Quantifying Patient Privacy in Public Disease Portals

Sean Simmons

Stanley Center, Broad Institute

1/10/2017

Acknowledgments

Portal Team:

- Jason Flannick
- Noel Burtt
- The rest of the team!

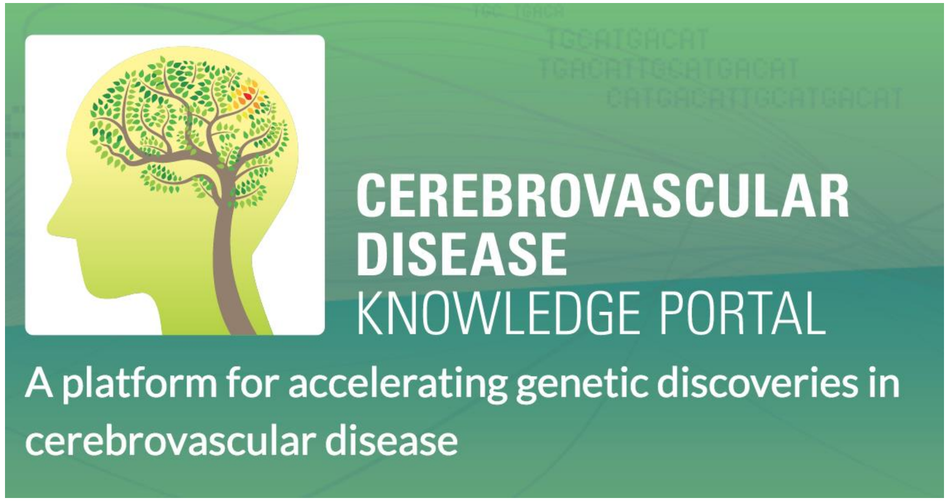
Sahinalp Lab:

- Cenk Sahinalp (see his talk later!)
- Many others!

Berger Lab:

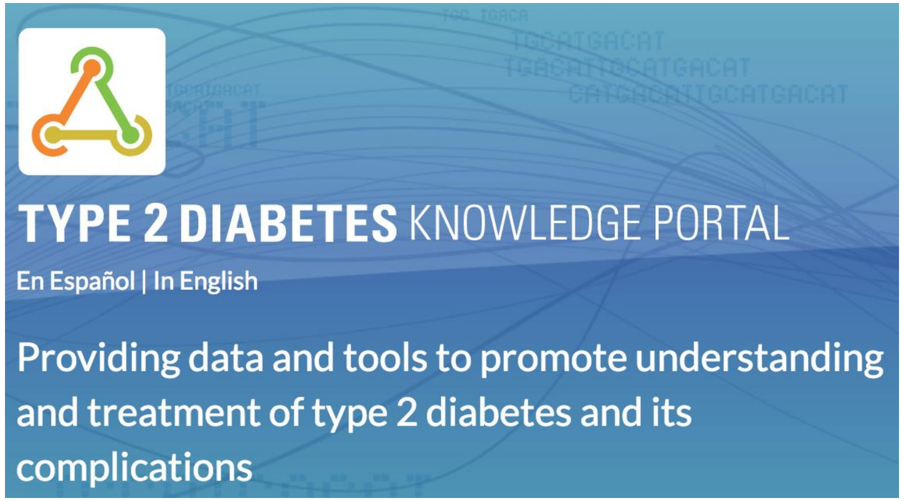
- Bonnie Berger
- Many others!

Knowledge Portals: A new way to share GWAS!



CEREBROVASCULAR DISEASE KNOWLEDGE PORTAL

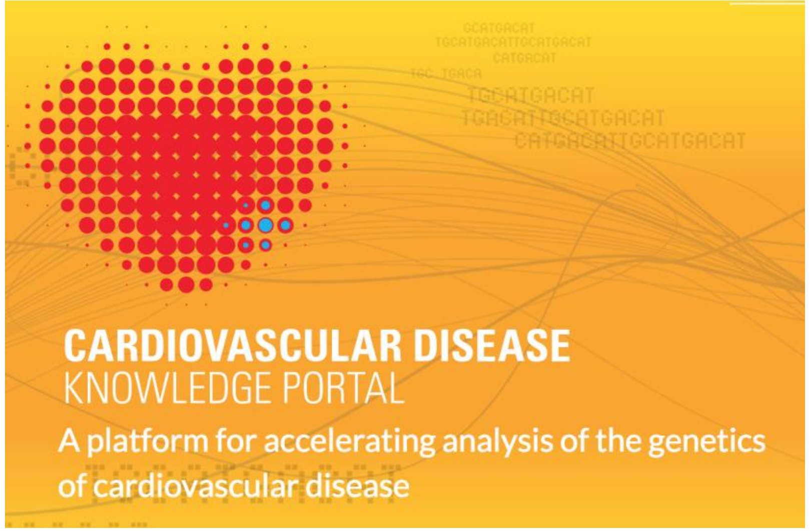
A platform for accelerating genetic discoveries in cerebrovascular disease



TYPE 2 DIABETES KNOWLEDGE PORTAL

En Español | In English

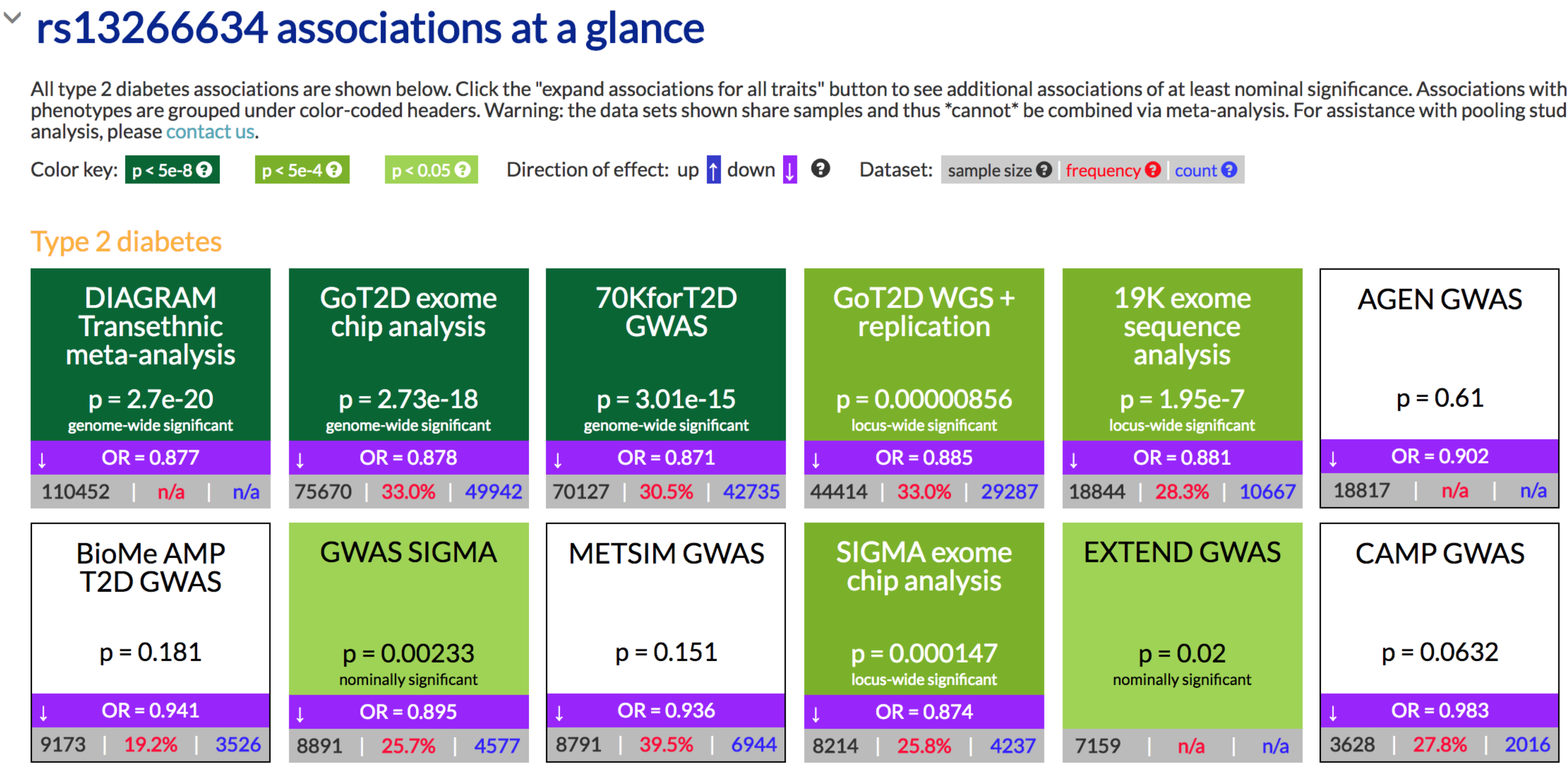
Providing data and tools to promote understanding and treatment of type 2 diabetes and its complications



CARDIOVASCULAR DISEASE KNOWLEDGE PORTAL

A platform for accelerating analysis of the genetics of cardiovascular disease

Knowledge Portals: Sharing GWAS results



rs13266634 associations at a glance

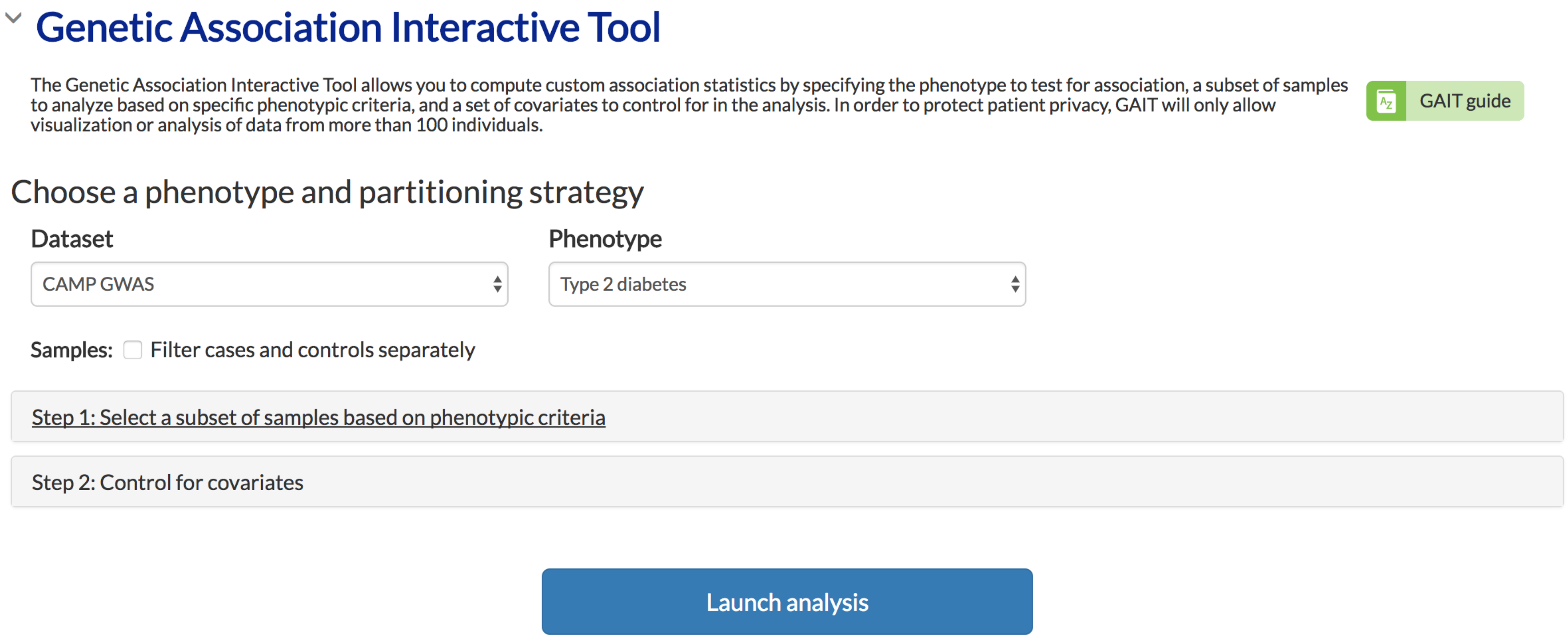
All type 2 diabetes associations are shown below. Click the "expand associations for all traits" button to see additional associations of at least nominal significance. Associations with related phenotypes are grouped under color-coded headers. Warning: the data sets shown share samples and thus *cannot* be combined via meta-analysis. For assistance with pooling studies for analysis, please contact us.

Color key: p < 5e-8 ? p < 5e-4 ? p < 0.05 ? Direction of effect: up ↑ down ↓ ? Dataset: sample size ? | frequency ? | count ?

Type 2 diabetes

Dataset	p	Significance	Direction	OR	Sample size	Frequency	Count
DIAGRAM Transethnic meta-analysis	p = 2.7e-20	genome-wide significant	↓	OR = 0.877	110452	n/a	n/a
GoT2D exome chip analysis	p = 2.73e-18	genome-wide significant	↓	OR = 0.878	75670	33.0%	49942
70KforT2D GWAS	p = 3.01e-15	genome-wide significant	↓	OR = 0.871	70127	30.5%	42735
GoT2D WGS + replication	p = 0.00000856	locus-wide significant	↓	OR = 0.885	44414	33.0%	29287
19K exome sequence analysis	p = 1.95e-7	locus-wide significant	↓	OR = 0.881	18844	28.3%	10667
AGEN GWAS	p = 0.61		↓	OR = 0.902	18817	n/a	n/a
BioMe AMP T2D GWAS	p = 0.181		↓	OR = 0.941	9173	19.2%	3526
GWAS SIGMA	p = 0.00233	nominally significant	↓	OR = 0.895	8891	25.7%	4577
METSIM GWAS	p = 0.151		↓	OR = 0.936	8791	39.5%	6944
SIGMA exome chip analysis	p = 0.000147	locus-wide significant	↓	OR = 0.874	8214	25.8%	4237
EXTEND GWAS	p = 0.02	nominally significant			7159	n/a	n/a
CAMP GWAS	p = 0.0632		↓	OR = 0.983	3628	27.8%	2016

Knowledge Portals: Interactive analysis!



Genetic Association Interactive Tool

The Genetic Association Interactive Tool allows you to compute custom association statistics by specifying the phenotype to test for association, a subset of samples to analyze based on specific phenotypic criteria, and a set of covariates to control for in the analysis. In order to protect patient privacy, GAIT will only allow visualization or analysis of data from more than 100 individuals.

Choose a phenotype and partitioning strategy

Dataset: CAMP GWAS

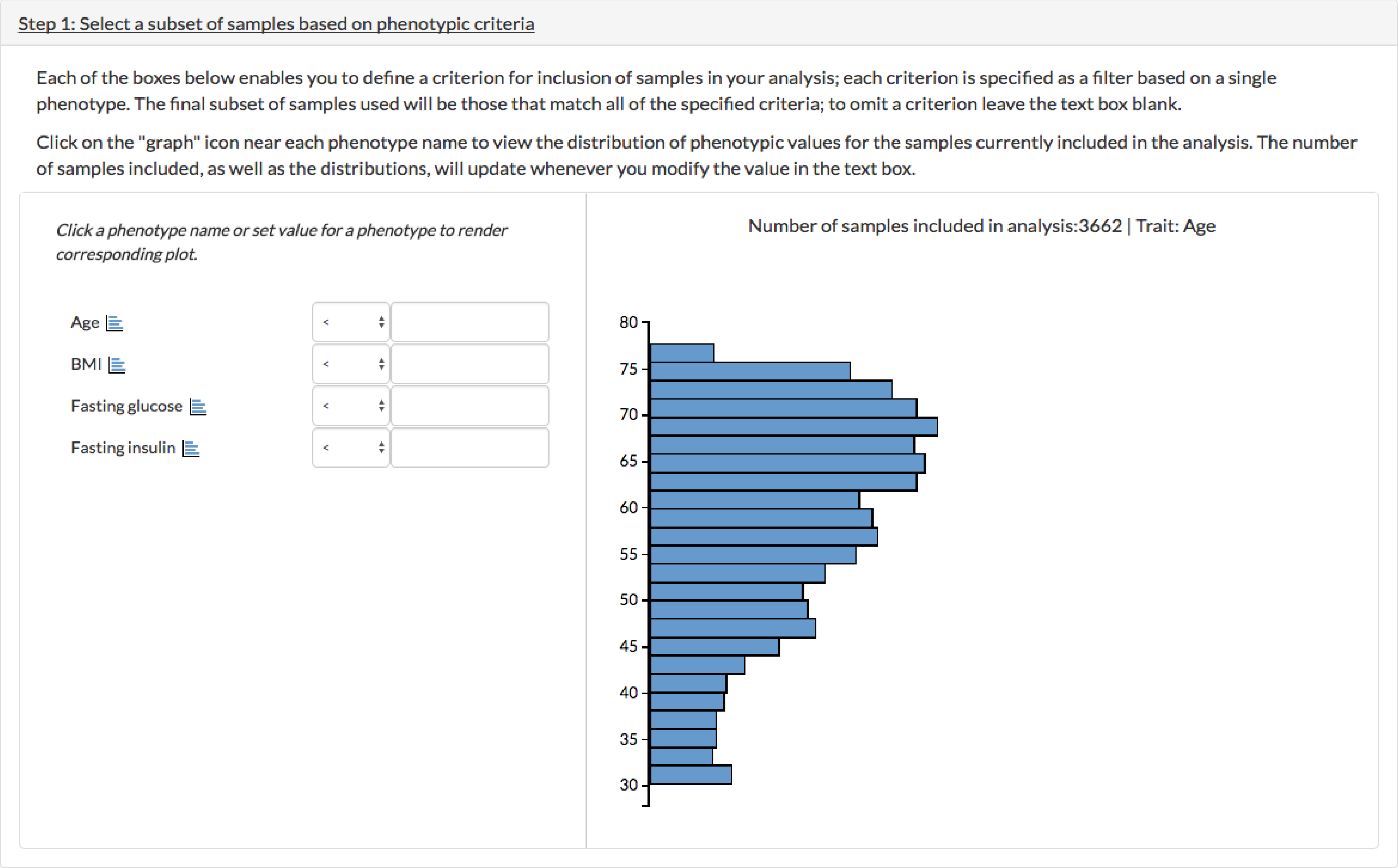
Phenotype: Type 2 diabetes

Samples: □ Filter cases and controls separately

Step 1: Select a subset of samples based on phenotypic criteria

Step 2: Control for covariates

Step 1: Choose who to include



Step 1: Select a subset of samples based on phenotypic criteria

Each of the boxes below enables you to define a criterion for inclusion of samples in your analysis; each criterion is specified as a filter based on a single phenotype. The final subset of samples used will be those that match all of the specified criteria; to omit a criterion leave the text box blank.

Click on the "graph" icon near each phenotype name to view the distribution of phenotypic values for the samples currently included in the analysis. The number of samples included, as well as the distributions, will update whenever you modify the value in the text box.

Click a phenotype name or set value for a phenotype to render corresponding plot.

Age <

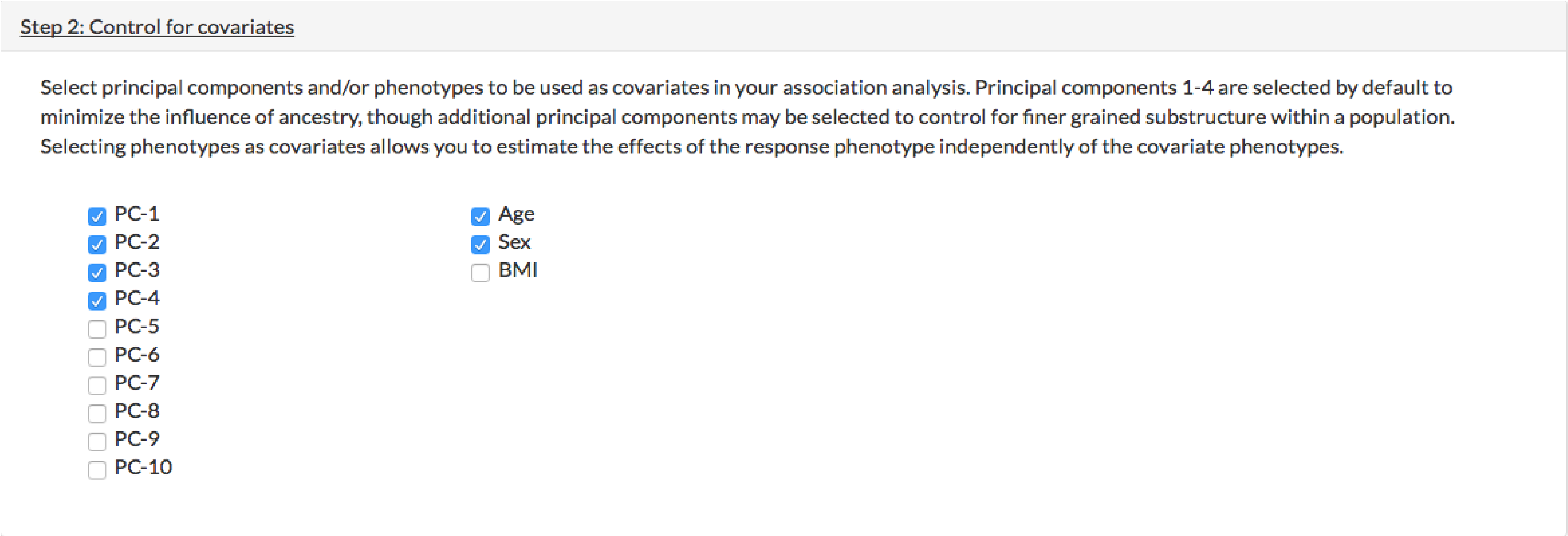
BMI <

Fasting glucose <

Fasting insulin <

Number of samples included in analysis:3662 | Trait: Age

Step 2: Choose what to correct for



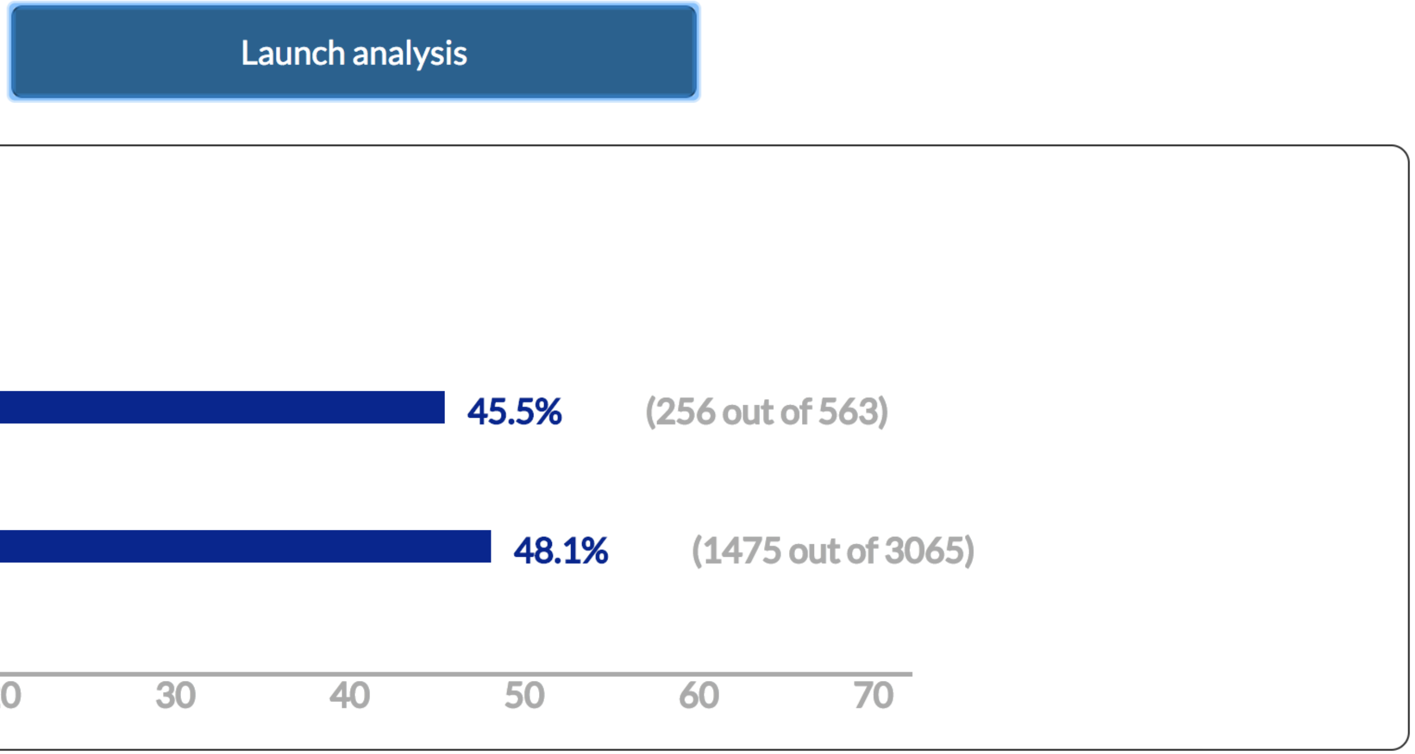
Step 2: Control for covariates

Select principal components and/or phenotypes to be used as covariates in your association analysis. Principal components 1-4 are selected by default to minimize the influence of ancestry, though additional principal components may be selected to control for finer grained substructure within a population. Selecting phenotypes as covariates allows you to estimate the effects of the response phenotype independently of the covariate phenotypes.

- ☑ PC-1
- ☑ PC-2
- ☑ PC-3
- ☑ PC-4
- ☐ PC-5
- ☐ PC-6
- ☐ PC-7
- ☐ PC-8
- ☐ PC-9
- ☐ PC-10

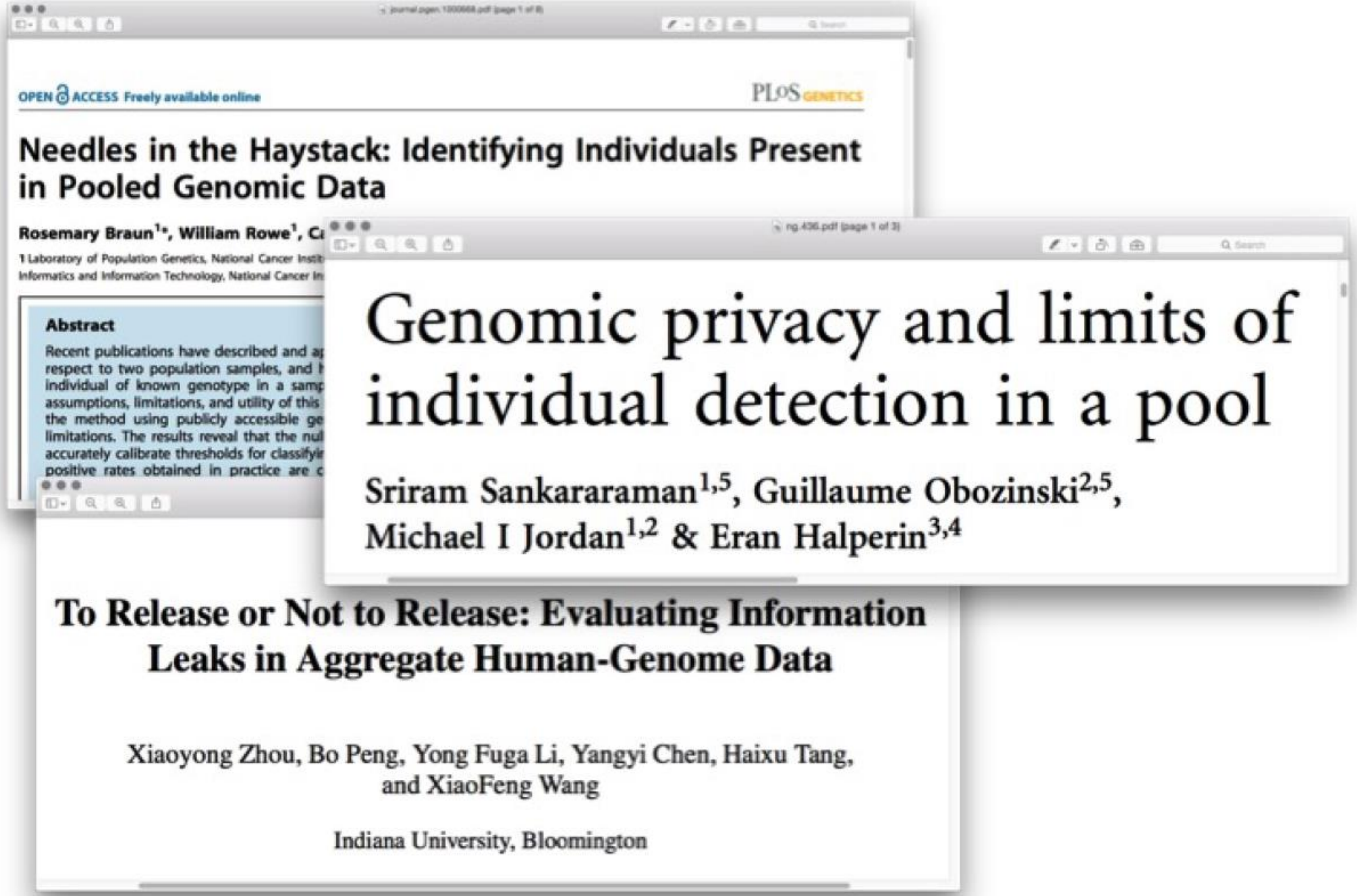
- ☑ Age
- ☑ Sex
- ☐ BMI

Step 3: Get results!



pValue = 0.0778
odds ratio = 0.873
95% CI: (0.751 to 1.02)

We know a lot about GWAS privacy…



journal.pgen.1000668.pdf (page 1 of 8)
OPEN ACCESS Freely available online
PLoS GENETICS
ng.436.pdf (page 1 of 3)

Needles in the Haystack: Identifying Individuals Present in Pooled Genomic Data

Rosemary Braun[1*], William Rowe[1], Ca

1 Laboratory of Population Genetics, National Cancer Instit
Informatics and Information Technology, National Cancer In

Abstract

Recent publications have described and ap
respect to two population samples, and h
individual of known genotype in a samp
assumptions, limitations, and utility of this
the method using publicly accessible ge
limitations. The results reveal that the nul
accurately calibrate thresholds for classifyi
positive rates obtained in practice are c

Genomic privacy and limits of individual detection in a pool

Sriram Sankararaman[1,5], Guillaume Obozinski[2,5], Michael I Jordan[1,2] & Eran Halperin[3,4]

To Release or Not to Release: Evaluating Information Leaks in Aggregate Human-Genome Data

Xiaoyong Zhou, Bo Peng, Yong Fuga Li, Yangyi Chen, Haixu Tang, and XiaoFeng Wang

Indiana University, Bloomington

Difference with standard GWAS portals

- Smaller datasets analysis/ subsampling
- Repeated queries per SNP
- Ability to remove/ include confounders

Difference with standard GWAS portals

- Smaller datasets analysis/ subsampling
- Repeated queries per SNP
- Ability to remove/ include confounders

All of these change the privacy landscape!

Quantifying privacy

- Use a model based approach

- Looked at different types of private data leakage

- No formal guarantees (yet!)

Quantifying privacy: Our models

What Leaks?	**Outside knowledge**	**Statistic released**	**Likely Risk** **(very subjective)**
Participation in study	Background MAF and individuals genotype	Minor allele frequency	Low
Diseases status	Background MAF and individuals genotype	Minor allele frequency	Low
Other phenotype data	Unspecified	Interactive histogram	Moderate
Genotype	Background MAF	GWAS statistics for one query	Low
Genotype	Background MAF	GWAS statistics for numerous queries	High

Quantifying privacy: Adversarial uncertainty

- Adversaries knowledge modelled probabilistically
- Prior knowledge specifies probability distribution used.
- Look at entropy (among other measures) of posterior probability to determine privacy loss.

Quantifying privacy: Conditional entropy!

$$Entropy_{y|X} = -\sum \log(\Pr(y|X))\Pr(y, X)$$

Where:

- y value of secret trait
- Pr probaility distribution representing adversaries belief
- X information being released

Quantifying privacy: Example!

$$Entropy_{genotype|MAF} = -\sum_{MAF,i} \log(\Pr(genotype = i|MAF))\Pr(genotype = i, MAF)$$

Quantifying privacy: Our models

What Leaks?	**Outside knowledge**	**Statistic released**	**Likely Risk (very subjective)**
Participation in study	Background MAF and individuals genotype	Minor allele frequency	Low
Diseases status	Background MAF and individuals genotype	Minor allele frequency	Low
Other phenotype data	Unspecified	Interactive histogram	Moderate
Genotype	Background MAF	GWAS statistics for one query	Low
Genotype	Background MAF	GWAS statistics for numerous queries	High

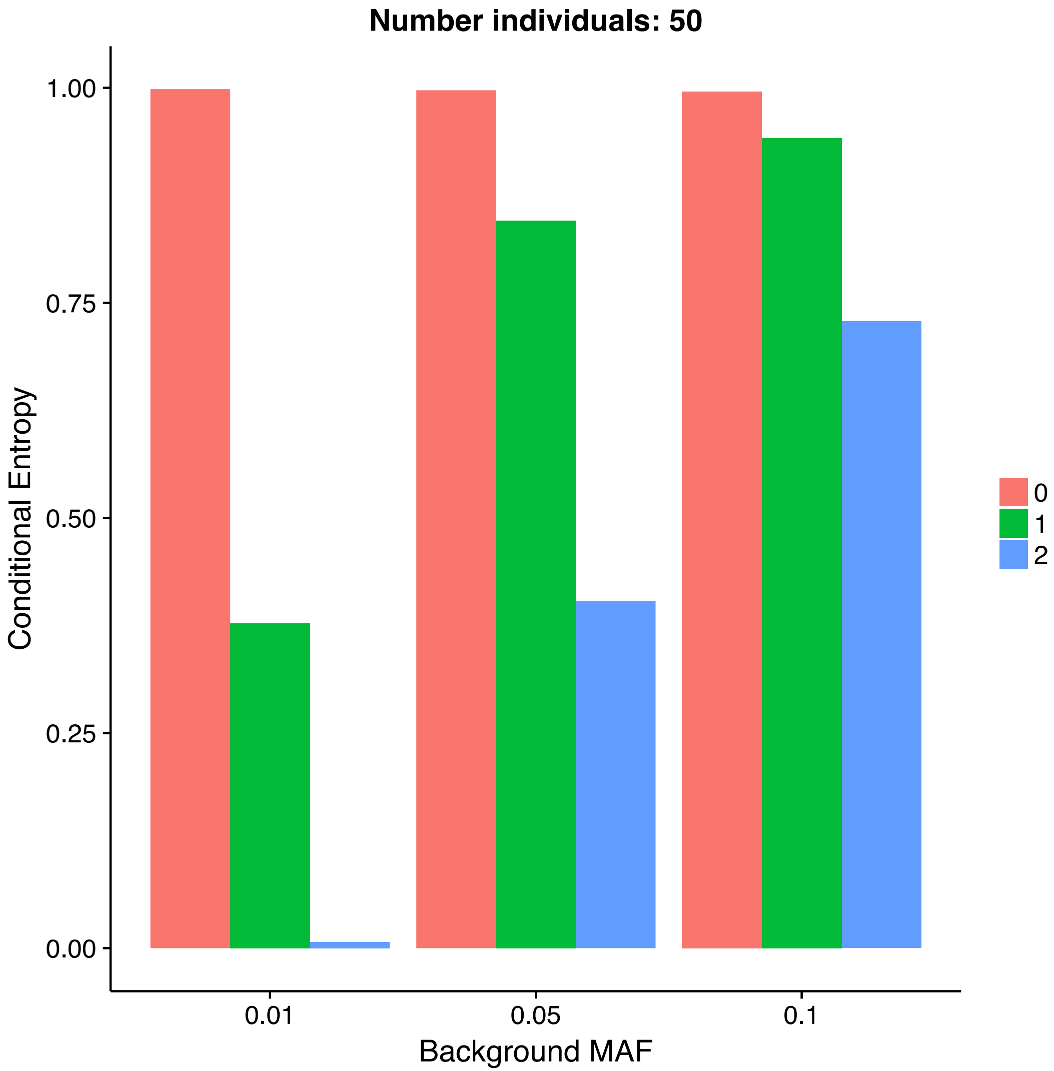
Quantifying privacy: First Analysis

What Leaks?	**Outside knowledge**	**Statistic released**	**Likely Risk (very subjective)**
Participation in study	Background MAF and individuals genotype	Minor allele frequency	Low
Diseases status	Background MAF and individuals genotype	Minor allele frequency	Low
Other phenotype data	Unspecified	Interactive histogram	Moderate
Genotype	Background MAF	GWAS statistics for one query	Low
Genotype	Background MAF	GWAS statistics for numerous queries	High

Leaking disease participation

- Apply to private disease status information
- Assume genotype known and MAF of background population known (similar to Homer et al.)
- Privacy leakage from minor allele frequencies!

Decrease in entropy of disease status



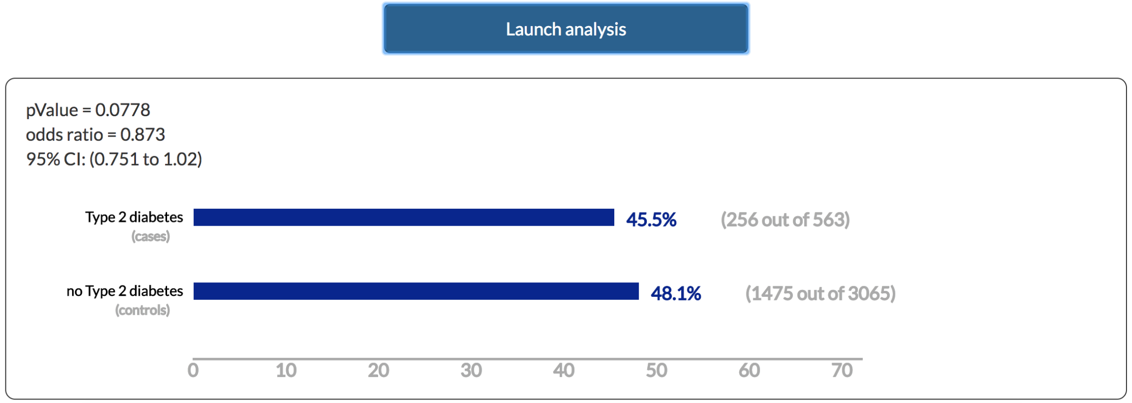
Quantifying privacy: First Analysis

What Leaks?	Outside knowledge	Statistic released	Likely Risk (very subjective)
Participation in study	Background MAF and individuals genotype	Minor allele frequency	Low
Diseases status	Background MAF and individuals genotype	Minor allele frequency	Low
Other phenotype data	Unspecified	Interactive histogram	Moderate
Genotype	Background MAF	GWAS statistics for one query	Low
Genotype	Background MAF	GWAS statistics for numerous queries	High

Quantifying privacy: Leaking genotype

What Leaks?	**Outside knowledge**	**Statistic released**	**Likely Risk (very subjective)**
Participation in study	Background MAF and individuals genotype	Minor allele frequency	Low
Diseases status	Background MAF and individuals genotype	Minor allele frequency	Low
Other phenotype data	Unspecified	Interactive histogram	Moderate
Genotype	Background MAF	GWAS statistics for one query	Low
Genotype	Background MAF	GWAS statistics for numerous queries	High

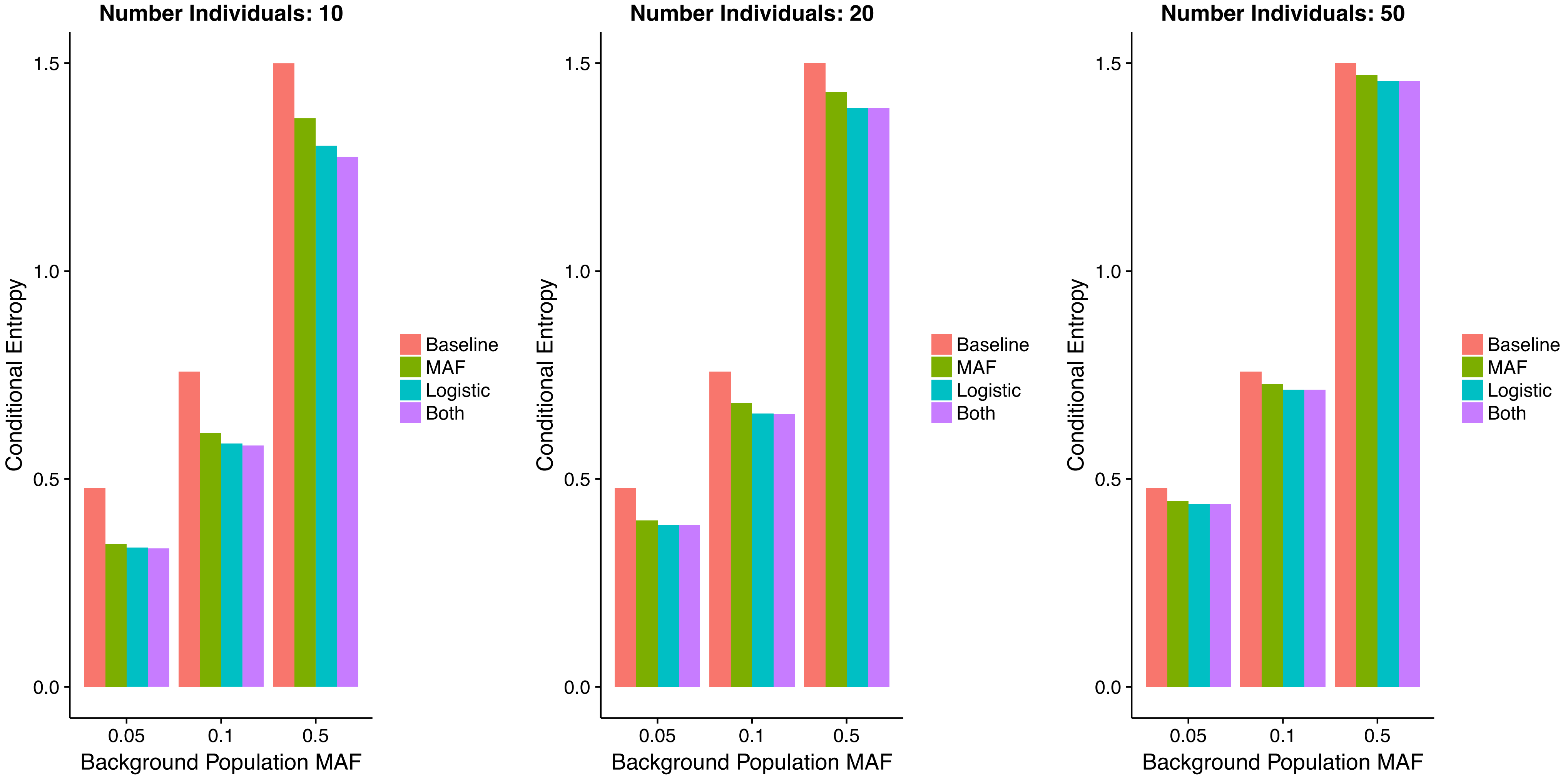
The concern:



pValue = 0.0778
odds ratio = 0.873
95% CI: (0.751 to 1.02)

ATGCAT
ATGGAT

Entropy of the genotype



Added concerns with interactive GWAS

- So far have focused on privacy concerns without taking advantage of the interactive aspect.

- Does interactive analysis add more?

Added concerns with interactive GWAS

- So far have focused on privacy concerns without taking advantage of the interactive aspect.

- Does interactive analysis add more?

Short answer? Yes

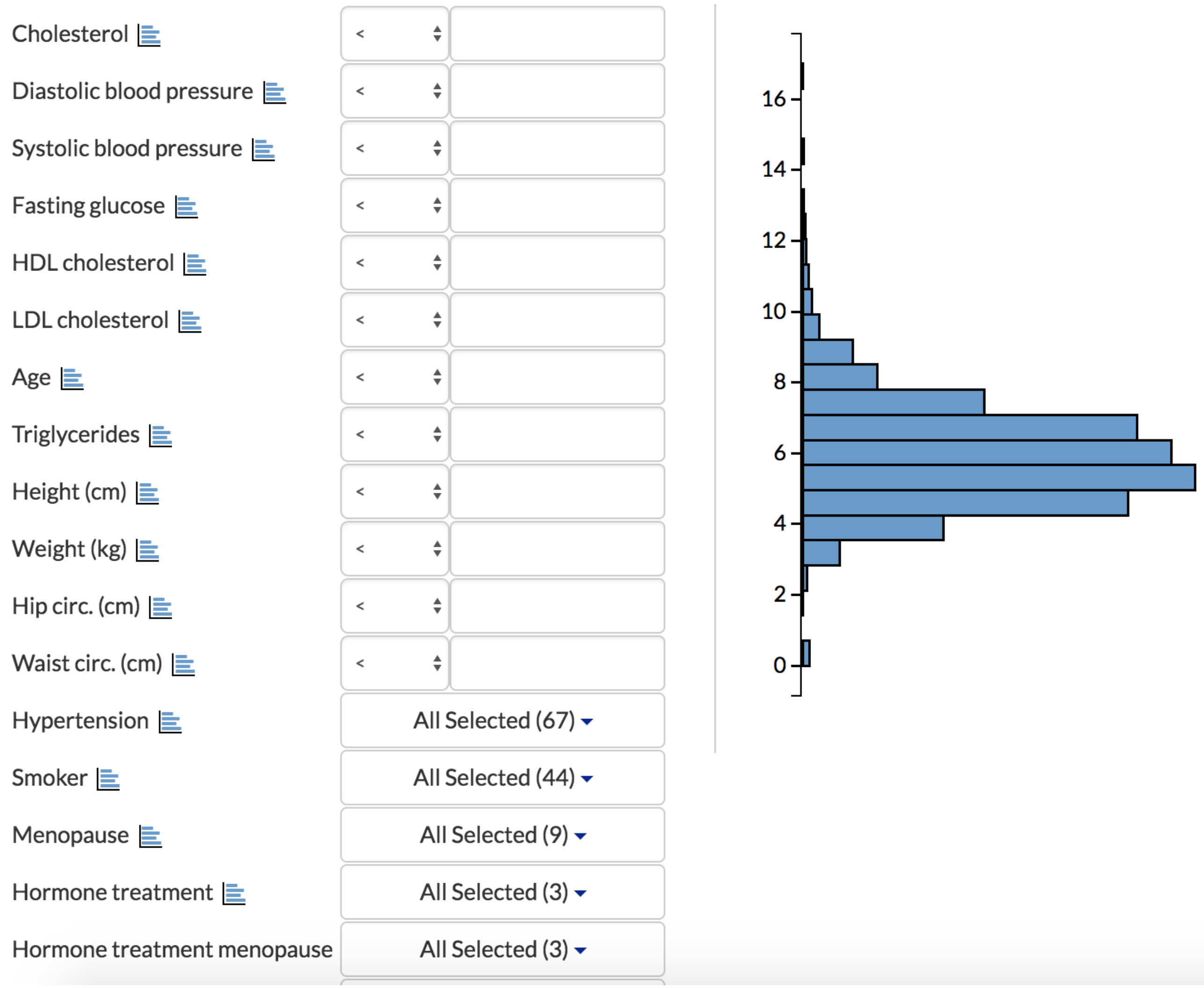
Quantifying privacy: Leaking genotype

What Leaks?	**Outside knowledge**	**Statistic released**	**Likely Risk (very subjective)**
Participation in study	Background MAF and individuals genotype	Minor allele frequency	Low
Diseases status	Background MAF and individuals genotype	Minor allele frequency	Low
Other phenotype data	Unspecified	Interactive histogram	Moderate
Genotype	Background MAF	GWAS statistics for one query	Low
Genotype	Background MAF	GWAS statistics for numerous queries	High

Quantifying privacy: Leaking health data

What Leaks?	**Outside knowledge**	**Statistic released**	**Likely Risk (very subjective)**
Participation in study	Background MAF and individuals genotype	Minor allele frequency	Low
Diseases status	Background MAF and individuals genotype	Minor allele frequency	Low
Other phenotype data	Unspecified	Interactive histogram	Moderate
Genotype	Background MAF	GWAS statistics for one query	Low
Genotype	Background MAF	GWAS statistics for numerous queries	High

Interactive histograms and phenotype data



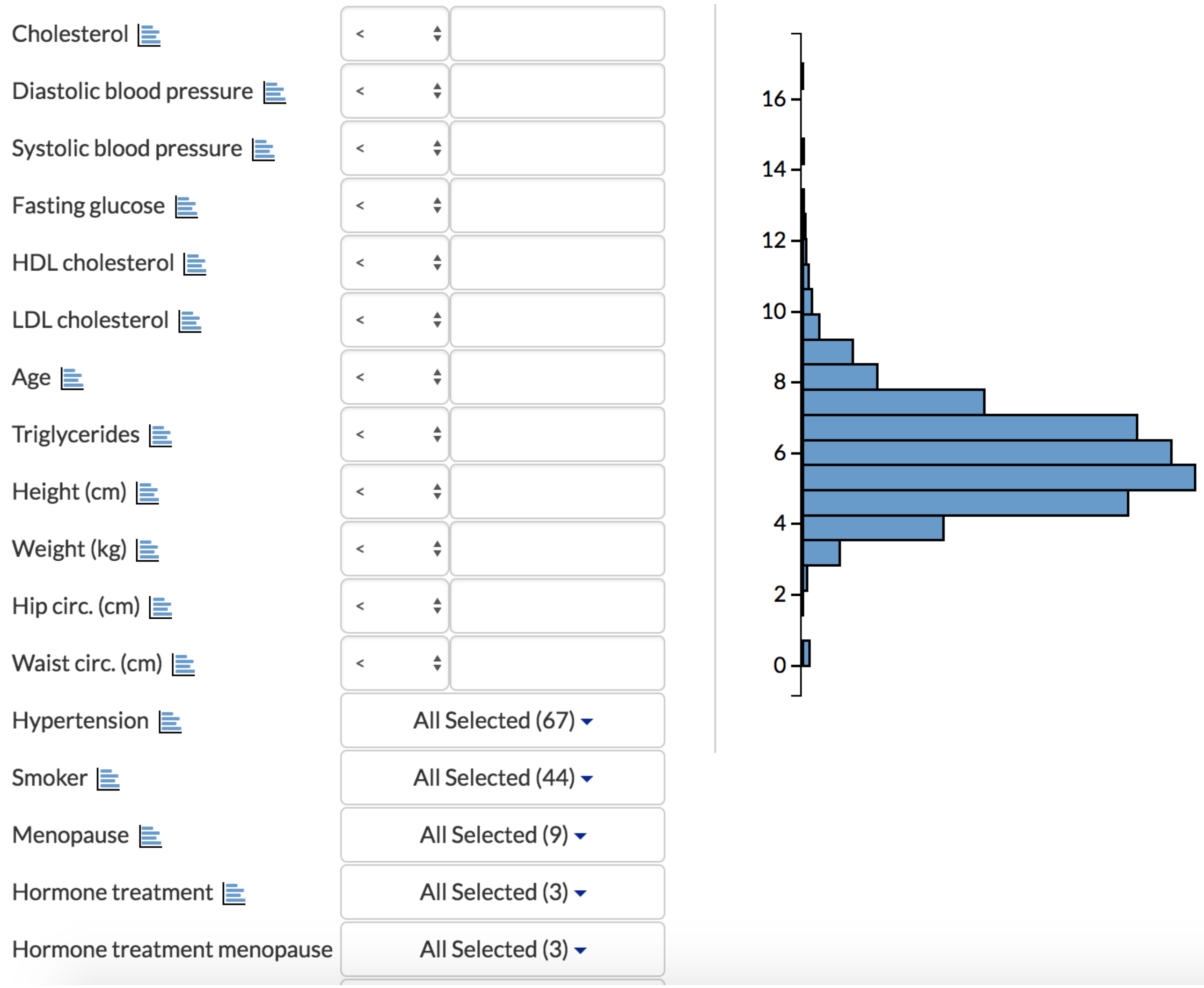
Cholesterol	<	
Diastolic blood pressure	<	
Systolic blood pressure	<	
Fasting glucose	<	
HDL cholesterol	<	
LDL cholesterol	<	
Age	<	
Triglycerides	<	
Height (cm)	<	
Weight (kg)	<	
Hip circ. (cm)	<	
Waist circ. (cm)	<	
Hypertension	All Selected (67)	
Smoker	All Selected (44)	
Menopause	All Selected (9)	
Hormone treatment	All Selected (3)	
Hormone treatment menopause	All Selected (3)	

Interactive histograms and phenotype data

- Interactive histograms allow users to understand how certain traits (age, smoking status, etc) differ in dataset

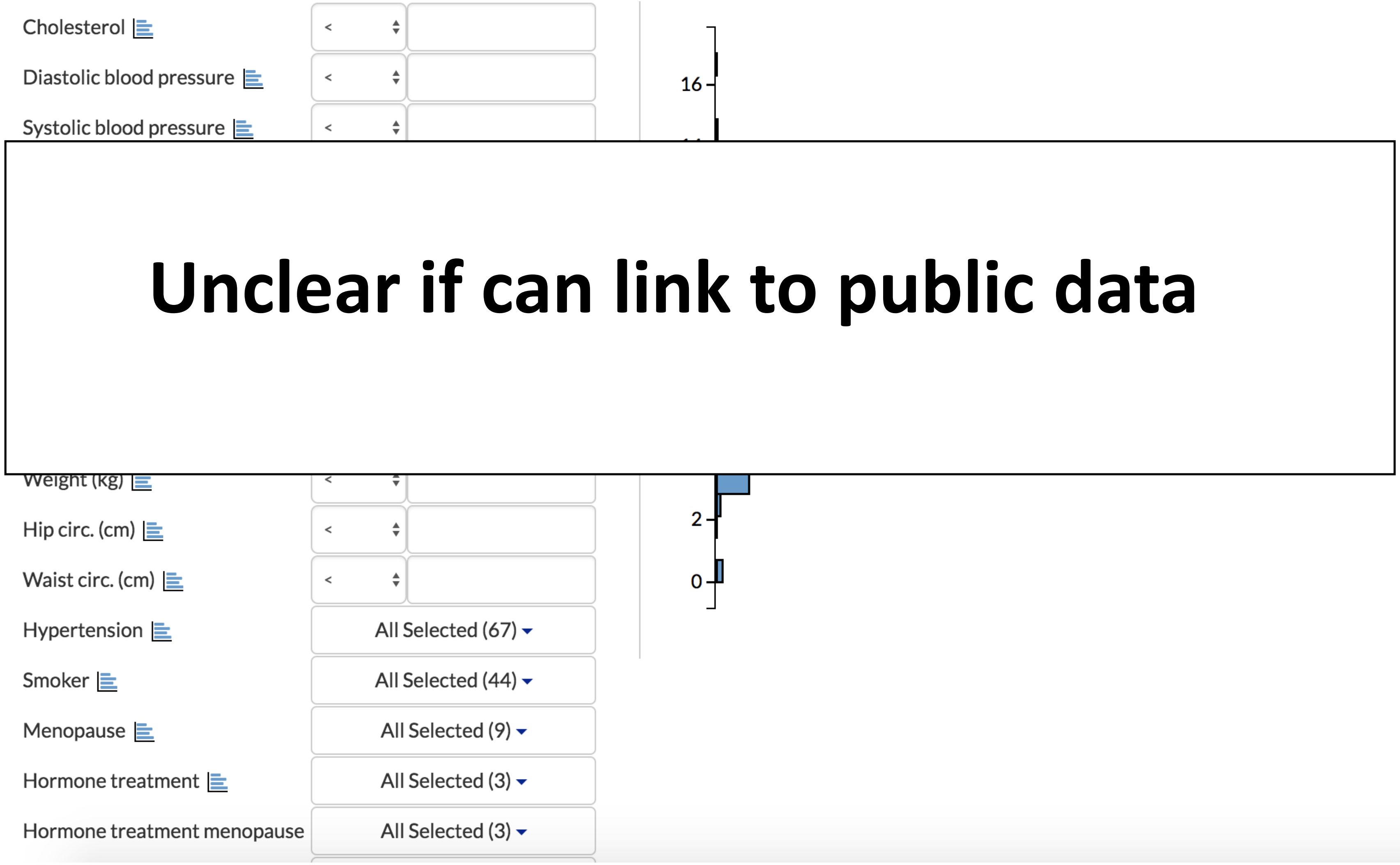
- Also allow for reconstruction of phenotype information!

Interactive histograms and phenotype data



Cholesterol	<	
Diastolic blood pressure	<	
Systolic blood pressure	<	
Fasting glucose	<	
HDL cholesterol	<	
LDL cholesterol	<	
Age	<	
Triglycerides	<	
Height (cm)	<	
Weight (kg)	<	
Hip circ. (cm)	<	
Waist circ. (cm)	<	
Hypertension	All Selected (67)	
Smoker	All Selected (44)	
Menopause	All Selected (9)	
Hormone treatment	All Selected (3)	
Hormone treatment menopause	All Selected (3)	

Interactive histograms and phenotype data



Cholesterol	<	
Diastolic blood pressure	<	
Systolic blood pressure	<	
Weight (kg)	<	
Hip circ. (cm)	<	
Waist circ. (cm)	<	
Hypertension	All Selected (67)	
Smoker	All Selected (44)	
Menopause	All Selected (9)	
Hormone treatment	All Selected (3)	
Hormone treatment menopause	All Selected (3)	

Unclear if can link to public data

Quantifying privacy: Leaking health data

What Leaks?	**Outside knowledge**	**Statistic released**	**Likely Risk (very subjective)**
Participation in study	Background MAF and individuals genotype	Minor allele frequency	Low
Diseases status	Background MAF and individuals genotype	Minor allele frequency	Low
Other phenotype data	Unspecified	Interactive histogram	Moderate
Genotype	Background MAF	GWAS statistics for one query	Low
Genotype	Background MAF	GWAS statistics for numerous queries	High

Quantifying privacy: Leaking genotype data

What Leaks?	Outside knowledge	Statistic released	Likely Risk (very subjective)
Participation in study	Background MAF and individuals genotype	Minor allele frequency	Low
Diseases status	Background MAF and individuals genotype	Minor allele frequency	Low
Other phenotype data	Unspecified	Interactive histogram	Moderate
Genotype	Background MAF	GWAS statistics for one query	Low
Genotype	Background MAF	GWAS statistics for numerous queries	High

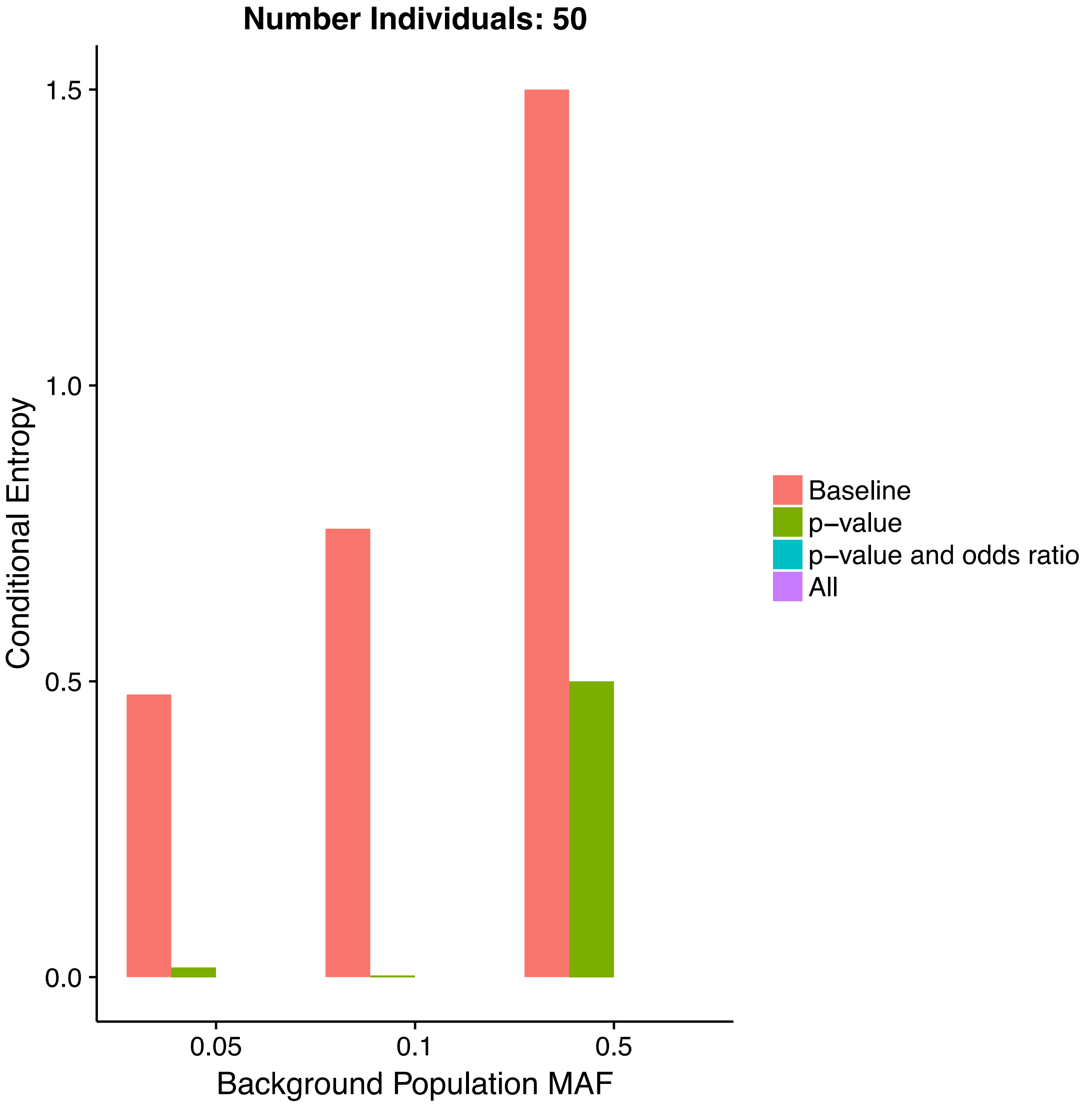
Interactive analysis and privacy

- Run an analysis on one group of individuals

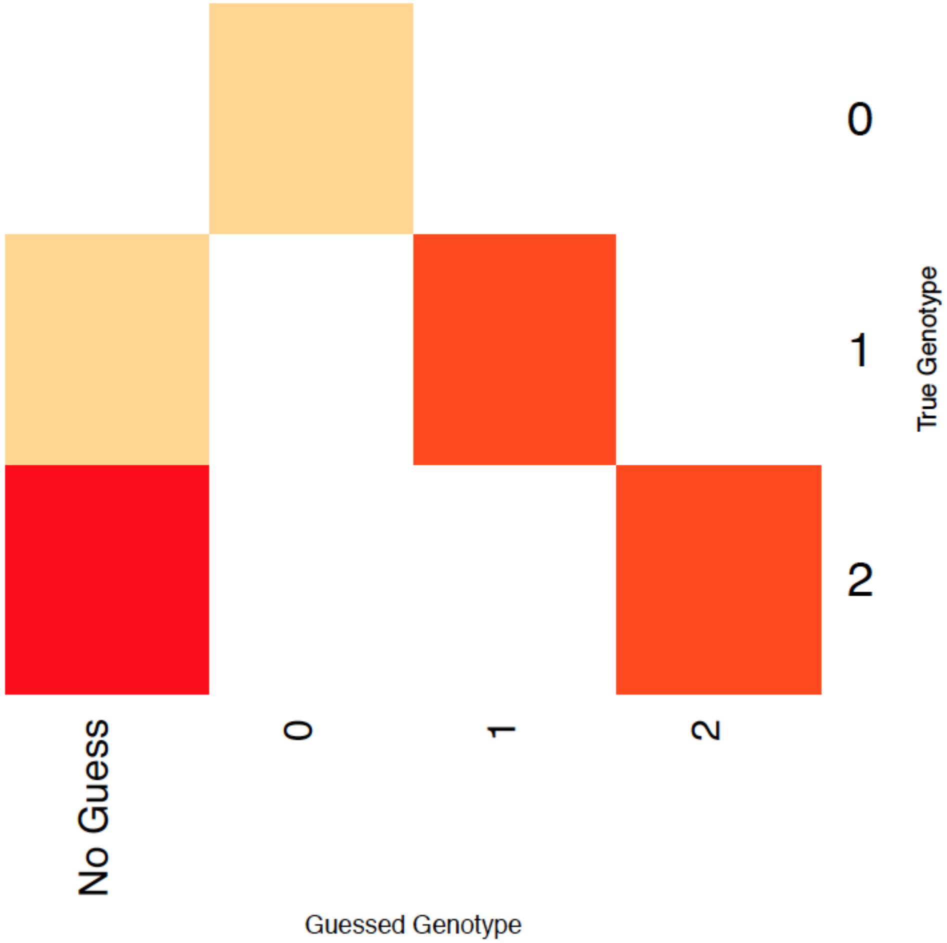
- Run an analysis on a slightly different group of individuals

- Combine to learn about genotype of one target!

Entropy in genotype data: Interactive analysis



Recovering genotype data



Mitigation?

- Larger minimum analysis size, less precision

- Remove some features

- Require PCs as covariates

- K-anonymity for histograms

- Remove Y chromosome

What else is needed?

- Doesn’t consider all possible queries, only a few select ones.

- What are we missing? No good way to measure!!

- How do covariates actually effect the privacy risk? Harder to answer.

- Can the information leaked here be linked to outside databases?