Use and Interpretation of LD Score Regression

Brendan Bulik-Sullivan

bulik@broadinstitute.org

PGC Stat Analysis Call

BROAD INSTITUTE

Outline of Talk

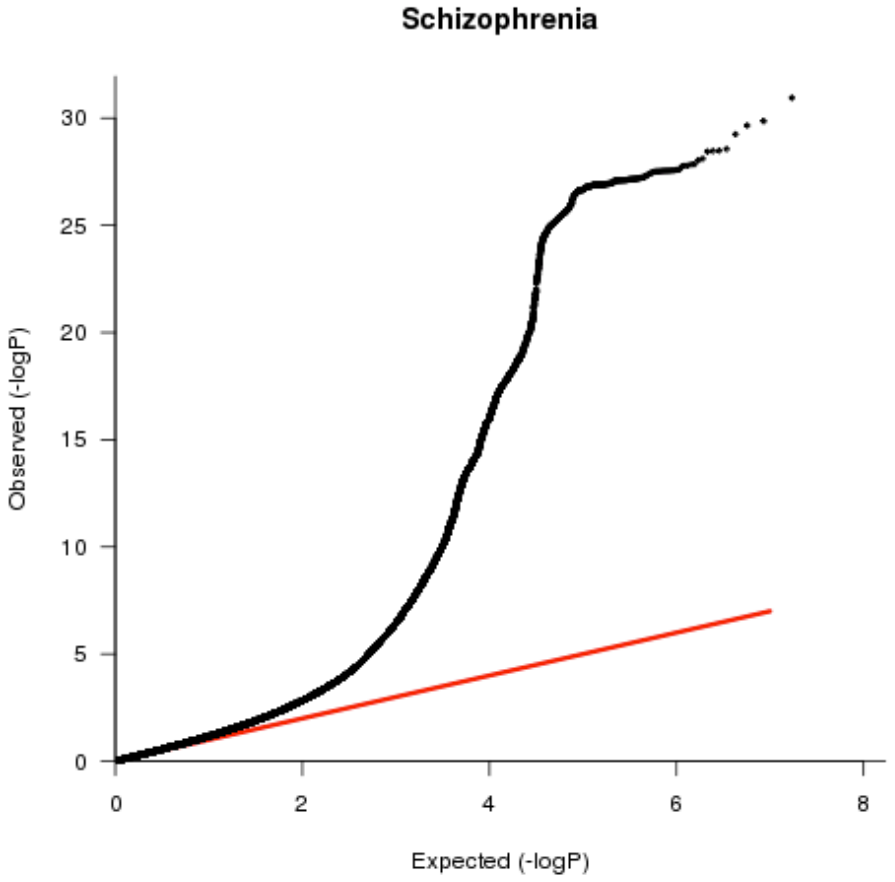
- Intuition, Theory, Results
  - LD Score regression intercept: distinguishing polygenicity from population stratification
  - Genetic correlation from summary statistics
- What can LD Score Regression do for *you*?
  - Practical advice on using LD Score in day-to-day GWAS analysis
- Useful links at the end

LD Score Regression Intercept

Distinguishing Polygenicity from Population Stratification

Test Statistic Inflation

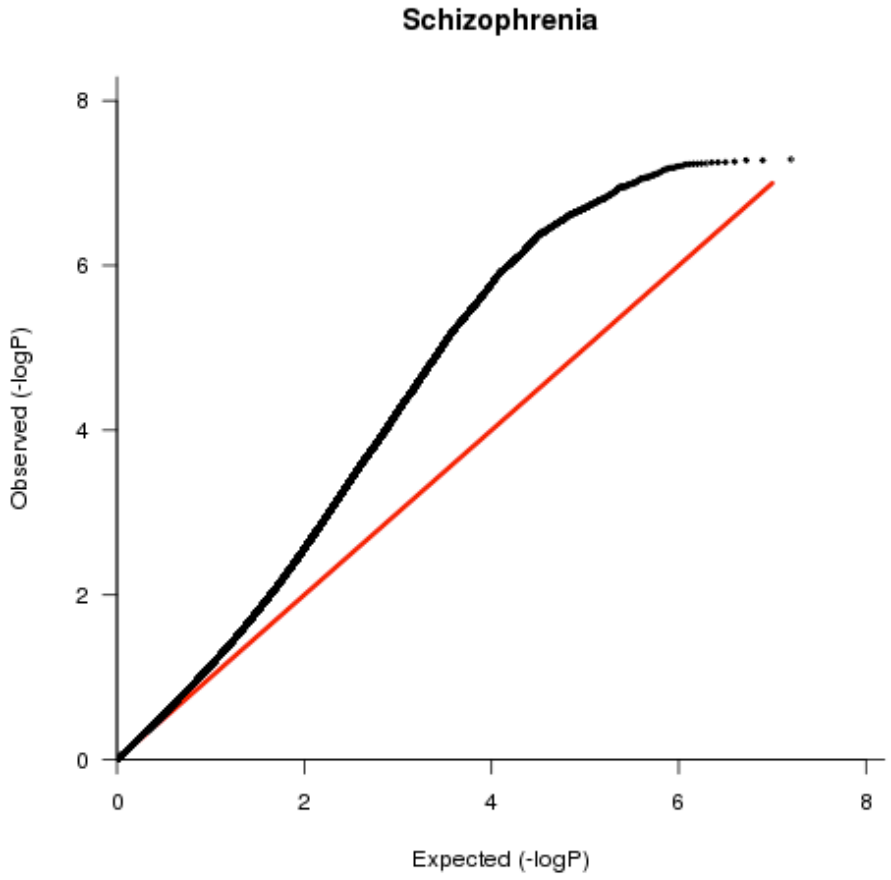
Genome-wide distribution of test statistics from large GWAS deviate strongly from the null



PGC SCZ, Nature, 2014

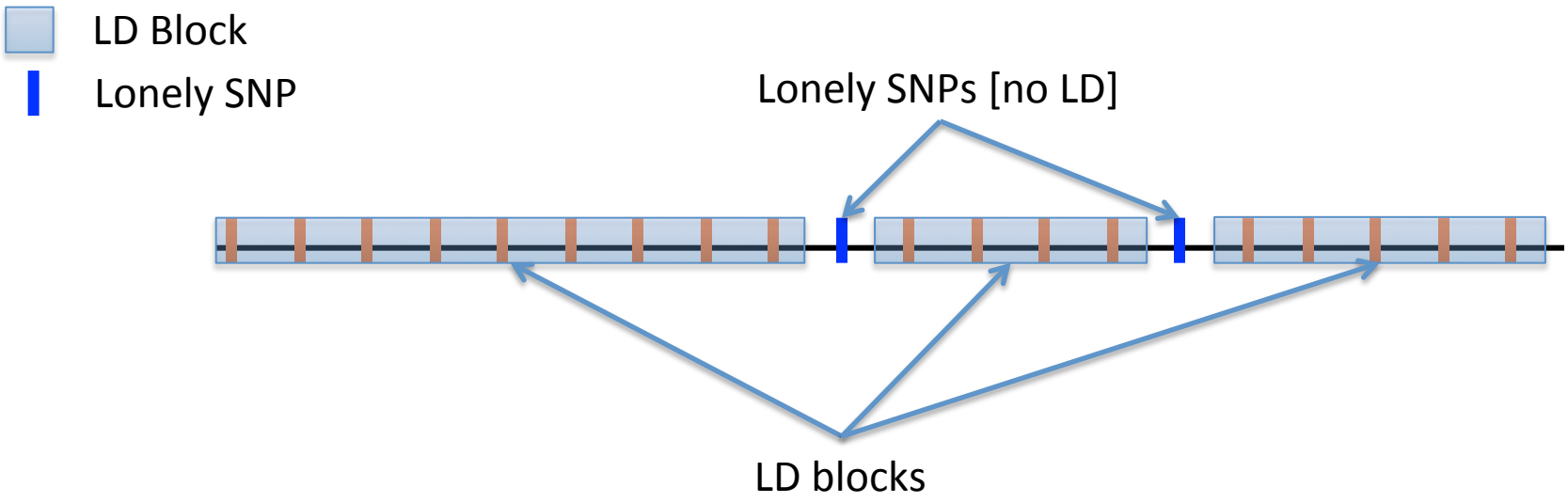
Test Statistic Inflation

Even when all gwas loci (+/- 1 MB, 10MB for MHC) removed

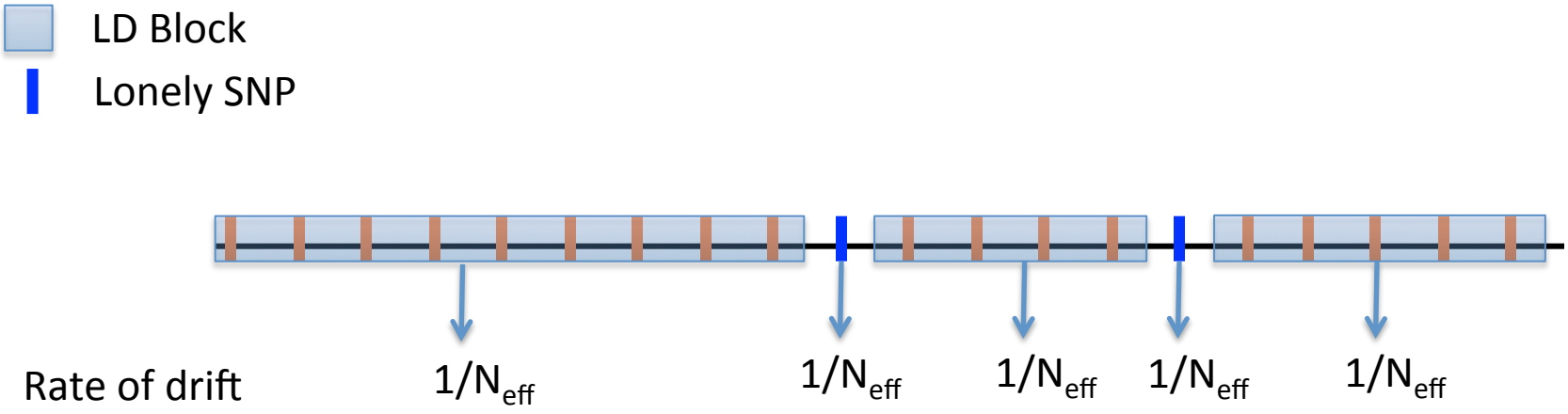


PGC SCZ, Nature, 2014

Toy Illustration of Genome



What happens under genetic drift?



Under pure drift, LD is uncorrelated to magnitude of allele frequency differences between populations

Simulation of a genetic signal in polygenic architecture

LD Block

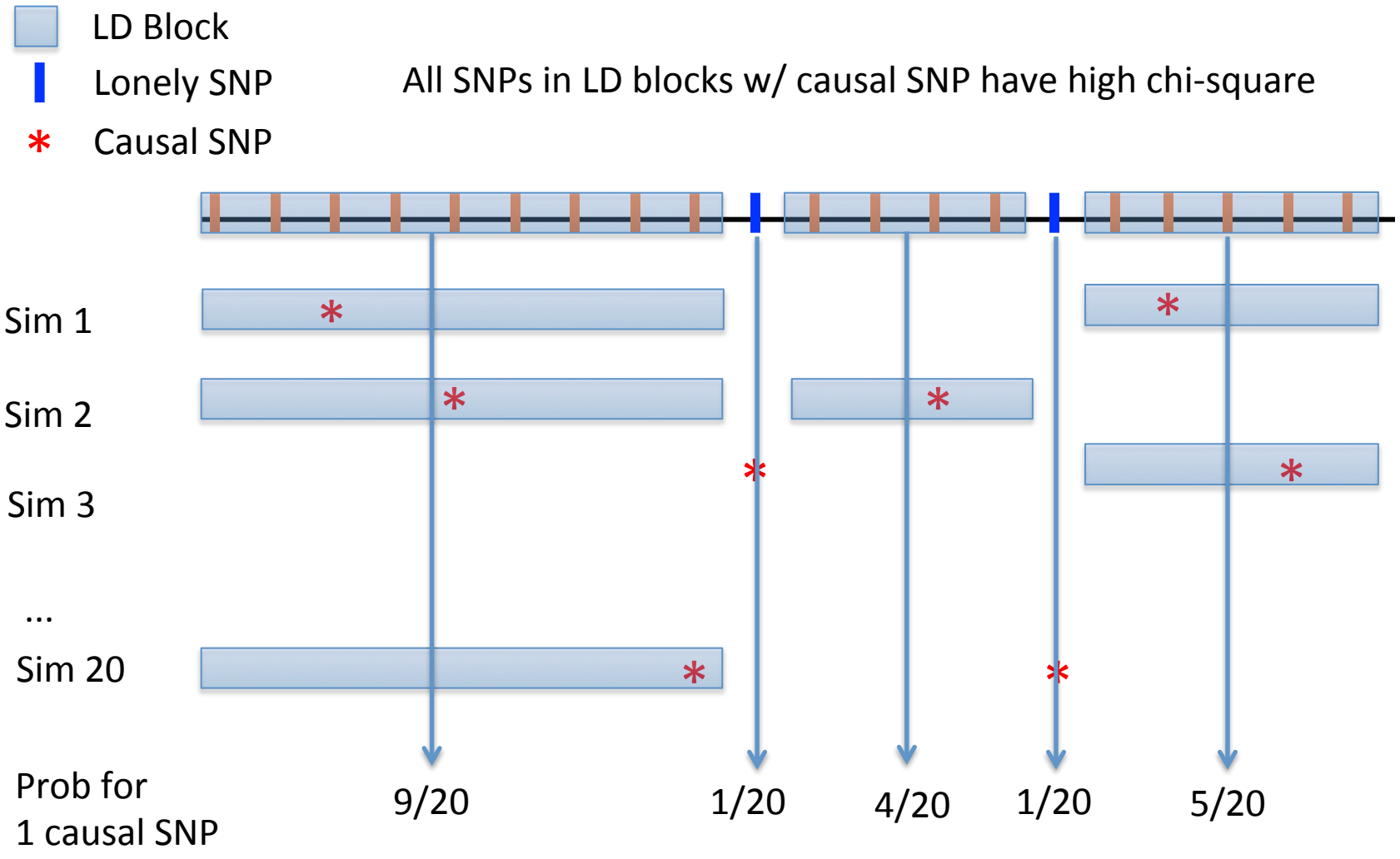
Lonely SNP

* Causal SNP

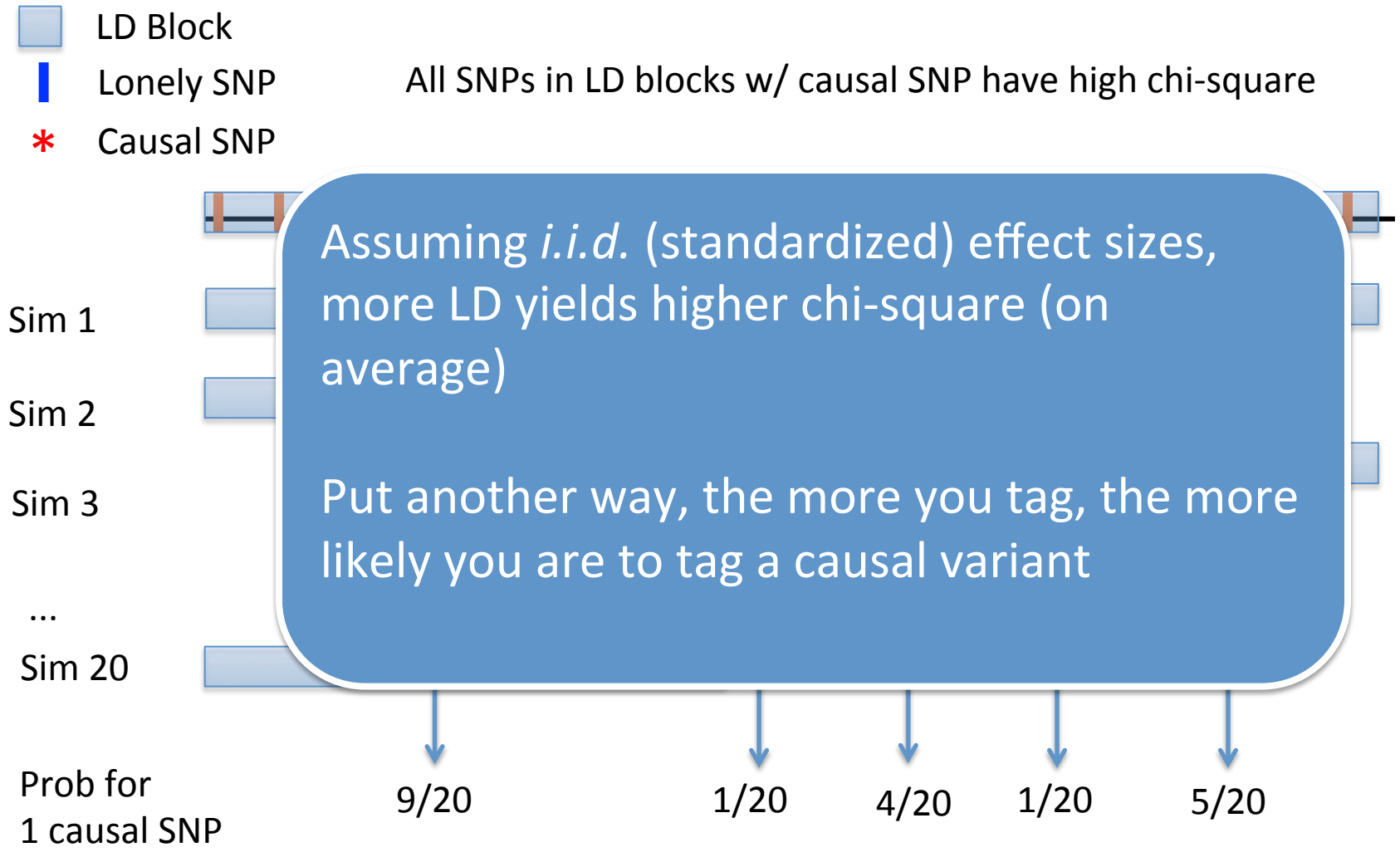
All SNPs in LD blocks w/ causal SNP have high chi-square

Sim 1

Simulation of a genetic signal in polygenic architecture

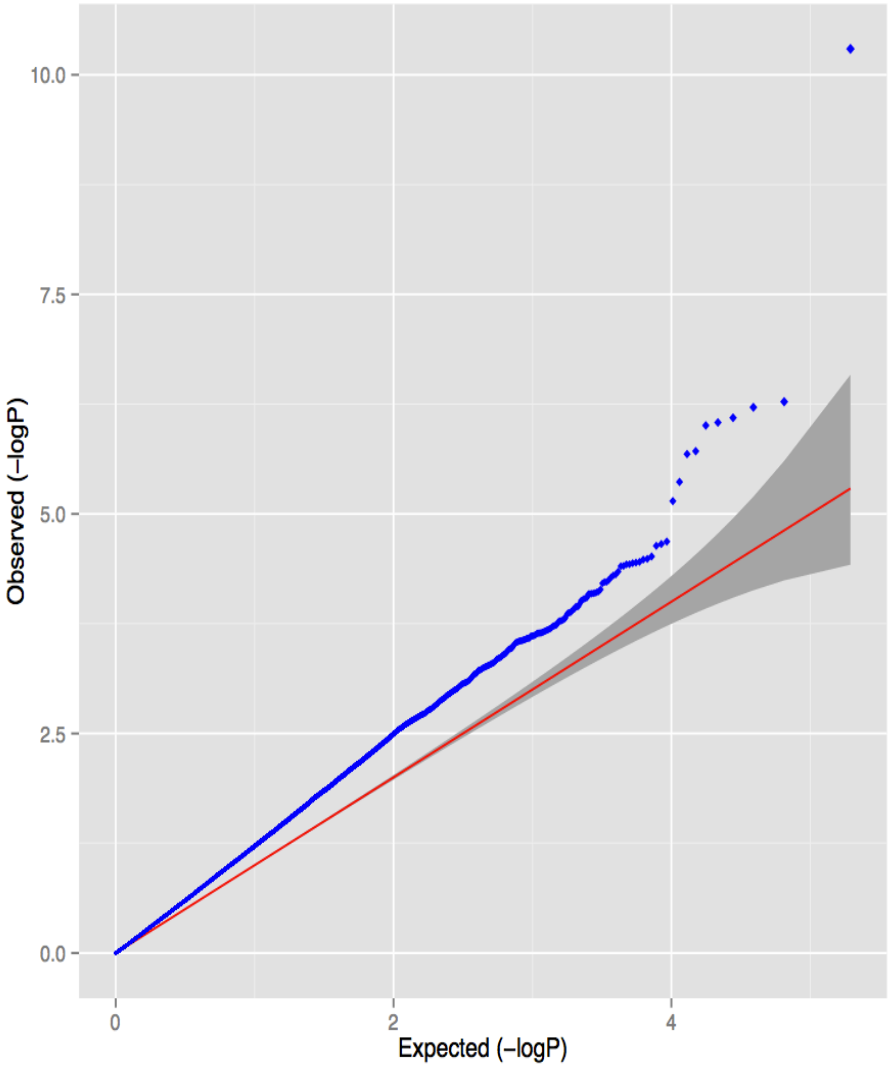


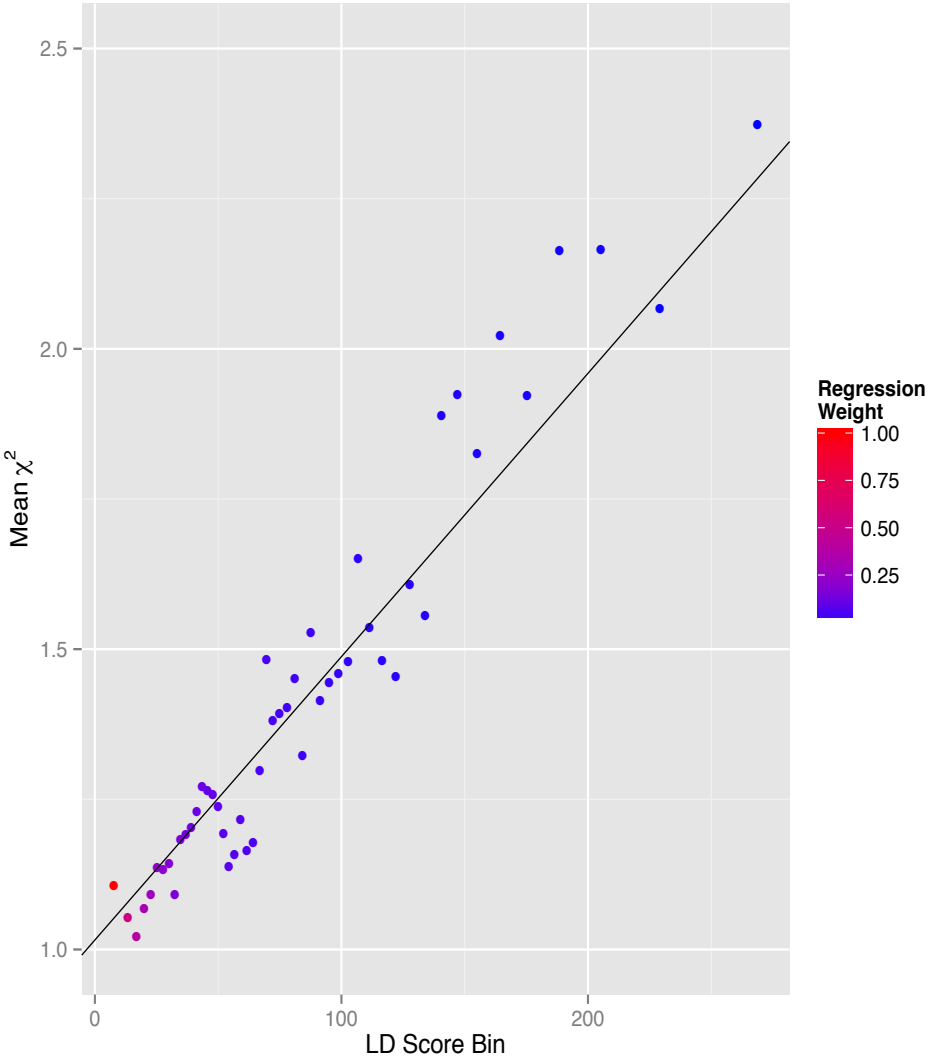
Simulation of a genetic signal in polygenic architecture



Simulated Polygenicity

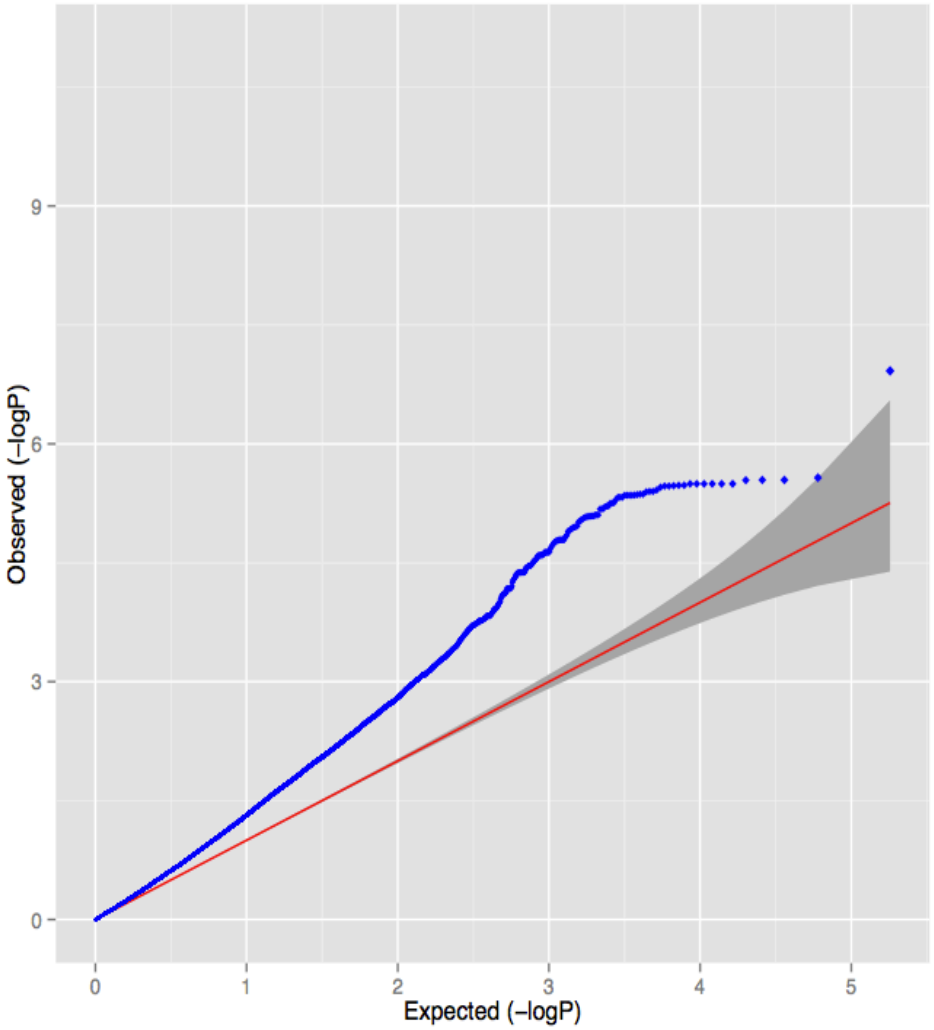
- λ_{GC} = 1.30; LD Score Regression intercept = 1.02

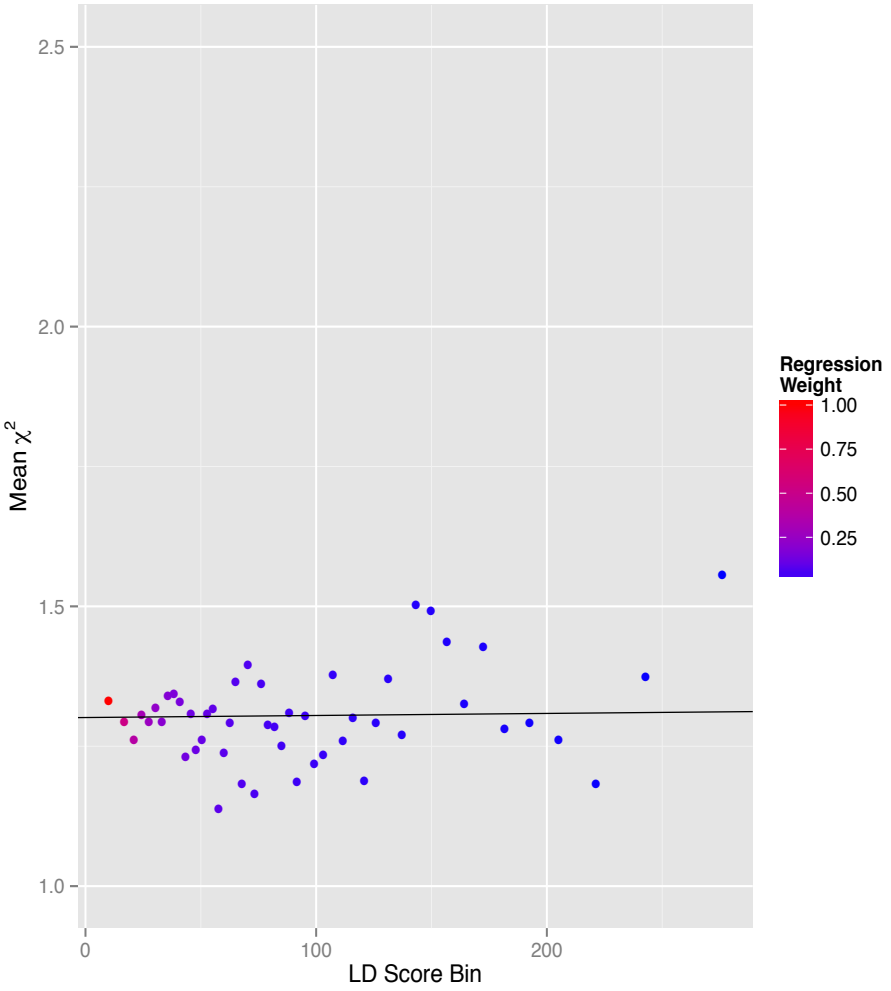




Simulated Pop Strat (Sweden vs UK)

- λ_{GC} = 1.30; LD Score Regression intercept = 1.32

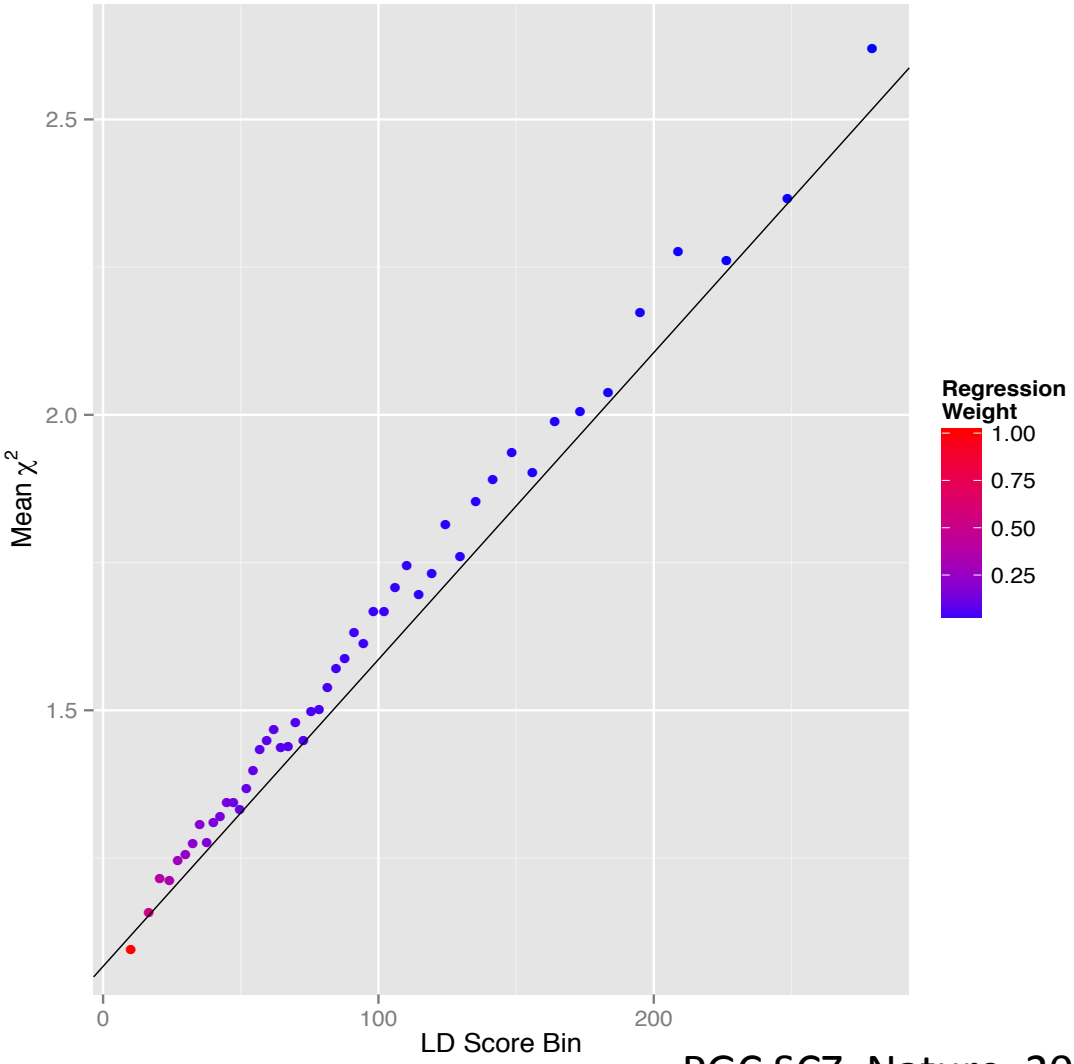




PGC Schizophrenia

- λ_{GC} = 1.48
- Intercept = 1.06
- *p*-value < 10^{-300}

Overwhelming majority of inflation is consistent with polygenic architecture



Bulik-Sullivan et al., Nat Genet, 2015

PGC SCZ, Nature, 2014

LD Score Regression

- Regress χ2 statistics against LD Score

$$E[\chi^2 \mid \ell_j] = Nh^2\ell_j / M + Na + 1$$

  - LD Score (L_j) is a property of SNP j, defined as sum r^2, estimated as sum r^2 w/ all other SNPs a 1cM window.
  - N is sample size.
  - M is # SNPs.
  - h^2 is SNP-heritability.
  - a is inflation from pop strat/cryptic relatedness.

Bulik-Sullivan et al., Nat Genet, 2015

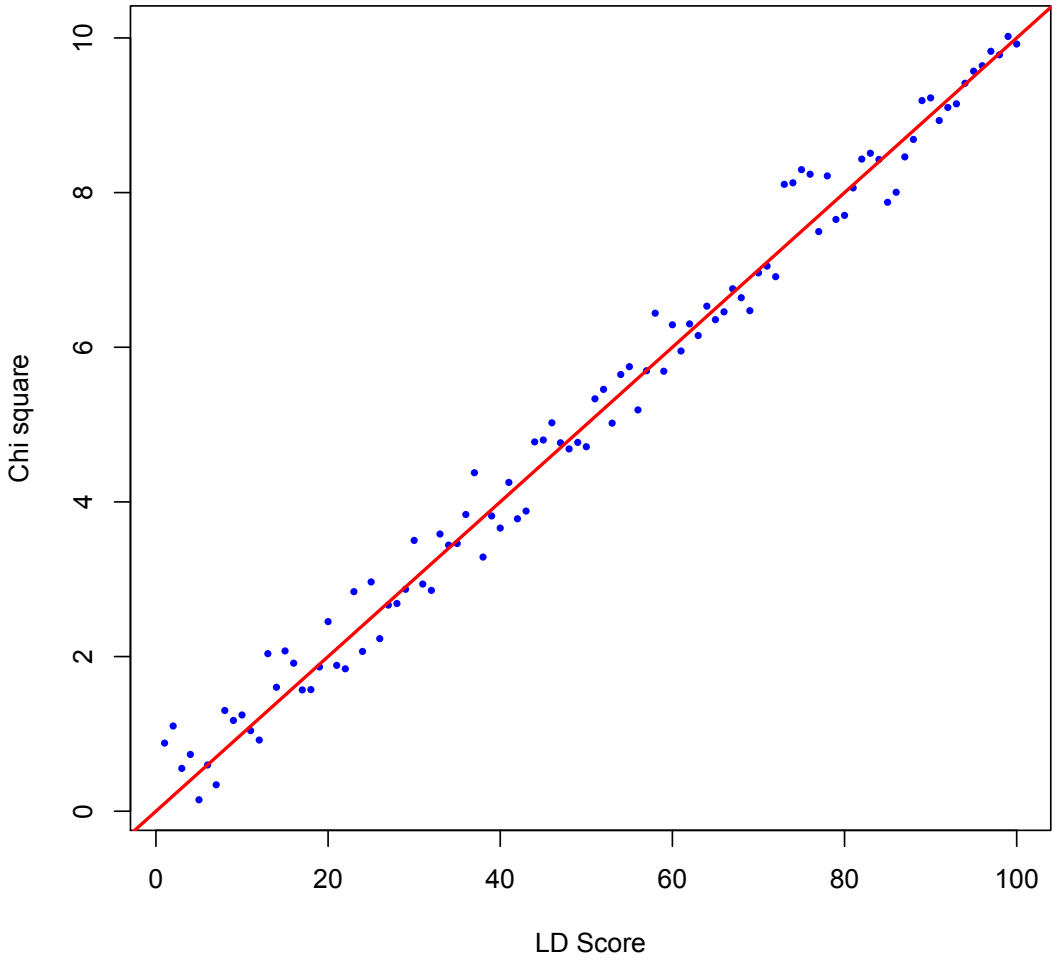
LD Score Results

- Applied to > 20 GWAS
  - Almost all inflation due to polygenicity.
  - LD Score intercept < λ_{GC} in all studies.
- Conclusions:
  - PCA / mixed models mostly appear to work.
  - Genomic control (dividing all $\chi 2$ statistics by λ_{GC}) is unnecessarily conservative.

Bulik-Sullivan et al., Nat Genet, 2015

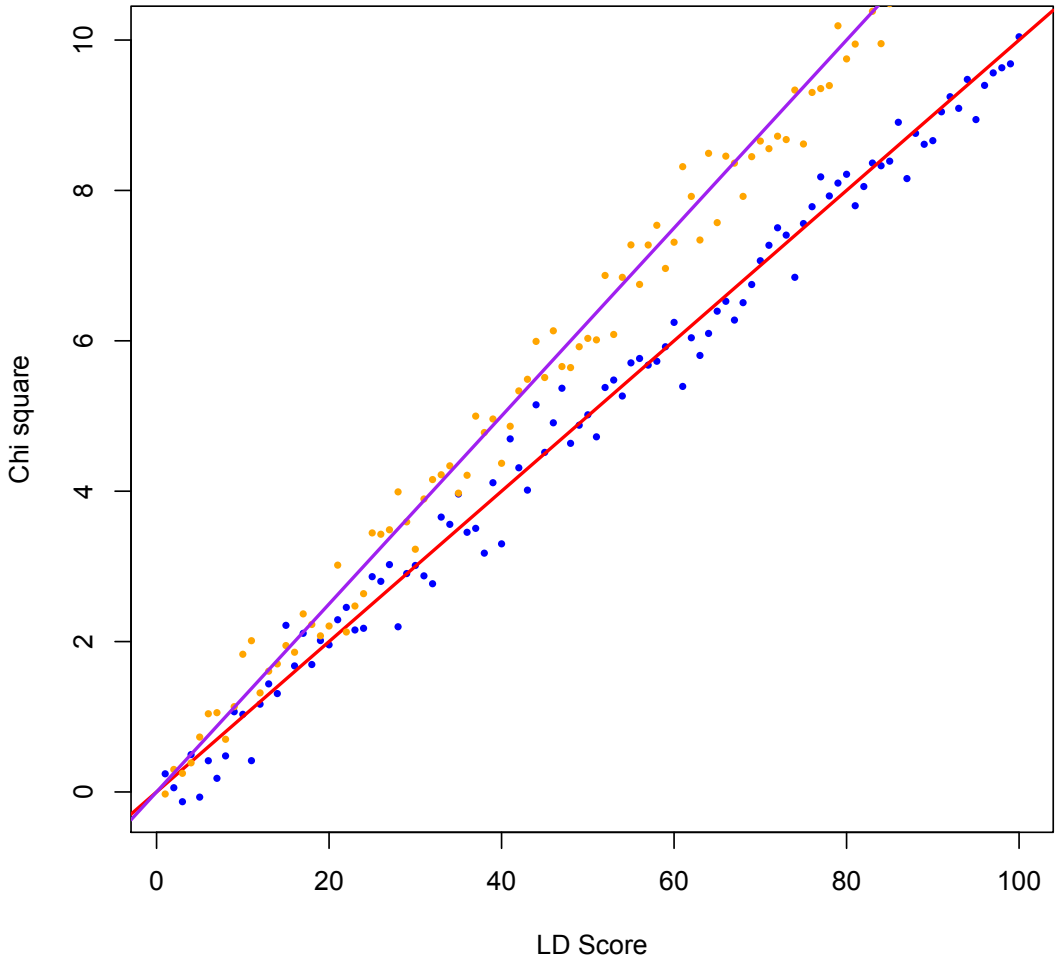
Genetic correlation

Reminder - univariate:

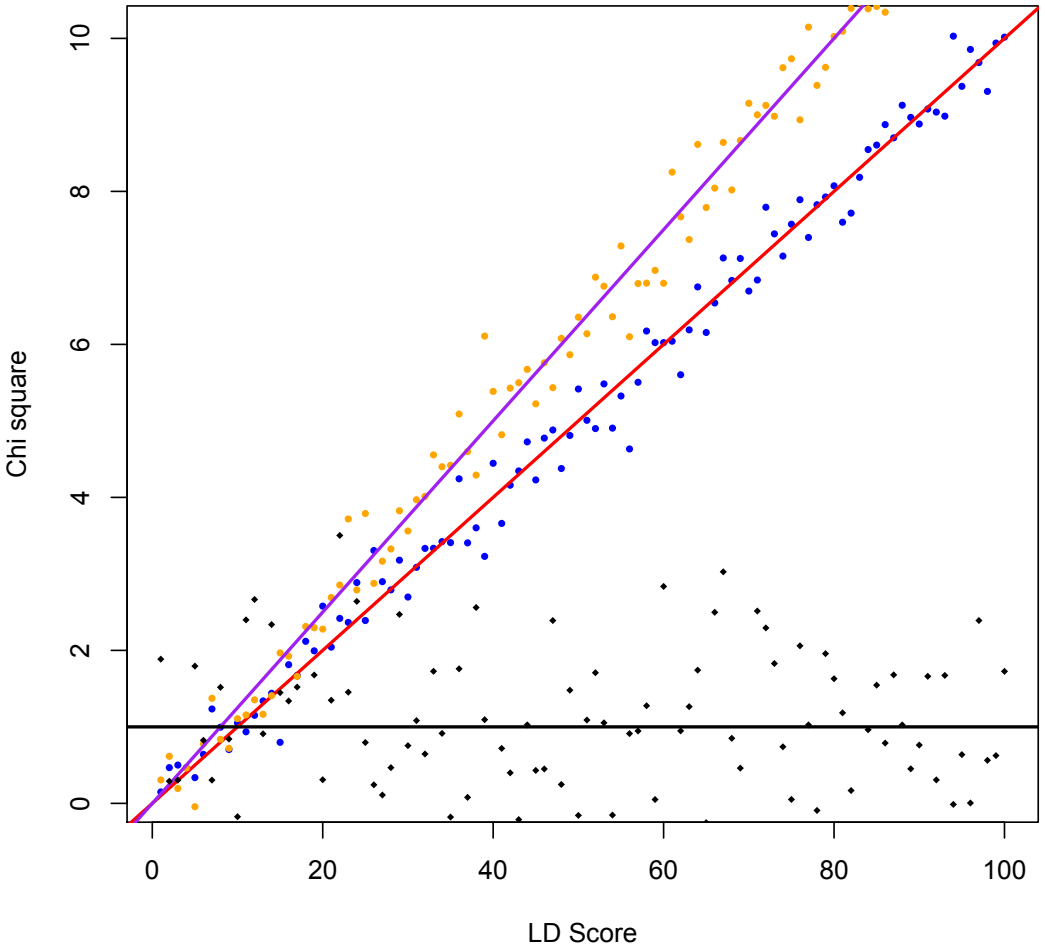


The slope of this regression line is an estimator of heritability

Add a second trait:

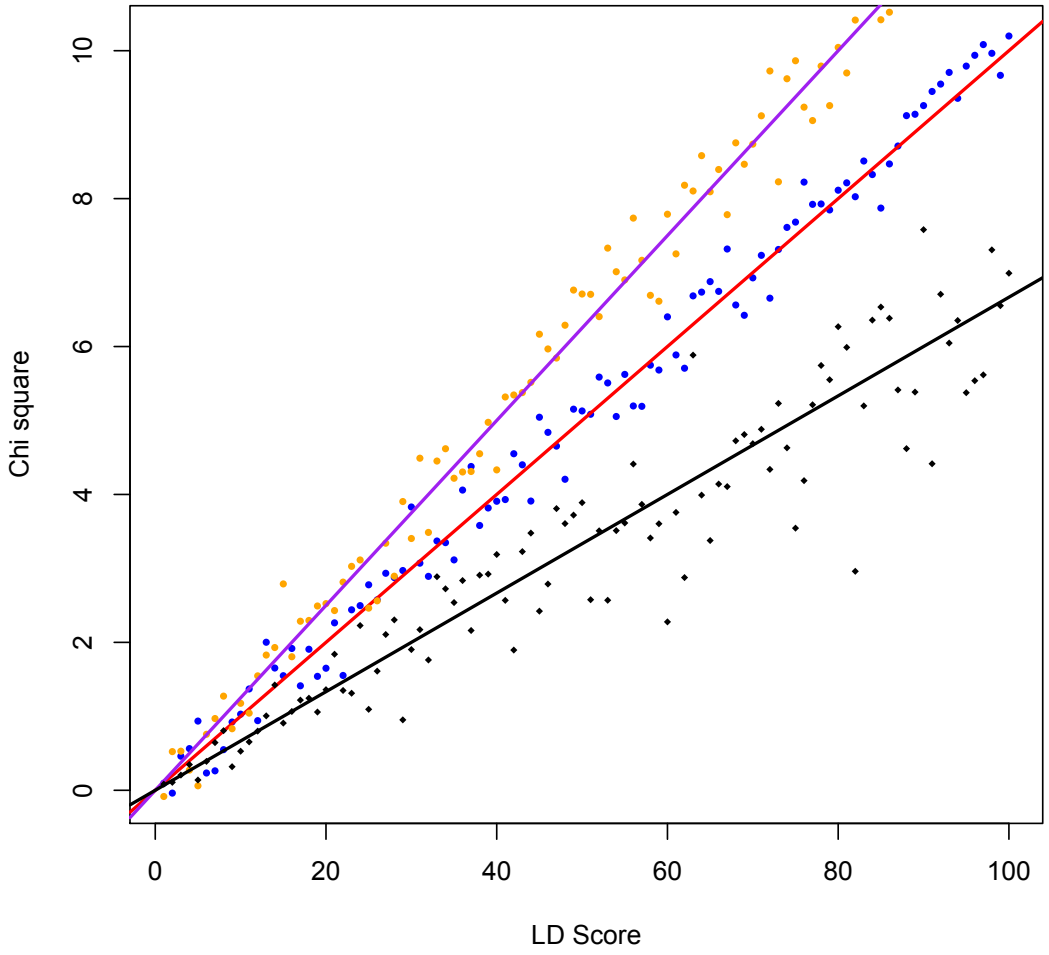


Genetic correlation = 0



Recall that $\chi 2 = Z^2$;
to estimate r_g, replace
$\chi 2$ with $Z_1 Z_2$.

Genetic correlation of ~0.5



The signed positive slope shows that genetic effects tend to be shared genome-wide

Formally

$$\mathbb{E}[z_{1j}z_{2j}] = \frac{\sqrt{N_1 N_2}\rho_g}{M}\ell_j + \frac{\rho N_s}{\sqrt{N_1 N_2}}$$

where N_1 and N_2 are the sample sizes for the two studies
p_g is the genetic correlation
l_j is the LD score
M is the total number of markers
p is the phenotypic correlation
N_s is the number of overlapping samples

Key point: *not biased by sample overlap*

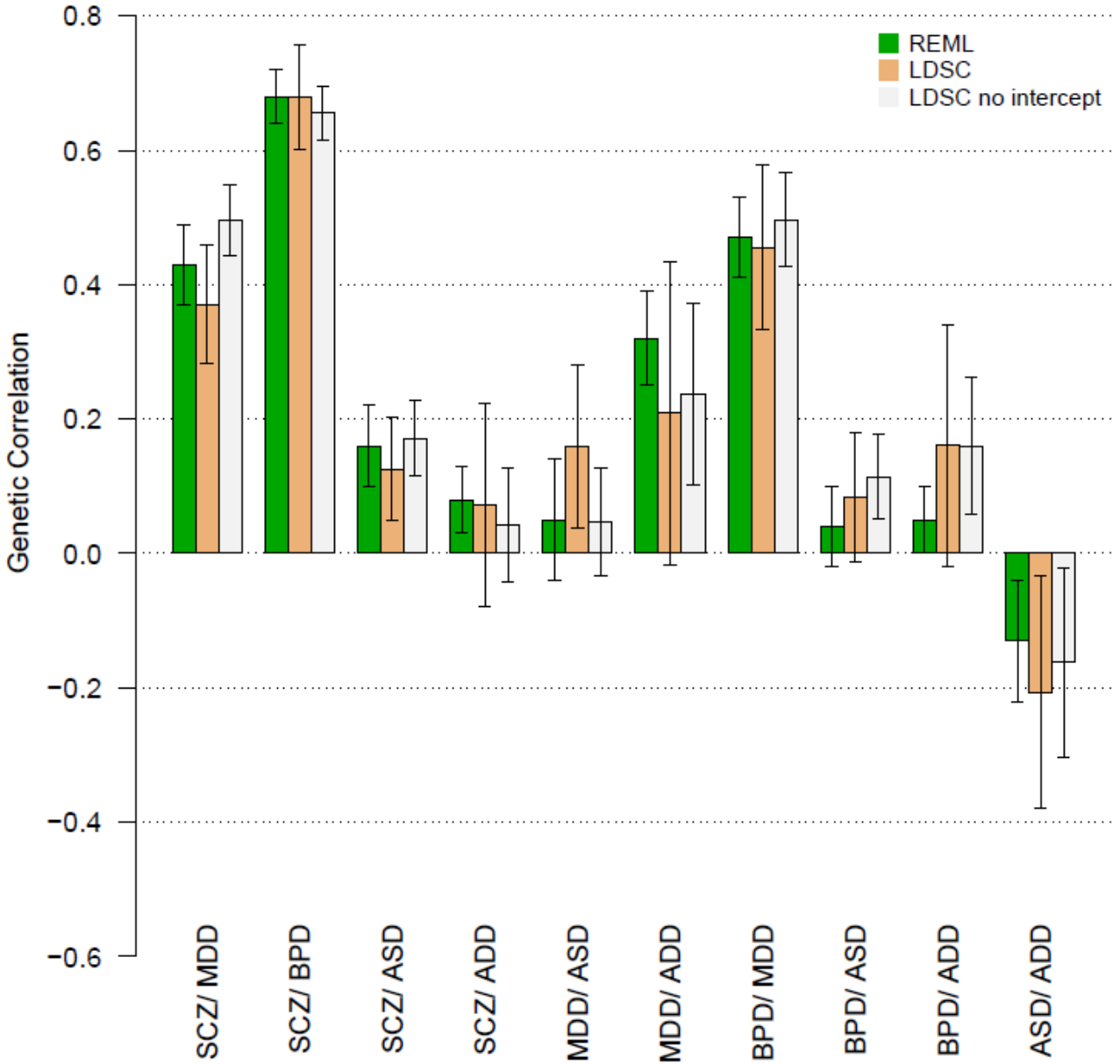
Bulik-Sullivan et al, bioRxiv

Proof of concept

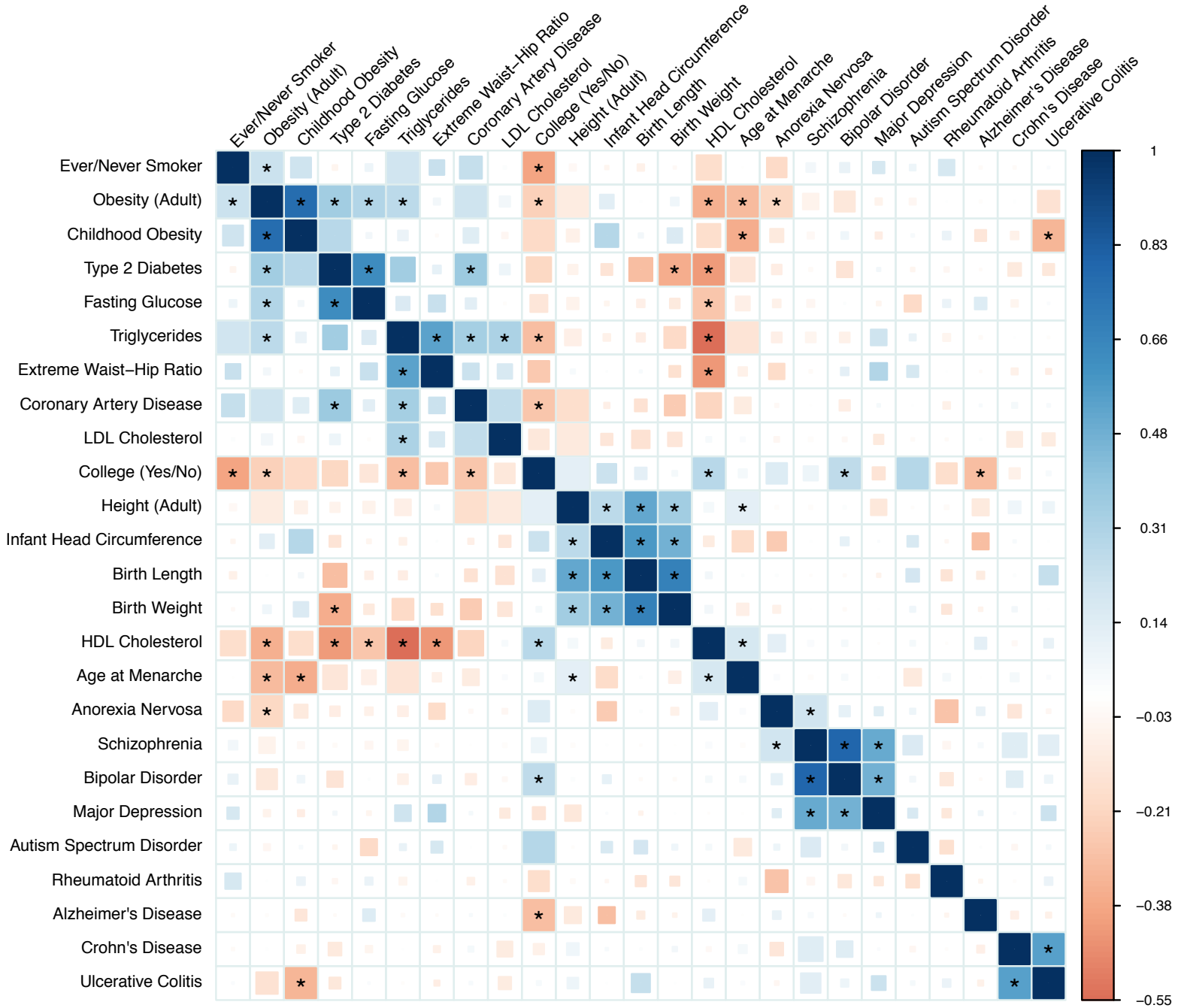
Supplementary Table 1. Bivariate analyses

	Trait 1/ Trait 2				
	SCZ/BPD	**SCZ/MDD**	**SCZ/ASD**	**SCZ/ADHD**	**BPD/MDD**
SNPs	909307	885448	896627	778235	938610
Cases	9032/6664	9051/8998	9111/3226	9013/4108	6665/8997
Controls	7980/5258	10385/7823	12146/3308	10115/9936	7408/7680
SNP-h^2 Trait 1[a]	0.22 (0.01)	0.21 (0.01)	0.23 (0.01)	0.23 (0.01)	0.23 (0.01)
SNP-h^2 Trait 2[a]	0.22 (0.01)	0.19 (0.02)	0.16 (0.02)	0.23 (0.02)	0.20 (0.02)
Covariance[b]	0.151 (0.010)	0.087 (0.011)	0.030 (0.011)	0.019 (0.011)	0.102 (0.013)
SNP-r_g (SE)	0.68 (0.04)	0.43 (0.06)	0.16 (0.06)	0.08 (0.05)	0.47 (0.06)
λ_{1st}-cov(SE)	1.7 (0.05)	1.2 (0.05)	1.2 (0.03)	1.1 (0.03)	1.2 (0.00)
λ_{1st}-r_g	4.7	1.6	1.5	1.2	1.6
p[c]	**<e-16**	**6.0e-15**	**0.0071**	0.072	**1.5e-14**
literature[d] **λ_{1st}**	M-A: 2.1[1], Offspring[2,e]: 2.4,5.2,4.5,6.0 Sib[2,e]: 3.9,3.7,3.9,5.0	M-A[f]: 1.5	Parent[3]: 2.9 Sibling[3]: 2.6	Parent[4,g]: > 1	M-A[5,h]: 3.1,2.7
			Sibling (ASD/ADHD)[6]: 2.4		
literature r_g	0.60[2,i]	N/A	N/A	N/A	0.65[7,j]

Lee et al, 2013 Nat Gen

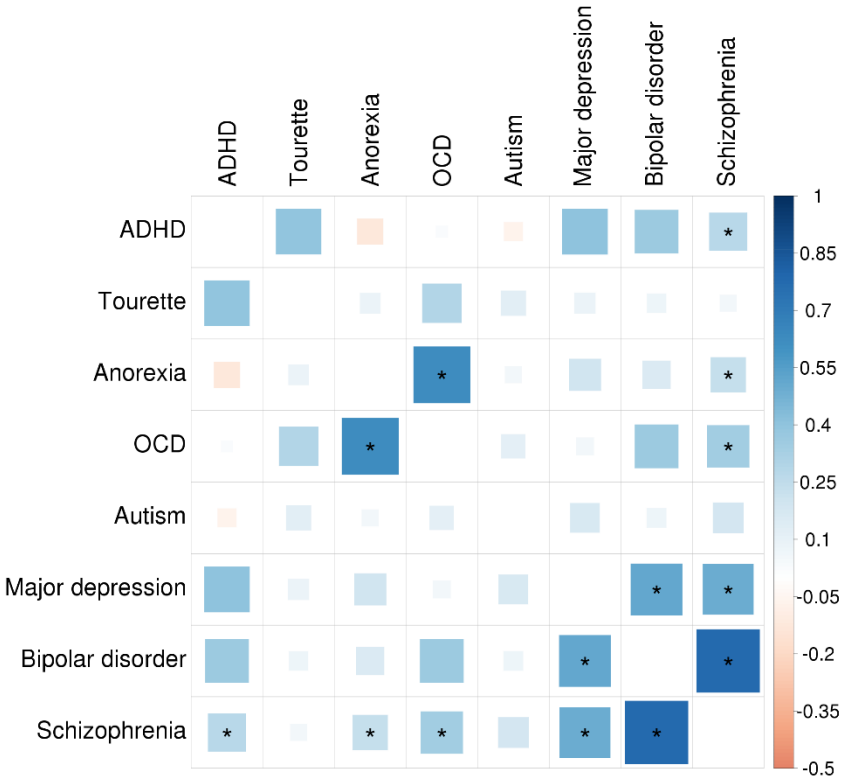


Bulik-Sullivan et al, bioRxiv



Bulik-Sullivan et al, bioRxiv

New Psychiatric r_g



In addition:
+20% rg between AN and BMI

V. Anttila, Brainstorm Project

Pause for questions …

What can LD Score do for *you?*

Practical advice on using LD Score in day-to-day GWAS analysis

Software

- LD Score regression implemented in free + open-source python command-line tool ldsc:
  - github.com/bulik/ldsc
- Tutorials & FAQ here:
  - github.com/bulik/ldsc/wiki
- Ask me questions on the google group!

LD Score is Fast and Easy

- Trivial run-time & memory (~15s, ~1GB for h^2).
- Automated data re-formatting and QC.
  - munge_sumstats.py included w/ ldsc.
  - No need for one-off perl scripts.
- Download pre-computed LD Scores.
  - broadinstitute.org/~bulik/eur_ldscores/
  - (European-only, for now)

Example: Estimating r_g(BIP, SCZ)

Automatically applies same MAF/INFO etc filters used in our papers + various sanity checks (e.g., log odds in OR column?)

```
python munge_sumstats.py
        --sumstats pgc.cross.SCZ17.2013-05.txt
        --N 17115
        --out scz
        --merge-alleles w_hm3.snplist
```

```
python munge_sumstats.py
        --sumstats pgc.cross.BIP11.2013-05.txt
        --N 11810
        --out bip
        --merge-alleles w_hm3.snplist
```

Automatically aligns strand + ref allele + filters out strand ambiguous SNPs

```
python ldsc.py
        --rg scz.sumstats.gz,bip.sumstats.gz
        --ref-ld-chr eur_w_ld_chr/
        --w-ld-chr eur_w_ld_chr/
        --out scz_bip
```

45 seconds on my MacBook Air

Basic QC with LD Score intercept

- QC Question: *have we adequately controlled for confounding from population stratification?*
- Solution: check LD Score intercept close to 1.
  - Caveat: only sensitive to sources of genome-wide inflation; can't tell you whether 10 suspect SNPs are OK.

QC with LD Score h^2

- QC Question: *do we see more or less inflation than we would expect given N and h^2?*
- Low inflation can mean phenotype problems.
  - Non-screened controls.
  - Bad phenotype def'n.
  - Data munging error, e.g., column swap in ped file.
- Solution: compare h^2(old data), h^2(new data).
  - Big + significant differences may indicate problems.

QC with LD Score r_g

- QC Question: *does phenotype definition in new data match older data?*
  - Coordinating pheno def'n across studies is hard.
  - Data munging error, e.g., column swap in ped file.
- Solution: compute r_g(new data, old data)
  - Particularly useful for summary-statistic meta-analysis consortia.

Streamlined PRS

- Statements about prediction R^2 from PRS analysis are often equivalent to statements about h^2 or r_g:
- PRS for X predicts[1] Y ***if and only if*** $r_g(X, Y)$!= 0.
- PRS for X predicts[1] X ***if and only if*** $h^2(X) > 0$.

[1]In independent samples

Dudbridge, PLoS Gen, 2013

Streamlined PRS

- LD Score r_g/h^2 often faster/easier than PRS
  - No LD pruning.
  - No individual-level genotype data.
  - Don't have to worry about sample overlap.
  - Don't have to split sample into train/test sets.
  - Caveat: GCTA and PRS have (slightly) better power than LD Score, possibly makes a big difference for small N.

Practical Advice

- LD Score is noisy at small N.
  - Rule of thumb: use GCTA for N < 3k.
- Partitioned h^2 requires very large N.
  - Rule of thumb: not worth trying for < 5k cases.

Practical Advice

- ldsc not presently applicable to admixed data.
  - LD structure in admixed samples is more complex.
- If no pop strat / no sample overlap, constrained intercept LD Score has lower SE
  - Equivalent to Haseman-Elston regression (Bulik-Sullivan, bioRxiv, 2015)

Notes for PGC users

- munge_sumstats.py --daner flag processes Ricopili-format data (daner* files)
- ldsc.py --samp-prev and --pop-prev flags convert to liability-scale h^2

Acknowledgements

- r_g + functional h^2 + ldsc joint
  work w/ Hilary Finucane
- Ben Neale
- Alkes Price
- Nick Patterson
- Po-Ru Loh
- Mark Daly
- Many others …

URLs

- ldsc
  - github.com/bulik/ldsc
  - Installation instructions
  - FAQ
- Tutorials / wiki
  - github.com/bulik/ldsc/wiki
- Pre-computed European LD Scores
  - broadinstitute.org/~bulik/eur_ldscores/
- ldsc_users google group:
  - groups.google.com/forum/?hl=en#!forum/ldsc_users

LD Score Papers

- LD Score regression distinguishes confounding from polygenicity in genome-wide association studies
- Partitioning heritability by functional category using GWAS summary statistics
- An Atlas of Genetic Correlations across Human Diseases and Traits
- Relationship between LD Score and Haseman-Elston Regression