Contrasting regional architectures of schizophrenia and other complex diseases using fast variance components analysis

10/7/15: ASHG 2015

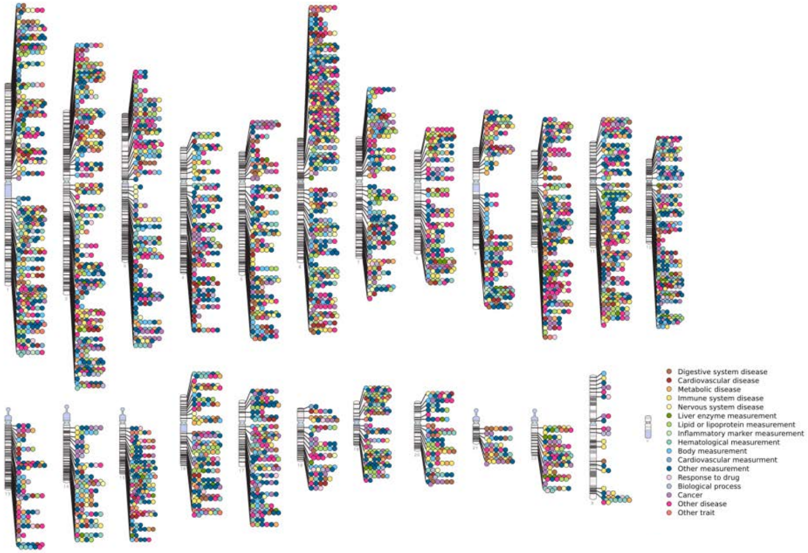
Po-Ru Loh
Harvard T.H. Chan School of Public Health

Heritability in the GWAS era:
Much is known – but much more is unknown

- GWAS have found thousands of associations between genes and traits…
- … but GWAS hits explain only a fraction of known heritability *Maher 2008 Nature*

The case of the missing heritability

Published Genome-Wide Associations through 12/2013
Published GWA at p≤5X10^{-8} for 17 trait categories

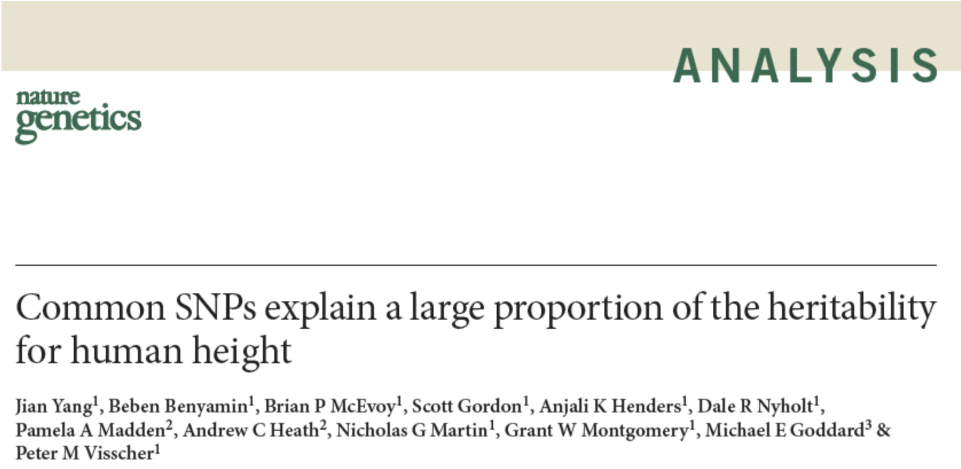


NHGRI GWAS catalog:
www.genome.gov/gwastudies/

Heritability in the GWAS era: How much is explained by genotyped SNPs?

- We now know that ***all*** genotyped SNPs ***together*** explain a large fraction of trait variance: $\boldsymbol{h_g^2}$
  - Note $h_g^2 < h^2$ (narrow-sense heritability); see ***ASHG 2015 platform talk 196, Bhatia***

Yang et al. 2010 Nat Genet

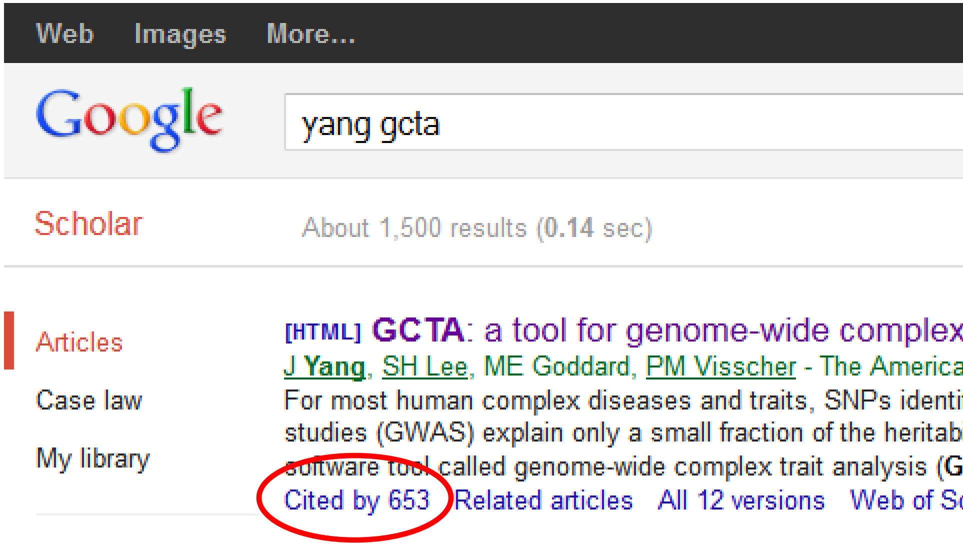

ANALYSIS
nature genetics

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

- Method: Variance components analysis (a.k.a. REML)
- GCTA software implementation is now widely used in genetics

Yang et al. 2011 AJHG



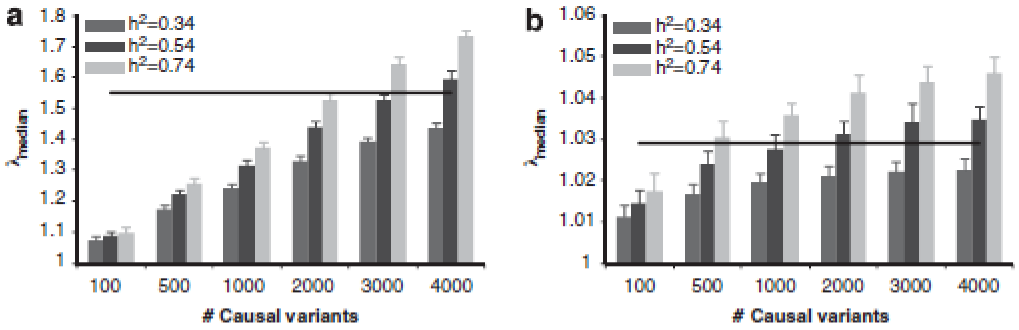
yang gcta

About 1,500 results (0.14 sec)

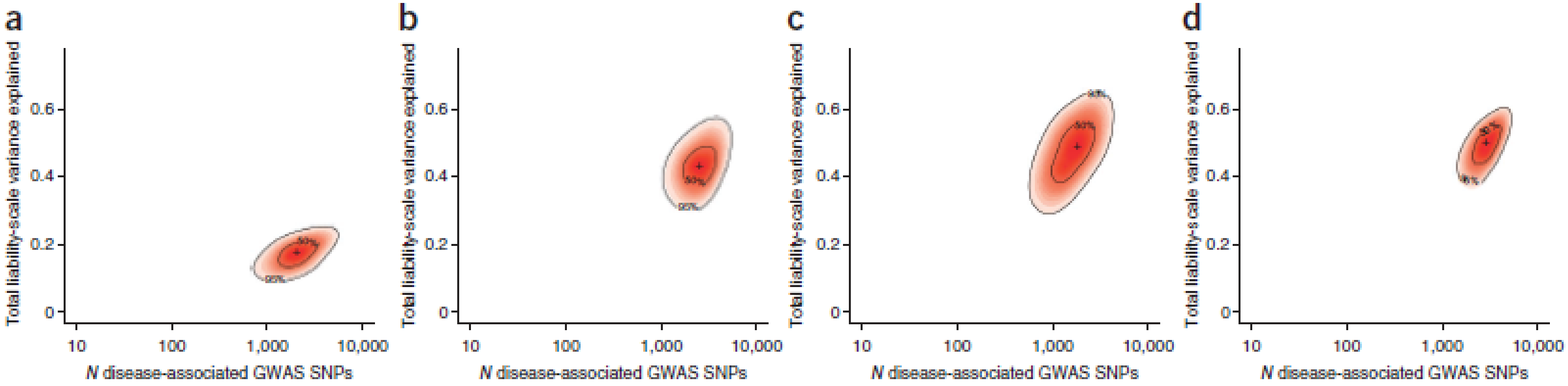
[HTML] GCTA: a tool for genome-wide complex
J Yang, SH Lee, ME Goddard, PM Visscher - The America
For most human complex diseases and traits, SNPs identi
studies (GWAS) explain only a small fraction of the heritab
software tool called genome-wide complex trait analysis (G
Cited by 653 Related articles All 12 versions Web of Sc

Beyond h_g^2: How many SNPs are causal?

- We know there are lots of causal SNPs explaining h_g^2 of the variance
- We still don't have power to find all the causal SNPs
- Can we say how many there are?

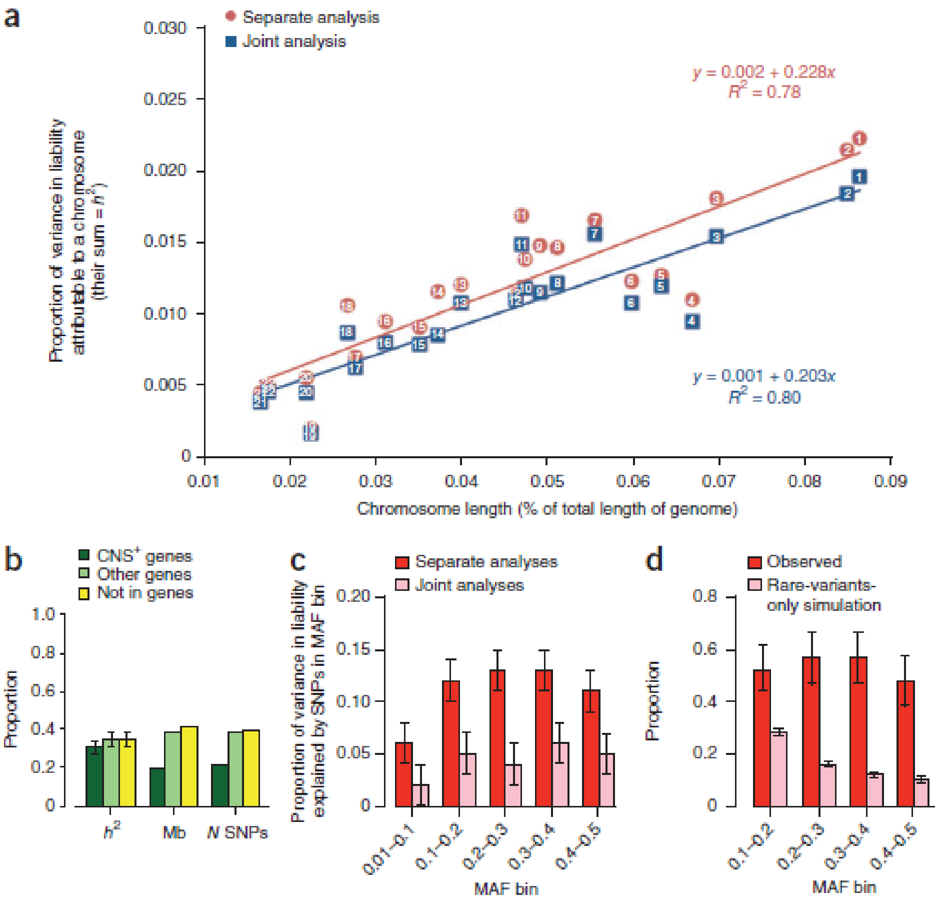


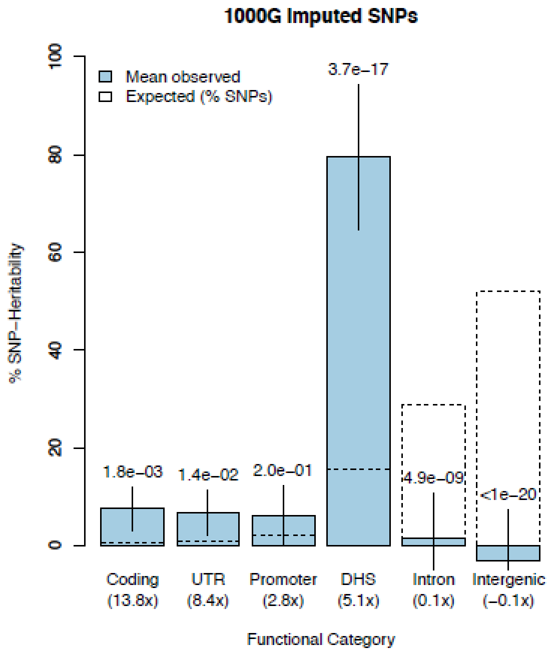
Yang et al. 2011 EJHG
Stahl et al. 2012 Nat Genet
Palla & Dudbridge 2015 AJHG



Beyond h_g^2: How is h_g^2 distributed across genomic elements?

- Partitioning heritability…
  - By chromosome
  - By MAF bin
  - By functional annotation





Lee et al. 2012 Nat Genet

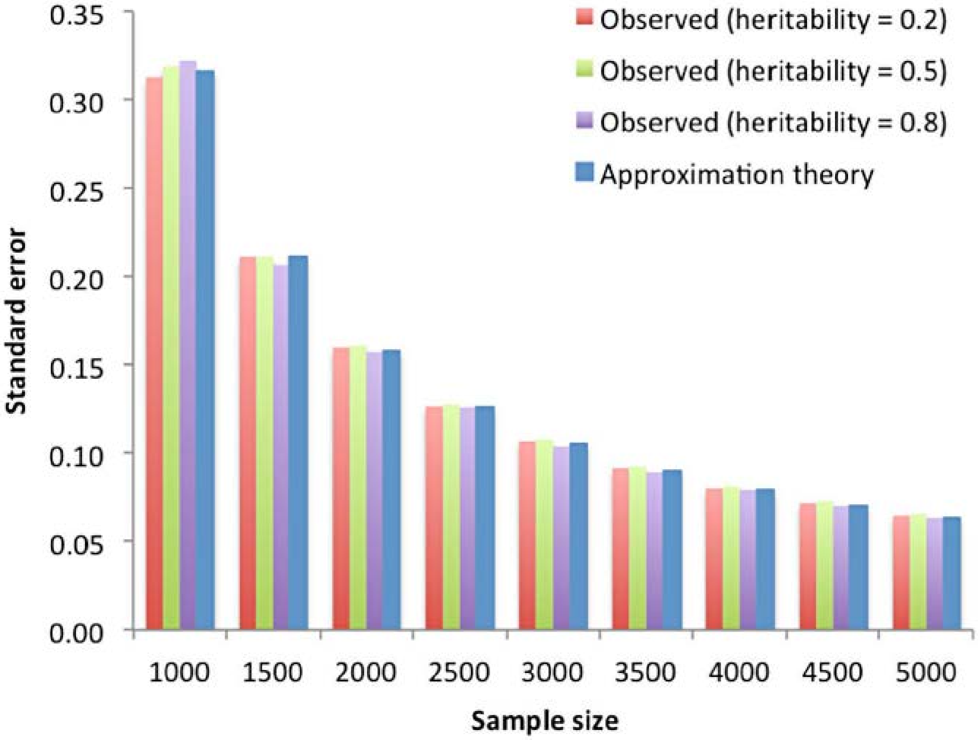
Gusev et al. 2014 AJHG

Finucane, Bulik-Sullivan* et al. 2015 Nat Genet*

Larger sample sizes are required to obtain further insights into h_g^2

- At a sample size of *N*=5000, h_g^2 estimates have standard errors of ≈0.06
  - Too large for precise inference

- **Problem**: For sample sizes above *N*=50K, standard variance components analysis is computationally intractable

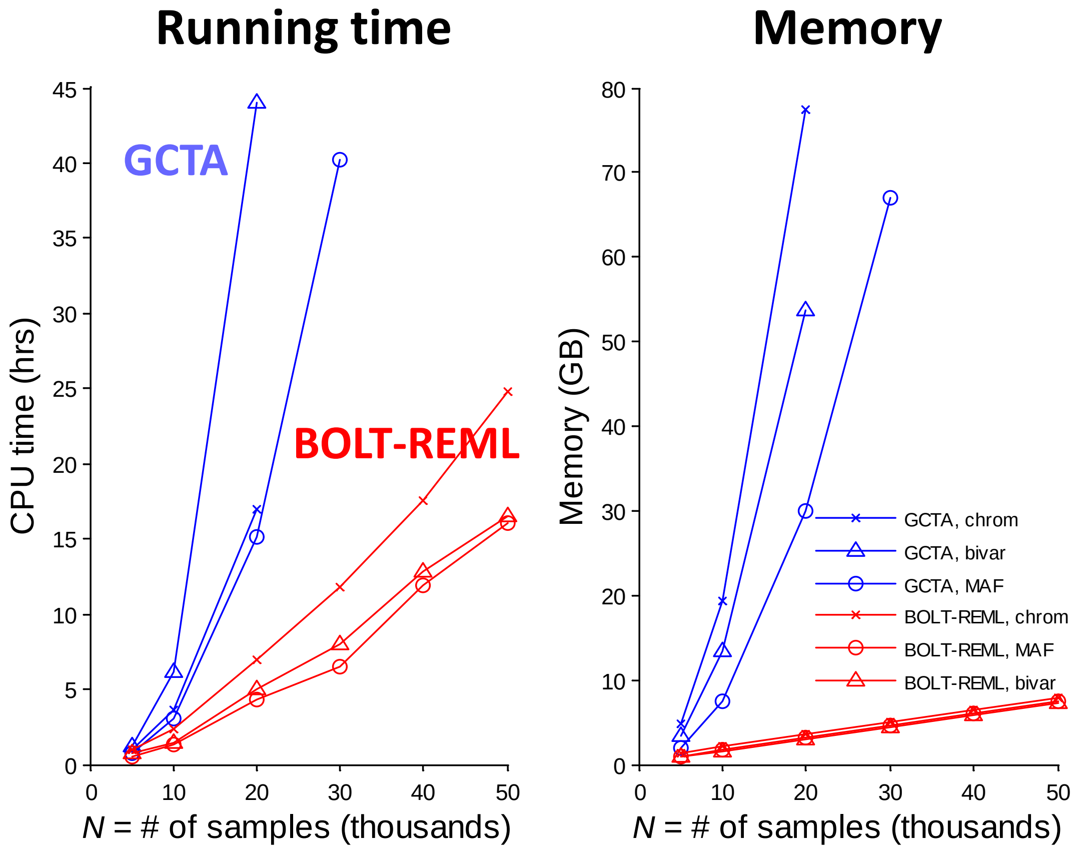


Visscher et al. 2014 PLOS Genet
Visscher & Goddard 2015 Genetics

New fast algorithm (BOLT-REML) performs h_g^2 analyses on *N*>50,000 samples

- Performs REML heritability parameter estimation
  - Multiple var. comps.: Partitioned h_g^2
  - Multiple phenotypes: Genetic correlation
- $\approx O(MN^{1.5})$ time, *MN*/4 memory (*M* = # SNPs) as in BOLT-LMM association analysis

  Loh et al. 2015a Nat Genet
- Much more efficient than GCTA at high *N*



Running time

Memory

Loh et al. 2015b Nat Genet (in press; bioRxiv)

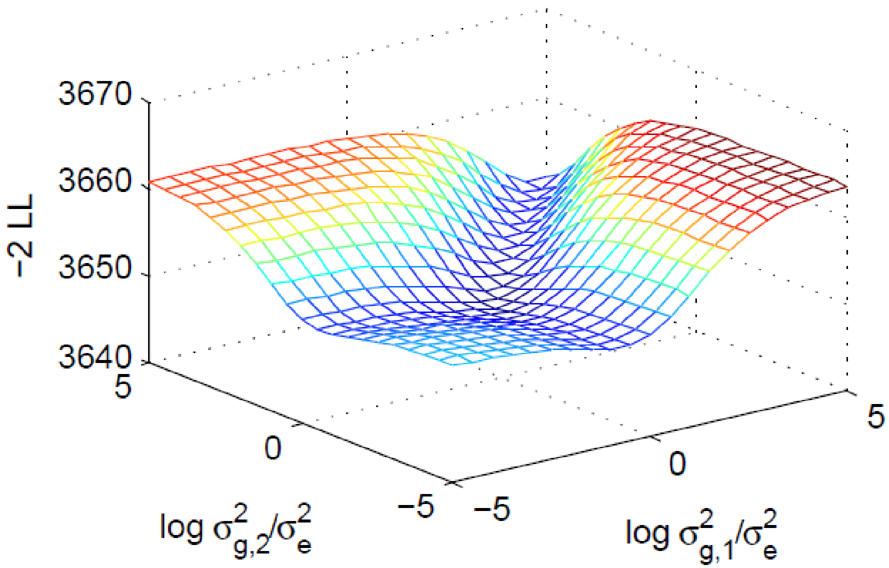
BOLT-REML algorithm

- Rapidly approximate gradient and Hessian of likelihood surface
  - Monte Carlo approximation => no need for $O(N^3)$-time matrix operations

    Garcia-Cortes et al. 1992 JABG

    Matilainen et al. 2013 PLOS ONE

  - Instead, just solve linear systems with $O(MN)$-time conjugate gradient iterations
- Ensure robustness using trust region optimization



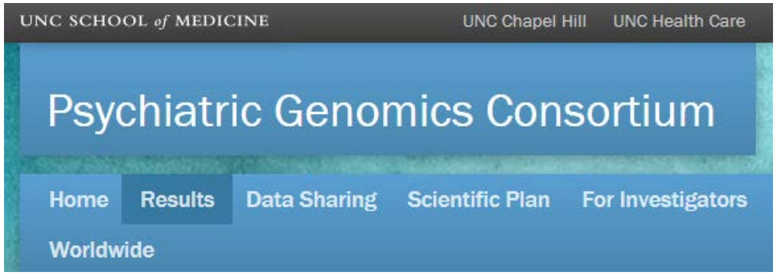
$$\overline{\frac{\partial \ell}{\partial \sigma_k^2}} = -\frac{1}{2}\left(\overline{y_V' V^{-1} Z_k Z_k' V^{-1} y_V} - y' V^{-1} Z_k Z_k' V^{-1} y\right)$$

$$\frac{\partial^2 \ell}{\partial \sigma_k^2 \partial \sigma_j^2} \approx -\frac{1}{2} y' V^{-1} Z_k Z_k' V^{-1} Z_j Z_j' V^{-1} y = \mathscr{I}_A$$

Application: two *N*>50K data sets

Psychiatric Genomics Consortium (PGC2)

- Largest schizophrenia data set ever collected
- Data size (after QC):
  - 22K schizophrenia cases + 28K controls (across 29 cohorts)
  - 472K well-imputed SNPs

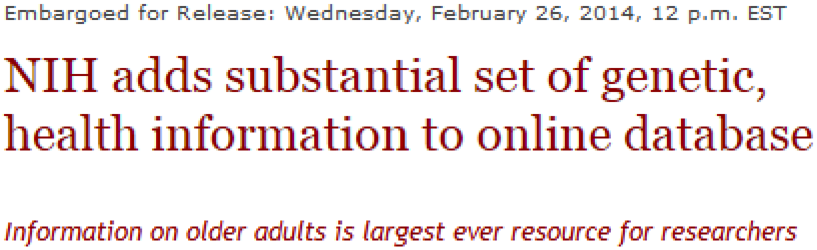


UNC SCHOOL *of* MEDICINE

Psychiatric Genomics Consortium

Genetic Epidemiology Research on Aging (GERA)

- 22 case-control diseases
  - Dyslipidemia, hypertension, type 2 diabetes, …
- Data size (after QC):
  - 54K European-ancestry samples (older adults in Kaiser Northern CA system)
  - 600K genotyped SNPs



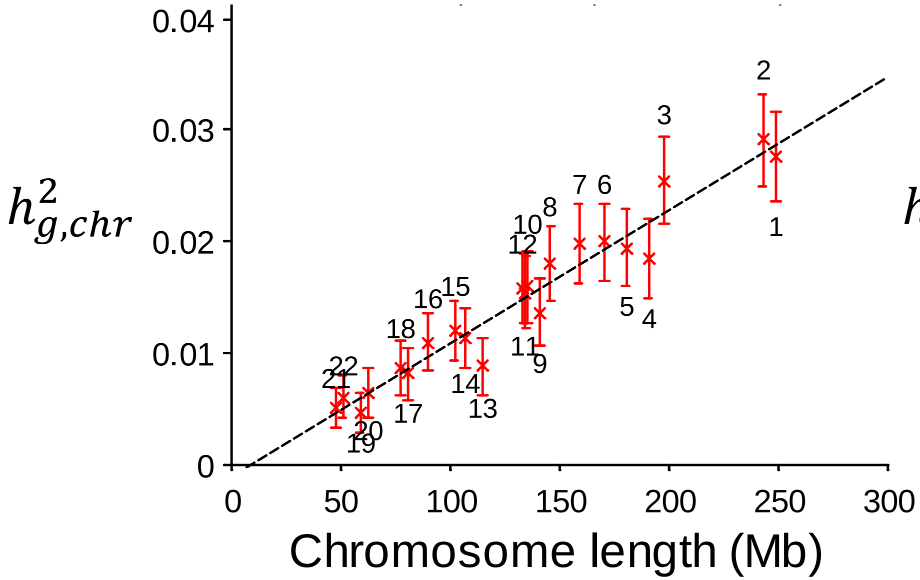
Embargoed for Release: Wednesday, February 26, 2014, 12 p.m. EST

NIH adds substantial set of genetic, health information to online database

Information on older adults is largest ever resource for researchers

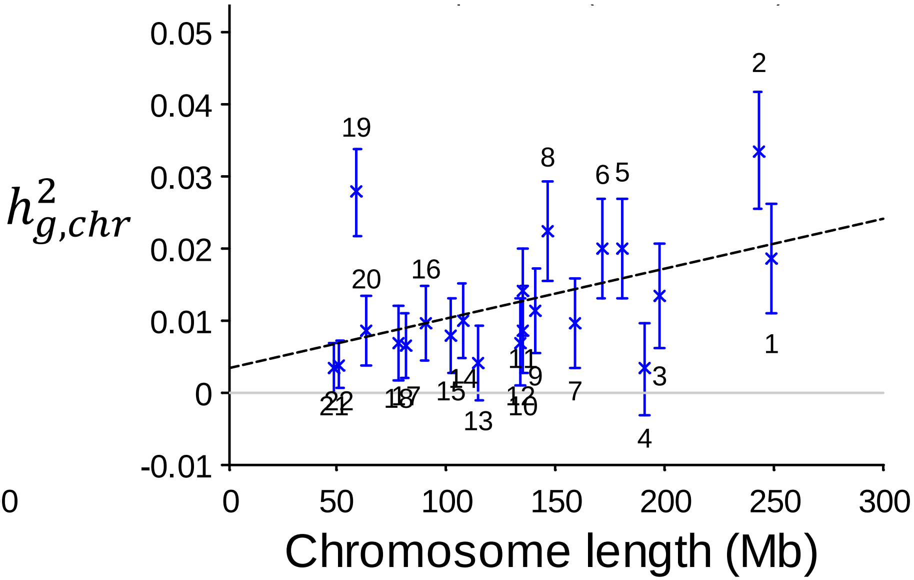
Chromosome-level polygenicity analysis: per-chrom h_g^2 scales strikingly linearly with length

Schizophrenia



(like Lee et al. 2012 Nat Genet but with less noise due to *N*=50K)

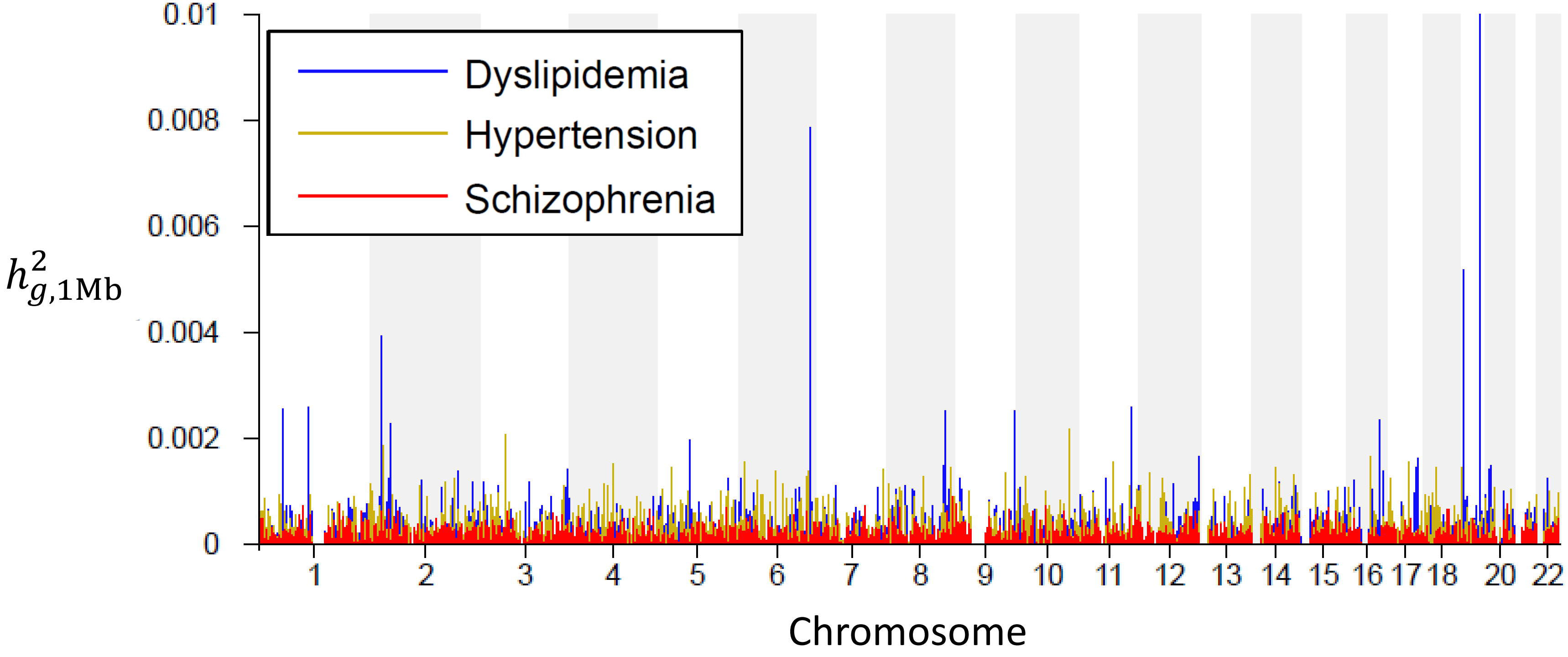
Dyslipidemia



(from *N*=54K GERA samples)

Yang et al. 2011 Nat Genet

BOLT-REML allows estimation of SNP-heritability explained per megabase



Details of BOLT-REML analysis:

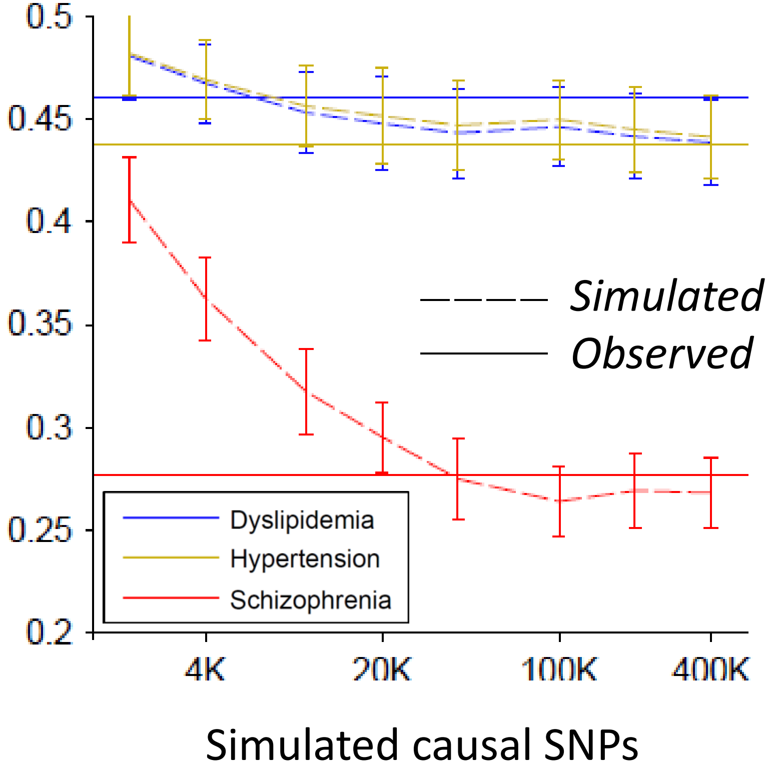
- Estimate $h^2_{g,1\mathrm{Mb}}$ for up to 100 regions at a time (100 VCs)
- 1 additional VC containing all remaining SNPs

Megabase-scale SNP-heritability estimates reveal extreme polygenicity of schizophrenia

How many SNPs are causal?

- Simulations to match observed distribution of per-megabase h_g^2 estimates suggest >20K causal SNPs
- Previous estimate (ABPA method, Ripke et al. 2013): ~8,300 causal SNPs
- Both methods are subject to assumed parameterizations of genetic architecture

Fraction of regions with $\widehat{h^2_{g,1\mathrm{Mb}}}$=0

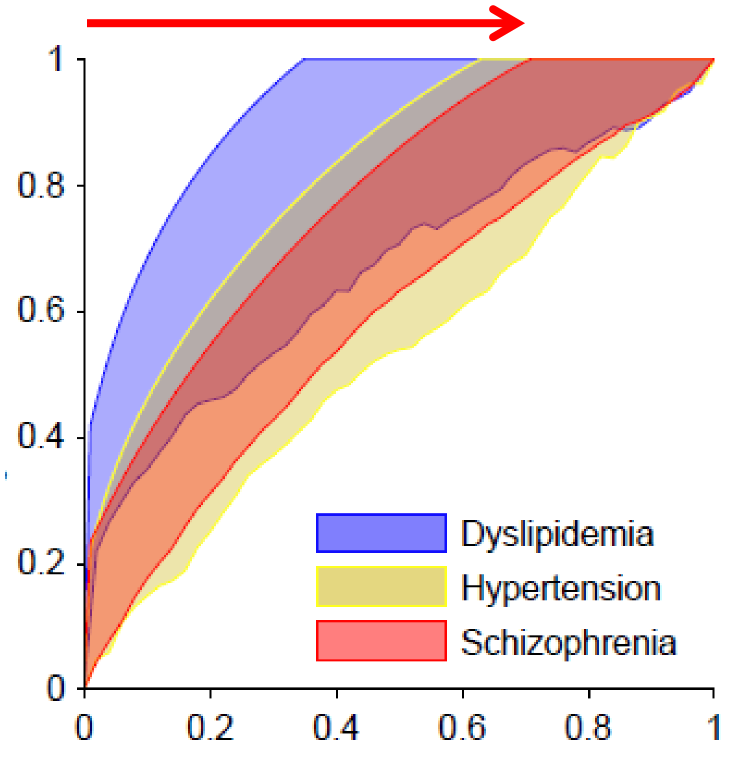


Megabase-scale SNP-heritability estimates reveal extreme polygenicity of schizophrenia

How much SNP-heritability do hottest ("top") 1Mb regions explain?

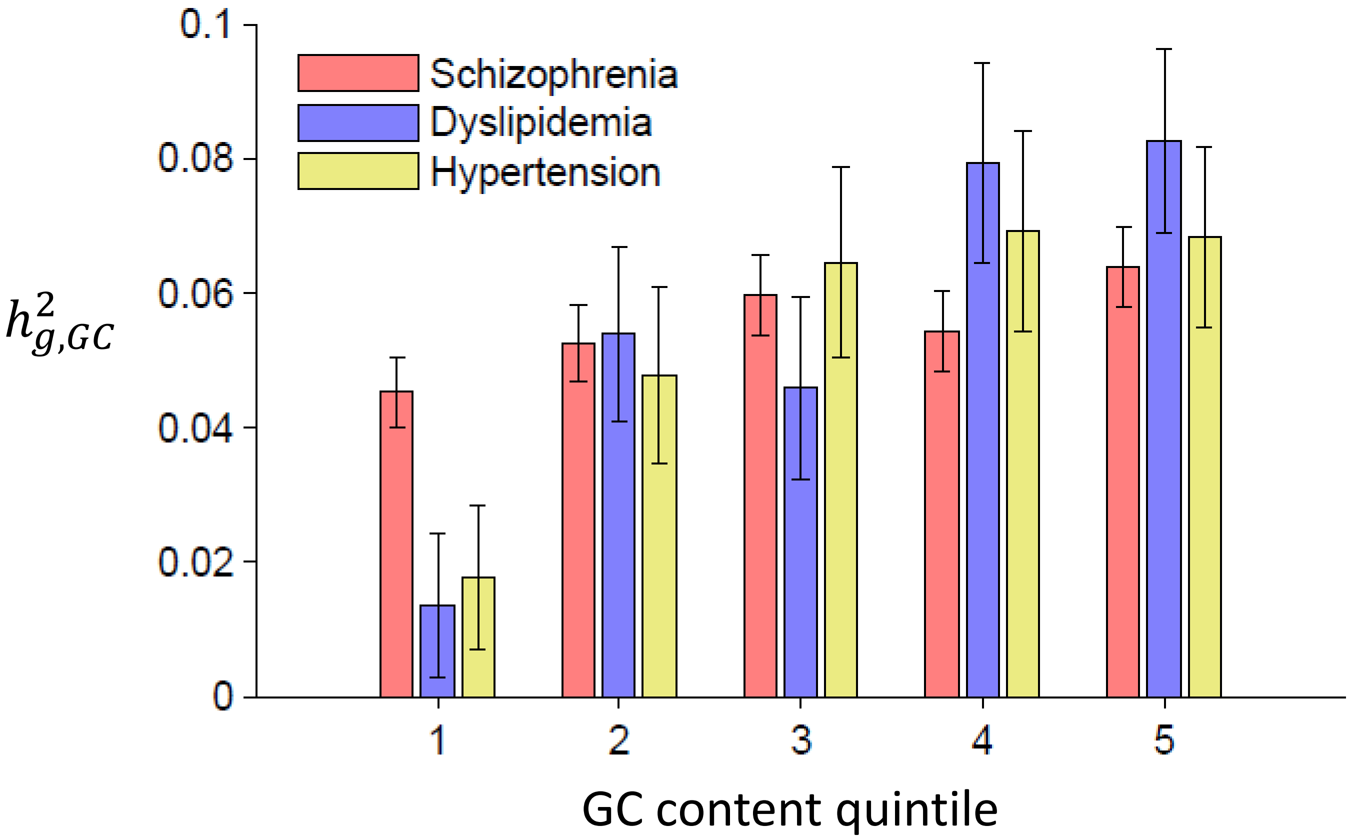
- We use a non-parametric method (i.e., robust to genetic architecture assumptions) to infer conservative 95% CIs from per-Mb estimates
- Inference: >71% of regions have loci

Fraction of h_g^2 explained by top regions



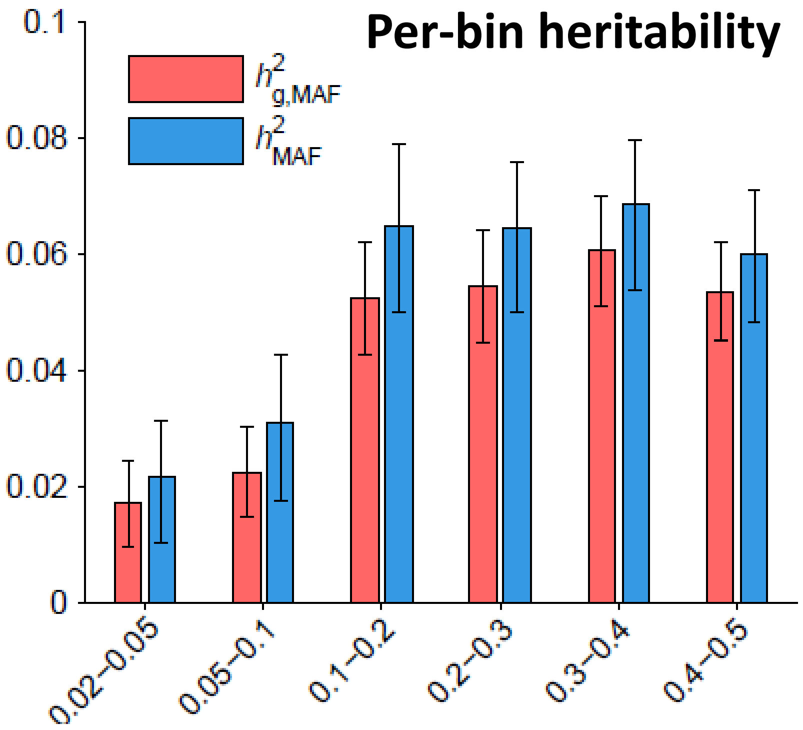
Fraction of 1Mb regions of genome

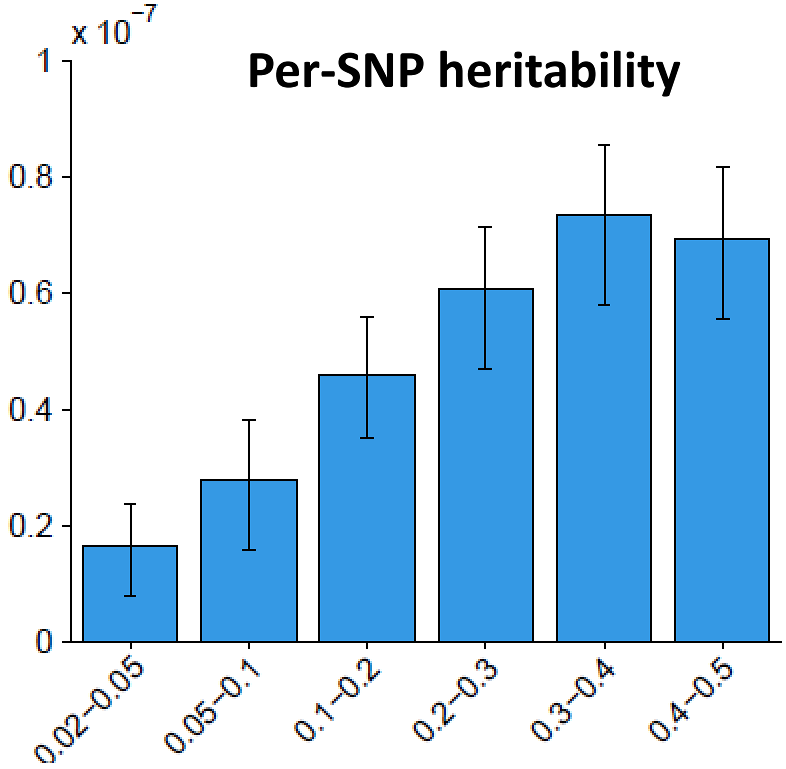
Heritability is enriched in GC-rich regions



- 1% increase in GC content => 1-4% increase in heritability explained
- Note: GC content is correlated with genic content, replication timing, etc.

Higher-frequency SNPs explain more schizophrenia liability (on average)

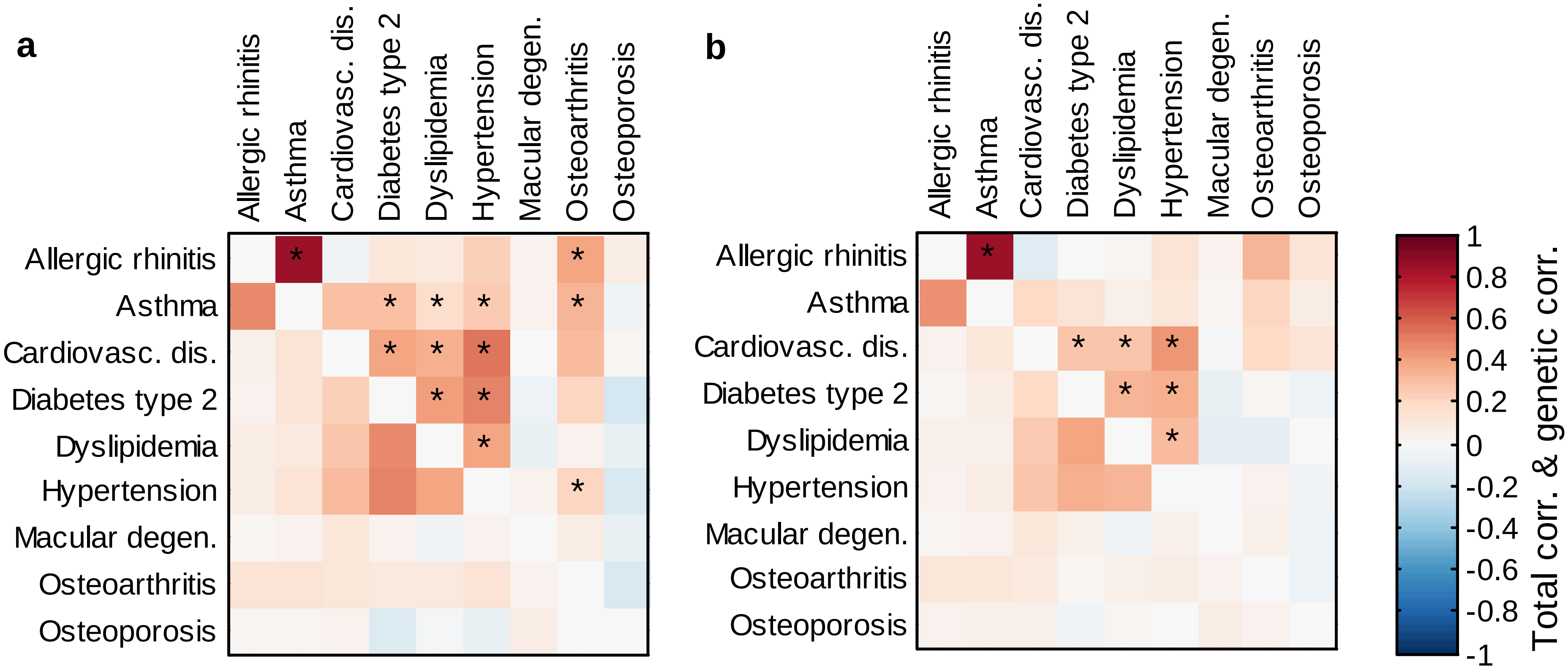




- MAF-partition h_g^2 using BOLT-REML
- Infer total narrow-sense h^2 per bin based on tagging ability (UK10K sequence data simulations)

- Divide by (# UK10K SNPs per bin) to estimate average heritability explained per SNP

Several pairs of GERA diseases exhibit significant genetic correlations, esp. asthma & allergic rhinitis (r_g=0.85)



Not adjusted for BMI

Adjusted for BMI

Consistent with LD Score regression-based analyses

Bulik-Sullivan, Finucane* et al. 2015 Nat Genet*

Conclusions

- BOLT-REML enables powerful heritability analyses of very large GWAS data sets
- Schizophrenia is extremely polygenic
- GC-rich regions contribute more heritability
- Higher-frequency SNPs contribute more heritability (per SNP)

Acknowledgments

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

- Alkes Price
- Gaurav Bhatia
- Sasha Gusev
- Hilary Finucane
- Samuela Pollack

BROAD INSTITUTE

- Nick Patterson
- Brendan Bulik-Sullivan
- Benjamin Neale

SURF SARA

UNC SCHOOL *of* MEDICINE UNC Chapel Hill UNC Health Care

Psychiatric Genomics Consortium

- Stephan Ripke
- Patrick Sullivan
- Michael O’Donovan
- S Hong Lee
- Naomi Wray
- Teresa de Candia
- Kenneth Kendler
- Daniella Posthuma

BOLT-REML software: http://hsph.harvard.edu/alkes-price/software/

Loh et al. 2015b Nat Genet (in press; bioRxiv)