Integrating GWAS with omic data

Wei Pan

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455
Joint work with Zhiyuan Xu (LinkedIn), Chong Wu (FSU), Haoran Xue, ...

March 1, 2019
CCBR

Outline

- Introduction: problem
- Review: Transcriptome-Wide Association Study (TWAS)/PrediXcan
- Our method: (weighted) aSPU test.
- GWAS + gene expression: application to the Lipid data
- GWAS + EPI + Methylation: application to the SCZ data
- On-going: Mendelian randomization (MR)
  determining causal direction between two variables: LDL/HDL vs CAD.

Introduction

- Problem: to detect SNP-disease associations in GWAS (or sequencing studies).
  Big question: mechanistic interpretation?
- Approach: integrating GWAS data with eQTL data
  1) to boost power;
  2) to enhance interpretation.
- Motivation: DNA $\Longrightarrow$ mRNA $\Longrightarrow$ ...$\Longrightarrow$ Disease.
  Known: many disease-associated SNPs are eQTL.

TWAS/PrediXcan: GWAS + eQTL

- Set-up: 2 independent data sets, one GWAS (main), one eQTL (smaller).
  GWAS: SNPs, Y (disease);
  eQTL: SNPs, X (gene expression).
- Conventional GWAS: $Y \sim SNP$.
- PrediXcan (Gamazon et al 2015, *Nat Genet*), TWAS (Gusev et al (2016, *Nat Genet*):
  1. eQTL: $E(X) = \text{SNPs}' * w \Longrightarrow \hat{w}$;
  2. GWAS: 1) $\hat{X} = SNPs' * \hat{w}$; 2) $Y \sim \hat{X}$
- Why? why not $Y \sim X$?
  Biologically: genetically regulated component of gene expression (GReX);
  X not available in the GWAS data, often.

Our views

- Statistically: (two-sample) 2SLS
  related to Mendelian randomization (MR); causal inference!
- Our **key** obs: PrediXcan/TWAS = weighted SPU(1)!
  weight each SNP j by $\hat{w}_j$, then ...
- Why not use aSPU (or other more powerful tests)? (Xu et al 2017, *Genetics*; Su et al 2018, *AJHG*)
- Why not use other weights derived from other omic or endophenotypes?
  brain imaging for AD (Xu et al 2017, *NeuroImage*);
  enhancer-promoter interactions (EPIs) (Wu and Pan 2018, *Genetics*);
  EPIs + meQTL (Wu and Pan 2019, *BI*).

Lipid GWAS + eQTL

- Discovery data: a large 2010 Lipid dataset (Teslovich et al 2010, *Nat Genet*).
  Summary stats of meta-analysis of 46 GWAS with
  $n \approx 100,000$ individuals;
  Lipid traits: LDL, HDL, TG, TC; use LDL here.
- Validation data: a **larger** 2013 dataset (GLGC 2013, *Nat Genet*).
  $n = 188,577$.
- Three eQTL datasets: NTR, YFS and METSIM;
  extracted the (optimal) weights constructed by Gusev et al (2016);
  containing 1264, 3555 and 2295 (and 1223 for a combined analysis) genes.
- As in Gusev et al (2016), for gene-based analysis, conservatively use
  genome-wide significance level=0.05/8500=5.88E-6.
- Applied our aSPU and TWAS=SPU(1).

Table: The numbers of the significant genes identified by analyzing the 2010 lipid data. a/b/c indicate the numbers of (a) the significant genes; (b) the significant genes that covered a genome-wide significant SNPs in the 2010 lipid data; (c) the significant genes that covered a genome-wide significant SNPs in the 2013 lipid data.

Trait	Test	NTR	YFS	METSIM	Combined
HDL	aSPU	19/16/17	29/27/29	22/19/22	21/17/17
	TWAS	16/14/15	25/22/24	19/15/19	20/16/17
LDL	aSPU	15/15/15	19/18/18	17/16/17	14/13/13
	TWAS	8/7/8	10/9/9	7/7/7	7/7/7
TG	aSPU	17/16/17	33/30/32	15/14/14	20/19/19
	TWAS	9/9/9	17/16/17	8/7/7	12/11/11
TC	aSPU	26/25/26	28/26/27	28/28/28	20/20/20
	TWAS	15/14/15	18/16/17	15/14/15	14/13/13

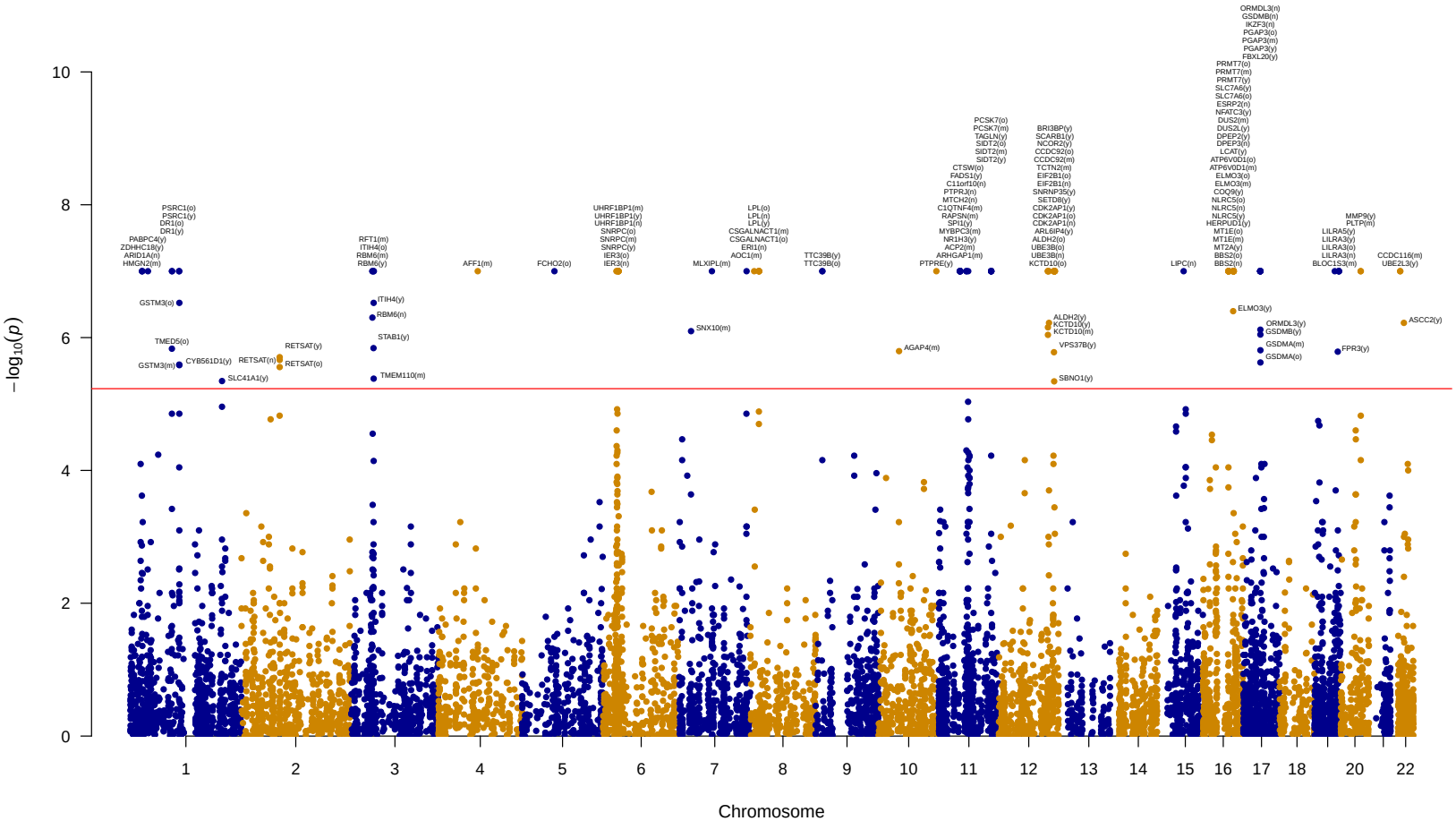


Figure: Manhattan plots for the pooled results of aSPU and aSPU-Ofor traits HDL based on the 2013 lipid data.

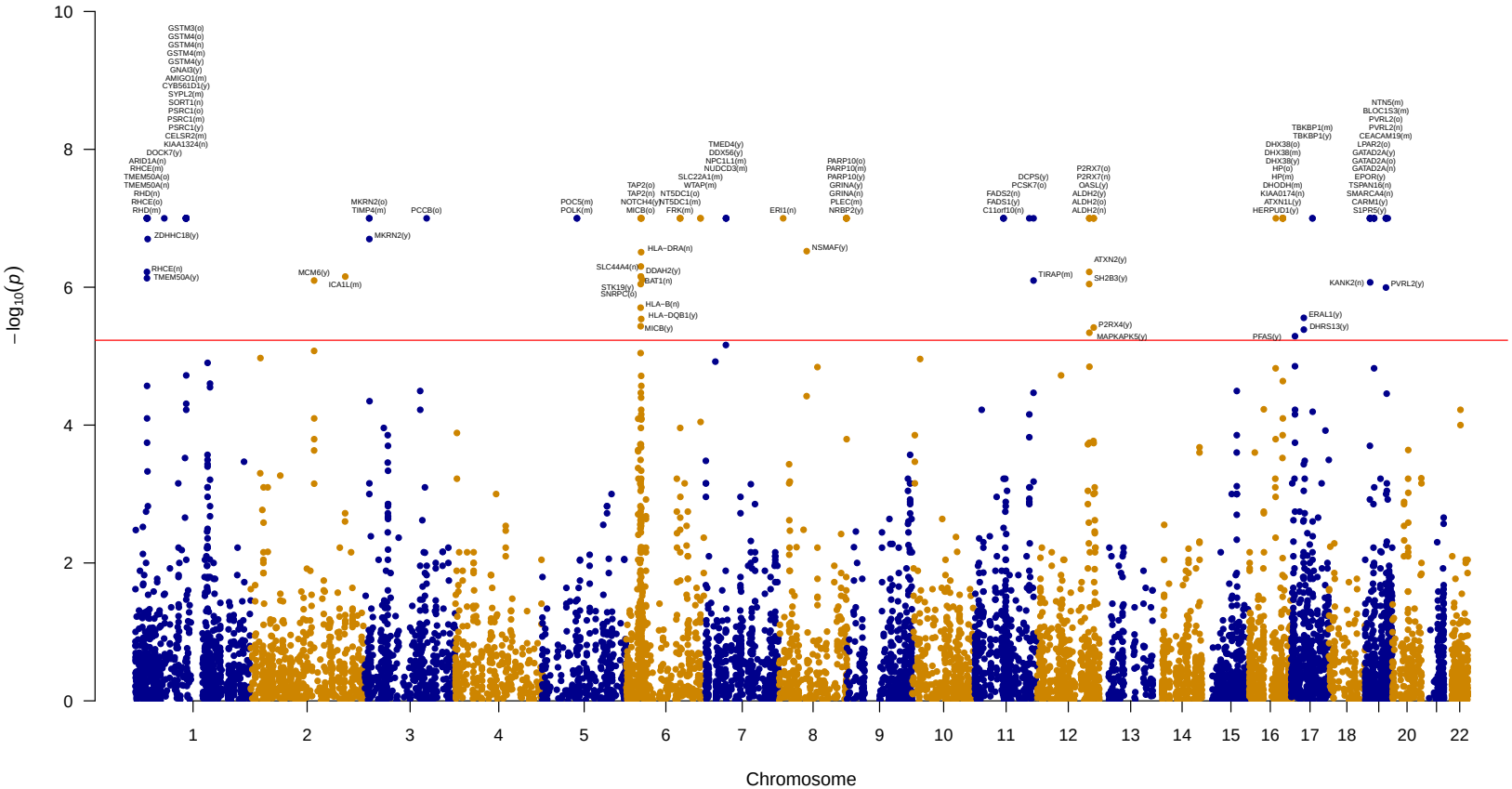
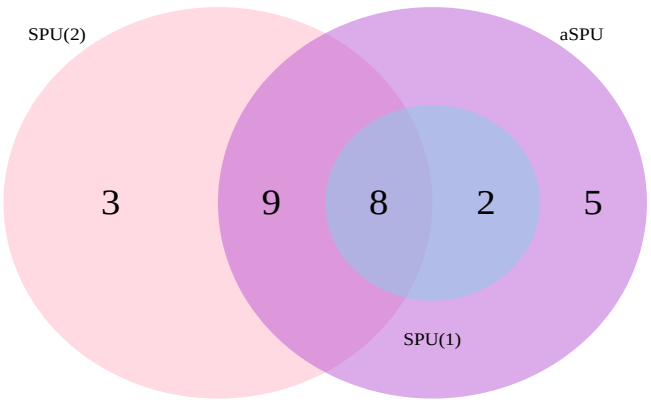


Figure: Manhattan plots for the pooled results of aSPU and aSPU-O for traits LDL based on the 2013 lipid data.

(PGC) SCZ GWAS + EPI + meQTL



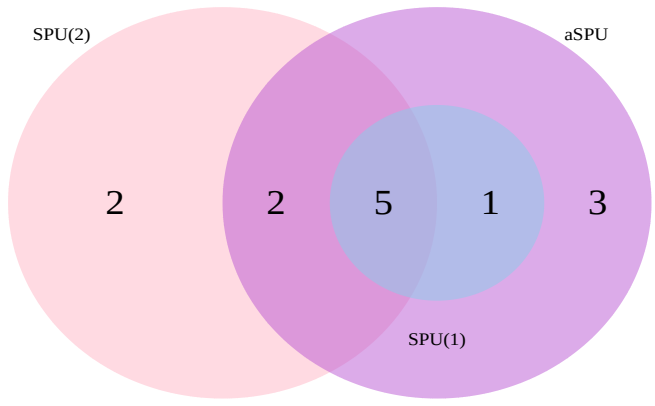
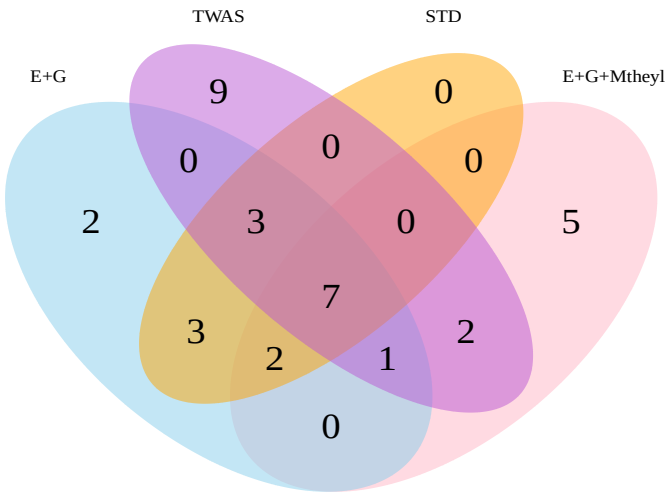


Figure: Left: significant genes; right: sigificant and novel ones for "EPI+meQTL".

(PGC) SCZ GWAS + EPI + meQTL



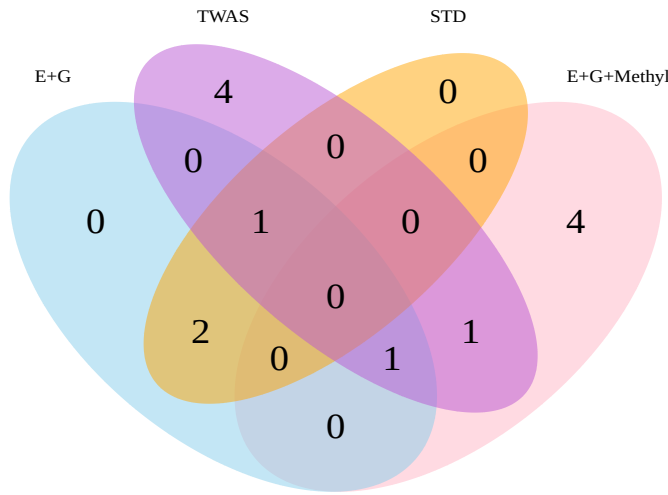


Figure: Left: significant genes; right: sigificant and novel ones. Combined SPU(1) and SPU(2).

EPI prediction

- TargetFinder: use epi-genomic features (from ENCODE and Roadmap); boosting ($>$ RF) (Whalen et al 2016, *Nat Genet*). Cao et al (2017, *Nat Genet*)
- SPEID: use DNA sequence (Singh et al 2016, bioRxiv); DL (Singh et al 2016, bioRxiv).
  RNN + CNN
- Ours: use DNA seq with two main contributions:
  1) only CNN;
  2) transfer learning: use all cell lines before cell line specific learning.

Our CNN

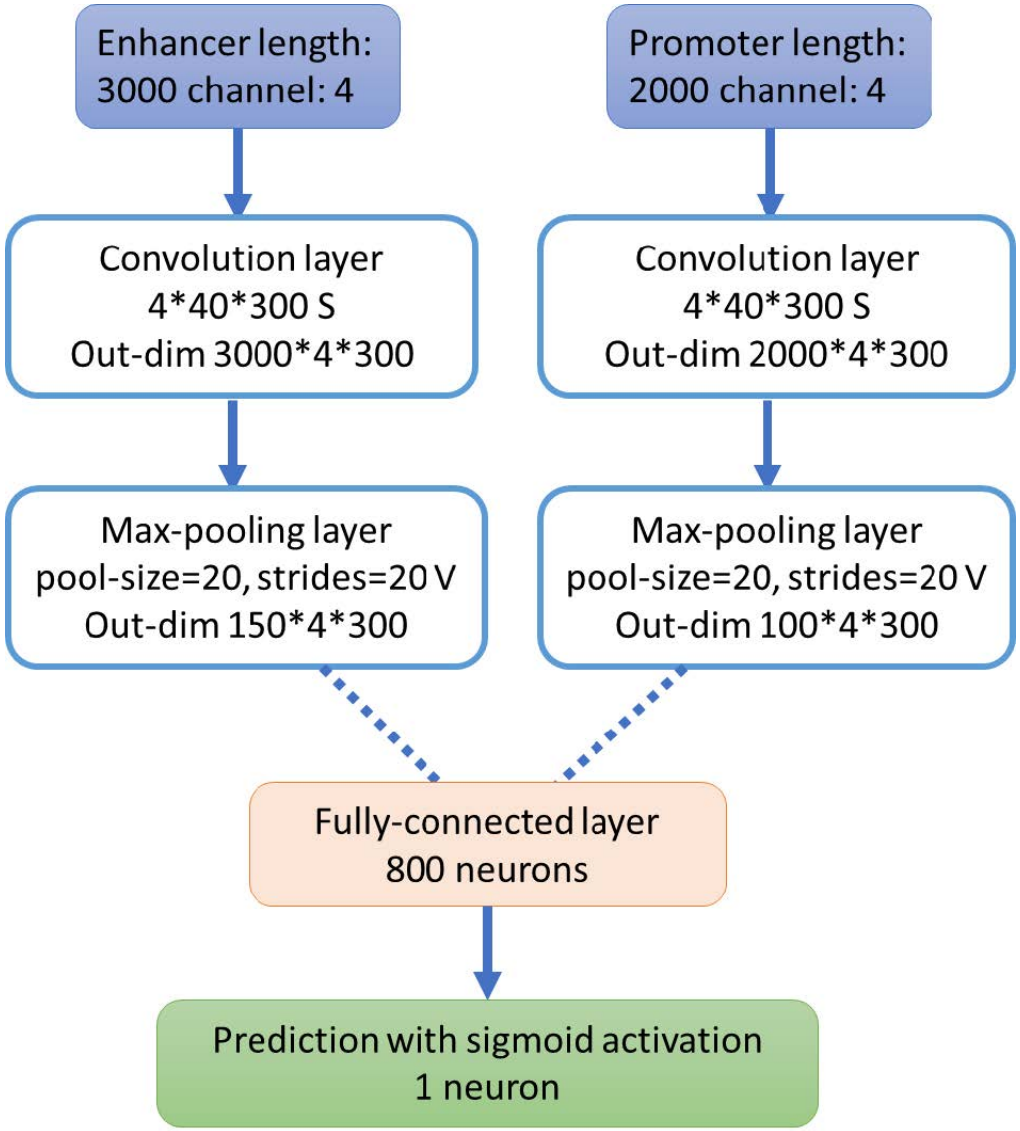


Figure: Our CNN architecture for EPI prediction.

Performance

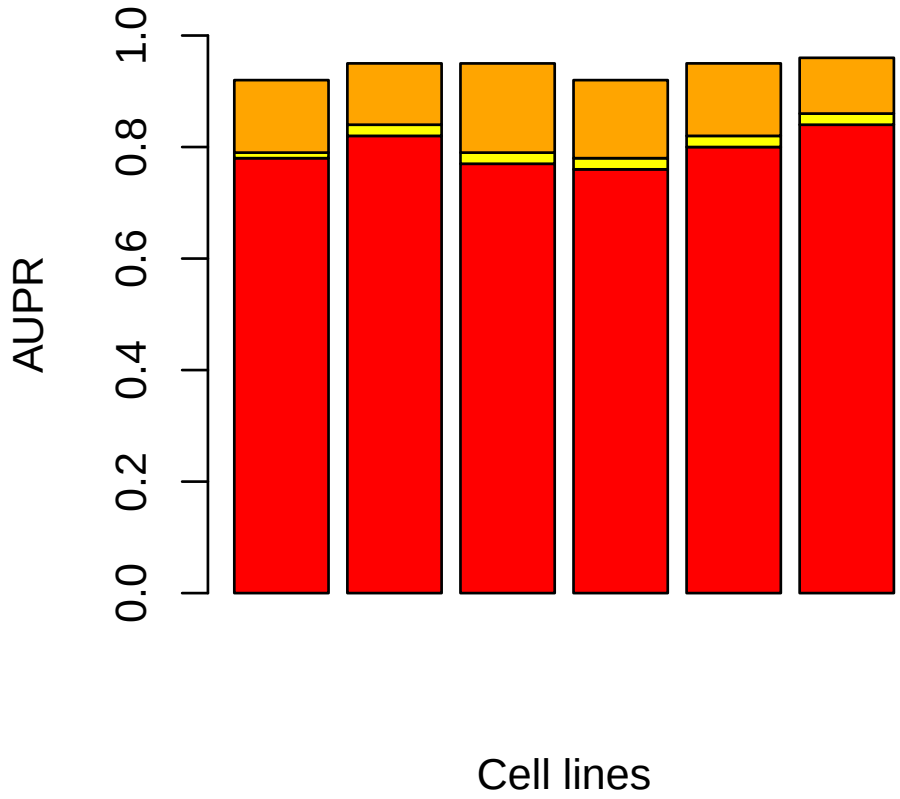


Figure: AUPR for the six cell lines based on SPEID (red/bottom bars), our CNN (yellow/middle bars) and our CNN+transfer learning (organge/top bars).

Orienting the causal relation

- Question: causal direction between X and Y?
  $X \Longrightarrow Y$, or $Y \Longrightarrow X$?
- Previous applications: gene expression $\Longrightarrow$ LDL (or SCZ)?
- Example: LDL $\Longrightarrow$ CAD?
  Statins; Mendelian randomization (MR) analyses.
- Example: HDL $\Longrightarrow$ CAD?
  Failed drug trials; MR analyses: inconclusive.
- Example: Education level $\Longrightarrow$ AD?
  A *Lancet* Commission (Livingston et al 2017): possible to prevent about 35% of dementia by controlling nine risk factors: education to a maximum of age 11-12 years, midlife hypertension, midlife obesity, hearing loss, late-life depression, diabetes, physical inactivity, smoking, and social isolation.

Genetics-based methods

- *Using genetic data to strengthen causal inference ...* (Pingault et al 2018, *Nat Rev Genet*).
- Use SNPs as anchors/instrumental variables (IVs) (Schadt et al 2005, *Nat Genet*; Chen et al 2007, *Genom Biol*; ...).
  SNPs $\Longrightarrow$ ...; not the reverse!
  SNPs: somewhat randomized.
  take advantage of many existing large-scale GWAS!
- Mediation analysis:
  Causal inference test (CIT) (Millstein et al 2009, *BMC Genet*)
  Limitations: 1) require data (SNP, X, Y);
  often have two samples: (SNP, X), (SNP, Y).
  2) less robust to measurement errors.
- MR: Steiger's test (Hemani et al 2017, *PLOS Genet*)
  Theory: If SNP $\Longrightarrow X \Longrightarrow Y$, then $|\rho_{gX}| > |\rho_{gY}|$!
  Main idea: test their difference!
  Limitation: based on a single SNP, thus low statistical efficiency **and** low robustness! ——our task here!
- Others: Pickrell's (2016, *Nat Genet*); bi-directional MR ...

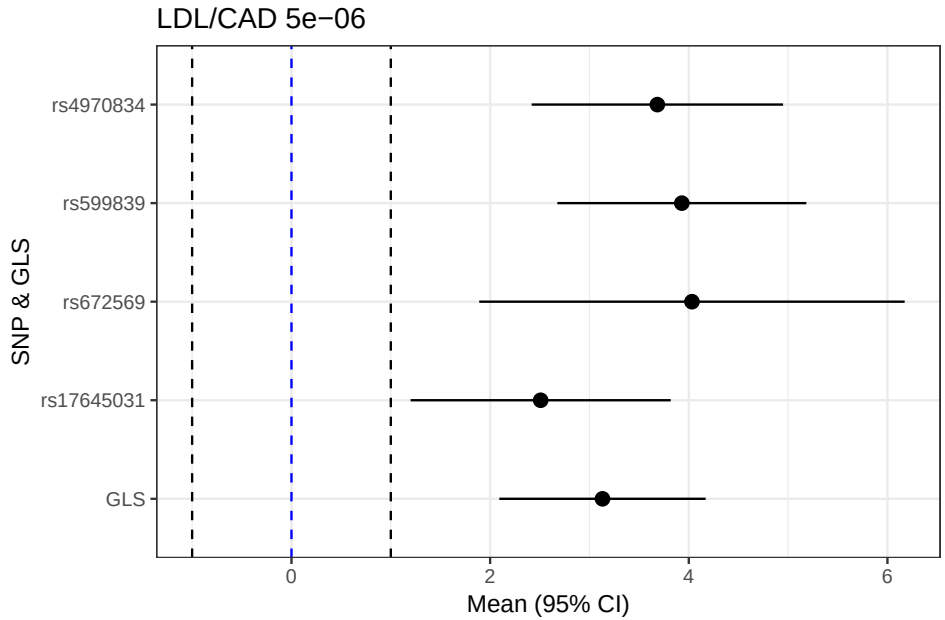
Our method

- Motivation: extending MR Steiger's method from using a single SNP to multiple SNPs.
  1) multiple correlated SNPs in a locus;
  2) multiple independent loci.
- Theory: If SNP $\Longrightarrow X \Longrightarrow Y$, then $\rho_{Yg} = \rho_{Xg}\rho_{YX}$.
  $\frac{\rho_{Yg}}{\rho_{Xg}} = \rho_{YX} := K$, $|K| < 1$,
  independent of g.
  Similarly, if SNP $\Longrightarrow Y \Longrightarrow X$, then ...
- Limitation: **cannot** distinguish $X \Longleftarrow$ SNP $\Longrightarrow Y$
- Main idea: combining multiple estimates r_{Yg}/r_{Xg} across g's...
  1) one locus: GLSE;
  2) multi-loci: IVW (meta-analysis).

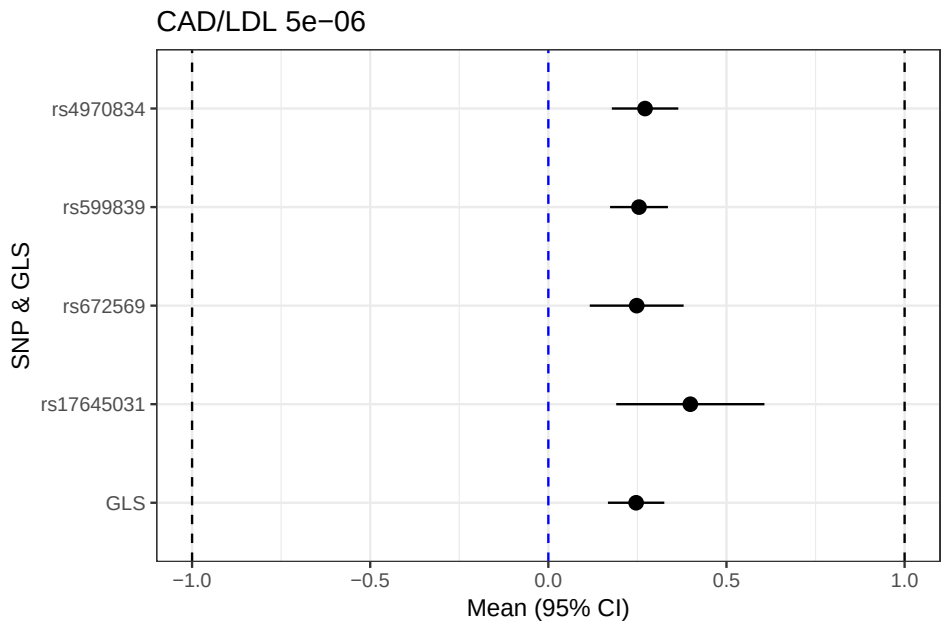
Example: LDL/HDL vs CAD

- Lipid GWAS summary data (Teslovich et al 2010, *Nat Genet*); CAD GWAS summary data (Schunkert et al 2011, *Nat Genet*); Reference panel: 489 individuals of EA in the 1000 Genomes Project.
- Partition the genome into 1703 (approximately) independent loci (Berisa and Pickrell 2016, *Bioinformatics*).
- Consider 8 (or 4) indep loci significant for both LDL (or HDL) and CAD (at $p < 5\text{E-}6$).
- In each locus, pruned out highly correlated SNPs with $|r| > 0.8$.

LDL vs CAD: Locus 1

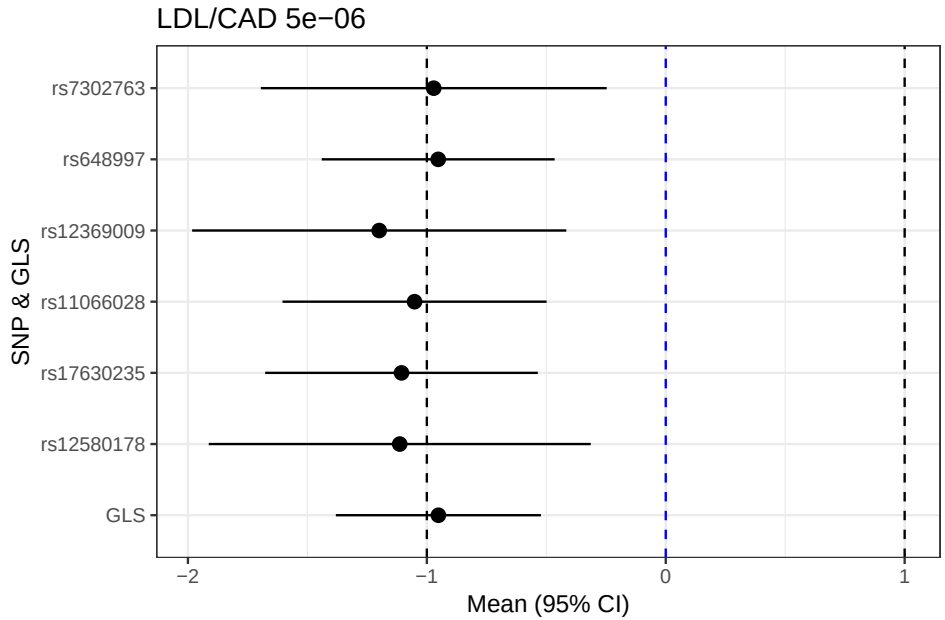


[]

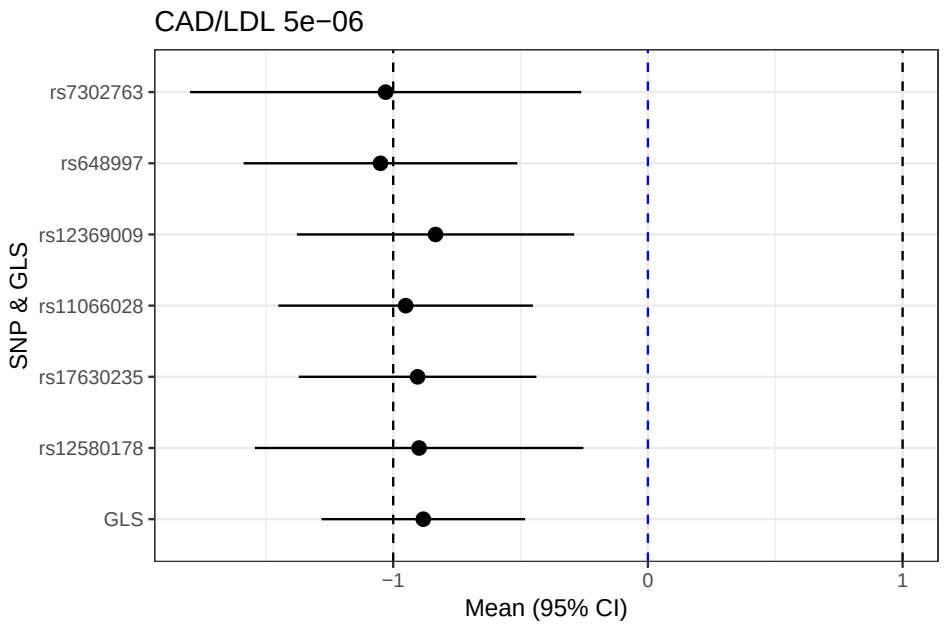


[]

LDL vs CAD: Locus 6

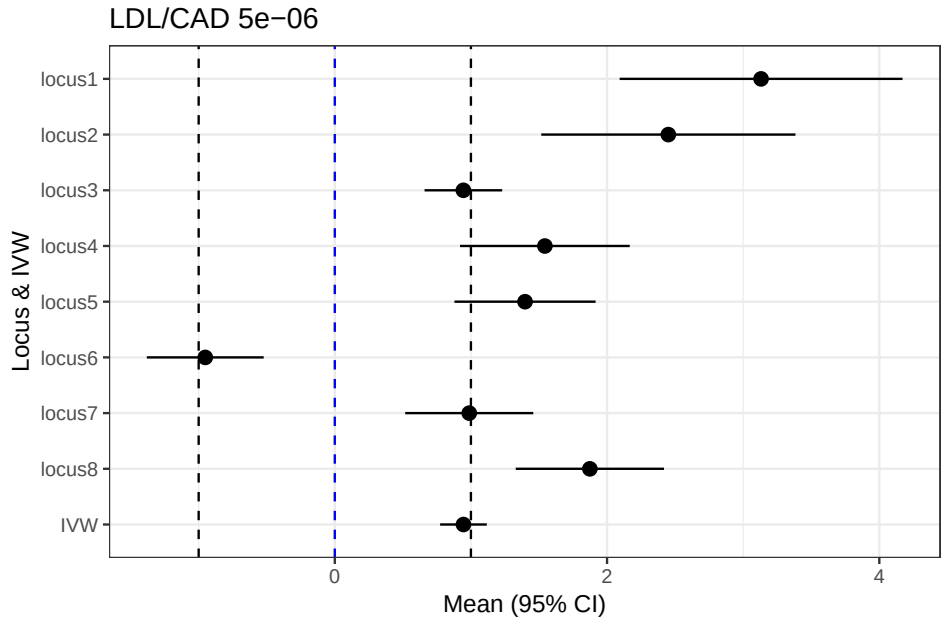


[]

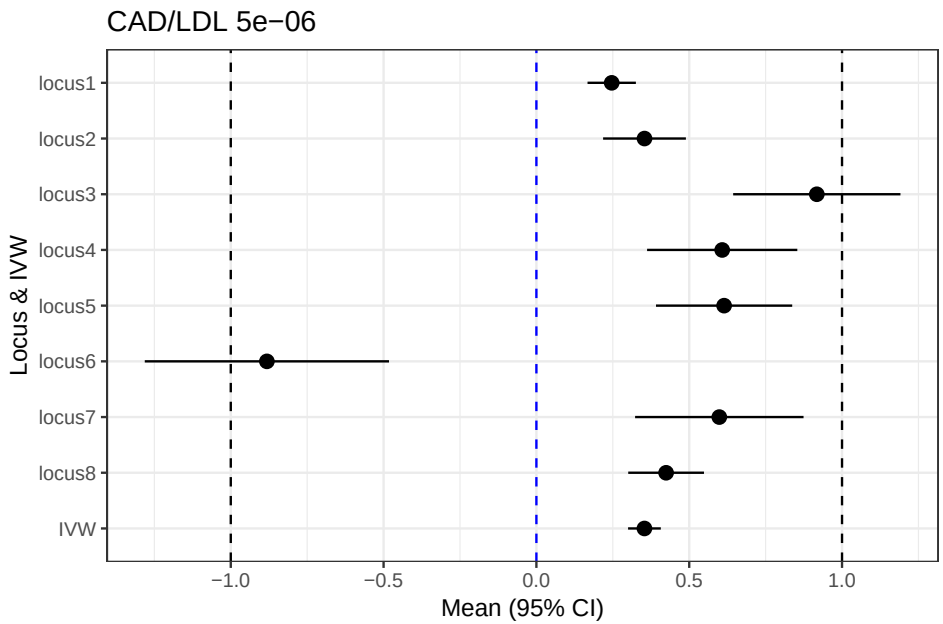


[]

LDL vs CAD: all 8 loci

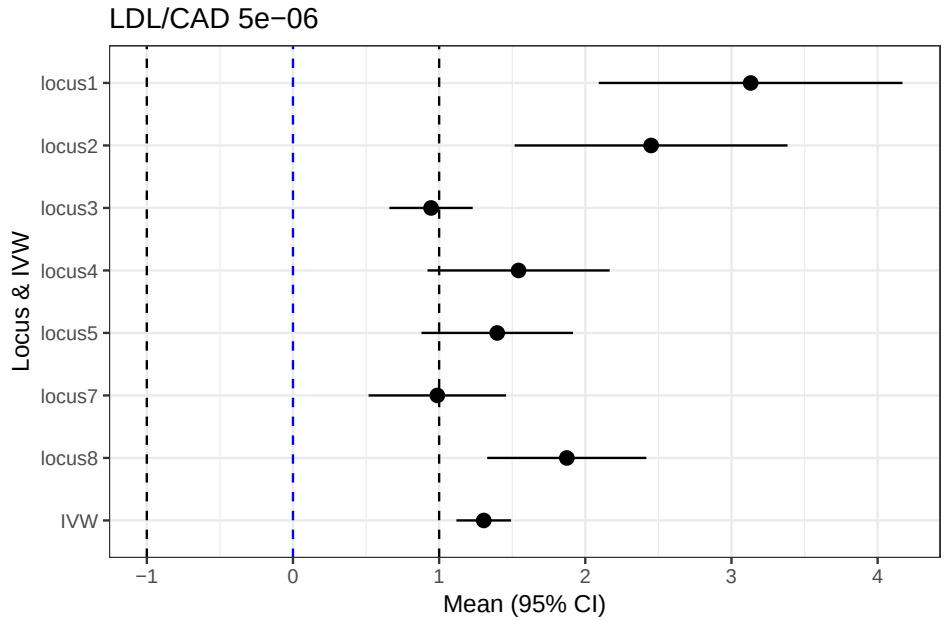


[]

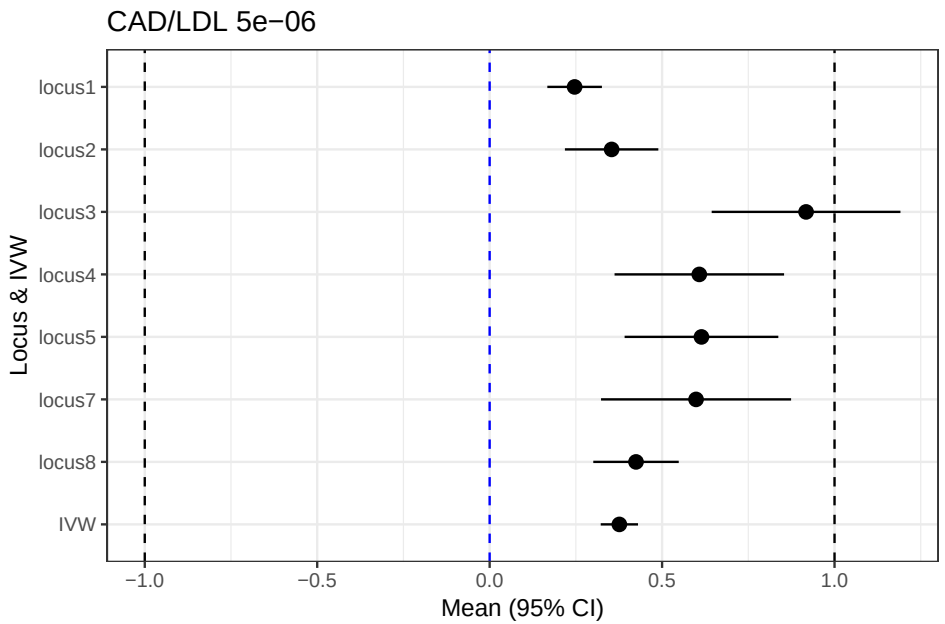


[]

LDL vs CAD: 7 loci

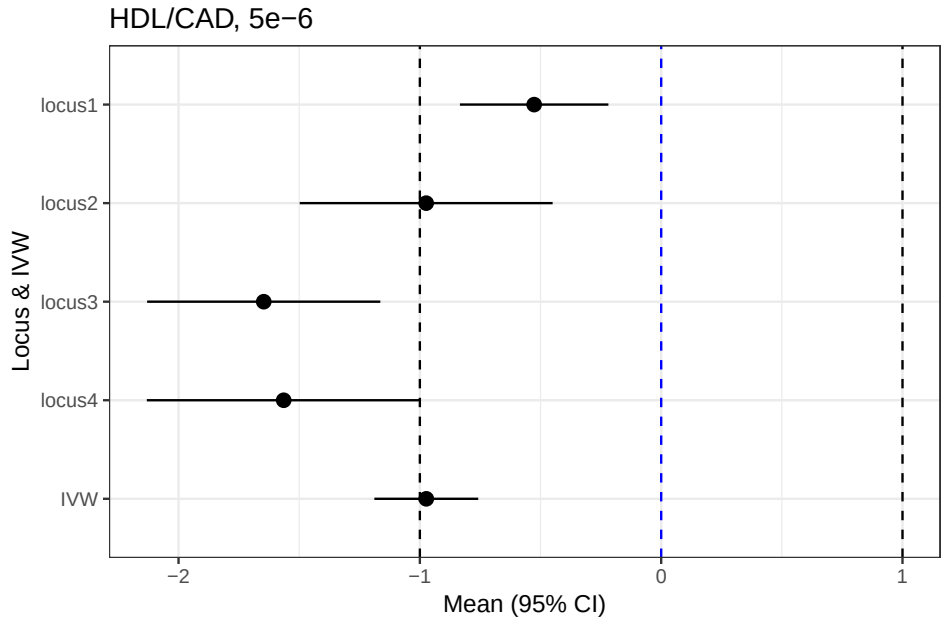


[]

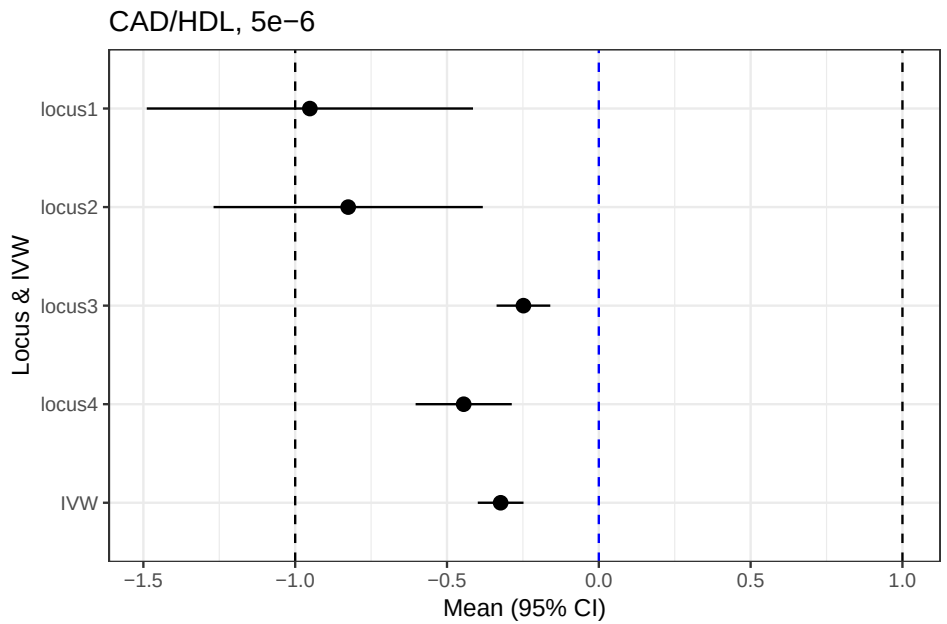


[]

HDL vs CAD: all loci



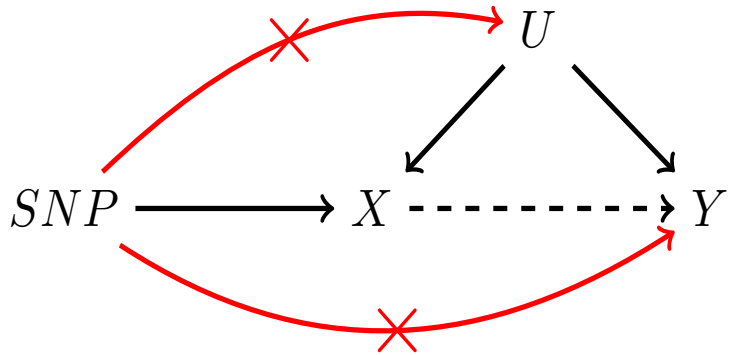
[]



[]

Estimating causal effects

- MR: using SNPs as IVs. IV assumptions:



- With a valid IV: $\beta_{YX} = \beta_{Yg}/\beta_{Xg}$.
  Wald ratios: $\hat{\beta}_{Yg}/\hat{\beta}_{Xg}$ for $g = 1, 2, ...$
  IVW (meta-analysis for multiple indep SNPs).
- More generally,
  $\hat{\beta}_{Yg} = \beta_{YX}\hat{\beta}_{Xg} + \epsilon_g$; IVW, PS (Zhao et al 2018)
  $\hat{\beta}_{Yg} = \beta_0 + \beta_{YX}\hat{\beta}_{Xg} + \epsilon_g$; Egger reg
  $\hat{\beta}_{Yg} = \beta_{0g} + \beta_{YX}\hat{\beta}_{Xg} + \epsilon_g$, $\beta_{0g} \sim_{iid} N(0, \tau^2)$; APS/RAPS
- Alternatively, use (weighted) median or mode of the Wald ratios ($\hat{\beta}_{Yg}/\hat{\beta}_{Xg}$'s).

On-going ...

- More applications:
  LDL/HDL vs CAD: larger datasets;
  brain imaging ROIs $\Longrightarrow$ AD?
- Co-localization testing
- Fine mapping
- TWAS/2SLS and MR: accounting for SNPs as invalid IVs
- Rare variants (RVs)
- .....

- `http://www.biostat.umn.edu/~weip`
  Code: `http://www.biostat.umn.edu/~weip/prog.html`
  R packages aSPU, `highmean`, GLMaSPU, GEEaSPU, POMaSPU, MiSPU; TLPglm; ...; all on CRAN.
  Websites with example code:
  `www.wuchong.org/IWAS.html`
  `www.wuchong.org/TWAS.html`
- This research was supported by NIH, NSF and MSI.
- Many collaborators and (former and current) **students**!

Thank you!