, But No Solution

COMMENT

DOI: 10.1038/s41467-018-07348-x

OPEN

Examining the current standards for genetic discovery and replication in the era of mega-biobanks

J.E. Huffman  ¹



False Positive Signals Often “Disguise” Themselves

Received: 5 August 2018 | Revised: 9 November 2018 | Accepted: 11 December 2018


DOI: 10.1002/gepi.22189

RESEARCH ARTICLE

WILEY Genetic
Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC
EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Spinning convincing stories for both true and false association signals

Richard J. Biedrzycki¹ | Ashley E. Sier^{1,2*} | Dongjing Liu^{1*} | Erika N. Dreikorn¹ |
Daniel E. Weeks^{1,3} 

A Behavioral Experiment

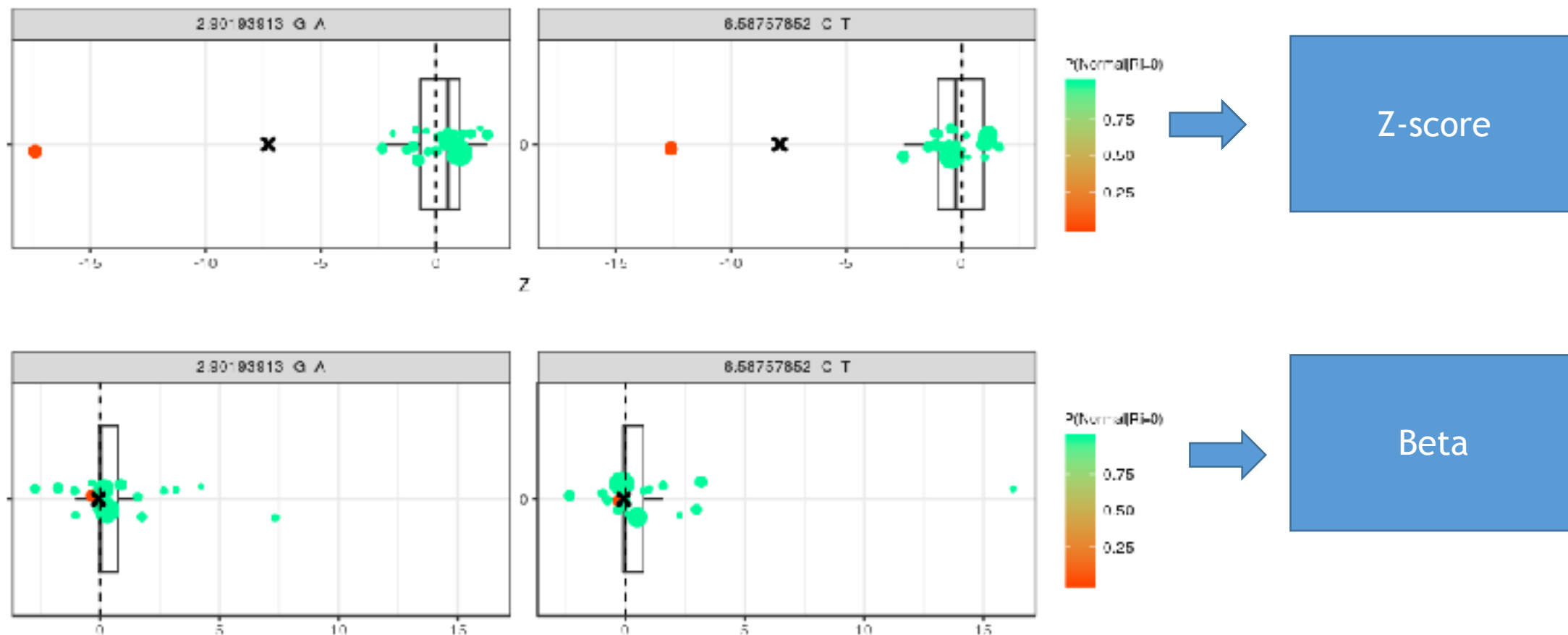
Abstract

When interpreting genome-wide association peaks, it is common to annotate each peak by searching for genes with plausible relationships to the trait. However, “all that glitters is not gold”—one might interpret apparent patterns in the data as plausible even when the peak is a false positive. Accordingly, we sought to see how human annotators interpreted association results containing a mixture of peaks from both the original trait and a genetically uncorrelated “synthetic” trait. Two of us prepared a mix of original and synthetic peaks of three significance categories from five different scans along with relevant literature search results and then we all annotated these regions. Three annotators also scored the strength of evidence connecting each peak to the scanned trait and the likelihood of further studying that region. While annotators found original peaks to have stronger evidence ($p_{\text{Bonferroni}} = 0.017$) and higher likelihood of further study ($p_{\text{Bonferroni}} = 0.006$) than synthetic peaks, annotators often made convincing connections between the synthetic peaks and the original trait, finding these connections 55% of the time. These results show that it is not difficult for annotators to make convincing connections between synthetic association signals and genes found in those regions.

How to Assess Replicability

- A motivating example

- Size of the point: Sample Size



Method for Assessing Replicability

- MAMBA: meta-analysis model-based assessment of replicability
- Observation:
 - Real signals are more likely to be significantly associated in multiple studies
 - Real signals are likely to be consistent across different studies
- Key idea:
 - Use mixture model to represent replicable and non-replicable SNPs
 - Based upon empirical data, non-replicable SNP is modeled as SNPs with disproportionately large variability in the effect sizes
 - Develop a coherent framework for robust meta-analysis and assessing replicability at the same time
 - Posterior probability of replicability can be estimated using ECM algorithms
 - Frequentist p-values can be calculated using parametric bootstrap based upon the posterior probability of replicability

Model Details

- Model input:
 - M SNPs
 - K different studies
 - Genetic effect across different studies
 - $\hat{\beta}_{jk}, j=1,\dots,M$ and $k=1,\dots,K$
 - Standard deviation of the genetic effects
 - $s_{jk}, j=1,\dots,M$ and $k=1,\dots,K$

$$\hat{\beta}_{jk} \sim N(\mu_j, s_{jk}^2)$$

Model Details

- Depending on the replicability

- Replicable SNPs:

- $\mu_j \sim N(0, \tau^2)$

- Non-replicable SNPs:

- $\mu_j = 0$

- Non-outlier

- $\hat{\beta}_{jk} \sim N(0, s_{jk}^2)$

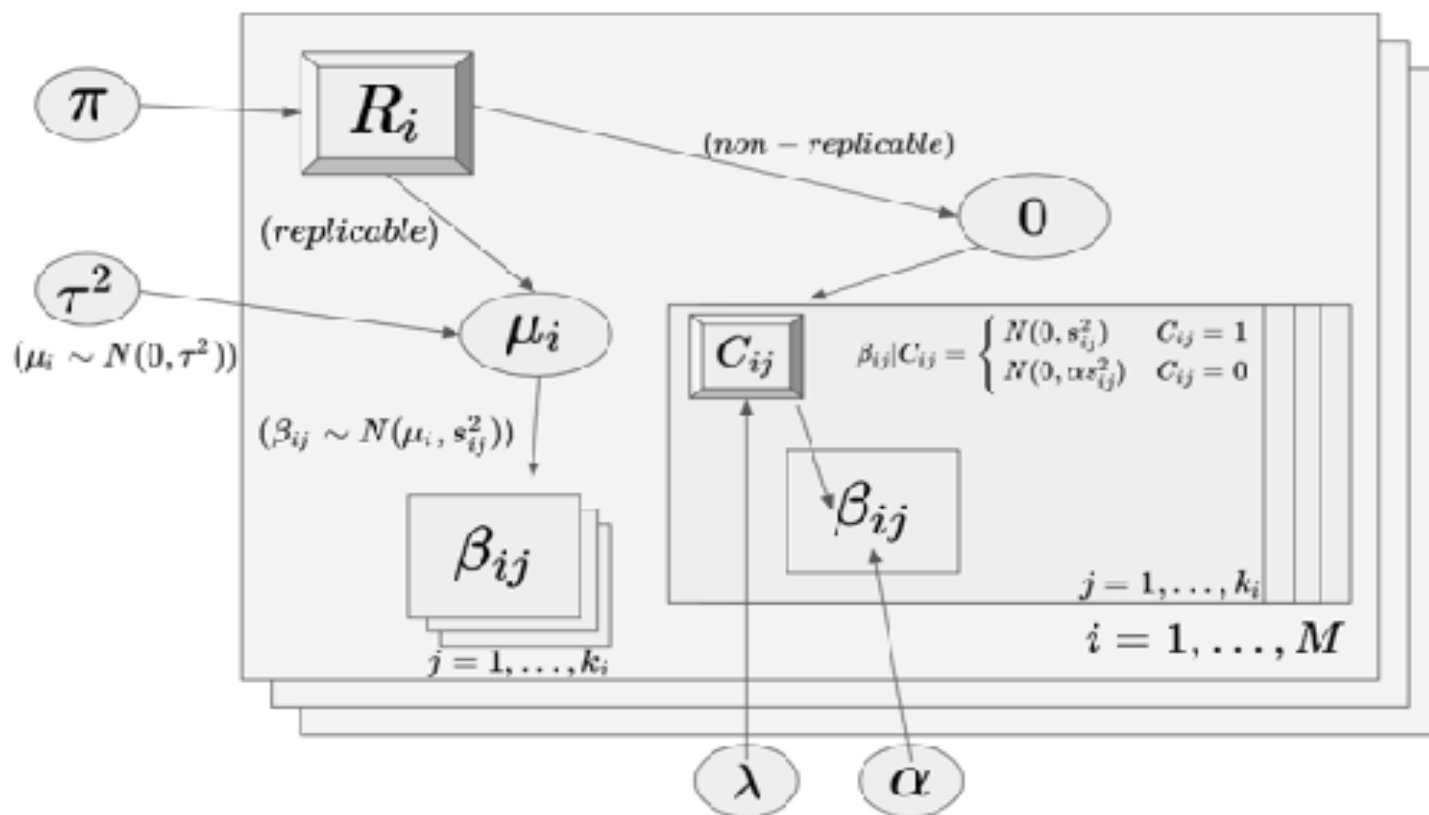
- Outlier

- $\hat{\beta}_{jk} \sim N(0, \alpha s_{jk}^2)$

- The full likelihood is the product of per-SNP likelihood

- Assume SNPs are independent

- Reasonable assumption for assessing the replicability of associated SNPs



Workflow of MAMBA

- **Step 0:** Do fixed-effects GWAMA first, to identify loci of interest we want to test with the MAMBA model.
- **Step 1:** Prune summary statistics and obtain approximately independent set of variants based upon a reference panel
- **Step 2:** Create dataset for MAMBA model by combining pruned variants
- **Step 3:** Fit MAMBA models (by chromosome)

Comparison with Alternative Methods

- Alternative methods considered
 - FE:
 - Fixed effect meta-analysis
 - RE:
 - Random effect meta-analysis (with maximum likelihood estimation)
 - RE2:
 - Random effect meta-analysis assuming no heterogeneity under the null

ARTICLE

Random-Effects Model Aimed at Discovering Associations
in Meta-Analysis of Genome-wide Association Studies

Buham Han¹ and Eleazar Eskin^{2,*}

• BE

- Binary effect model assuming the effect is a mixture of 0 and a non-zero fixed effect

OPEN ACCESS Freely available online

PLOS GENETICS

**Interpreting Meta-Analyses of Genome-Wide Association
Studies**

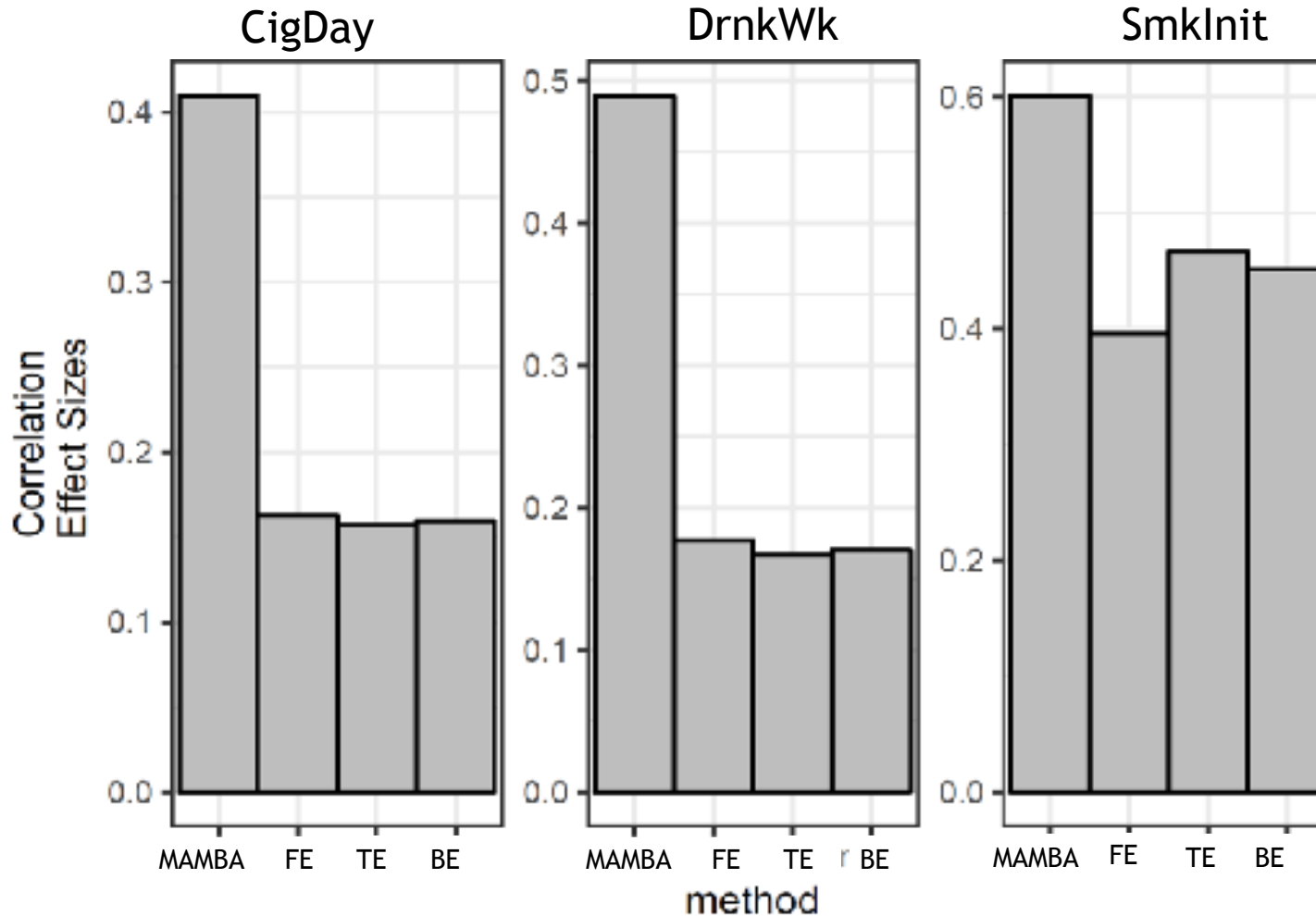
Buham Han¹, Eleazar Eskin^{1,2*}

Real Data Evaluation

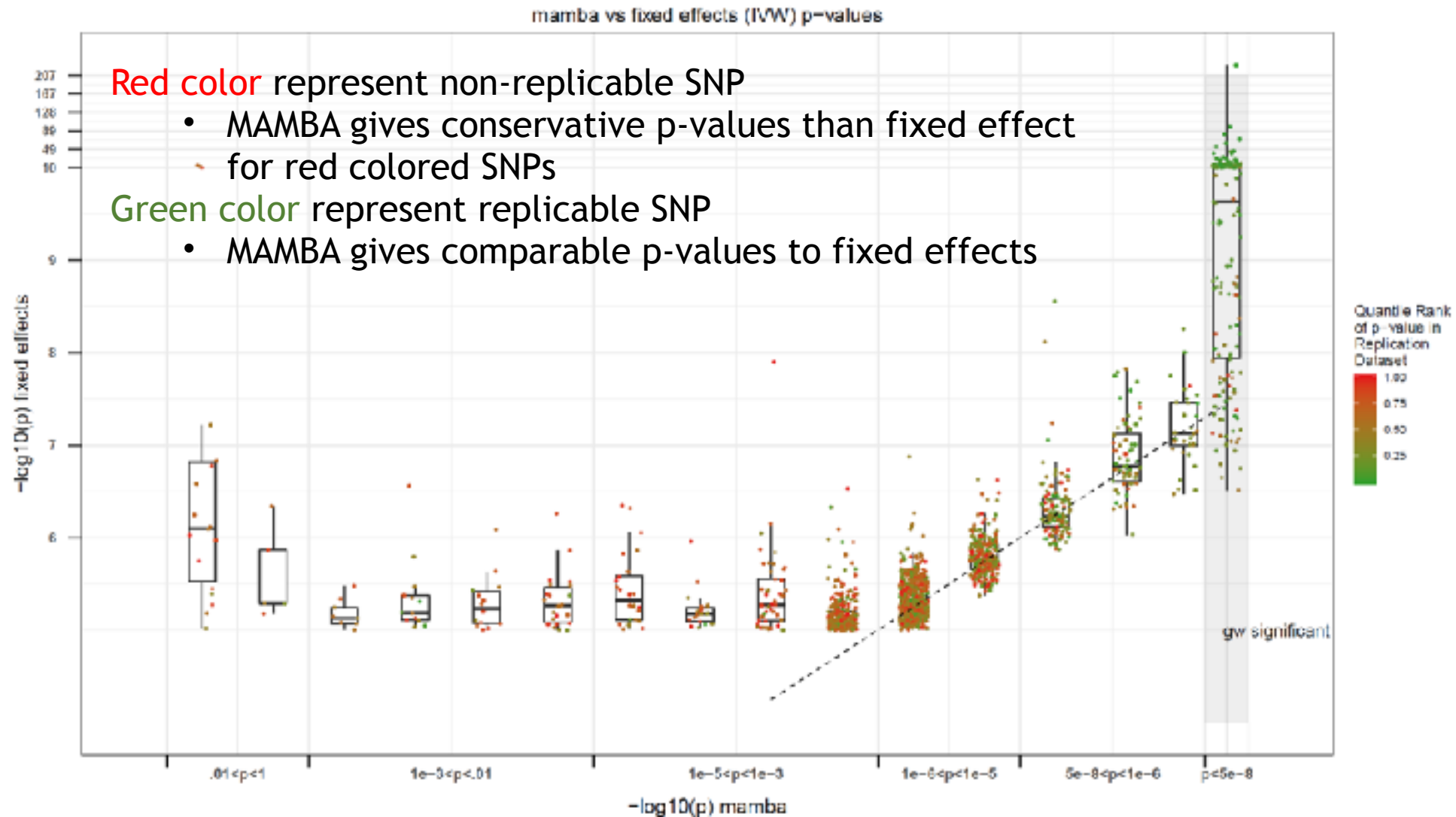
- Discovery cohort
 - GSCAN - 23andMe data
- Replication:
 - 23andMe

Phenotype	Discovery N (# Contributing Studies)	Replication N
Cigarettes Per Day	263,954 (34)	73,380
Drinks Per Week	537,349 (33)	403,931
Smoking Initiation	651,337 (35)	599,289

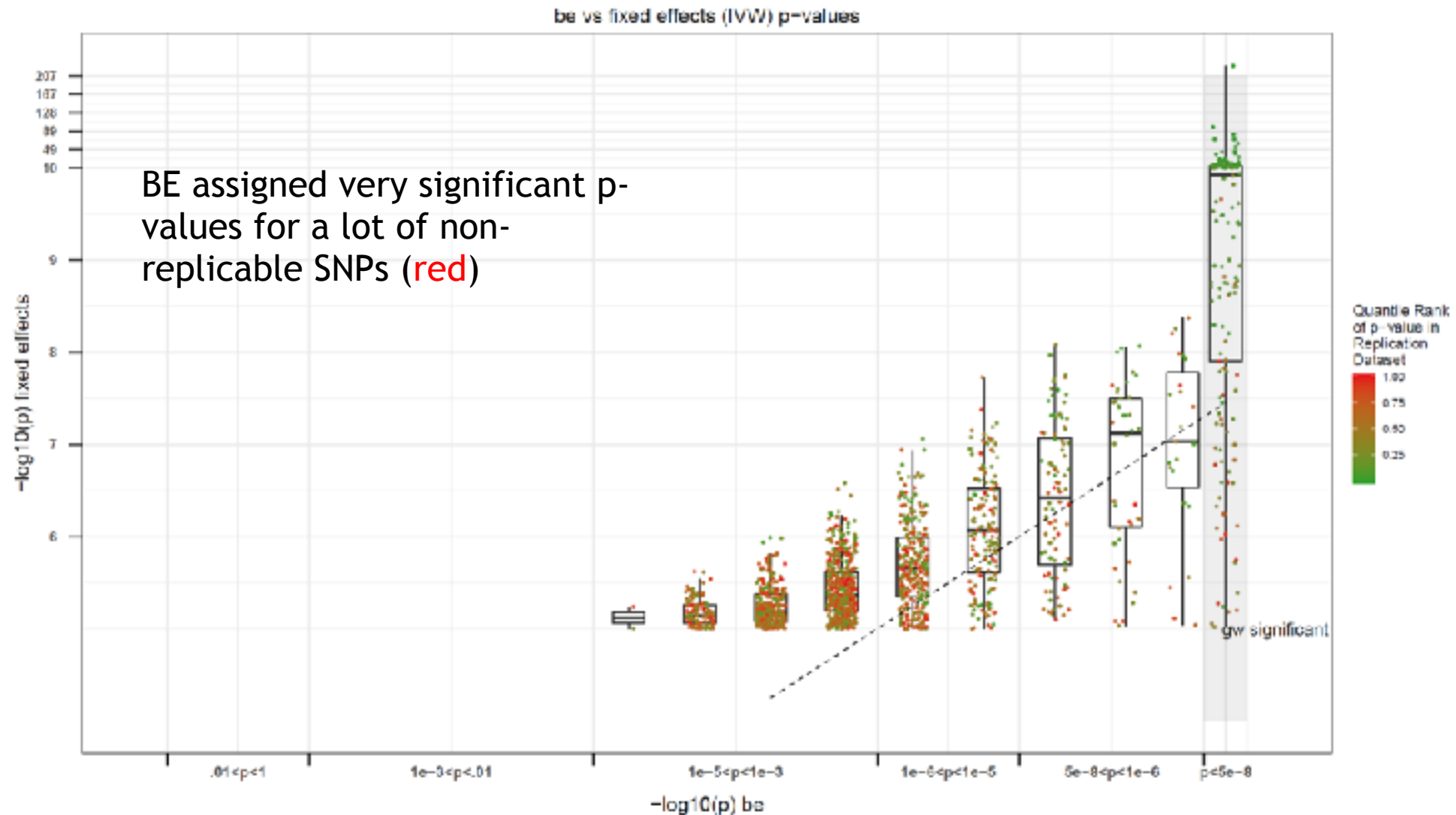
Rank Correlation of Effect Sizes Between Discovery and Replication Data



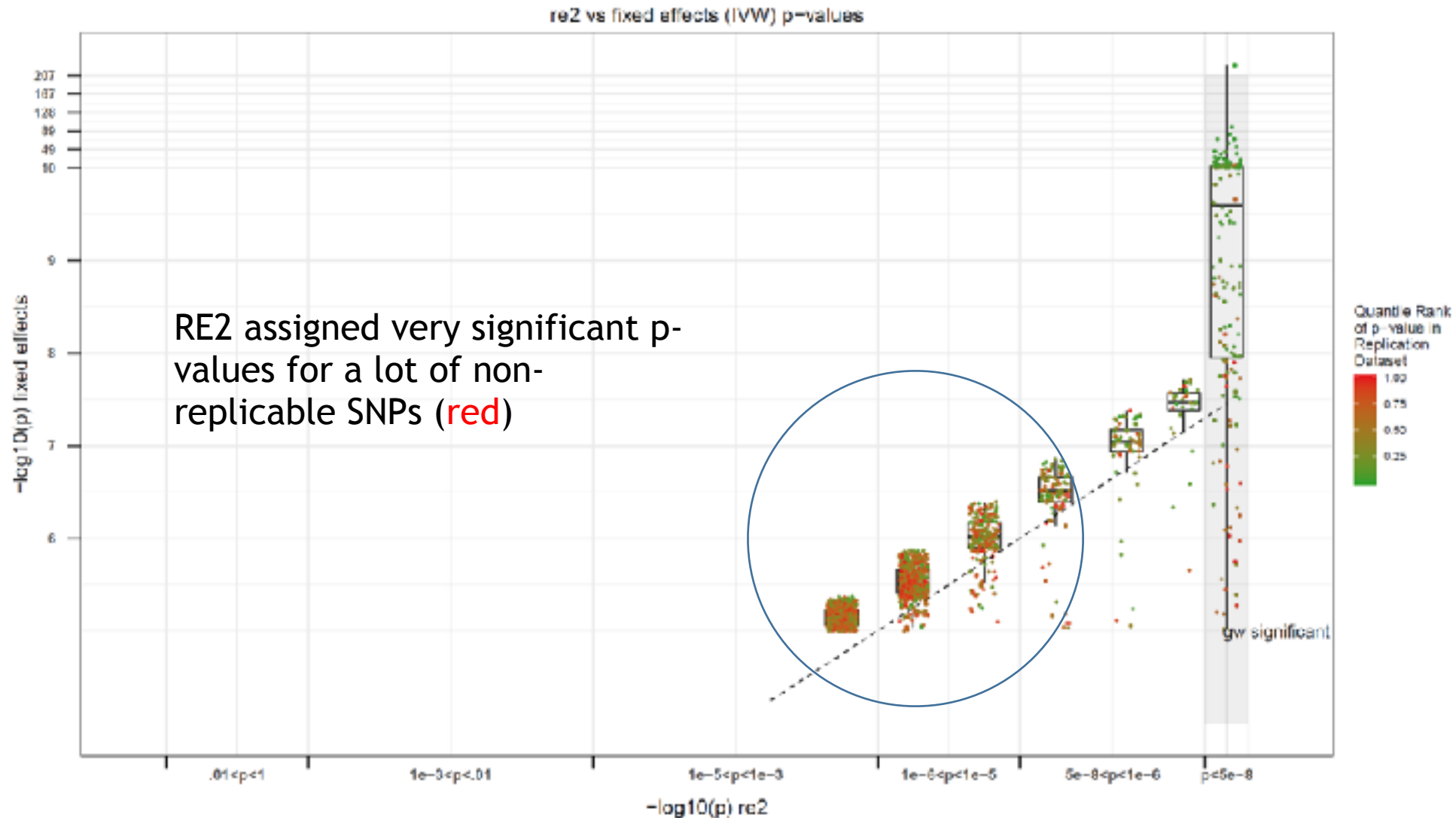
P-Value Comparison in GSCAN Data (MAMBA)



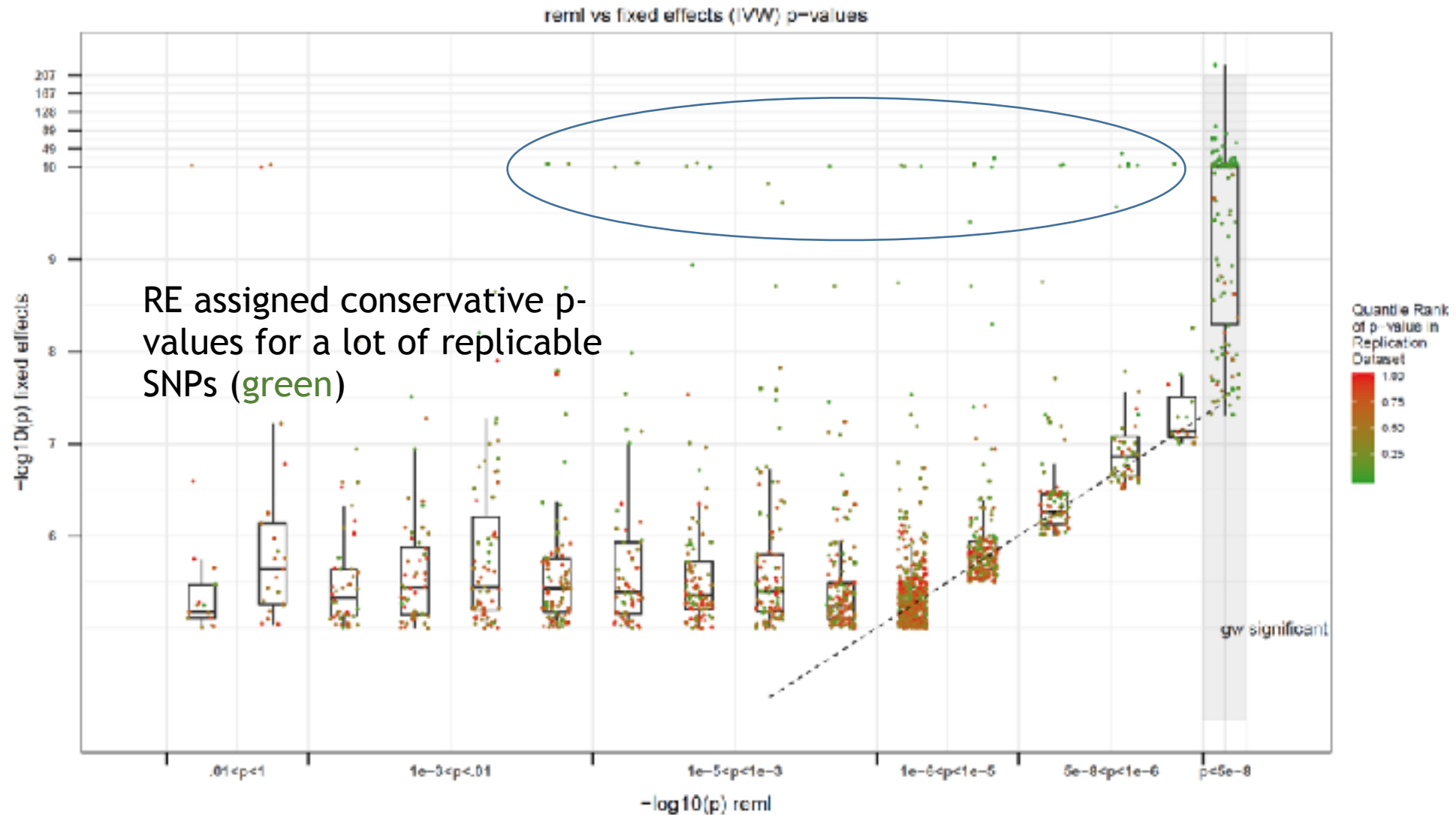
P-Value Comparison in GSCAN Data (BE)



P-Value Comparison in GSCAN Data (RE2)



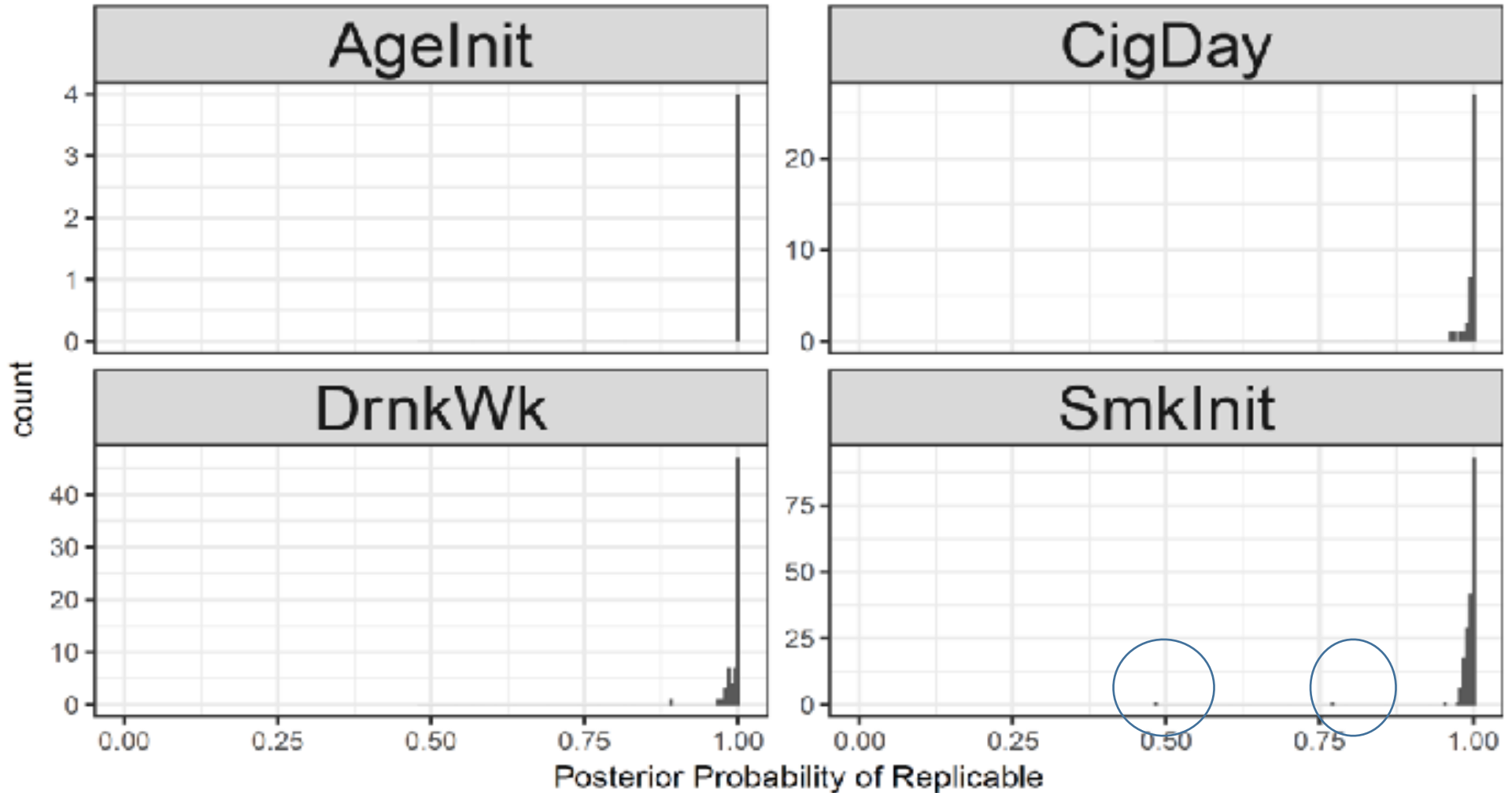
P-Value Comparison in GSCAN Data (RE)



Examples of Non-replicable Associations Before Filtering

Trait	Variant	PVALUE in 23andMe Replication	MAF	PVALUE (MAMBA)	PVALUE (FE)	PVALUE (RE2)	PVALUE (BE)	PVALUE (RE)
SmkInit	6:58757852	0.31	9.9×10^{-4}	1.0	3.0×10^{-15}	2.3×10^{-32}	5.1×10^{-27}	0.10
SmkInit	2:90193913	0.25	4.7×10^{-4}	1.0	3.7×10^{-13}	1.5×10^{-62}	3.3×10^{-56}	0.43
CigDay	13:89700202	0.46	2.0×10^{-5}	0.15	4.6×10^{-7}	9.0×10^{-7}	7.3×10^{-6}	9.0×10^{-7}
CigDay	6:150560926	0.39	7.1×10^{-6}	0.078	7.1×10^{-6}	4.3×10^{-6}	7.6×10^{-7}	0.030
DrnkWk	5:113421490	0.75	1.2×10^{-5}	0.051	3.3×10^{-6}	6.9×10^{-6}	8.1×10^{-6}	3.8×10^{-4}
DrnkWk	11:73222943	0.29	1.1×10^{-5}	0.11	5.1×10^{-6}	9.3×10^{-6}	3.6×10^{-6}	0.012

Posterior Probability of Replicability for GSCAN Result



Acknowledgement

- Dr. Qunhua Li
- GSCAN Consortium Collaborator
- Lab members
 - Yu Jiang
 - Fang Chen
 - **Dan McGuire**
 - **Jordan Hughey**
 - Dylan Weissenkampen
 - Lina Yang
 - Scott Eckert
 - Renan Sauteraud
 - Xingyan Wang
 - Joe Cirilo
- Funding support
 - NIH/NIDA R21DA040177 (Liu)
 - NIH/NHGRI R01HG008983 (Liu)
 - NIH/NIGMS R01GM126479 (Liu)
 - NIH/NIDA R01DA037904

