Estimating Genetic Effects Across Lipid Traits Jointly

Pradeep Natarajan

Sarah Urbut

The Problem: Interrogate 2.4 million SNPS across 4 traits jointly

- MLE summary statistics available across LDL, HDL, TG and TC from N = 89K
- The goal: to combine information across SNPs and across phenotypes to improve power and reveal underlying patterns of sharing
- Avoid simply looking for intersection underestimates sharing, misses power
- Report an Effect size rather than Simply Significance
  - existing methods typically focus only on testing for **significant effects** in each condition, and not **on estimating effect sizes**
  - Effect Size: **Quantitative** Heterogeneity
- Capture systematic heterogeneity (structured effects)
  - identify new *characteristic patterns of sharing: not same size or sign in all tissues*

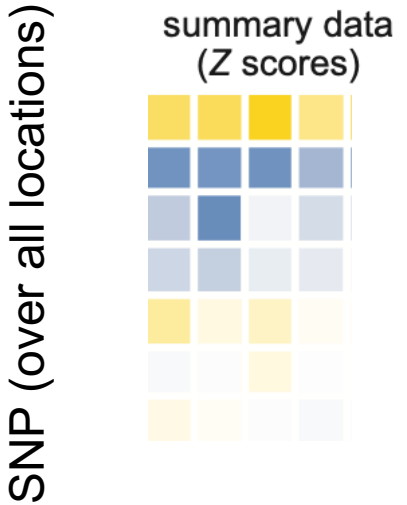
Patterns of sharing

- All SNPs belong to a finite number of classes
- Each class is characterized by a continuous patterns of sharing across conditions (but not necessarily limited by 'on/off' distinction
- Posterior estimates on any SNP are nudged towards the globalpatterns of patterns of sharing according to their best pattern 'fit'

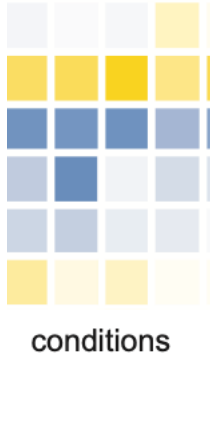
$$p(\boldsymbol{b};\boldsymbol{\pi},\boldsymbol{U}) = \sum_{k=1}^{K}\sum_{l=1}^{L} \pi_{kl} \quad N_R(\boldsymbol{b};\boldsymbol{0},\omega_l U_k),$$

- U_k captures *pattern 'shape'* or direction, ω_l captures scale
- π_{kl} to represent the (unknown) prior weight on prior covariance matrix U_{kl}:

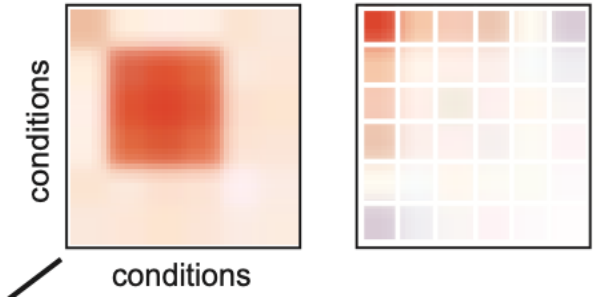
Overview



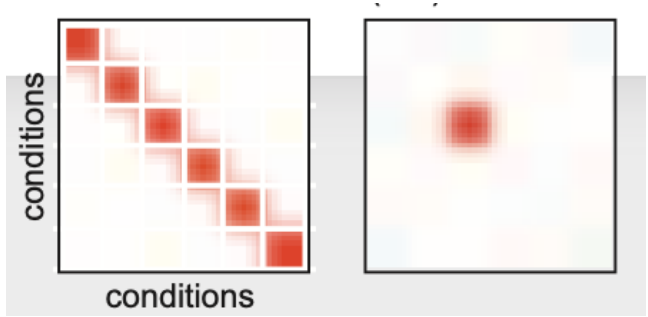
Select Strongest Signal
(max across conditions, 1 per LD Block)



Compute data-drive estimates of 'sharing'
Covariance matrices, U_k

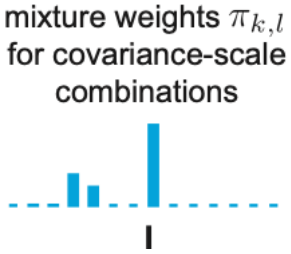


Add in canonical covariance matrices, U_k



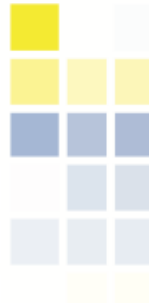
Expand by a grid of scaling factor, omega
Return relative weights,

$\pi_{k,l}$



Adapted from Urbut et al, NG 2019

Compute posterior estimates on
ANY snp (e.g. effect size, lfsr)



That are jointly 'shrunken' by larger data set to exploit sharing and increased precision

Power

- First considering all 2.4 million x 4 conditions possible associations (9.74 M)
- Univariate Shrinkage methods double power over Naiive P Val threshold of 5 e -8
- Joint Approach 2.5 fold increase

	Over All associations	SNPs significant in at least one condition
Bonferroni	14005	8595
UnivariateAsh	29301	18099
Mash	80539	32176

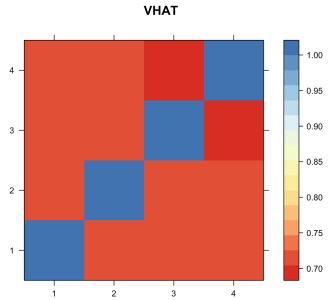
- Then consider each of 1704 LD blocks to see if a jointly shrunk SNP contains a significant (LFSR < 0.05) association in each condition
- Choose lowest association in each condition, maximum number of 1704 per condition
- Return list of 'best SNP' per condition, per block to avoid LD

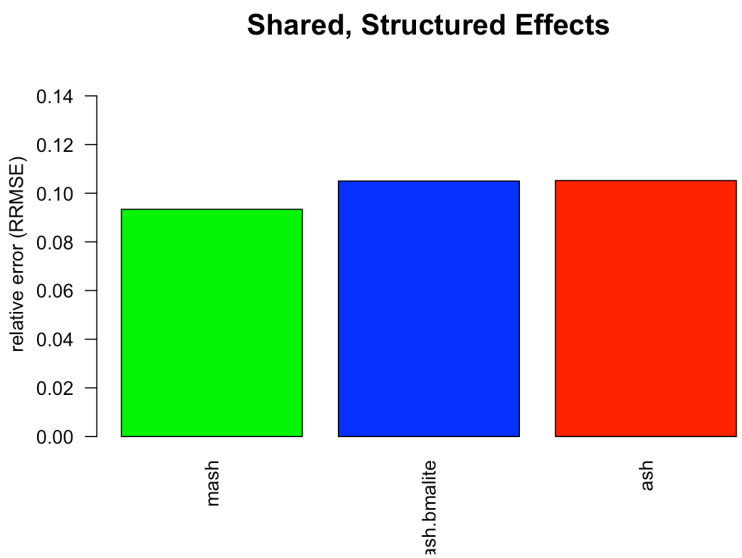
HDL	1081
LDL	490
TG	1031
TC	643

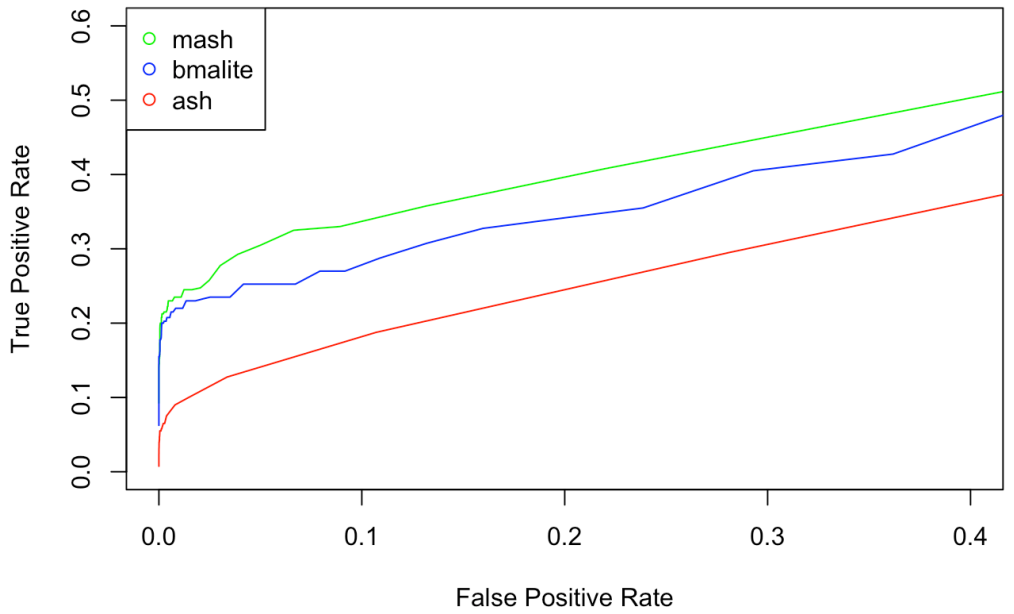
- 119 SNPS significant and max in their block and shared between LDL and TG

But are these real?

- Simulate data from empirical covariance matrices, 1% signal, true effects on same scale as observed
- Use correlated error matrix (rho ~ 0.8))
- Beats univariate and configuration approaches in both accuracy and power

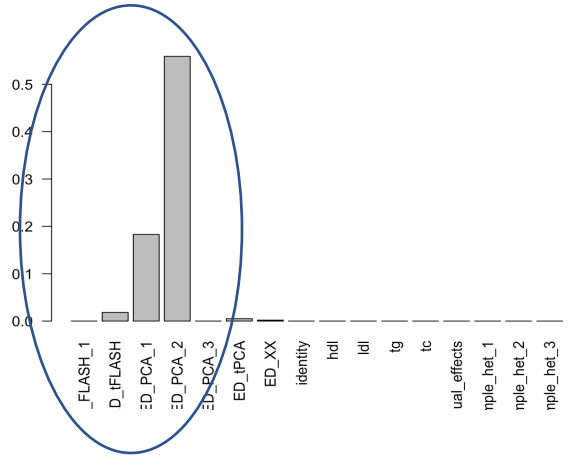


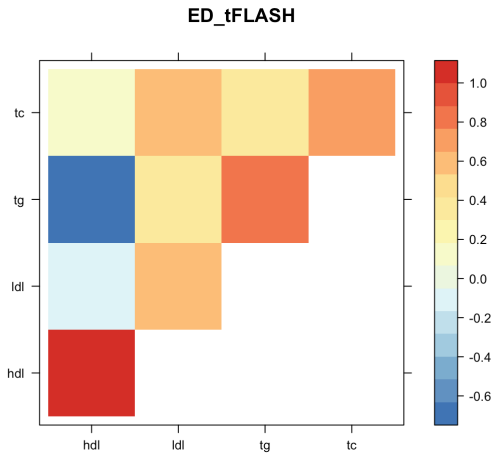


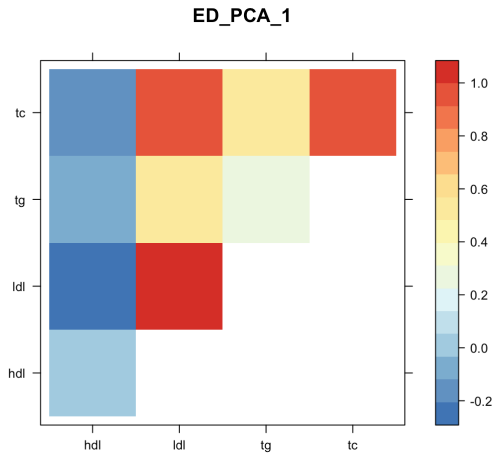


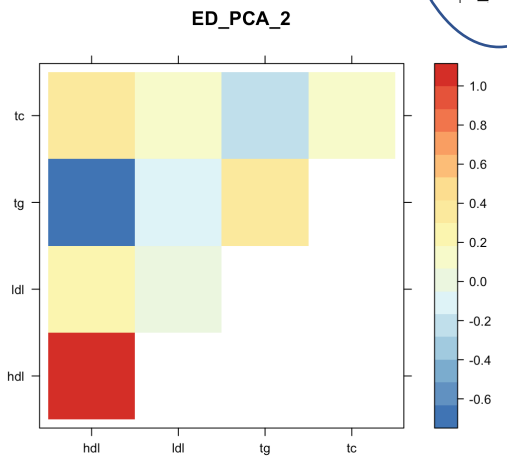
Patterns of Sharing

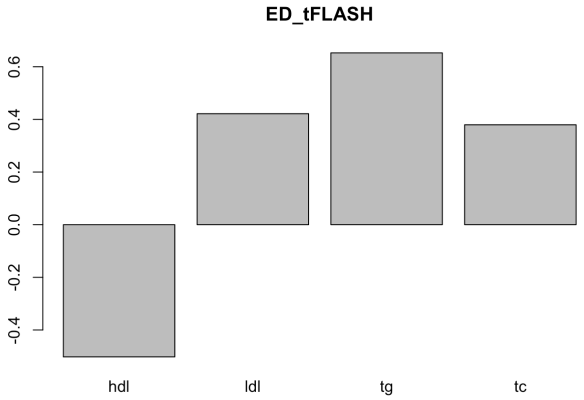
Majority loading on data-driven covariance matrices that reflect a high degree of sharing

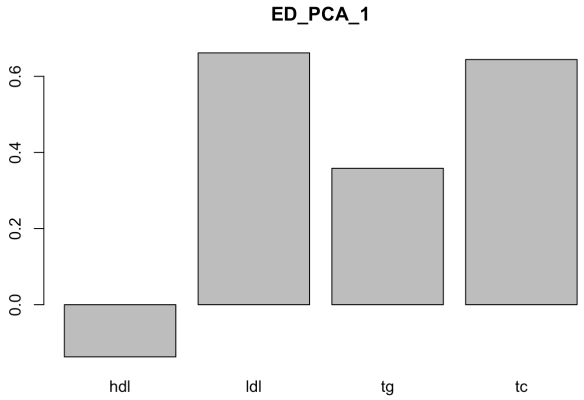


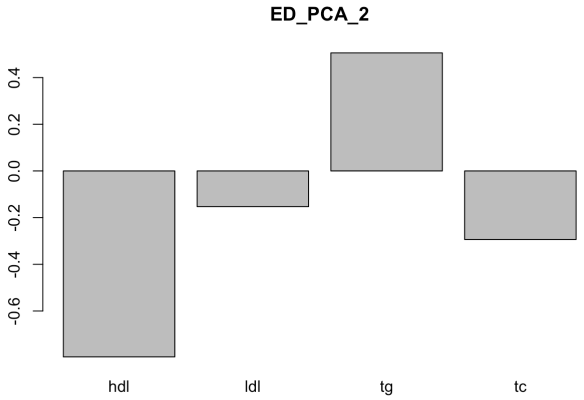








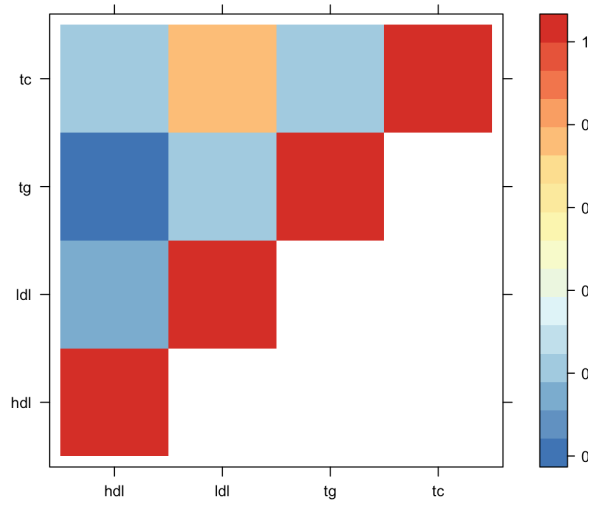




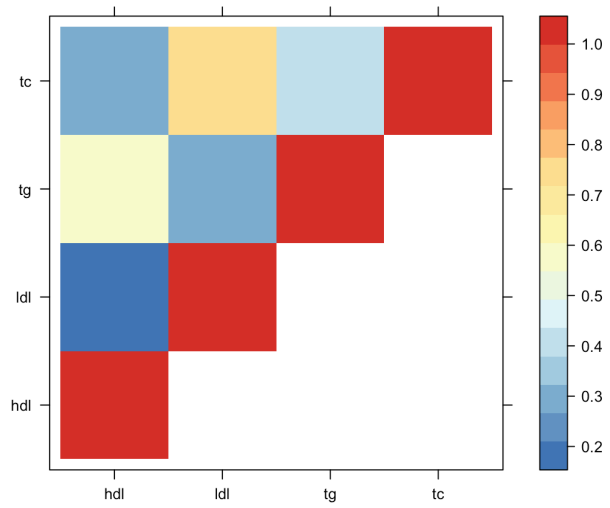
Sharing by Magnitude

Proportion of effects that are significant in at least one and within 2 fold magnitude (or absolute mag)

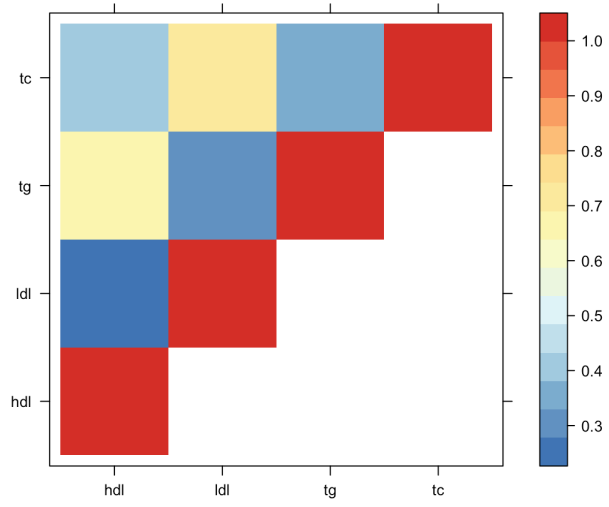
Shared by Magnitude



Shared by |Magnitude|

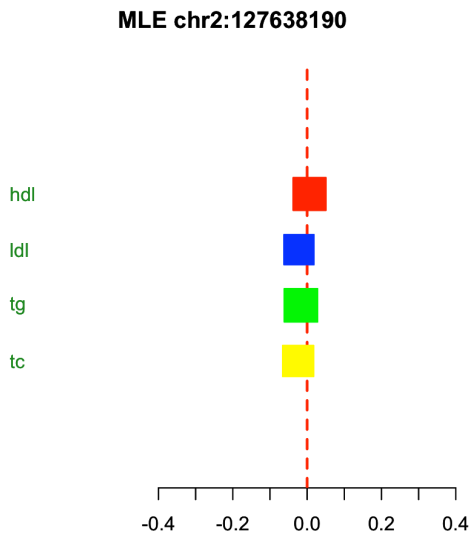


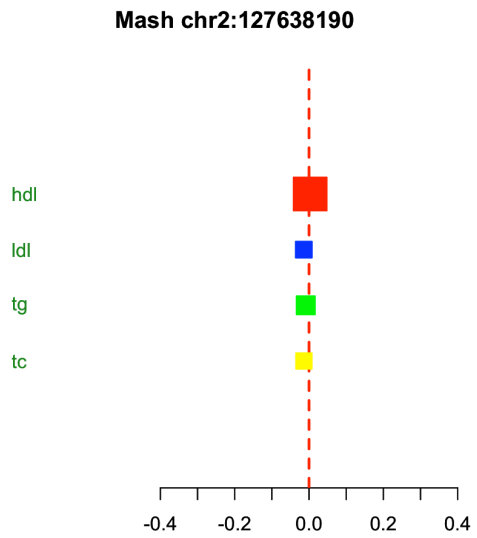
Shared by Significance

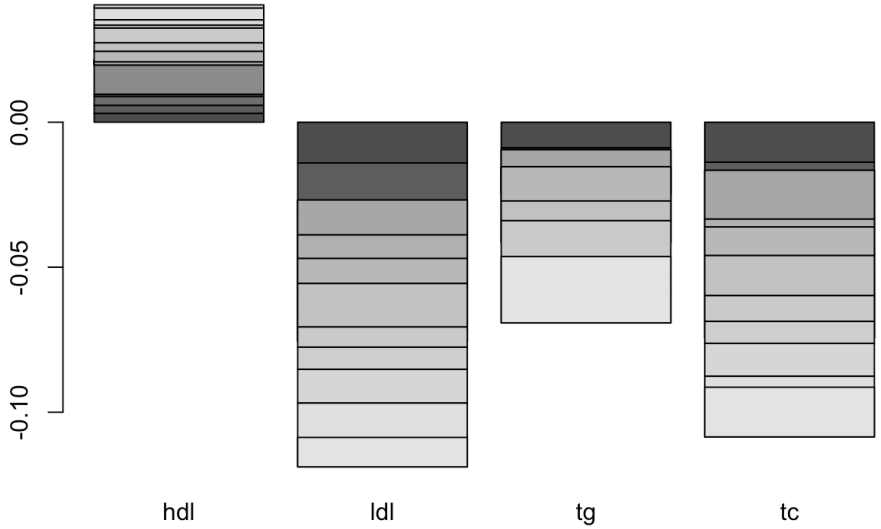


Candidate List

- SNPS that are max within block and significant at least at lfsr < 0.05, and shared by LDL and TG, and not HDL
- Intersect with CAD risk
- Downstream bitrait MR?







Shrink Error, nudge towards significance in LDL and TG due to heavy weight on sharing