

REVERSE ENGINEERING FOR PRIVACY-PRESERVING VALIDATION OF GWAS OUTCOMES

Yuzhou Jiang

yxj466@case.edu

Case Western Reserve University

Tina Xu

yxx914@case.edu

Case Western Reserve University

Erman Ayday

exa208@case.edu

Case Western Reserve University

Research outcomes should be verifiable

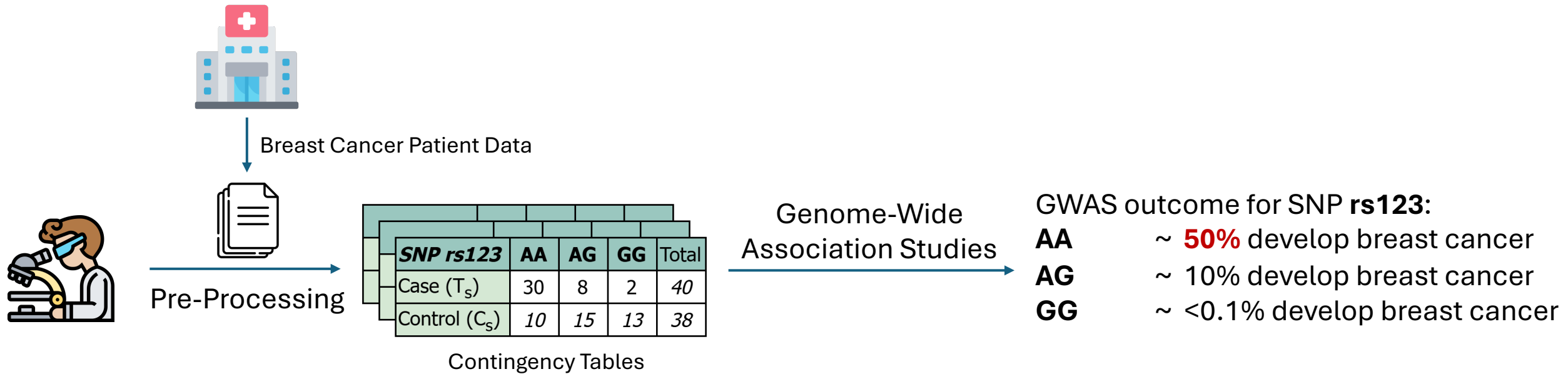
- However, reproducibility can be challenging
- Data sharing is often restricted due to privacy risks
 - Especially true for genomic area

Why is reproducibility still worth fighting for?
What could go wrong if we skip it?



Reproducibility – Cancer Example

Researcher wants to know which part of DNA is associated with breast cancer.



Reproducibility – Cancer Example Cont.

GWAS outcome for SNP **rs123**:

AA	~ 50% develop breast cancer
AG	~ 10% develop breast cancer
GG	~ <0.1% develop breast cancer

E.g., different populations/environments

What we can do:

- Warn **AA** patients about potential risks
- Recommend preventive measures
- Regular screenings (detect in early stages)

But what if the research is **wrong** or **misapplied**?

E.g., rs234's CC patients ~ 50% breast cancer instead

- High-risk individuals get no warning
- Low-risk people panic unnecessarily
- Misleading future research directions

It already happened!

- ! Alzheimer's study with fabricated images [1]
- ! Longevity gene study misled due to sequencing errors [2]
- ! ...

[1] Piller, C. (2022). Blots on a field? Science, 377(6604), 358–363.

[2] Ledford, H. (2011). Paper on genetics of longevity retracted. Nature. <https://www.nature.com/articles/news.2011.429>



- ! Alzheimer's study with fabricated images
- ! Misleading longevity gene study due to sequencing errors
- ! ...

Why not just share the data for reproducibility?

- It enables re-identification if not protected properly.
- If re-identified
 - Insurance bias and medical discrimination
 - Family linkage
 - ...

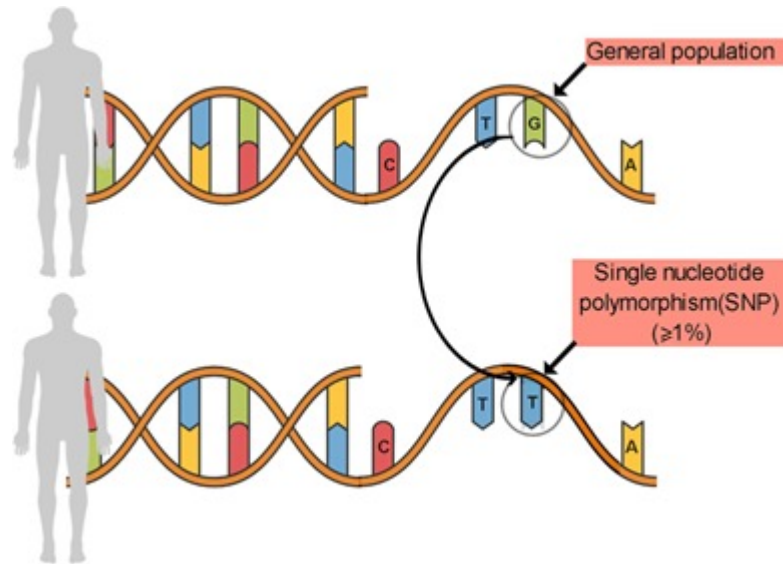
Problem

Can we verify GWAS outcome without exposing privacy risks?



Background

Single Nucleotide Polymorphism (SNP)



SNP Value (**A** major, **G** minor): # of minor alleles
AA 0: homozygous for the major allele
AG 1: heterozygous, carrying one copy of the minor allele
GG 2: homozygous for the minor allele

Lima, L. S., Galiciolli, M. E. A., Pereira, M. E., Felisbino, K., Machado-Souza, C., de Oliveira, C. S., & Guiloski, I. C. (2022). Modification by genetic polymorphism of lead-induced IQ alteration: a systematic review. *Environmental Science and Pollution Research*, 29(29), 43435-43447.



Modeling GWAS Reproducibility

What we have

- The control row in the contingency table for each SNP
- MAFs for each SNP from the research findings
- Number of samples in the case row

<i>SNP rs123</i>	AA	AG	GG	Total
Case (T_s)	?	?	?	40
Control (C_s)	10	15	13	38

→ $\chi^2 = 8.36$

What if we

- Try to find the case row that generates the reported GWAS statistic?



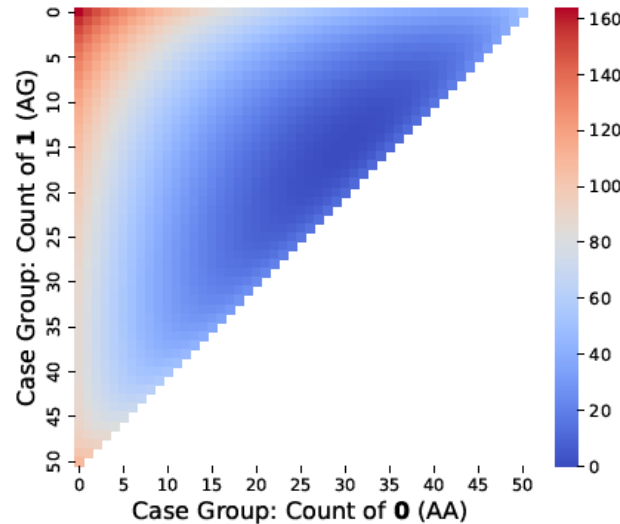
A Brute Force Solution to Find Case Row

- Try all possible combos for each SNP
 - Time complexity is an issue for large-sample datasets

Can we speed up?



A Visualization of Chi-Square Statistic



Observations:

- Statistically significant outcomes (in dark blue) are regional
- Almost concave (with slight local fluctuations)

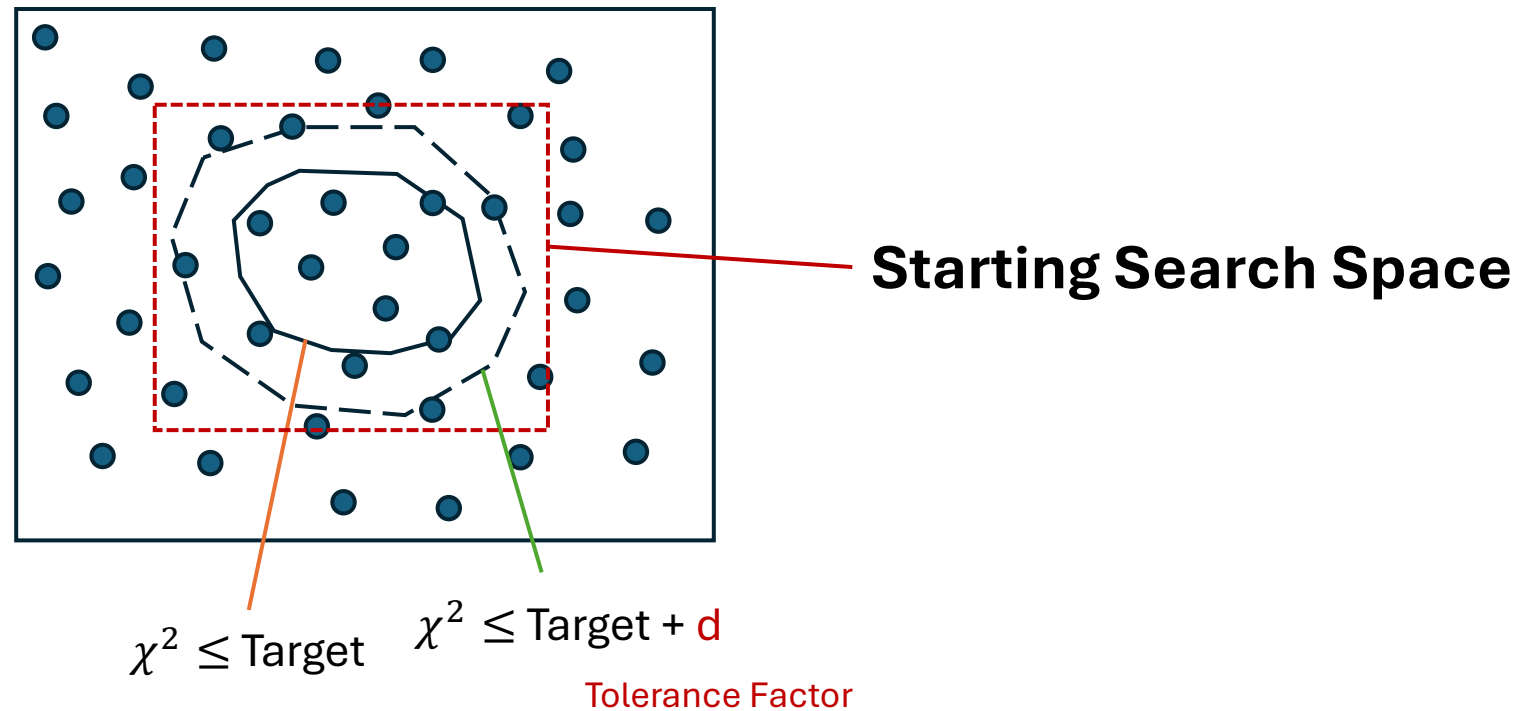
Optimizations

Monte Carlo Sampling

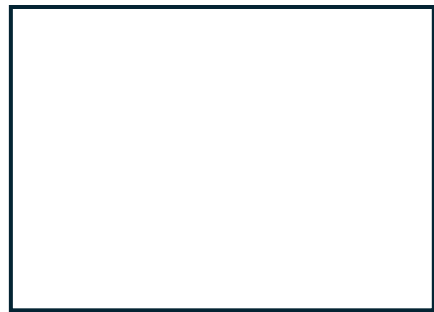
Divide and Refine



Monte Carlo Sampling



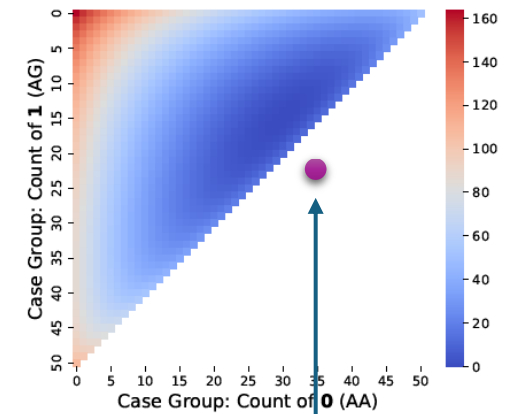
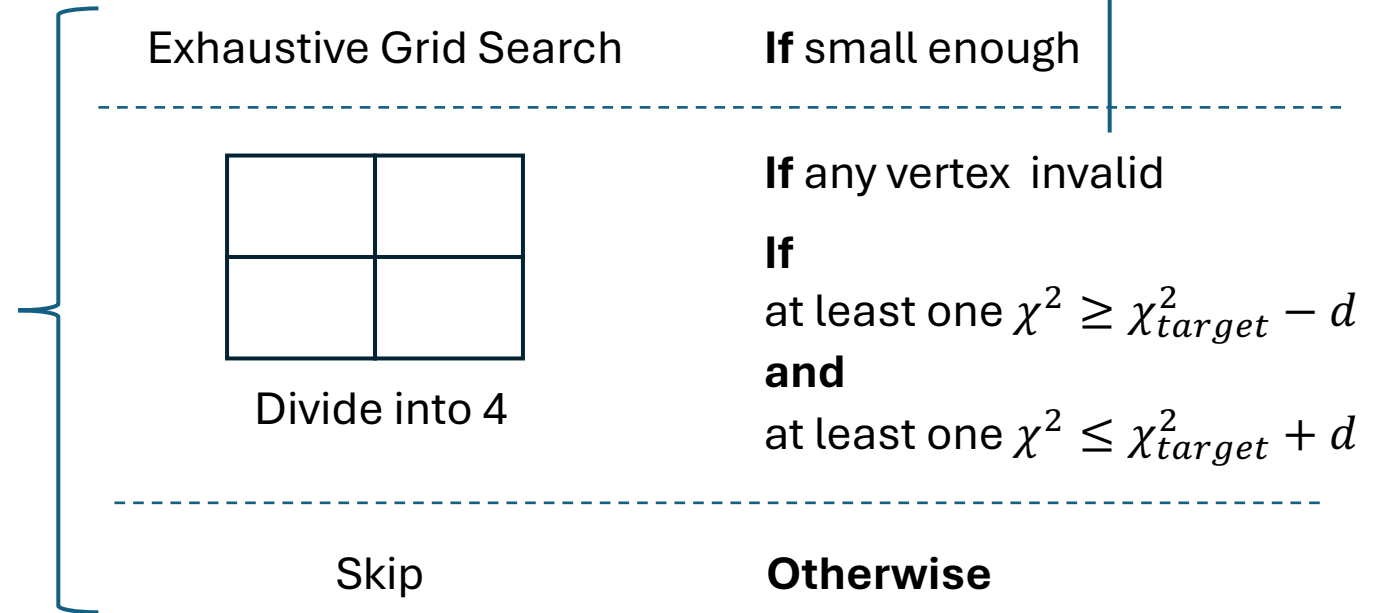
Divide and Refine



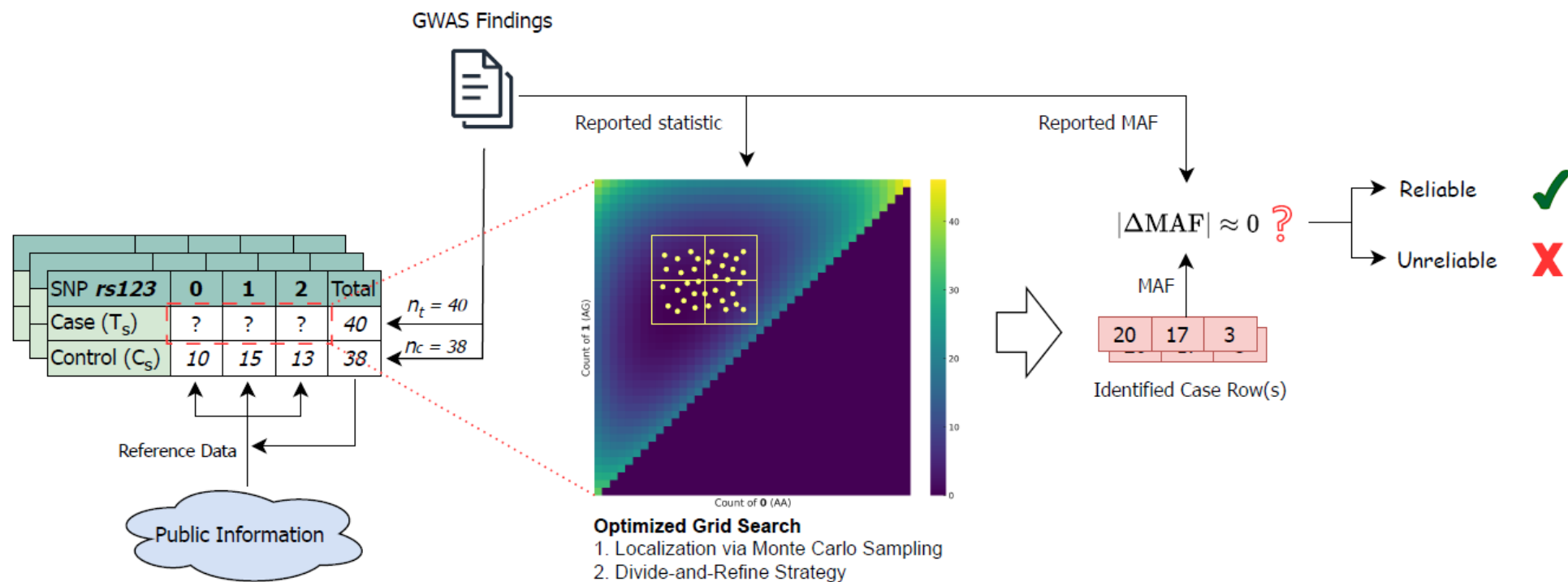
Search Region



Check 4 vertices

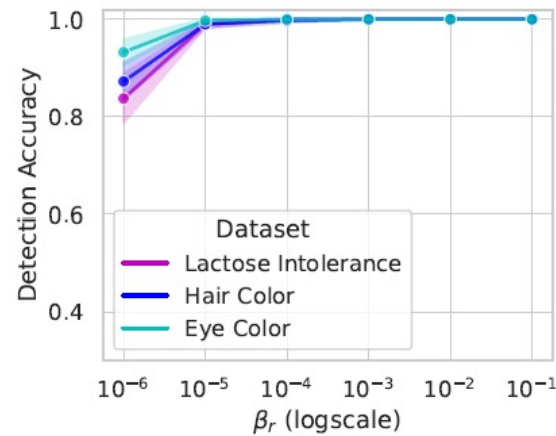


REGenOV Workflow

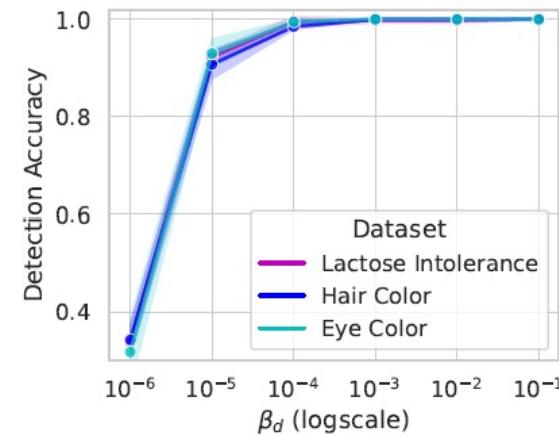


Performance: Detecting Errors During GWAS

- Reporting Error
 - Report a chi-square statistic that deviates uniformly within a bounded range (parameterized by β_r)
- Deviation Error
 - Report a chi-square statistic **perturbed by Gaussian noise** with standard deviation β_d



(a) Vs. Reporting Error



(b) Vs. Deviation Error



Performance: Time Complexity

n_t	100	1,000	10,000	100,000	1,000,000
t (s)	0.5746	6.8334	171.2239	1348.3858	7565.7778

Table 6.1: Execution time across varying numbers of case samples.



Future Directions

- Expand to more types of errors in more GWAS stages
- Leverage Zero-Knowledge Proof (ZKP) to validate outcomes



Thank You!

