EPI 511, Advanced Population and Medical Genetics

Week 6:

- Mixed model association
- Rare variant analysis

Alkes Price
Harvard School of Public Health
February 28 & March 2, 2017

EPI511, Advanced Population and Medical Genetics

Week 6:

- **Mixed model association**
- Rare variant analysis

Final project: due date is officially Mar 10 at 5pm, but anytime before Mar 13 at 6am is ok.

Outline

1. Introduction / review of mixed model association
2. Inclusion/exclusion of candidate marker in the GRM
3. BOLT-LMM: improving speed
4. BOLT-LMM: improving power

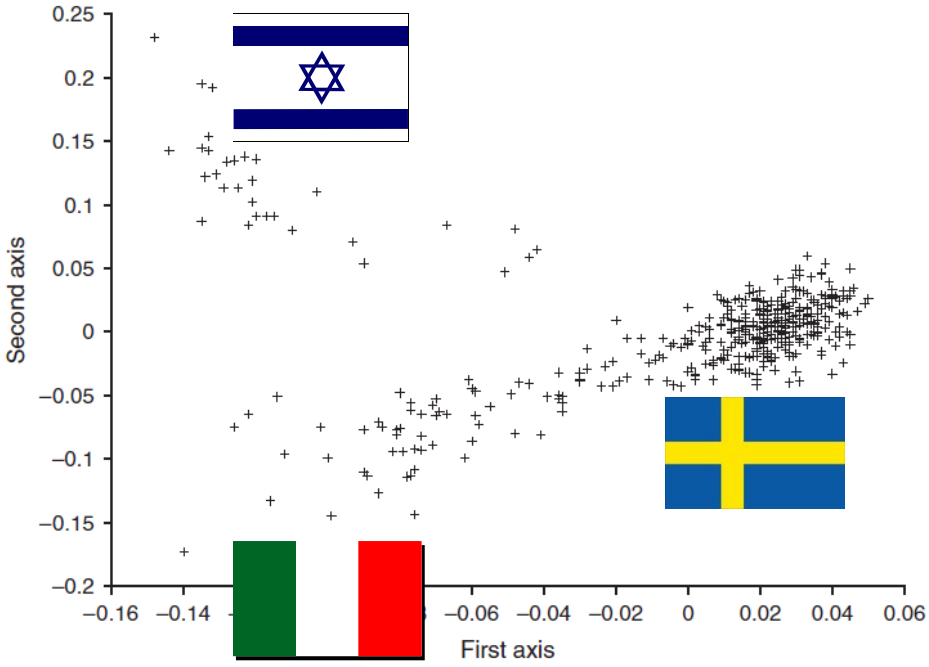
Outline

1. **Introduction / review of mixed model association**
2. Inclusion/exclusion of candidate marker in the GRM
3. BOLT-LMM: improving speed
4. BOLT-LMM: improving power

PCA: a solution for population stratification

Principal components analysis corrects for stratification in genome-wide association studies

Alkes L Price[1,2], Nick J Patterson[2], Robert M Plenge[2,3], Michael E Weinblatt[3], Nancy A Shadick[3] & David Reich[1,2]



Price et al. 2006 Nat Genet

PCA: a solution for population stratification

	λ_{GC}	PCA
• Population stratification	✘	✓

Price et al. 2006 Nat Genet

… does not correct for cryptic relatedness

	λ_{GC}	PCA
• Population stratification	✗	✓
• Cryptic relatedness	✓	✗
• Family relatedness	✓	✗

reviewed in Price et al. 2010 Nat Rev Genet, Yang et al. 2014 Nat Genet

Mixed model association saves the day

	λ_{GC}	PCA	MLM
• Population stratification	✗	✓	✓
• Cryptic relatedness	✓	✗	✓
• Family relatedness	✓	✗	✓

Variance component model to account for sample structure in genome-wide association studies

Hyun Min Kang[1,2,8], Jae Hoon Sul[3,8], Susan K Service[4], Noah A Zaitlen[5], Sit-yee Kong[4], Nelson B Freimer[4], Chiara Sabatti[6] & Eleazar Eskin[3,7]

Kang et al. 2010 Nat Genet

reviewed in Price et al. 2010 Nat Rev Genet, Yang et al. 2014 Nat Genet

(from Thu of Week 3)

Mixed model = Fixed effects + Random effects

$$\mathbf{Y} = \underbrace{\mathbf{XB}}_{\text{fixed effects}} + \underbrace{u}_{\text{random effects}} + \varepsilon$$

$\mathbf{Y}$ = N x 1 vector of phenotypes

$\mathbf{X}$ = N x (1+c) matrix of genotypes at candidate SNP + c covariates

$\mathbf{B}$ = (1+c) x 1 vector of effect sizes of candidate SNP + c covariates

$u \sim \mathrm{N}(0, \sigma_g^2 \mathbf{A})$ is residual variance due to genetic effects

$\varepsilon \sim \mathrm{N}(0, \sigma_e^2 \mathbf{I})$ is residual variance due to environmental effects

$\mathbf{V} = \mathrm{Var}(u + \varepsilon) = \sigma_g^2 \mathbf{A} + \sigma_e^2 \mathbf{I}$

"heritability" $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$

Kang et al. 2010 Nat Genet; also see Listgarten et al. 2012 Nat Methods, Zhou & Stephens 2012 Nat Genet, Yang et al. 2014 Nat Genet

(modified from Thu of Week 3)

Mixed model = Fixed effects + Random effects

$$\mathbf{Y} = \underbrace{\mathbf{XB}}_{\text{fixed effects}} + \underbrace{u}_{\text{random effects}} + \varepsilon$$

$\mathbf{Y}$ = N x 1 vector of phenotypes

$\mathbf{X}$ = N x (1+c) matrix of genotypes at candidate SNP + c covariates

$\mathbf{B}$ = (1+c) x 1 vector of effect sizes of candidate SNP + c covariates

$u \sim \mathrm{N}(0,\sigma_g^2\mathbf{A})$ is residual variance due to genetic effects

$\varepsilon \sim \mathrm{N}(0,\sigma_e^2\mathbf{I})$ is residual variance due to environmental effects

$\mathbf{V} = \mathrm{Var}(u + \varepsilon) = \sigma_g^2\mathbf{A} + \sigma_e^2\mathbf{I}$

"heritability" $h_g^2 = \sigma_g^2/(\sigma_g^2+ \sigma_e^2)$

Yang et al. 2010 Nat Genet, reviewed in Zaitlen & Kraft 2012 Hum Genet
also see Kang et al. 2010 Nat Genet, Zaitlen et al. 2013 PLoS Genet

Estimating h_g^2 using max likelihood

$V = h_g^2 A + (1 - h_g^2) I$ (after normalization) implies that the likelihood of h_g^2 given the data (normalized Y) is

$$L(h_g^2 \mid Y) = \frac{1}{\sqrt{\det(V)}} \exp\left(-\frac{1}{2} Y^T V^{-1} Y \right).$$

Estimate h_g^2 using maximum likelihood (ML) (or restricted maximum likelihood, REML)

(from Tue of Week 5)

Yang et al. 2010 Nat Genet

also see Loh, Bhatia et al. 2015 Nat Genet (faster BOLT-REML algorithm)

Estimating h_g^2 using max likelihood

$V = \sigma_g^2 A + \sigma_e^2 I$ implies that the likelihood of σ_g^2 and σ_e^2 given the data (phenotypes Y) is

$$L(\sigma_g^2, \sigma_e^2 \mid Y) = \frac{1}{\sqrt{\det(V)}} \exp\left(-\frac{1}{2} Y^T V^{-1} Y\right).$$

Estimate σ_g^2, σ_e^2 using maximum likelihood (ML) (or restricted maximum likelihood, REML)

(from Tue of Week 5)

Yang et al. 2010 Nat Genet

also see Loh, Bhatia et al. 2015 Nat Genet (faster BOLT-REML algorithm)

(from Tue of Week 5)

$$h_g^2 < h^2$$

h^2 (total narrow-sense heritability):

- Related individuals
- Use IBD matrix **K**
- $\mathbf{V} = h^2\mathbf{K} + (1 - h^2)\mathbf{I}$

h_g^2 (heritability explained by genotyped SNPs):

- Unrelated individuals
- Use IBS matrix **A** (GRM)
- $\mathbf{V} = h_g^2\mathbf{A} + (1 - h_g^2)\mathbf{I}$
- $h_g^2 < h^2$

Yang et al. 2010 Nat Genet, Yang et al. 2011 Am J Hum Genet
also see Purcell et al. 2009 Nature, Zhou et al. 2013 PLoS Genet

(from Thu of Week 3)

Correcting for stratification: EMMAX

1. Use genotype data to estimate genetic relationship matrix **A**.

Kang et al. 2010 Nat Genet; also see Listgarten et al. 2012 Nat Methods, Zhou & Stephens 2012 Nat Genet, Svishcheva et al. 2012 Nat Genet, Yang et al. 2014 Nat Genet, Loh, Tucker et al. 2015 Nat Genet

(from Thu of Week 3)

Correcting for stratification: EMMAX

1. Use genotype data to estimate genetic relationship matrix **A**.
2. Use genetic relationship matrix **A** and phenotype vector **Y** to estimate the parameters σ_g^2 and σ_e^2 of the model $\mathbf{Y} = u + \varepsilon$ where $u \sim N(0,\sigma_g^2\mathbf{A})$, $\varepsilon \sim N(0,\sigma_e^2\mathbf{I})$.

Kang et al. 2010 Nat Genet; also see Listgarten et al. 2012 Nat Methods, Zhou & Stephens 2012 Nat Genet, Svishcheva et al. 2012 Nat Genet, Yang et al. 2014 Nat Genet, Loh, Tucker et al. 2015 Nat Genet

(from Thu of Week 3)

Correcting for stratification: EMMAX

1. Use genotype data to estimate genetic relationship matrix **A**.
2. Use genetic relationship matrix **A** and phenotype vector **Y** to estimate the parameters σ_g^2 and σ_e^2 of the model $\mathbf{Y} = u + \varepsilon$ where $u \sim N(0,\sigma_g^2\mathbf{A})$, $\varepsilon \sim N(0,\sigma_e^2\mathbf{I})$.
3. Test for non-zero effect size at candidate SNP in the model $\mathbf{Y} = \mathbf{XB} + u + \varepsilon$ where $u \sim N(0,\sigma_g^2\mathbf{A})$, $\varepsilon \sim N(0,\sigma_e^2\mathbf{I})$.

Kang et al. 2010 Nat Genet; also see Listgarten et al. 2012 Nat Methods, Zhou & Stephens 2012 Nat Genet, Svishcheva et al. 2012 Nat Genet, Yang et al. 2014 Nat Genet, Loh, Tucker et al. 2015 Nat Genet

(from Thu of Week 3)

Correcting for stratification: EMMAX

1. Use genotype data to estimate genetic relationship matrix **A**.
2. Use genetic relationship matrix **A** and phenotype vector **Y** to estimate the parameters σ_g^2 and σ_e^2 of the model $\mathbf{Y} = u + \varepsilon$ where $u \sim N(0,\sigma_g^2\mathbf{A})$, $\varepsilon \sim N(0,\sigma_e^2\mathbf{I})$.
3. Test for non-zero effect size at candidate SNP in the model $\mathbf{Y} = \mathbf{XB} + u + \varepsilon$ where $u \sim N(0,\sigma_g^2\mathbf{A})$, $\varepsilon \sim N(0,\sigma_e^2\mathbf{I})$.

   Note: if there are no covariates, an appropriate statistic is $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y})^2 / (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})$, where $\mathbf{V} = \sigma_g^2\mathbf{A} + \sigma_e^2\mathbf{I}$, generalizing the Armitage trend test (Armitage 1955 Biometrics)

Kang et al. 2010 Nat Genet; also see Listgarten et al. 2012 Nat Methods, Zhou & Stephens 2012 Nat Genet, Svishcheva et al. 2012 Nat Genet, Yang et al. 2014 Nat Genet, Loh, Tucker et al. 2015 Nat Genet

Mixed model association saves the day

	λ_{GC}	PCA	MLM
• Population stratification	✗	✓	✓
• Cryptic relatedness	✓	✗	✓
• Family relatedness	✓	✗	✓

Variance component model to account for sample structure in genome-wide association studies

Hyun Min Kang[1,2,8], Jae Hoon Sul[3,8], Susan K Service[4], Noah A Zaitlen[5], Sit-yee Kong[4], Nelson B Freimer[4], Chiara Sabatti[6] & Eleazar Eskin[3,7]

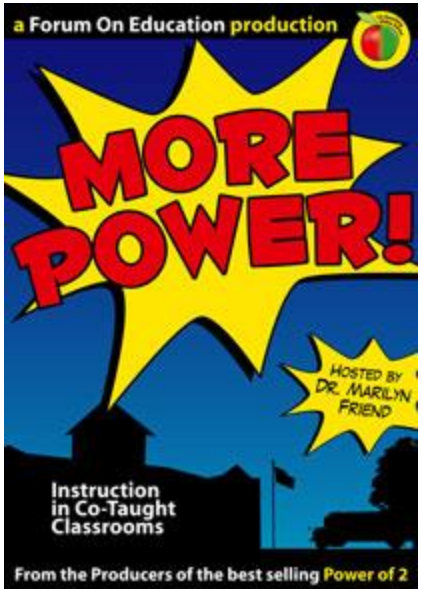
Kang et al. 2010 Nat Genet

reviewed in Price et al. 2010 Nat Rev Genet, Yang et al. 2014 Nat Genet

MLM increases power in association studies **without** cryptic or family relatedness

The reason: MLM implicitly conditions on other markers, reducing noise variance and increasing signal to noise.

This effect increases as sample size *N* increases (relative to effective # independent markers *M*).



MLM = association mapping on BLUP residual!

Mixed model association $\chi^2 = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y})^2 / (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})$

Fact 1: BLUP prediction $\hat{u} = \sigma_g^2 \mathbf{A}\,\mathbf{V}^{-1}\,\mathbf{Y}$ (see Thu of Week 5)

Note: this is BLUP prediction of *in-sample* genetic values

Henderson 1975 Biometrics
reviewed in de los Campos et al. 2010 Nat Rev Genet

MLM = association mapping on BLUP residual!

Mixed model association $\chi^2 = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y})^2 / (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})$

Fact 1: BLUP prediction $\hat{u} = \sigma_g^2 \mathbf{A} \mathbf{V}^{-1} \mathbf{Y}$ (see Thu of Week 5)

Fact 2: BLUP residual $\mathbf{Y}_{\text{resid}} = \mathbf{Y} - \hat{u} = \mathbf{Y} - \sigma_g^2 \mathbf{A} \mathbf{V}^{-1} \mathbf{Y}$

$= \mathbf{Y} - (\mathbf{V} - \sigma_e^2 \mathbf{I})\mathbf{V}^{-1} \mathbf{Y} \sim \mathbf{V}^{-1}\mathbf{Y}$

Henderson 1975 Biometrics
reviewed in de los Campos et al. 2010 Nat Rev Genet

MLM = association mapping on BLUP residual!

Mixed model association $\chi^2 = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y})^2 / (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})$

Fact 1: BLUP prediction $\hat{u} = \sigma_g^2 \mathbf{A} \mathbf{V}^{-1} \mathbf{Y}$ (see Thu of Week 5)

Fact 2: BLUP residual $\mathbf{Y}_{resid} = \mathbf{Y} - \hat{u} = \mathbf{Y} - \sigma_g^2 \mathbf{A} \mathbf{V}^{-1} \mathbf{Y}$

$= \mathbf{Y} - (\mathbf{V} - \sigma_e^2 \mathbf{I})\mathbf{V}^{-1} \mathbf{Y} \sim \mathbf{V}^{-1}\mathbf{Y}$

Thus, numerator of χ^2 statistic = $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y})^2 \sim (\mathbf{X}^T\mathbf{Y}_{resid})^2$

MLM ⇔ Association mapping on BLUP residual $\mathbf{Y}_{resid}$

Henderson 1975 Biometrics

reviewed in de los Campos et al. 2010 Nat Rev Genet

BLUP coefficients = mixed model coefficients

Predictions: $\underset{K \times 1}{\hat{Y}_{test}} = \underset{K \times N}{\sigma_g^2 \, A} \; \underset{N \times N}{V^{-1}} \; \underset{N \times 1}{Y}$

de los Campos et al. 2010 Nat Rev Genet

Coefficients: $\underset{M \times 1}{\hat{\beta}} = \underset{M \times N}{(\sigma_g^2/M) \, X} \; \underset{N \times N}{V^{-1}} \; \underset{N \times 1}{Y}$

Yang et al. 2011 Am J Hum Genet

- Same as "mixed model" coefficients (Kang et al. 2010 Nat Genet)

(from Thu of Week 5)

reviewed in de los Campos et al. 2010 Nat Rev Genet

Mixed model association has advantages and pitfalls

Pitfalls: Standard mixed model association methods can suffer from

- suboptimal power due to inclusion of candidate marker in GRM
- suboptimal power due to not modeling sparse polygenic architectures
- a loss in power in ascertained case-control studies

Yang et al. 2014 Nat Genet

Outline

1. Introduction / review of mixed model association
2. **Inclusion/exclusion of candidate marker in the GRM**
3. BOLT-LMM: improving speed
4. BOLT-LMM: improving power

Defining MLMi vs. MLMe

MLMi approach:
Build genetic relationship matrix A including the candidate SNP
e.g. EMMAX (Kang et al. 2010 Nat Genet). Also see
Zhou & Stephens 2012 Nat Genet, Svishcheva et al. 2012 Nat Genet

reviewed in Yang et al. 2014 Nat Genet

Defining MLMi vs. MLMe

MLMi approach:
Build genetic relationship matrix A including the candidate SNP
e.g. EMMAX (Kang et al. 2010 Nat Genet). Also see
Zhou & Stephens 2012 Nat Genet, Svishcheva et al. 2012 Nat Genet

MLMe approach:
Build genetic relationship matrix A excluding the candidate SNP
e.g. EMMA (Kang et al. 2008 Genetics) uses a different
genetic relationship matrix (GRM) for each candidate SNP.
BUT this is extremely computationally intensive: $O(MN^3)$

reviewed in Yang et al. 2014 Nat Genet

Defining MLMi vs. MLMe

MLMi approach:
Build genetic relationship matrix A including the candidate SNP
e.g. EMMAX (Kang et al. 2010 Nat Genet). Also see
Zhou & Stephens 2012 Nat Genet, Svishcheva et al. 2012 Nat Genet

MLMe approach:
Build genetic relationship matrix A excluding the candidate SNP
e.g. EMMA (Kang et al. 2008 Genetics) uses a different
genetic relationship matrix (GRM) for each candidate SNP.
e.g. FaST-LMM (Listgarten et al. 2012 Nat Methods) uses a different
GRM for each candidate SNP (excluding nearby SNPs).
with a speedup to avoid repeating work for each GRM.

reviewed in Yang et al. 2014 Nat Genet

Defining MLMi vs. MLMe

MLMi approach:
Build genetic relationship matrix A <u>including</u> the candidate SNP
e.g. EMMAX (Kang et al. 2010 Nat Genet). Also see
Zhou & Stephens 2012 Nat Genet, Svishcheva et al. 2012 Nat Genet

MLMe approach:
Build genetic relationship matrix A <u>excluding</u> the candidate SNP
e.g. EMMA (Kang et al. 2008 Genetics) uses a different genetic relationship matrix (GRM) for each candidate SNP.
e.g. FaST-LMM (Listgarten et al. 2012 Nat Methods) uses a different GRM for each candidate SNP (excluding nearby SNPs).
e.g. GCTA-**L**eave **O**ne **C**hromosome **O**ut (GCTA-LOCO) uses a different GRM excluding each chromosome in turn.

reviewed in Yang et al. 2014 Nat Genet

Average χ^2 of ATT vs. MLMi vs. MLMe if there is no stratification or relatedness

Let N = # samples, M = effective # independent markers

	All markers	Null markers
ATT: Average χ^2 statistic:	$1+h_g^2 N/M$	1
MLMi: (if $N<M$, $r^2 \approx h_g^2 N/M$)	1	$\frac{1-r^2 h_g^2}{h_g^2 N/M+1-r^2 h_g^2}$
MLMe: (if $N<M$, $r^2 \approx h_g^2 N/M$)	$1+\frac{h_g^2 N/M}{1-r^2 h_g^2}$	1

see Yang et al. 2014 Nat Genet for detailed derivations + simulations

ATT: Average $\chi^2 > 1$ for large N

Let N = # samples, M = effective # independent markers

ATT:	All markers	Null markers
Average χ^2 statistic:	$\boxed{1 + h_g^2 N / M}$	1

see Yang et al. 2014 Nat Genet for detailed derivations + simulations

How much inflation in λ_{GC} is OK ??

- Long answer: $\lambda_{GC} \leq 1.05$ is usually considered OK
  BUT λ_{GC} scales with sample size.

  $\lambda_{GC} = 1.05$ @ N=1,000 implies a more severe effect than
  $\lambda_{GC} = 1.05$ @ N=100,000

  At very large sample sizes, $\lambda_{GC} > 1$ can be expected due to true polygenic effects (Yang et al. 2011 Eur J Hum Genet, Bulik-Sullivan, Loh et al. 2015 Nat Genet; Tue of Week 3 + Tue of Week 5)

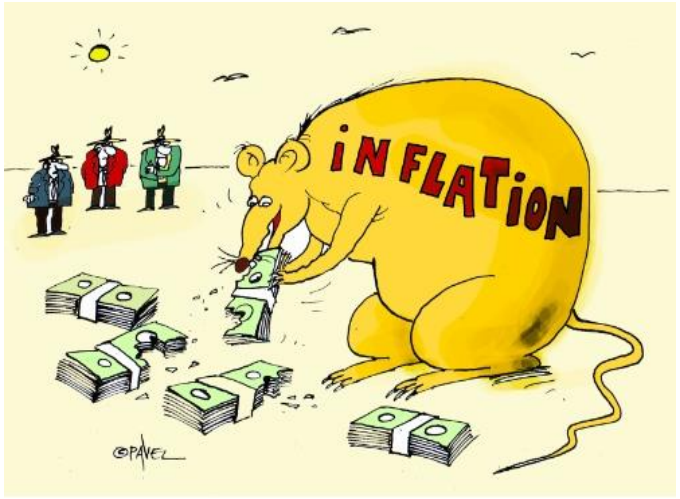
(from Tue of Week 3)

$h_g^2 > 0$ => Average $\chi^2 > 1$ for large N !!!

Assuming N individuals and M unlinked markers,
and no confounding due to population stratification:

Average value of ATT χ^2 statistic = $1 + h_g^2 N / M$

Inflation is not a bad thing!



(from Tue of Week 5)

CONFOUNDING

Yang et al. 2011 Eur J Hum Genet

$h_g^2 > 0 \Rightarrow \lambda_{GC} > 1$ for large N !!!

Assuming N individuals and M unlinked markers,
and no confounding due to population stratification:

Average value of ATT χ^2 statistic = $1 + h_g^2 N / M$

$1 < \lambda_{GC} \leq$ Average value of ATT χ^2 statistic
(depending on the number of causal markers)

(from Tue of Week 5)

Yang et al. 2011 Eur J Hum Genet

Average $\chi^2 > 1$ in WTCCC $\Leftrightarrow$ confounding?

Table 3 Comparison of genomic control inflation factor obtained with different models in seven WTCCC phenotypes

	Genomic control inflation factor		
Phenotype	Uncorrected	ES100	EMMAX
BD	1.105	1.071	0.998
CAD	1.063	1.048	1.006
CD	1.098	1.055	1.000
HT	1.055	1.051	0.997
RA	1.028	1.031	0.965 (0.989[a])
T1D	1.043	1.028	0.946 (0.991[a])
T2D	1.065	1.042	0.996

Inflation? Or polygenicity?

Well-calibrated? Or deflated?

(from Thu of Week 3)

Kang et al. 2010 Nat Genet

Average $\chi^2 > 1$ in WTCCC  confounding

Disease	h_g^2 on liab. scale	h_g^2 on obs. scale	$1 + h_g^2 N/M_{eff}$ obs. scale	λ_{GC} (LR)	λ_{GC} (ES100)
BD	0.38	0.76	1.061	1.105	1.071
CD	0.22	0.61	1.049	1.098	1.055

h_g^2 values from Lee et al. 2011 Am J Hum Genet

λ_{GC} values from Kang et al. 2010 Nat Genet

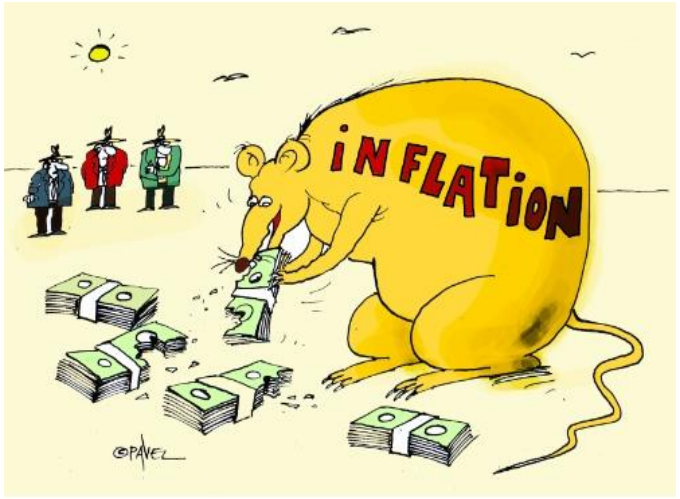
WTCCC 2007 Nature

ATT: Average $\chi^2 > 1$ for large N

Let N = # samples, M = effective # independent markers

ATT:	All markers	Null markers
Average χ^2 statistic:	$1 + h_g^2 N / M$	1

Inflation is not a bad thing!



CONFOUNDING

Yang et al. 2011 Eur J Hum Genet

MLMi is deflated and decreases power

Let N = # samples, M = effective # independent markers

	All markers	Null markers
Linear Regression: Average χ^2 statistic:	$1 + h_g^2 N / M$	1
MLMi: (if $N < M$, $r^2 \approx h_g^2 N / M$)	1	$\frac{1 - r^2 h_g^2}{h_g^2 N / M + 1 - r^2 h_g^2}$

Power loss / miscalibration! Including candidate SNP in GRM **A** in null model inflates the null likelihood and deflates χ^2 statistics.

NULL MODEL: $\mathbf{Y} = u + \varepsilon,\ u \sim N(0, \sigma_g^2 \mathbf{A}),\ \varepsilon \sim N(0, \sigma_e^2 \mathbf{I})$.

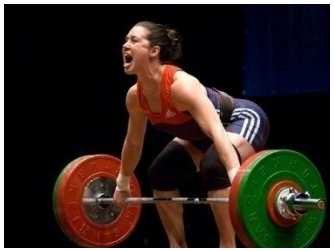
CAUSAL MODEL: $\mathbf{Y} = \mathbf{XB} + u + \varepsilon,\ u \sim N(0, \sigma_g^2 \mathbf{A}),\ \varepsilon \sim N(0, \sigma_e^2 \mathbf{I})$.

Listgarten et al. 2012 Nat Methods; also see Yang et al. 2014 Nat Genet

MLMe is well-calibrated and increases power!

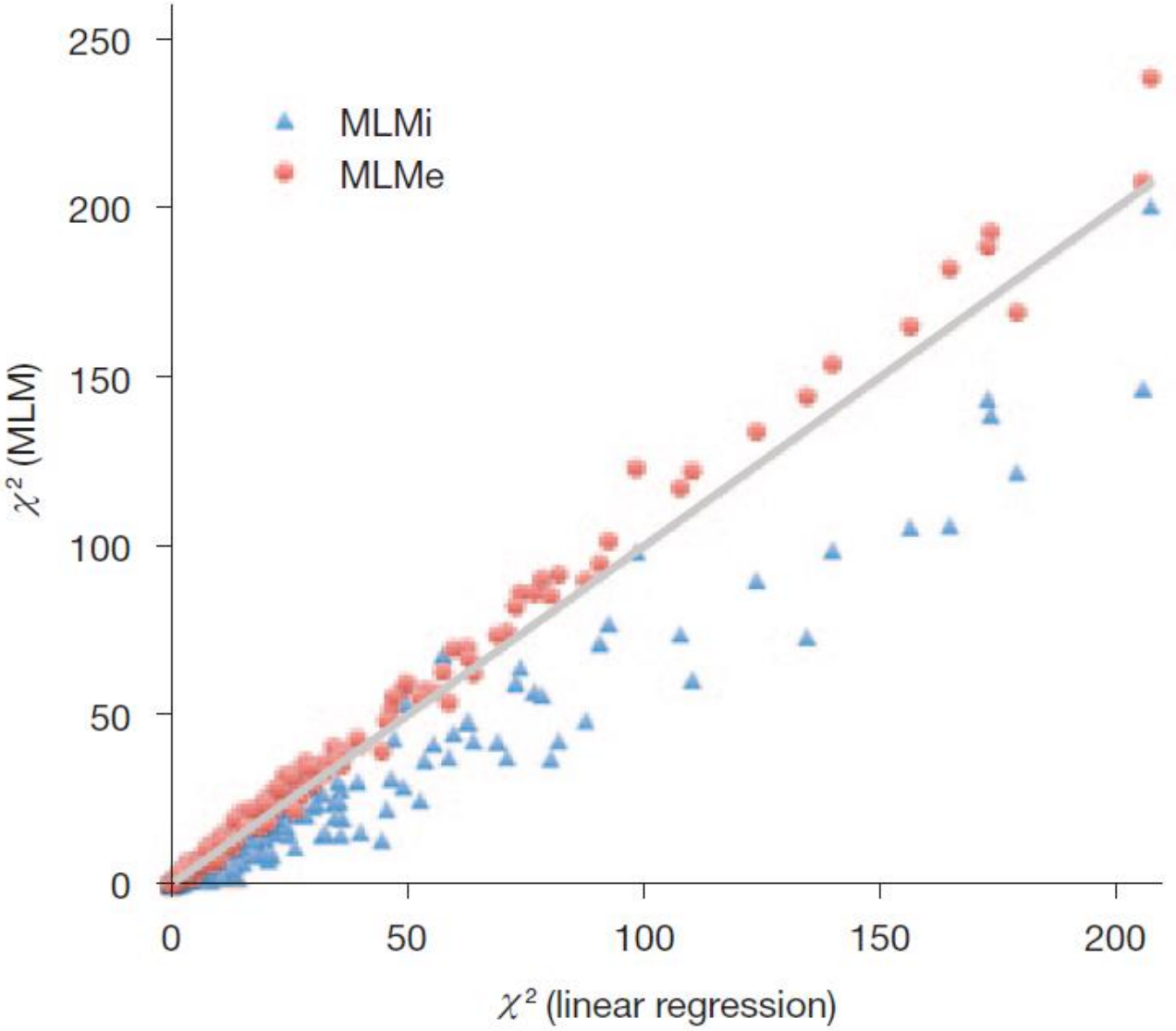
Let N = # samples, M = effective # independent markers

	All markers	Null markers
Linear Regression: Average χ^2 statistic:	$1+h_g^2 N/M$	1
MLMi: (if $N<M$, $r^2 \approx h_g^2 N/M$)	1	$\frac{1-r^2 h_g^2}{h_g^2 N/M+1-r^2 h_g^2}$
MLMe: (if $N<M$, $r^2 \approx h_g^2 N/M$)	$\boxed{1+\frac{h_g^2 N/M}{1-r^2 h_g^2}}$	1



see Yang et al. 2014 Nat Genet for detailed derivations + simulations

Power simulations: MLMe > ATT > MLMi



ATT

Yang et al. 2014 Nat Genet

Real data: MLMi is deflated and decreases power, but MLMe increases power

WTCCC2 MS data: 10,204 cases + 5,429 controls, 360,557 SNPs
WTCCC2 UC data: 2,697 cases + 5,652 controls, 458,560 SNPs
Average χ^2 statistics for all markers & for known associated SNPs

	ATT	PCA	MLMi	MLMe
MS, 360,557 SNPs	3.95	1.25	0.99	1.23
MS, 75 published SNPs	18.50	10.20	8.90	11.30
UC, 458,560 SNPs	1.16	1.11	1.00	1.10
UC, 24 published SNPs	14.06	13.63	12.11	13.43

MS data from Sawcer et al. 2011 Nature
UC data from Jostins et al. 2012 Nature

Yang et al. 2014 Nat Genet

MLMi vs. MLMe: recommendations

If $N << M$ (e.g. N<10K; note typically $M \approx$ 60K), MLMi is ok.

Otherwise, run MLMe instead of MLMi to avoid loss in power.

Implementations of MLMe in O(MN^2) time:

- FaST-LMM software (Listgarten et al. 2012 Nat Methods)
- GCTA software (GCTA-LOCO):
  http://www.complextraitgenomics.com/software/gcta/mlmassoc.html

Outline

1. Introduction / review of mixed model association

2. Inclusion/exclusion of candidate marker in the GRM

3. BOLT-LMM: improving speed

Loh, Tucker et al. 2015 Nat Genet

Building GRM + fitting variance components for each candidate SNP is computationally intensive

EMMA method (Kang et al. 2008 Genetics):
Build GRM and fit variance components for each candidate SNP
Time cost $O(MN^3)$ where M = #SNPs, N = #samples

also see Yu et al. 2006 Nat Genet, Zhao et al. 2007 PLoS Genet

Building GRM + fitting variance components once for all SNPs is much faster

EMMA method (Kang et al. 2008 Genetics):
Build GRM and fit variance components for each candidate SNP
Time cost O(MN^3) where M = #SNPs, N = #samples

EMMAX method (Kang et al. 2010 Nat Genet):
Build GRM and fit variance components once for all SNPs
Time cost O(MN^2) where M = #SNPs, N = #samples. Much faster!

Kang et al. 2010 Nat Genet, reviewed in Yang et al. 2014 Nat Genet; also see Listgarten et al. 2012 Nat Methods, Segura et al. 2012 Nat Genet, Zhou/Stephens 2012 Nat Genet, Svishcheva et al. 2012 Nat Genet, Lippert et al. 2013 Sci Rep

But building GRM still takes time $O(MN^2)$

All previous mixed model methods rely on building the GRM

$$A_{jk} = \frac{1}{M}\sum_i \frac{(g_{ij} - 2p_i)(g_{ik} - 2p_i)}{2p_i(1-p_i)} \ ,$$

which takes time $O(MN^2)$ where M = #markers, N = #samples.

Kang et al. 2010 Nat Genet, reviewed in Yang et al. 2014 Nat Genet; also see Listgarten et al. 2012 Nat Methods, Segura et al. 2012 Nat Genet, Zhou/Stephens 2012 Nat Genet, Svishcheva et al. 2012 Nat Genet, Lippert et al. 2013 Sci Rep

But building GRM still takes time $O(MN^2)$

All previous mixed model methods rely on building the GRM

$$A_{jk} = \frac{1}{M}\sum_i \frac{(g_{ij} - 2p_i)(g_{ik} - 2p_i)}{2p_i(1-p_i)} ,$$

which takes time $O(MN^2)$ where M = #markers, N = #samples.

e.g. M = 9 million, N = 80K (PGC2) => 2,000 days of CPU time.

Kang et al. 2010 Nat Genet, reviewed in Yang et al. 2014 Nat Genet; also see Listgarten et al. 2012 Nat Methods, Segura et al. 2012 Nat Genet, Zhou/Stephens 2012 Nat Genet, Svishcheva et al. 2012 Nat Genet, Lippert et al. 2013 Sci Rep

Is linear-time mixed model association possible?

Mixed model association $\chi^2 = \frac{(X_m{}^T V^{-1} Y)^2}{X_m{}^T V^{-1} X_m}$

BOLT-LMM-inf:

i. Compute the numerator in linear time

ii. Use a constant denominator for calibration

Loh, Tucker et al. 2015 Nat Genet

Is linear-time mixed model association possible?

Mixed model association $\chi^2 = \frac{(X_m{}^T V^{-1} Y)^2}{X_m{}^T V^{-1} X_m}$

BOLT-LMM-inf:

i. Compute the numerator in linear time

ii. Use a constant denominator for calibration

Loh, Tucker et al. 2015 Nat Genet

MLM = association mapping on BLUP residual!

Mixed model association $\chi^2 = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y})^2 / (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})$

Fact 1: BLUP prediction $\hat{u} = \sigma_g^2 \mathbf{A} \mathbf{V}^{-1} \mathbf{Y}$ (see Thu of Week 5)

Fact 2: BLUP residual $\mathbf{Y}_{resid} = \mathbf{Y} - \hat{u} = \mathbf{Y} - \sigma_g^2 \mathbf{A} \mathbf{V}^{-1} \mathbf{Y}$

$= \mathbf{Y} - (\mathbf{V} - \sigma_e^2 \mathbf{I})\mathbf{V}^{-1} \mathbf{Y} \sim \mathbf{V}^{-1}\mathbf{Y}$

Thus, numerator of χ^2 statistic = $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y})^2 \sim (\mathbf{X}^T\mathbf{Y}_{resid})^2$
MLM ⇔ Association mapping on BLUP residual $\mathbf{Y}_{resid}$

Goal: Compute BLUP residual Y_{resid} in linear time

Henderson 1975 Biometrics
reviewed in de los Campos et al. 2010 Nat Rev Genet

Iterative algorithm computes Y_{resid} in linear time

Initialize $\beta_m = 0$ for each SNP m

Initialize $Y_{resid} = Y$

At each iteration

For each SNP m

Step 1. Unresidualize Y_{resid} for SNP m: $Y_m = Y_{resid} + X_m\beta_m$

Step 2. Re-estimate $\beta_m = \left[\frac{h^2/M}{h^2/M + (1-h^2)/N}\right]\frac{X_m{}^T Y_m}{N}$

Step 3. Residualize Y_{resid} for SNP m: $Y_{resid} = Y_m - X_m\beta_m$

Converges to correct BLUP residual Y_{resid} in 10-20 iterations!!

Legarra & Misztal 2008 J Dairy Sci, Meuwissen et al. 2009 Gen Sel Evol, Carbonetto & Stephens 2012 Bayesian Analysis, Logsdon et al. 2012 Bioinformatics

Iterative algorithm computes Y_{resid} in linear time

Initialize $\beta_m = 0$ for each SNP m

Initialize $Y_{resid} = Y$

At each iteration

For each SNP m

LOCO (leave-one-chromosome-out) approach is used to avoid "proximal contamination" (Lippert et al. 2011; reviewed in Yang et al. 2014 Nat Genet)

Step 1. Unresidualize Y_{resid} for SNP m: $Y_m = Y_{resid} + X_m\beta_m$

Step 2. Re-estimate $\beta_m = \left[\frac{h^2/M}{h^2/M + (1-h^2)/N}\right]\frac{X_m{}^T Y_m}{N}$

Step 3. Residualize Y_{resid} for SNP m: $Y_{resid} = Y_m - X_m\beta_m$

Converges to correct BLUP residual Y_{resid} in 10-20 iterations!!

Legarra & Misztal 2008 J Dairy Sci, Meuwissen et al. 2009 Gen Sel Evol, Carbonetto & Stephens 2012 Bayesian Analysis, Logsdon et al. 2012 Bioinformatics

Is linear-time mixed model association possible?

Mixed model association $\chi^2 = \frac{(X_m{}^T V^{-1} Y)^2}{X_m{}^T V^{-1} X_m}$

BOLT-LMM-inf:

i. Compute the numerator in linear time

ii. Use constant denominator for calibration

Denominator is known to be approximately constant
(Svishcheva et al. 2012 Nat Genet)

Loh, Tucker et al. 2015 Nat Genet

Outline

1. Introduction / review of mixed model association
2. Inclusion/exclusion of candidate marker in the GRM
3. BOLT-LMM: improving speed
4. **BOLT-LMM: improving power**

Loh, Tucker et al. 2015 Nat Genet

MLM = association mapping on BLUP residual!

Mixed model association $\chi^2 = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y})^2 / (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})$

Fact 1: BLUP prediction $\hat{u} = \sigma_g^2 \mathbf{A} \mathbf{V}^{-1} \mathbf{Y}$ (see Thu of Week 5)

Fact 2: BLUP residual $\mathbf{Y}_{resid} = \mathbf{Y} - \hat{u} = \mathbf{Y} - \sigma_g^2 \mathbf{A} \mathbf{V}^{-1} \mathbf{Y}$

$= \mathbf{Y} - (\mathbf{V} - \sigma_e^2 \mathbf{I})\mathbf{V}^{-1} \mathbf{Y} \sim \mathbf{V}^{-1}\mathbf{Y}$

Thus, numerator of χ^2 statistic = $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y})^2 \sim (\mathbf{X}^T\mathbf{Y}_{resid})^2$
MLM ⇔ Association mapping on BLUP residual $\mathbf{Y}_{resid}$

Goal: Compute BLUP residual Y_{resid} in linear time

Henderson 1975 Biometrics
reviewed in de los Campos et al. 2010 Nat Rev Genet

… but BLUP is not the best predictor!

- Predictors that are based on non-infinitesimal prior distributions attain higher prediction accuracy
  (Meuwissen et al. 2001 Genetics, de los Campos et al. 2010 Nat Rev Genet, Erbe et al. 2012 J Dairy Sci, Zhou et al. 2013 PLoS Genet. Also see Carbonetto/Stephens 2012 Bayesian Analysis, Lippert et al. 2013 Sci Rep)

… but BLUP is not the best predictor!

- Predictors that are based on non-infinitesimal prior distributions attain higher prediction accuracy
  (Meuwissen et al. 2001 Genetics, de los Campos et al. 2010 Nat Rev Genet, Erbe et al. 2012 J Dairy Sci, Zhou et al. 2013 PLoS Genet. Also see Carbonetto/Stephens 2012 Bayesian Analysis, Lippert et al. 2013 Sci Rep)

Idea: Association mapping on more accurate prediction residual (more accurate than BLUP) will increase power!

Iterative algorithm computes Y_{resid} in linear time

Specify *prior* $\beta_m \sim$ mixture of two normal distributions.

Initialize $\beta_m = 0$ for each SNP m

Initialize $Y_{resid} = Y$

At each iteration

For each SNP m

Step 1. Unresidualize Y_{resid} for SNP m: $Y_m = Y_{resid} + X_m\beta_m$

Step 2. Re-estimate $\beta_m = E\left(\beta_m \middle| prior, \frac{X_m{}^T Y_m}{N}\right)$

Step 3. Residualize Y_{resid} for SNP m: $Y_{resid} = Y_m - X_m\beta_m$

Converges to more accurate prediction residual Y_{resid} in small #iterations!

Legarra & Misztal 2008 J Dairy Sci, Meuwissen et al. 2009 Gen Sel Evol, Carbonetto & Stephens 2012 Bayesian Analysis, Logsdon et al. 2012 Bioinformatics

Powerful linear-time mixed model association!

Mixed model association $\chi^2 = \dfrac{(X_m{}^T Y_{resid})^2}{(Y_{resid})^T A^* (Y_{resid})}$

BOLT-LMM:

i. Compute the numerator in linear time

ii. Use constant denominator for calibration (LDscore regression)

Loh, Tucker et al. 2015 Nat Genet

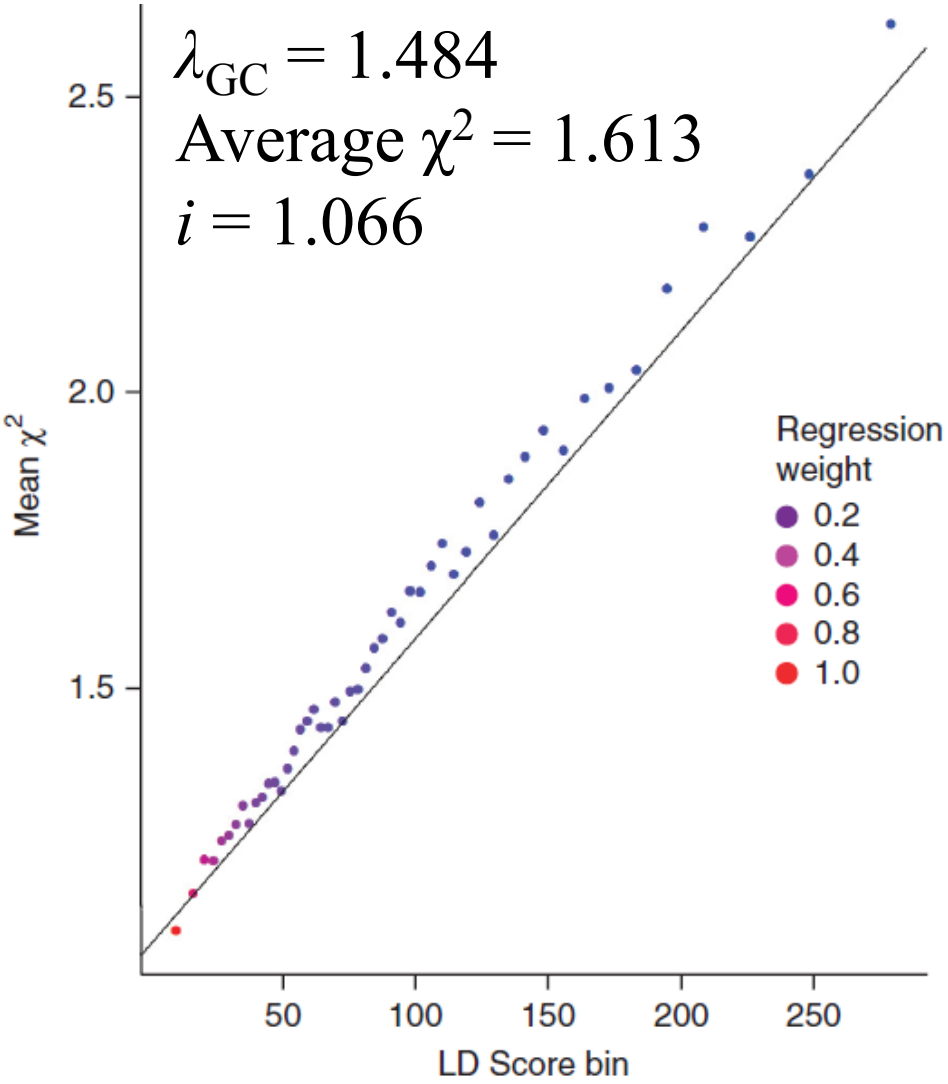
LD score regression: calibrating test statistics

LDscore(SNP m)
$= \sum_{m'} r^2(m, m')$

Regress
$\chi^2 = i + s$*LDscore

Intercept i should be
= 1 if no confounding,
> 1 if confounding

(from Tue of Week 3)



Bulik-Sullivan, Loh et al. 2015 Nat Genet; SCZ data from PGC-SCZ 2014 Nature

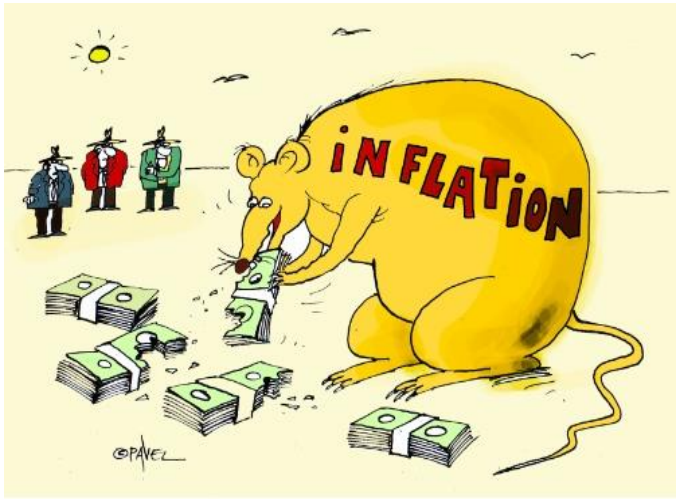
Average $\chi^2 > 1$ does not imply confounding

Linear Regression: E(χ^2 statistic) = 1 + ($h_g^2 N/M$)LDscore, where

M = #markers, N = #samples, LDscore(SNP m) = $\sum_{m'} r^2(m, m')$

Thus, average χ^2 = 1 + ($h_g^2 N/M$)LDscore$_{avg}$ = 1 + $h_g^2 N/M_{eff}$

Inflation is not a bad thing! (see Yang et al. 2011 Eur J Hum Genet)



CONFOUNDING

also see Yang et al. 2014 Nat Genet

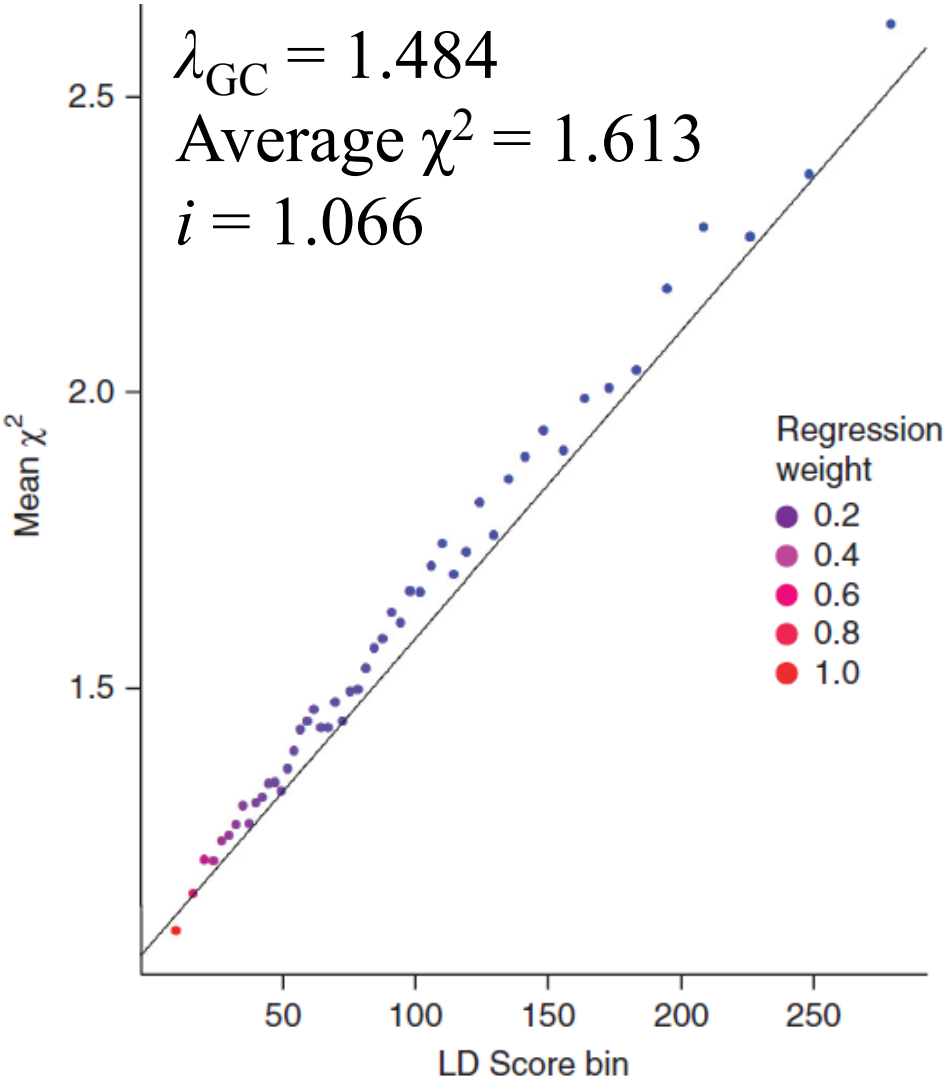
LD score regression: calibrating test statistics

LDscore(SNP m)
$= \sum_{m'} r^2(m, m')$

Regress
$\chi^2 = i + s$*LDscore

Intercept i should be
= 1 if no confounding,
> 1 if confounding

(from Tue of Week 3)



Bulik-Sullivan, Loh et al. 2015 Nat Genet; SCZ data from PGC-SCZ 2014 Nature

Powerful linear-time mixed model association!

Mixed model association $\chi^2 = \dfrac{(X_m{}^T Y_{resid})^2}{(Y_{resid})^T A^* (Y_{resid})}$

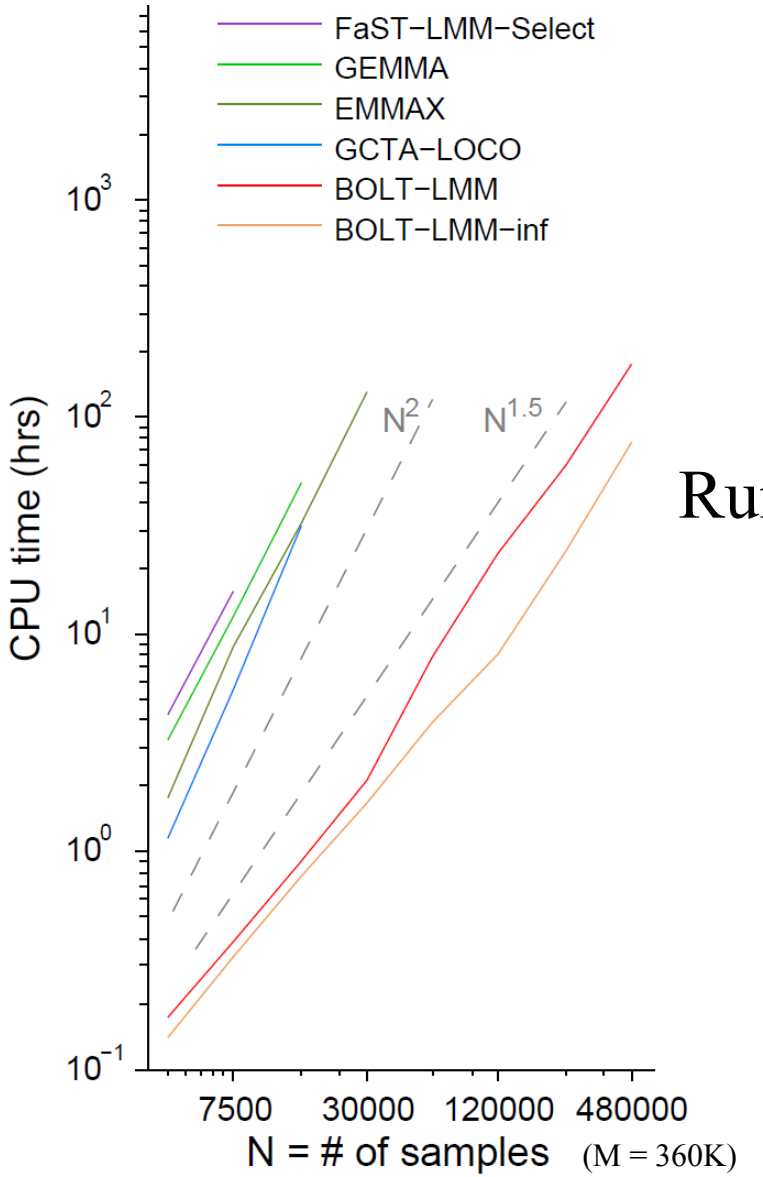
BOLT-LMM:

i. Compute the numerator in linear time

ii. Use constant denominator for calibration (LDscore regression)

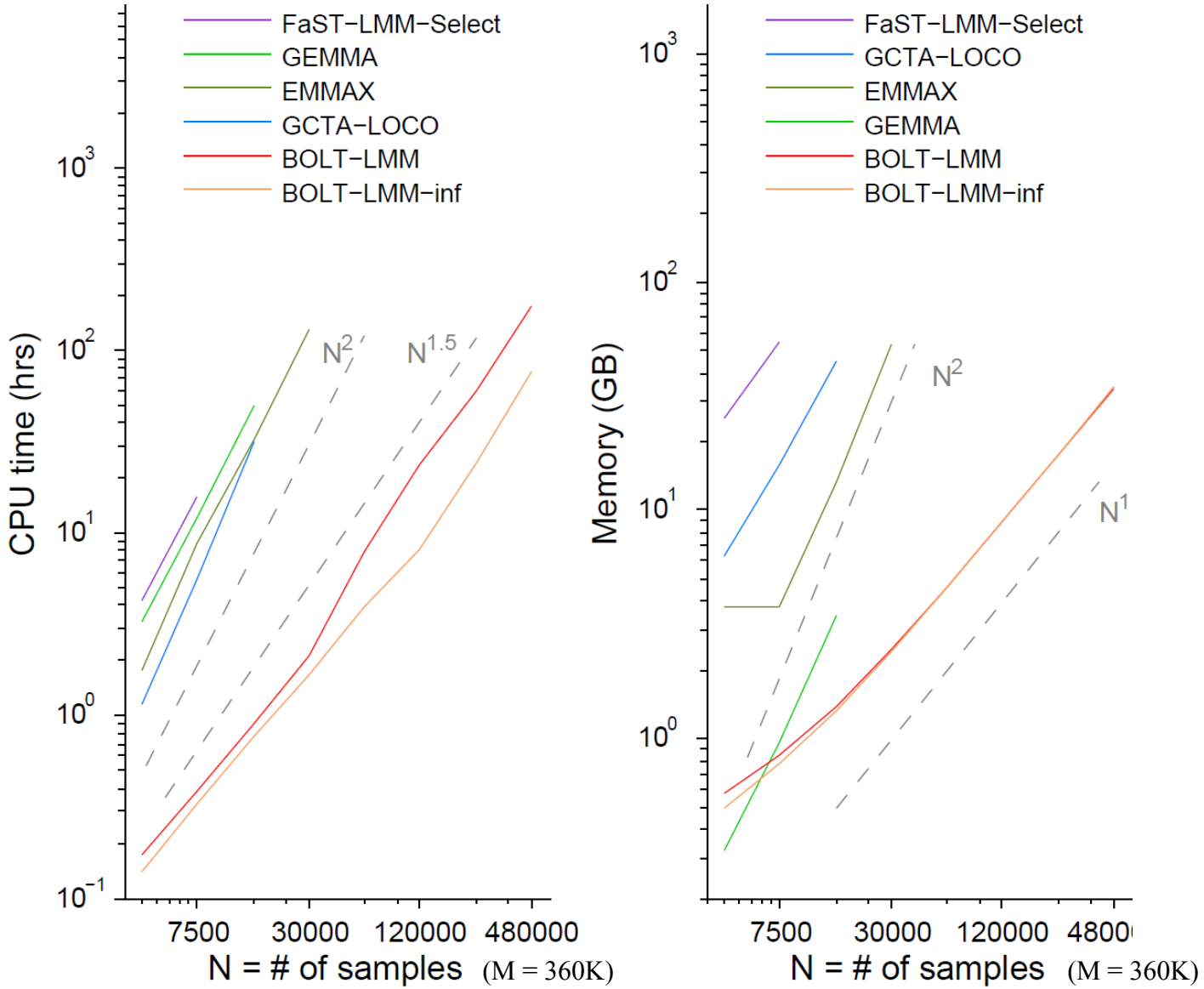
Loh, Tucker et al. 2015 Nat Genet

Simulations: BOLT-LMM is fast …

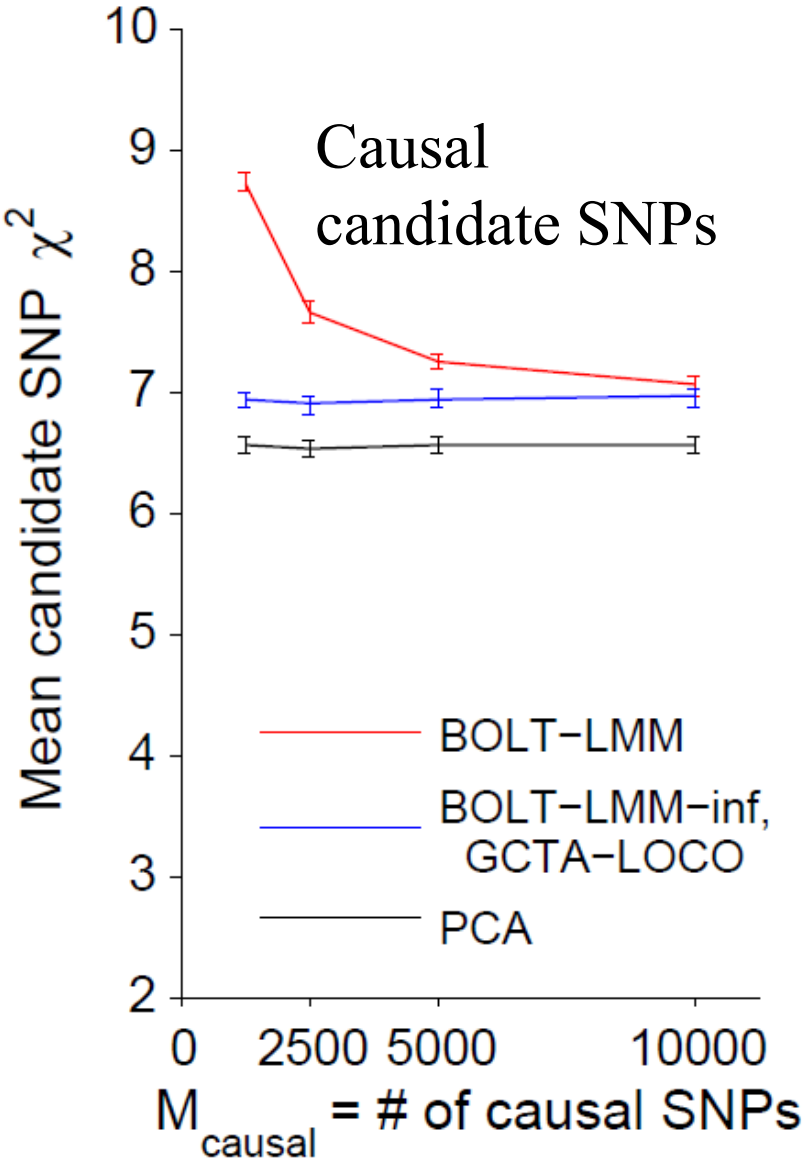


Running time ≈ O($MN^{1.5}$)

Simulations: BOLT-LMM is fast, and has low memory usage

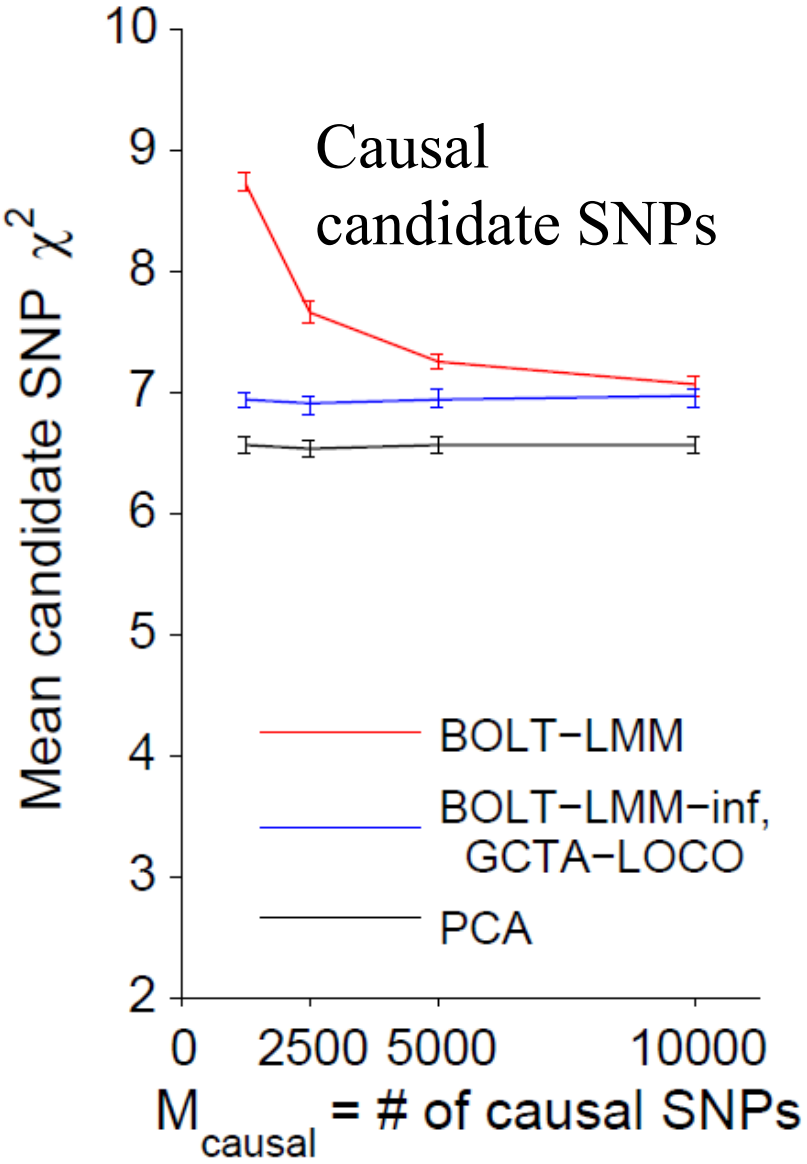


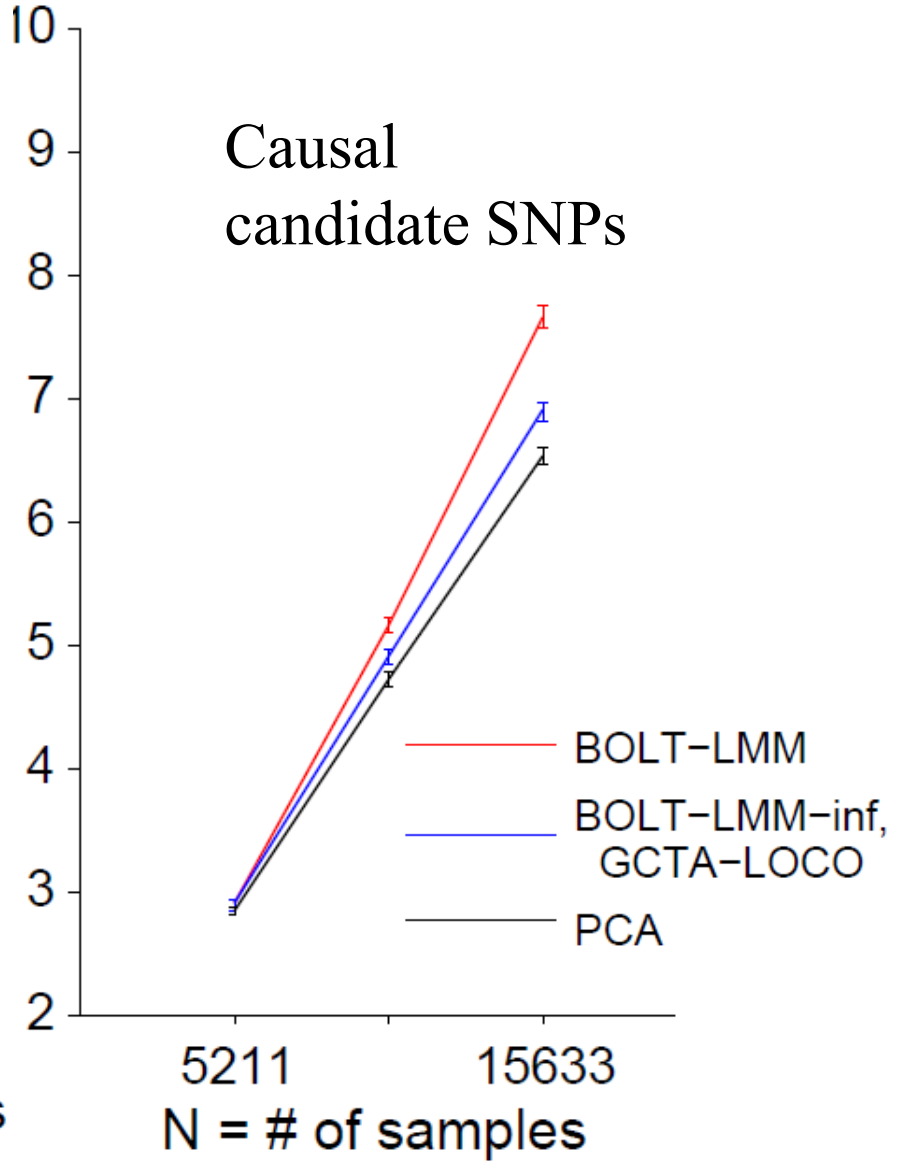
Simulations: BOLT-LMM is powerful …



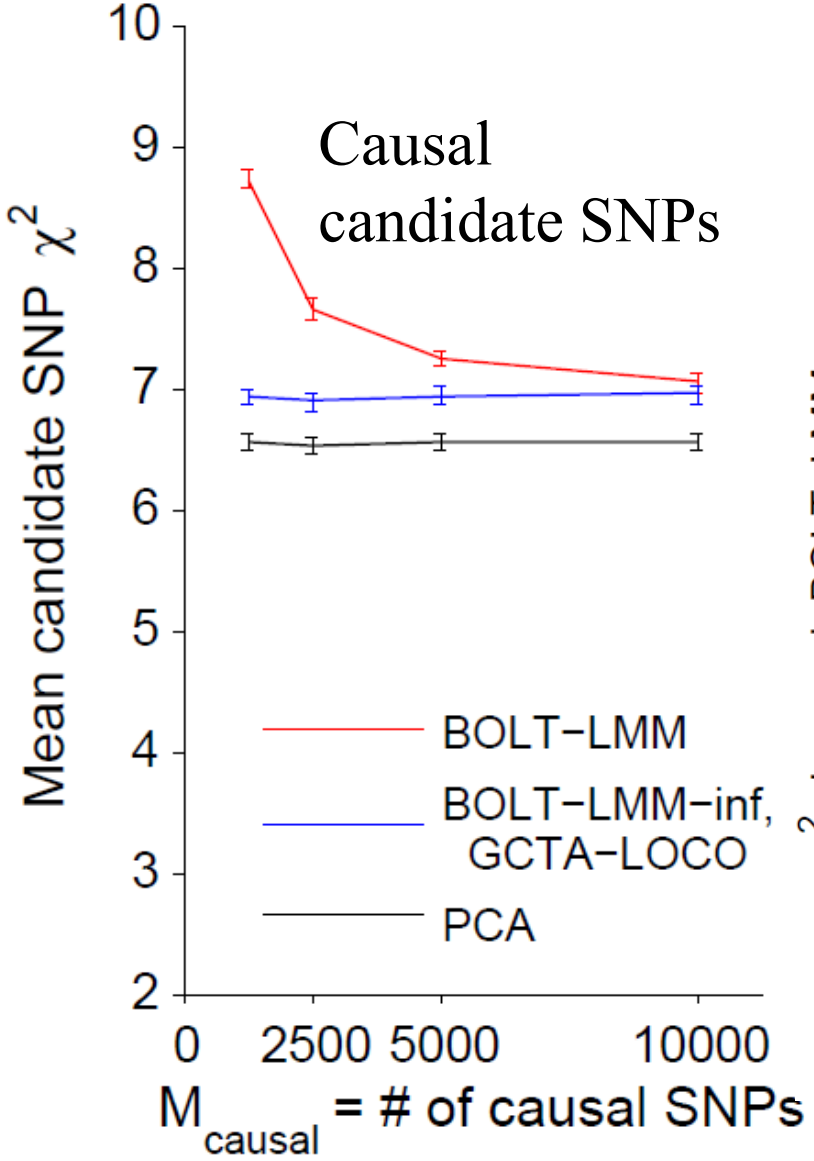
(M = 360K, N = 16K)

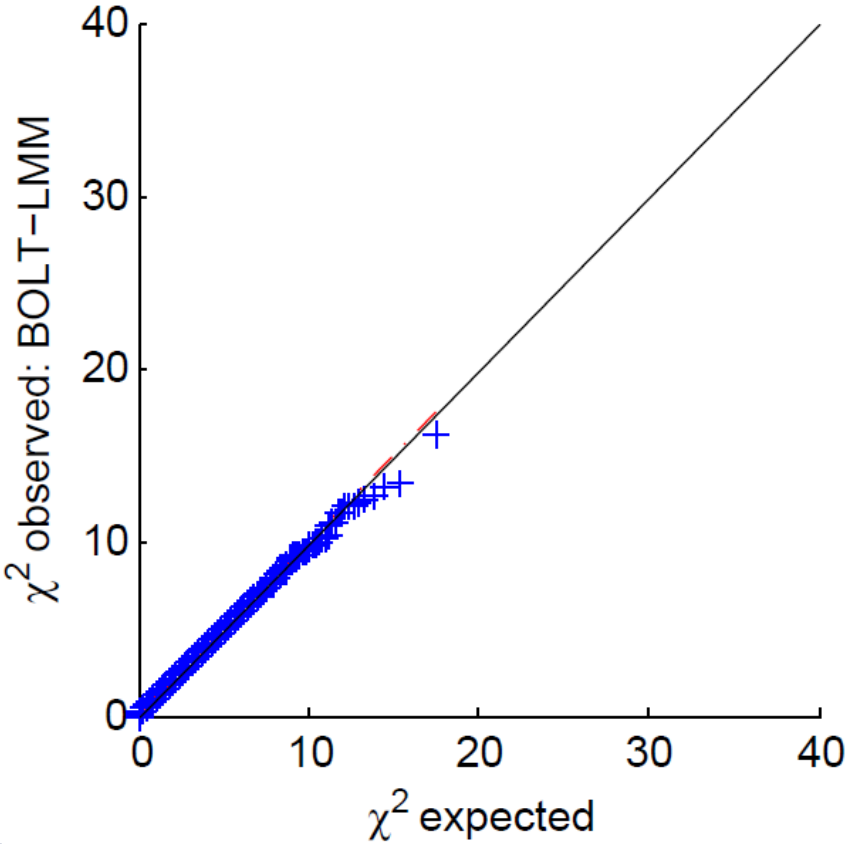
Simulations: BOLT-LMM is powerful …





Simulations: BOLT-LMM is powerful, and correctly calibrated at null SNPs





(M = 360K, N = 16K)

Approaches to Scientific Research

Just Dance.

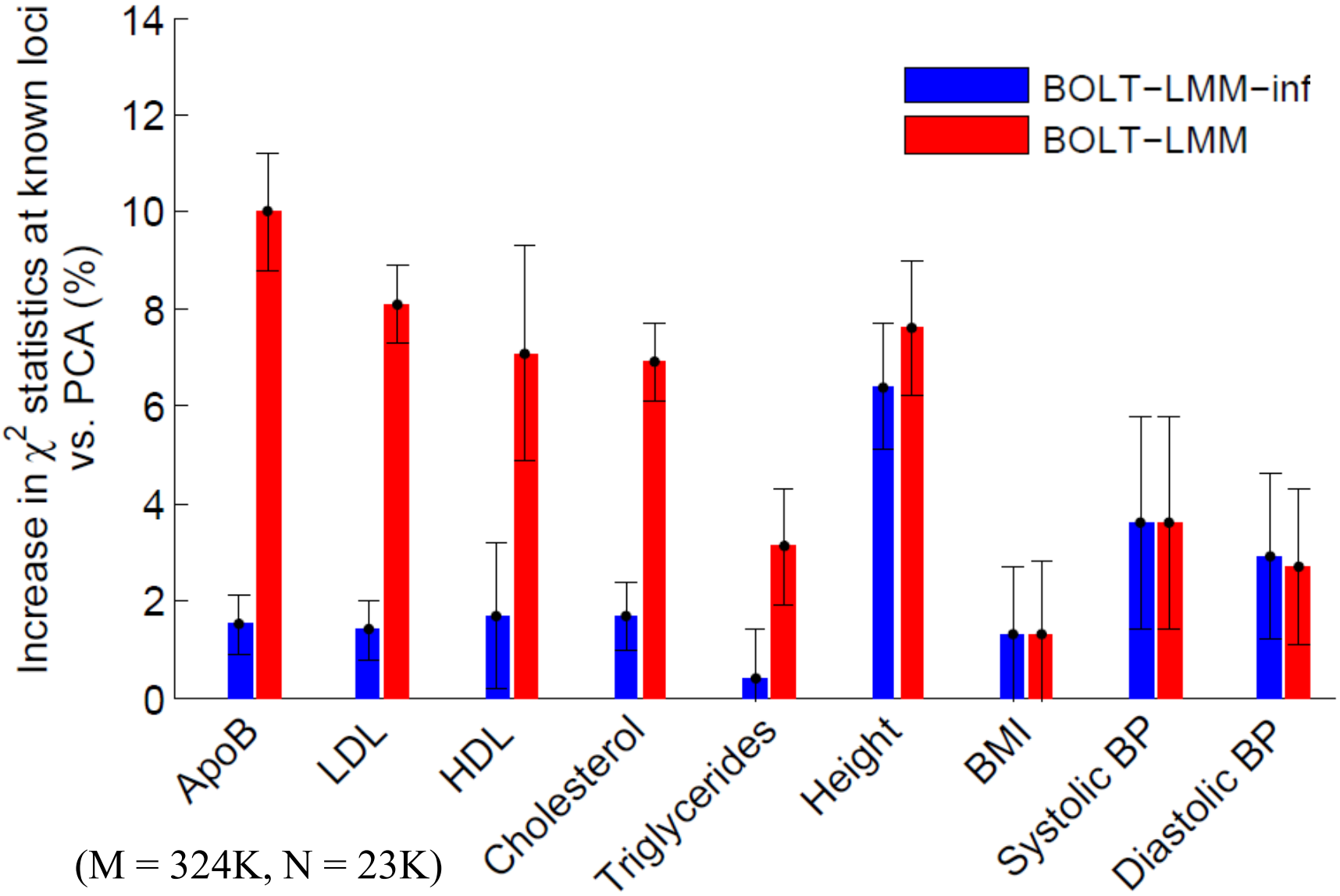
-- Gaga

Just Data.

-- Alkes

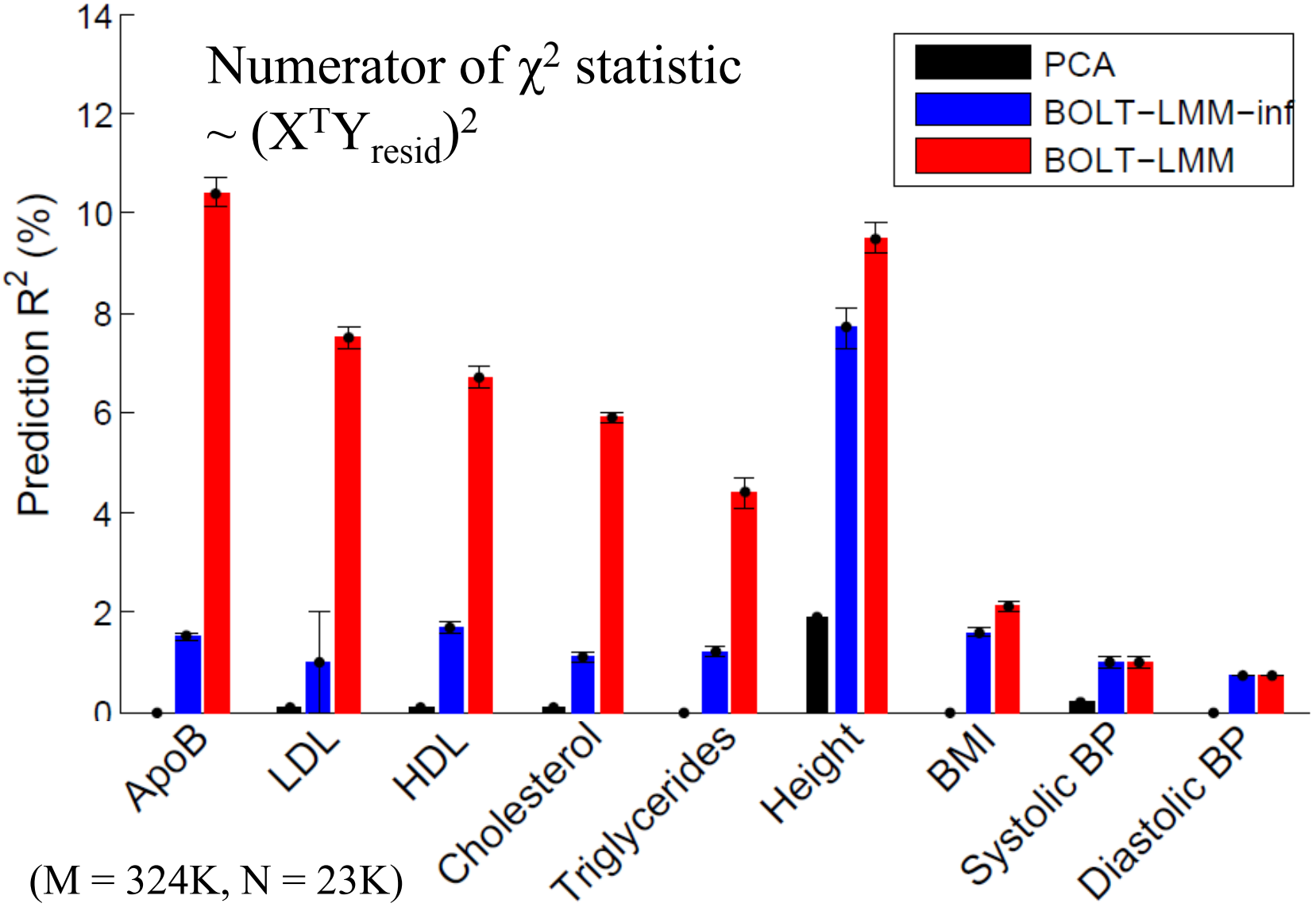
http://www.youtube.com/watch?v=Fl4L4M8m4d0

WGHS phenotypes: BOLT-LMM increases power at known associated loci



(M = 324K, N = 23K)

WGHS phenotypes: BOLT-LMM increases power ⇔ prediction r^2



(M = 324K, N = 23K)

Mixed model association has advantages and pitfalls

Pitfalls: Standard mixed model association methods can suffer from

- suboptimal power due to inclusion of candidate marker in GRM
- suboptimal power due to not modeling sparse polygenic architectures
- **a loss in power in ascertained case-control studies**

Yang et al. 2014 Nat Genet

Modeling case-control ascertainment increases mixed model association power

Mixed Model with Correction for Case-Control Ascertainment Increases Association Power

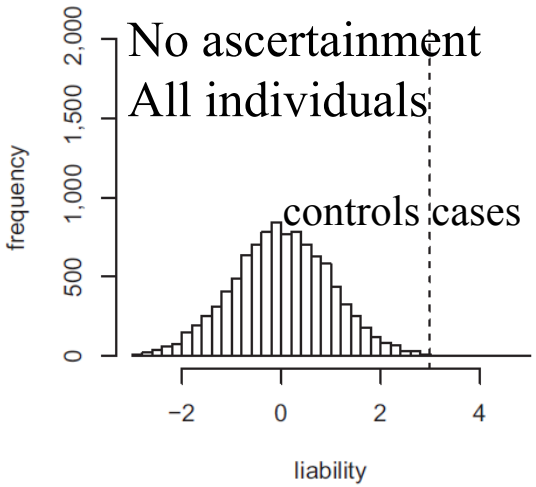
Tristan J. Hayeck,[1,2,*] Noah A. Zaitlen,[3] Po-Ru Loh,[2,4] Bjarni Vilhjalmsson,[2,4] Samuela Pollack,[2,4] Alexander Gusev,[2,4] Jian Yang,[5,6] Guo-Bo Chen,[5] Michael E. Goddard,[7] Peter M. Visscher,[5,6] Nick Patterson,[2] and Alkes L. Price[1,2,4,*]

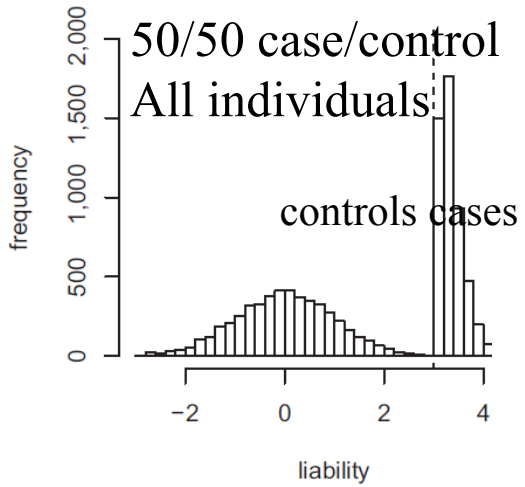
Accurate liability estimation improves power in ascertained case-control studies

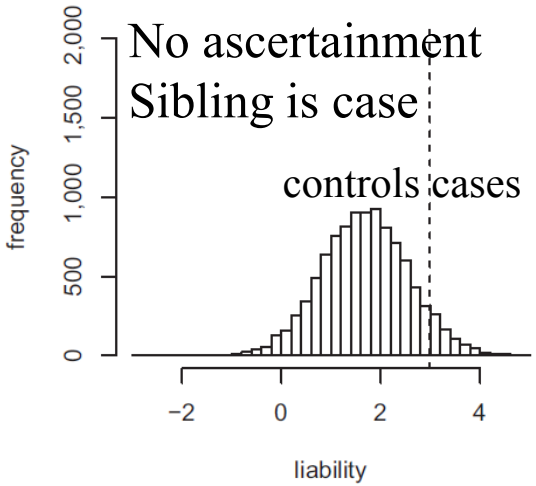
Omer Weissbrod[1], Christoph Lippert[2], Dan Geiger[1] & David Heckerman[2]

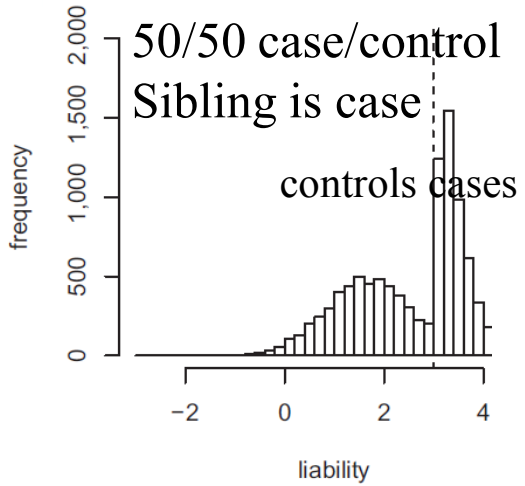
Hayeck et al. 2015 Am J Hum Genet, Weissbrod et al. 2015 Nat Methods
also see Hayeck et al. 2017 Am J Hum Genet

Modeling case-control ascertainment increases mixed model association power









Under liability threshold model: distribution of an individual's liability depends on relatedness to other cases/controls.

Hayeck et al. 2015 Am J Hum Genet, Weissbrod et al. 2015 Nat Methods
also see Hayeck et al. 2017 Am J Hum Genet

Modeling case-control ascertainment increases mixed model association power

- Estimate posterior mean liabilities (PML) conditional on liability-scale phenotypic covariance matrix V_{liab}, via MCMC.
- Compute χ^2 statistic proportional to $(X^T V_{liab}^{-1} PML)^2$
  (X = candidate SNP genotypes, PML = posterior mean liabilities)
- Retrospective score statistic enables appropriate treatment of case-control ascertainment, increasing power.

Simulations: 2%-26% improvement in χ^2 statistics, depending on sample size (5K-50K) and disease prevalence (0.1%-1%).

WTCCC2 MS data set (*N*=10K): 4% improvement, consistent with simulations (larger % improvement expected at larger sample sizes).

Hayeck et al. 2015 Am J Hum Genet, Weissbrod et al. 2015 Nat Methods
also see Hayeck et al. 2017 Am J Hum Genet

Conclusions

- Mixed model association methods are a promising approach for correcting for confounding and increasing power.
- At large sample sizes, to maximize power it is important to exclude the candidate SNP from the GRM (MLMe) instead of including the candidate SNP from the GRM (MLMi).
- Mixed model association can be performed in $\approx MN^{1.5}$ time by using an iterative scheme to compute BLUP residuals, since mixed model association = association on BLUP residuals.
- Using a normal-mixture prior on effect sizes increases power while preserving $\approx MN^{1.5}$ running time.

EPI511, Advanced Population and Medical Genetics

Week 6:

- **Mixed model association**
- Rare variant analysis

Final project: due date is officially Mar 10 at 5pm, but anytime before Mar 13 at 6am is ok.

EPI511, Advanced Population and Medical Genetics

Week 6:

- Mixed model association
- **Rare variant analysis**

Outline

1. Properties of rare and low-frequency variants
2. Rare variant association tests: methods
3. Rare variant association tests: results
4. Rare variant heritability

Outline

1. Properties of rare and low-frequency variants

2. Rare variant association tests: methods

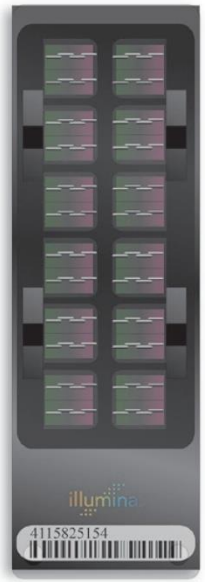
3. Rare variant association tests: results

4. Rare variant heritability

Common Disease/Common Variant hypothesis

"For common diseases, there will be one or a few predominating disease alleles with relatively high frequencies at each of the major underlying disease loci"

(from Tue of Week 1)



Lander 1996 Science; Reich & Lander 2001 Trends Genet
reviewed in Gibson 2012 Nat Rev Genet, Visscher et al. 2012 Am J Hum Genet

Are rare and low-frequency variants important?

Five Years of GWAS Discovery

Peter M. Visscher,[1,2,*] Matthew A. Brown,[1] Mark I. McCarthy,[3,4] and Jian Yang[5]

Introduction: Have GWASs Been a Failure?

From McCLellan and King, *Cell* 2010[1]: If common alleles influenced common diseases, many would have been found by now. The issue is not how to develop still larger studies, or how to parse the data still further, but rather whether the common disease–common variant hypothesis has now been tested and found not to apply to most complex human diseases."

(from Tue of Week 1)

Visscher et al. 2012 Am J Hum Genet

Are rare and low-frequency variants important?

GENOME-WIDE ASSOCIATION STUDIES

Rare and common variants: twenty arguments

Greg Gibson

Abstract | Genome-wide association studies have greatly improved our understanding of the genetic basis of disease risk. The fact that they tend not to identify more than a fraction of the specific causal loci has led to divergence of opinion over whether most of the variance is hidden as numerous rare variants of large effect or as common variants of very small effect. Here I review 20 arguments for and against each of these models of the genetic basis of complex traits and conclude that both classes of effect can be readily reconciled.

(from Tue of Week 1)

Gibson 2012 Nat Rev Genet

The 1000 Genomes (1000G) Project

Sequence the entire genomes of 1,092 individuals:
379 of European ancestry (Europe and USA)
286 of East Asian ancestry (Asia)
246 of African ancestry (Africa and USA)
181 of Latino ancestry (Latin America and USA)

Use next-generation sequencing technologies (~4x coverage):
e.g. Illumina, 454, SOLiD (read lengths 25-400bp)

(Metzker 2010 Nat Rev Genet, Davey et al. 2011 Nat Rev Genet,
also see Nielsen et al. 2011 Nat Rev Genet)

(from Tue of Week 1)

1000 Genomes Project Consortium 2012 Nature

1000G project: Summary of main results

- 38 million SNPs discovered and successfully genotyped. Most of these are rare and low-frequency variants.

- The 38 million SNPs include
  99.7% of all SNPs with minor allele frequency 5%
  98% of all SNPs with minor allele frequency 1% ***
  50% of all SNPs with minor allele frequency 0.1%
  based on an independent UK European sample.

  ***: stated goal to identify >95% of SNPs with frequency 1% was successfully achieved.

  (from Tue of Week 1)

1000 Genomes Project Consortium 2012 Nature

1000G project: the final phase

Sequence the entire genomes of 2,504 individuals:
503 of European ancestry (Europe and USA)
504 of East Asian ancestry (Asia)
661 of African ancestry (Africa and USA)
347 of Latino ancestry (Latin America and USA)
489 of South Asian ancestry (South Asia and USA)

Use next-generation sequencing technologies (~7x coverage):
Illumina only (read lengths 70-400bp only)
85 million SNPs, of which 64 million have MAF<0.5%

(from Tue of Week 1)

1000 Genomes Project Consortium 2015 Nature; also see UK10K Consortium 2015 Nature, Gudbjartsson et al. 2015 Nat Genet, McCarthy et al. 2016 Nat Genet

Other recent WGS reference panels

UK10K project (UK10K Consortium 2015 Nature):

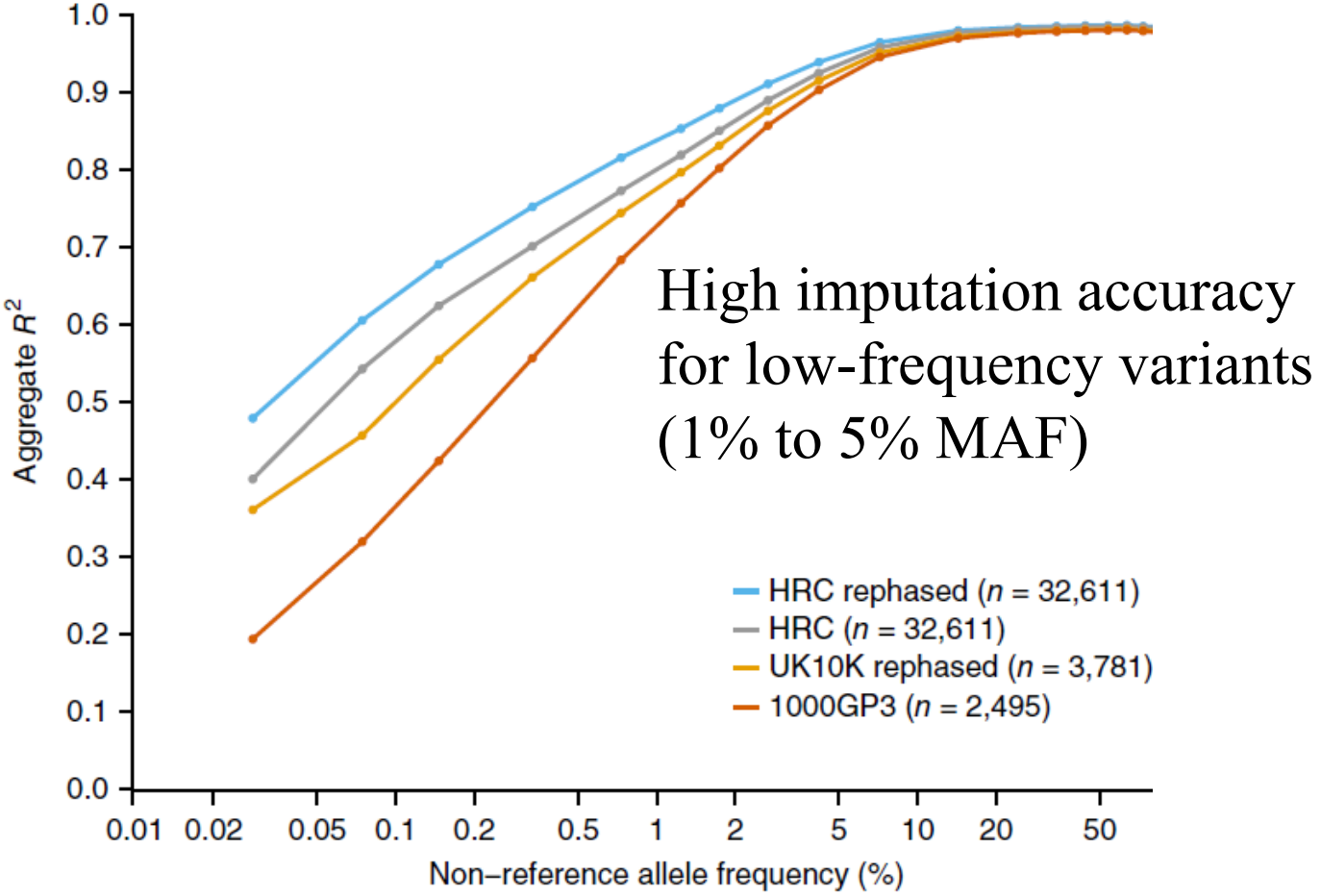
7x WGS of 3,781 UK samples

Improved imputation accuracy vs. 1000G (Huang et al. 2015 Nat Commun)

(also see Genome of Netherlands Consortium 2014 Nat Genet)

Other recent WGS reference panels

Haplotype Reference Consortium (McCarthy et al. 2016 Nat Genet):
4-8x WGS of 32,488 mostly European samples from 20 studies
Available for imputation via Michigan and Sanger imputation servers



High imputation accuracy
for low-frequency variants
(1% to 5% MAF)

Other recent WGS reference panels

UK10K project (UK10K Consortium 2015 Nature):
7x WGS of 3,781 UK samples
Improved imputation accuracy vs. 1000G (Huang et al. 2015 Nat Commun)
(also see Genome of Netherlands Consortium 2014 Nat Genet)

Haplotype Reference Consortium (McCarthy et al. 2016 Nat Genet):
4-8x WGS of 32,488 mostly European samples from 20 studies
Available for imputation via Michigan and Sanger imputation servers

deCODE Genetics WGS data set (Gudbjartsson et al. 2015 Nat Genet):
20x WGS of 2,636 Icelanders
Accurate long-range phasing of WGS reference panel enables accurate imputation down to 0.1% MAF in the Icelandic population

What about rare variants?

- The 1000G project has identified most low-frequency variants (minor allele frequency 1%-5%). These variants can be placed on genotyping arrays or imputed (see Thu of Week 1)

- Rare variants: most have not been identified by 1000 Genomes!
  Must sequence disease samples directly.
  Past focus has been mostly on exome sequencing, but now shifting to whole-genome sequencing.

  (**from Tue of Week 1**)

Kiezun et al. 2012 Nat Genet, Tennessen et al. 2012 Science, Pasaniuc et al. 2012 Nat Genet, Purcell et al. 2014 Nature, Do et al. 2015 Nature, Cai et al. 2015 Nature. Reviewed in Goldstein et al. 2013 Nat Rev Genet, Lee et al. 2014 Am J Hum Genet, Zuk et al. 2014 PNAS

What about rare variants?

- The 1000G project has identified most low-frequency variants (minor allele frequency 1%-5%). These variants can be placed on genotyping arrays or imputed (see Thu of Week 1)

- Rare variants: most have not been identified by 1000 Genomes! Must sequence disease samples directly. **Or impute?** Past focus has been mostly on exome sequencing, but now shifting to whole-genome sequencing.

  (**from Tue of Week 1**)

Kiezun et al. 2012 Nat Genet, Tennessen et al. 2012 Science, Pasaniuc et al. 2012 Nat Genet, Purcell et al. 2014 Nature, Do et al. 2015 Nature, Cai et al. 2015 Nature. Reviewed in Goldstein et al. 2013 Nat Rev Genet, Lee et al. 2014 Am J Hum Genet, Zuk et al. 2014 PNAS

African populations have more genetic diversity

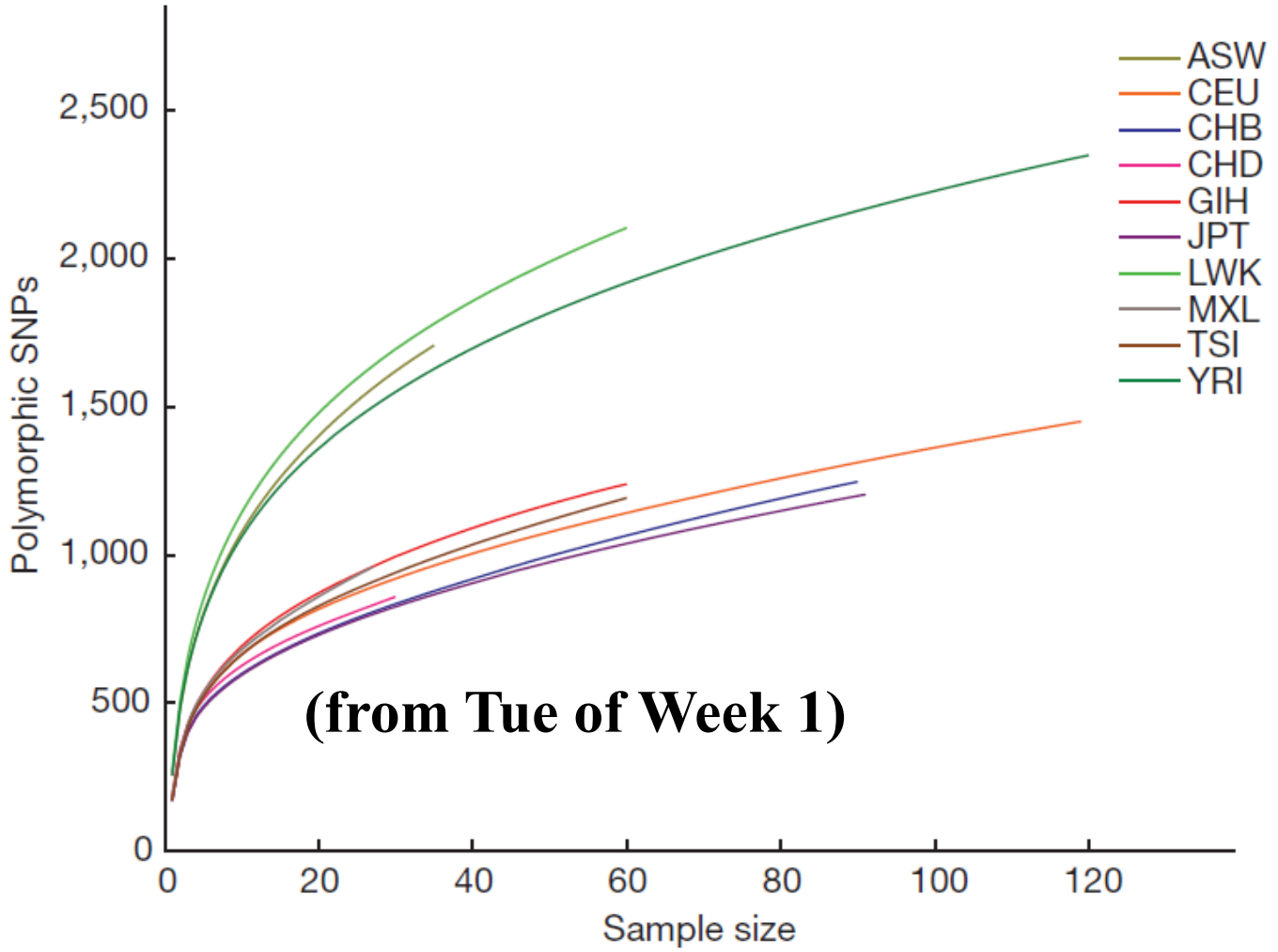
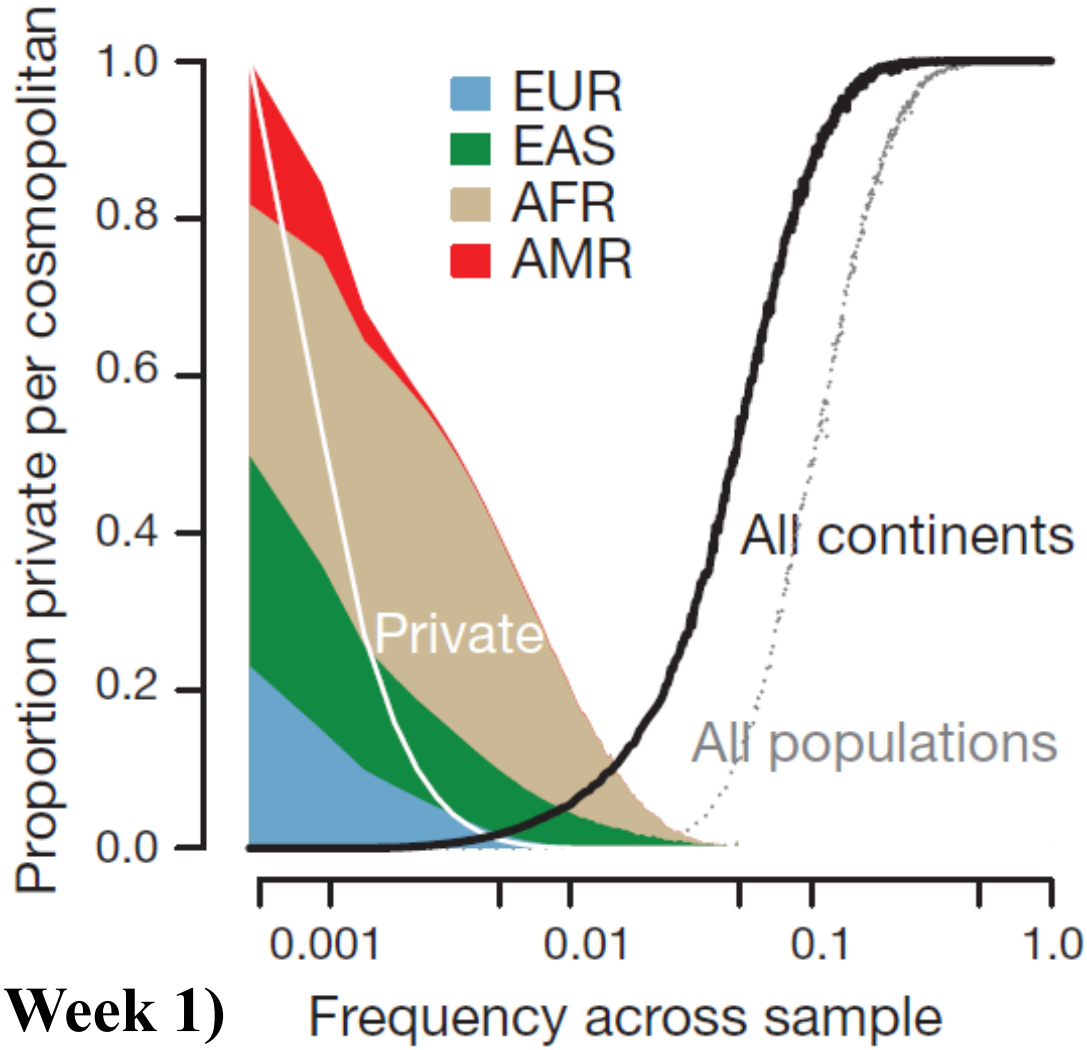


Figure 3 | Effect of sample size on SNP ascertainment.

International HapMap3 Consortium 2010 Nature

Common variants are shared across populations, but rare variants are often population-private



(from Tue of Week 1)

1000 Genomes Project Consortium 2012 Nature

African populations have more genetic diversity, and rare variants are often population-private

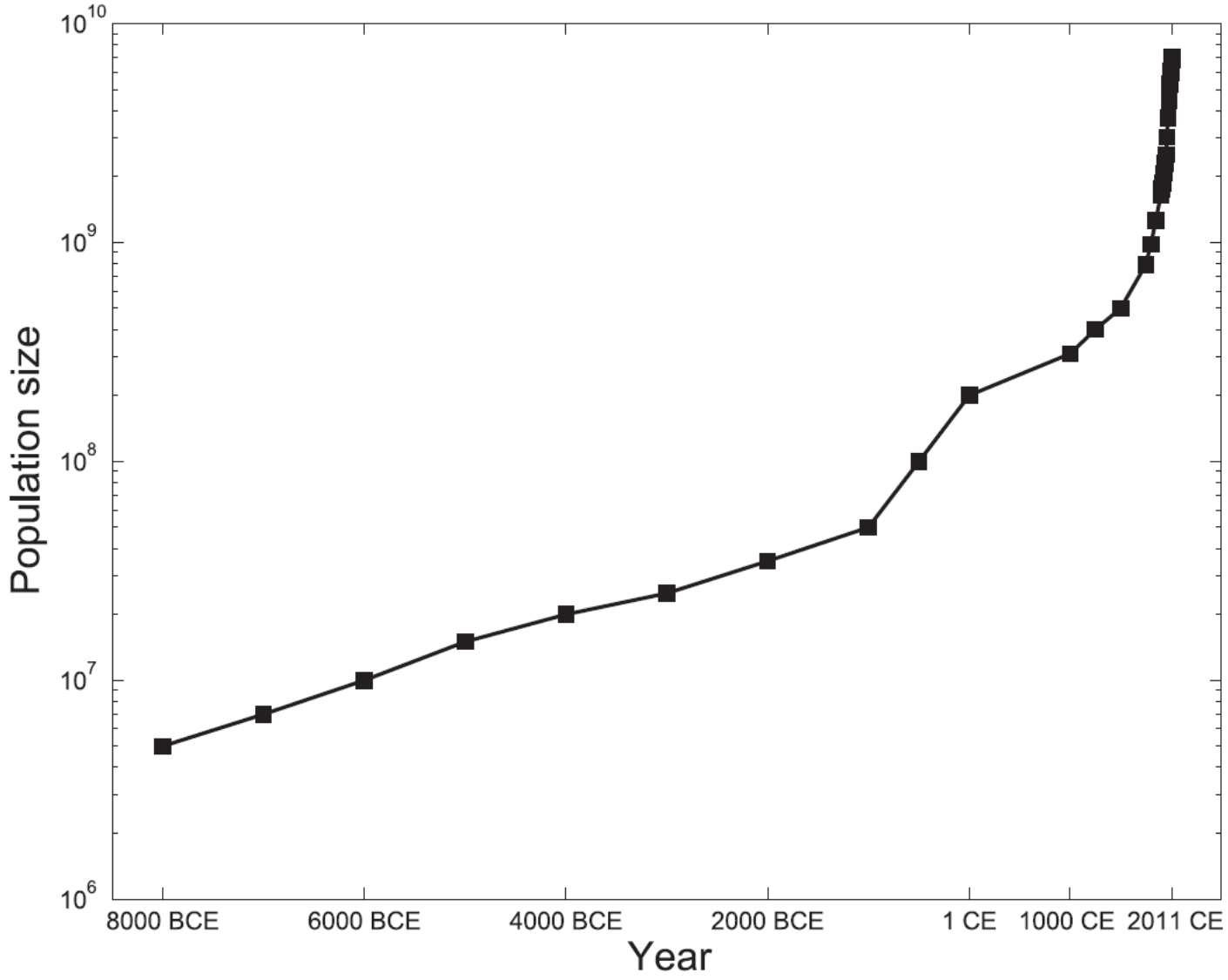
Whole-exome sequencing (NHLBI Exome Sequencing Project) of 1,351 European Americans (EA) + 1,088 African Americans (AA):
[Note: African Americans inherit both African and European ancestry]

- 503,481 SNPs total: 86% rare (MAF<0.5%), 57% singleton*
  18% observed in both EA and AA
  35% EA-specific
  47% AA-specific
- 217,624 non-singleton SNPs
  42% observed in both EA and AA
  15% EA-specific
  43% AA-specific

*Greater than expected under standard model; consistent with recent population growth

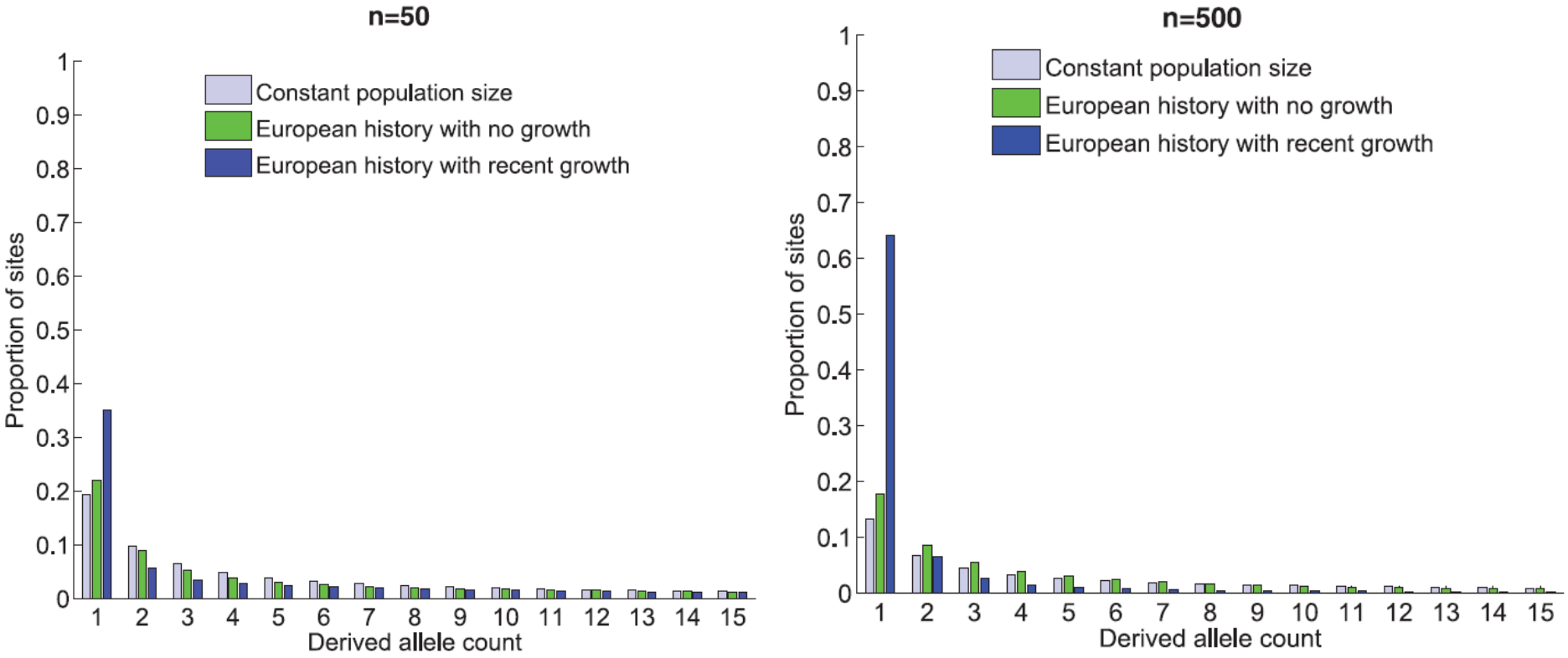
Tennessen et al. 2012 Science; also see Fu et al. 2013 Nature (6,515 exomes), Lek et al. 2016 Nature (60,706 exomes), Dewey et al. 2016 Science (50,726 exomes)

Recent population growth is accelerating



Keinan & Clark 2012 Science

Recent population growth implies more rare variants

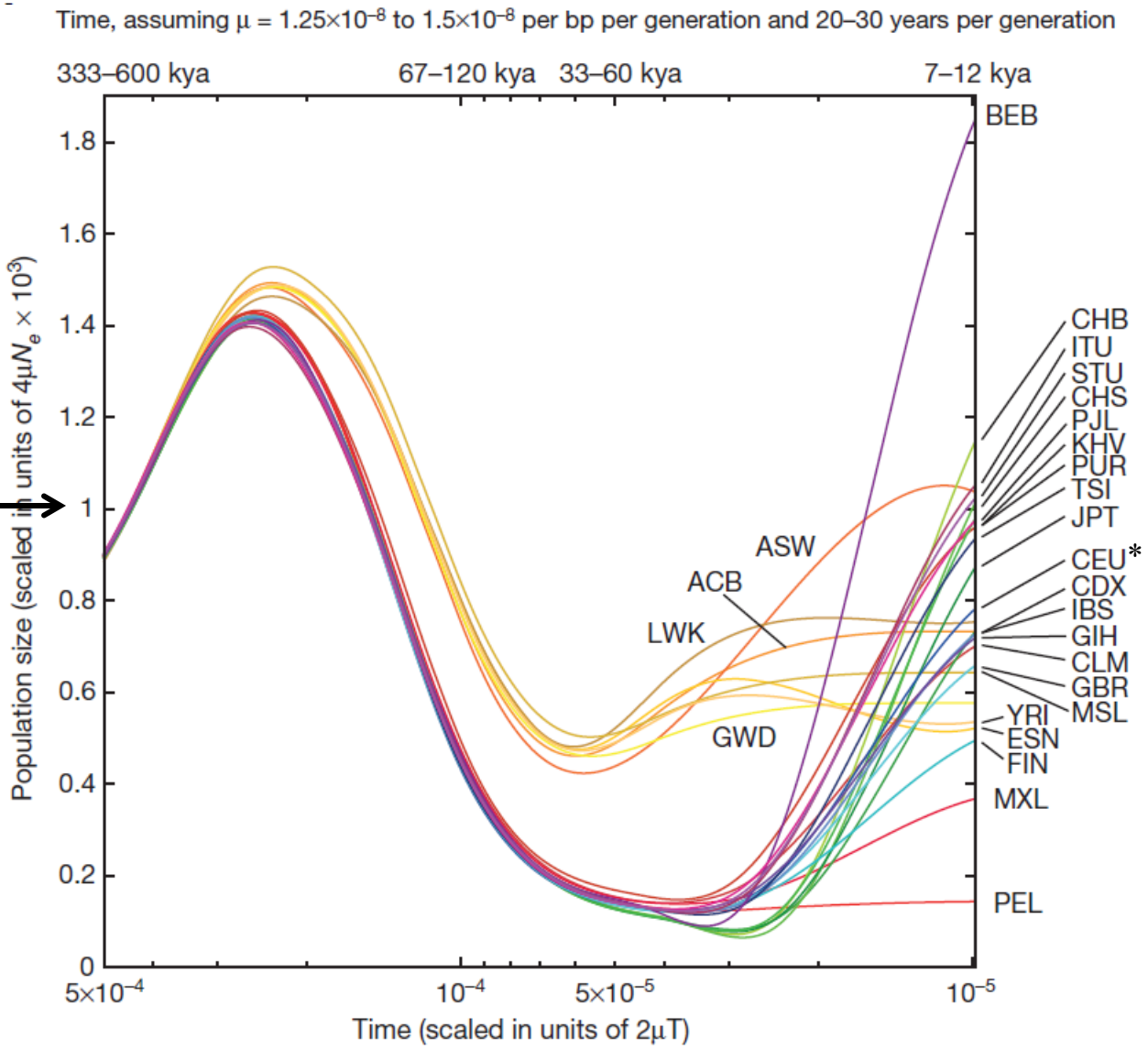


"European history with no growth" population sizes from Keinan et al. 2007 Nat Genet; also see Li & Durbin 2011 Nature, Gravel et al. 2011 PNAS,
1000 Genomes Project Consortium 2015 Nature, Mallick et al. 2016 Nature

Keinan & Clark 2012 Science

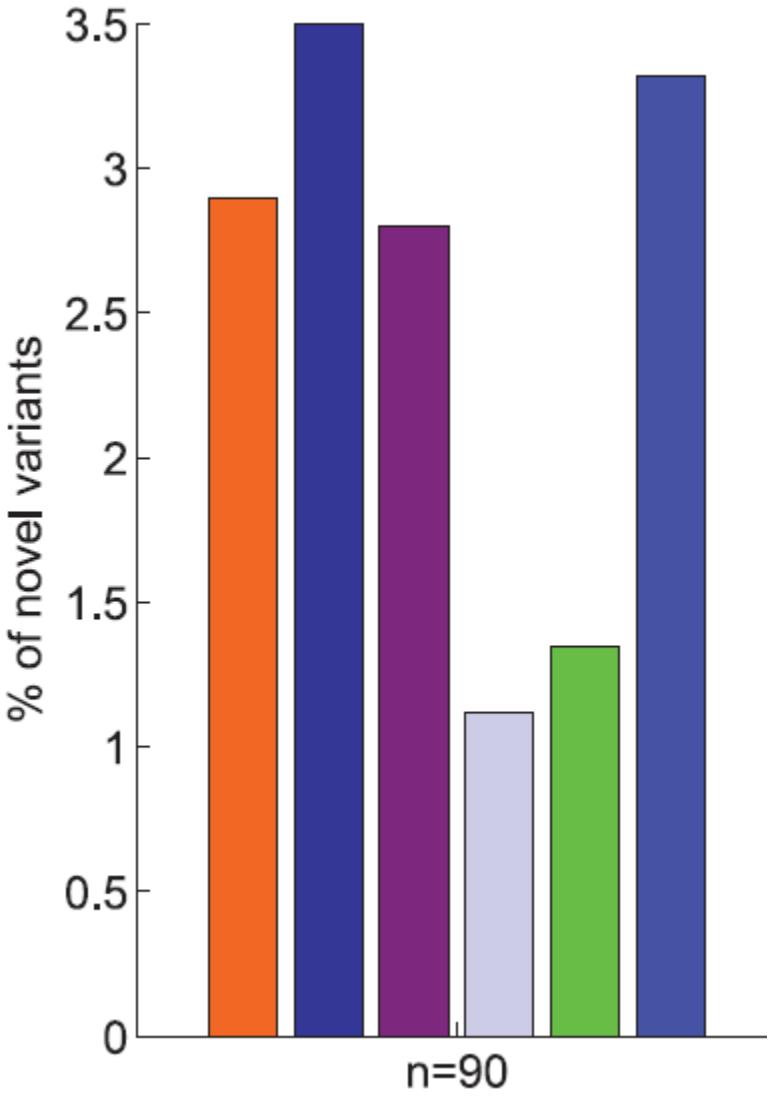
What does "European history with no growth" mean?

$N_e \approx 15K$ →



1000 Genomes Project Consortium 2015 Nature (PSMC method; Li & Durbin 2011 Nature)

Recent population growth implies more rare variants



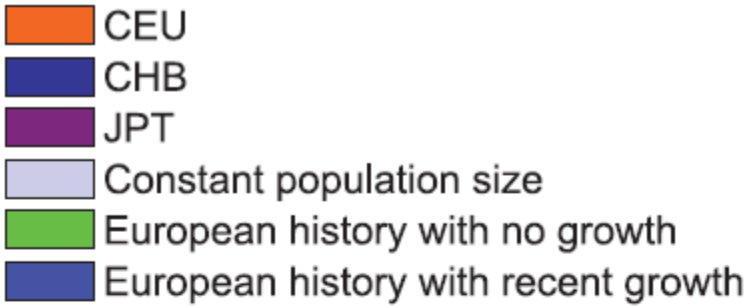
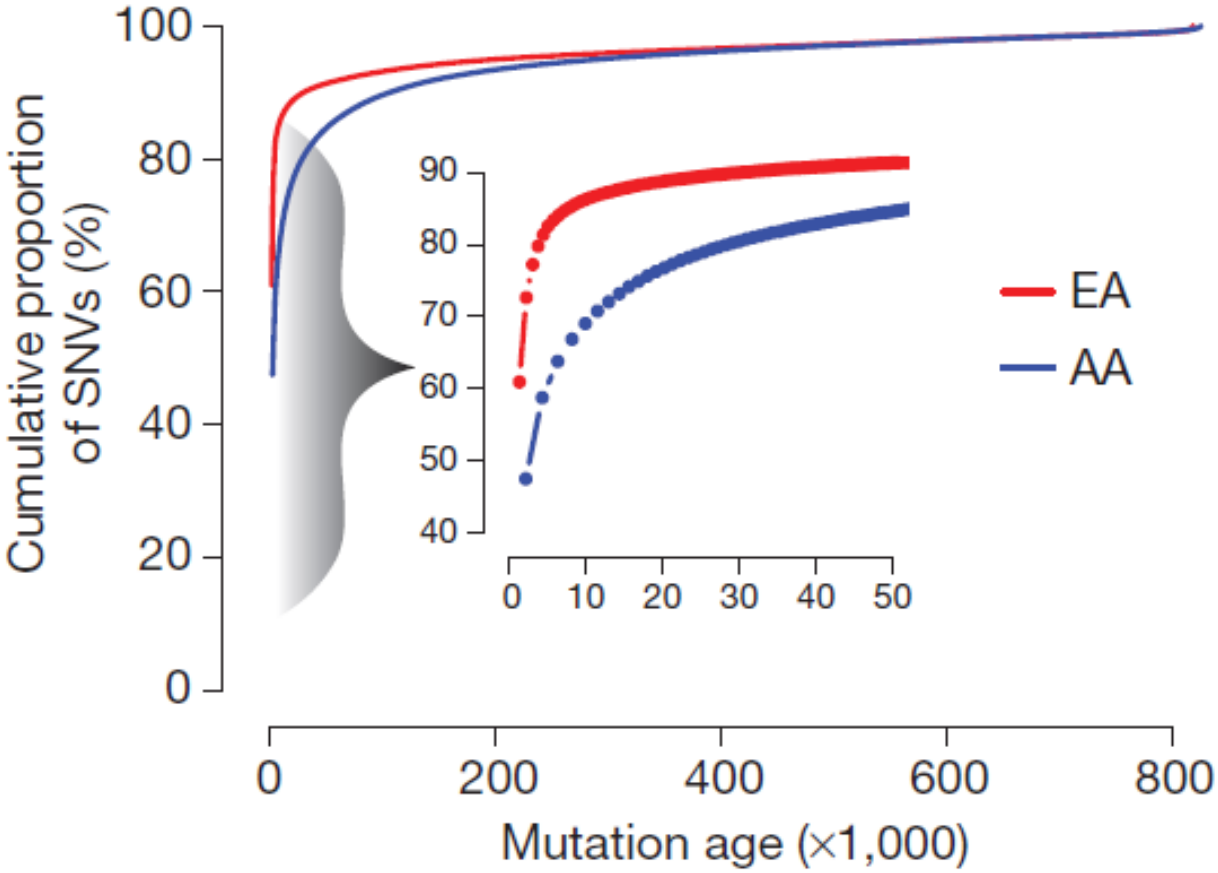


Fig. 3. Percentage of novel variants in a sequenced individual. Bars denote the fraction, out of all sites for which a given individual is heterozygous, that are also monomorphic in a separate sample of 90 individuals from the same population.

Keinan & Clark 2012 Science

Recent population growth implies more rare variants, most of which have arisen in the past 5,000 years

Whole-exome sequencing (NHLBI Exome Sequencing Project) of 4,298 European Americans (EA) + 2,217 African Americans (AA): 1,146,401 SNPs total. 73% are predicted to be <5,000 years old.



Fu et al. 2013 Nature

African populations have more genetic diversity

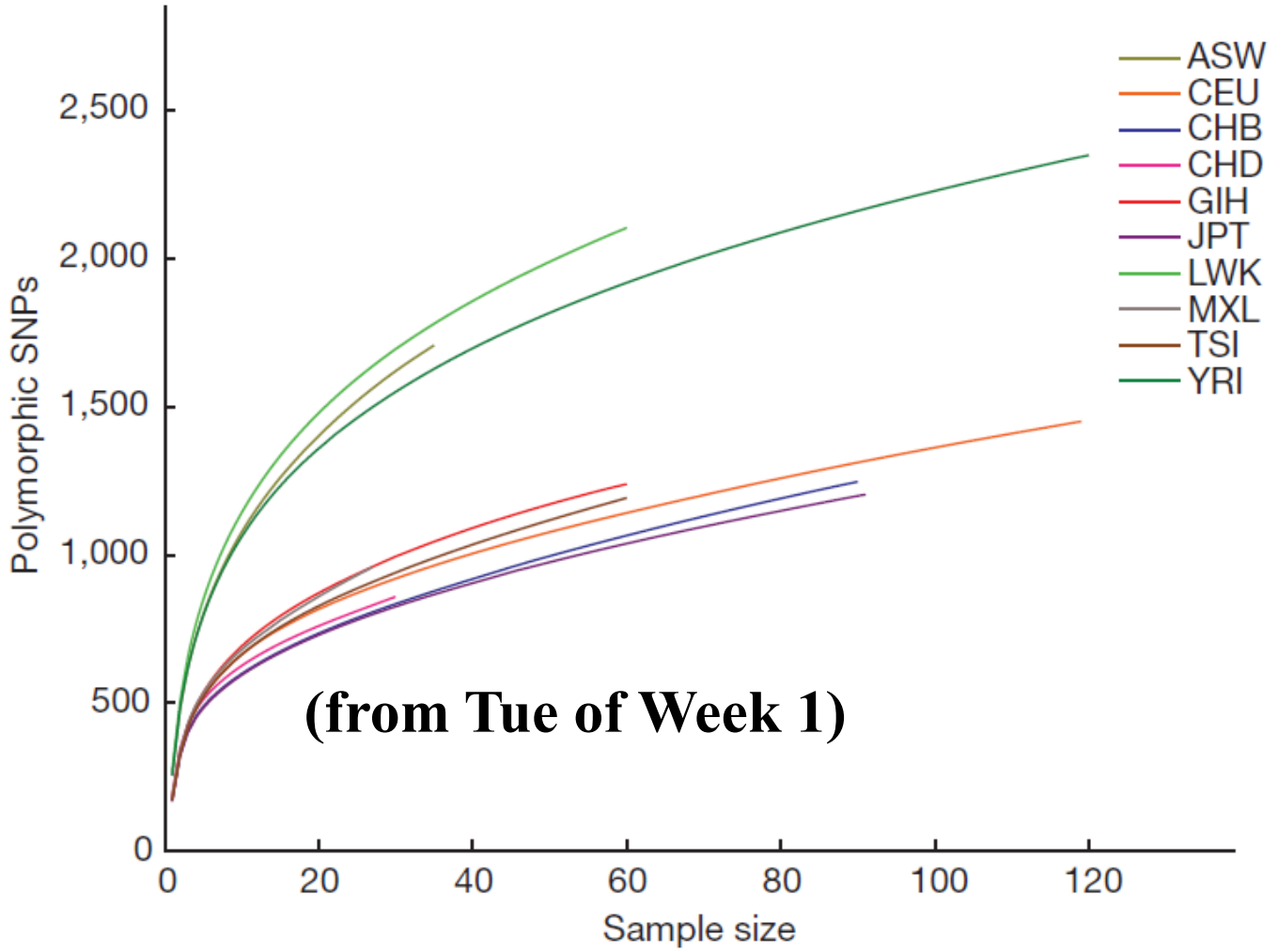


Figure 3 | Effect of sample size on SNP ascertainment.

International HapMap3 Consortium 2010 Nature

Populations **within** Europe have varying genetic diversity

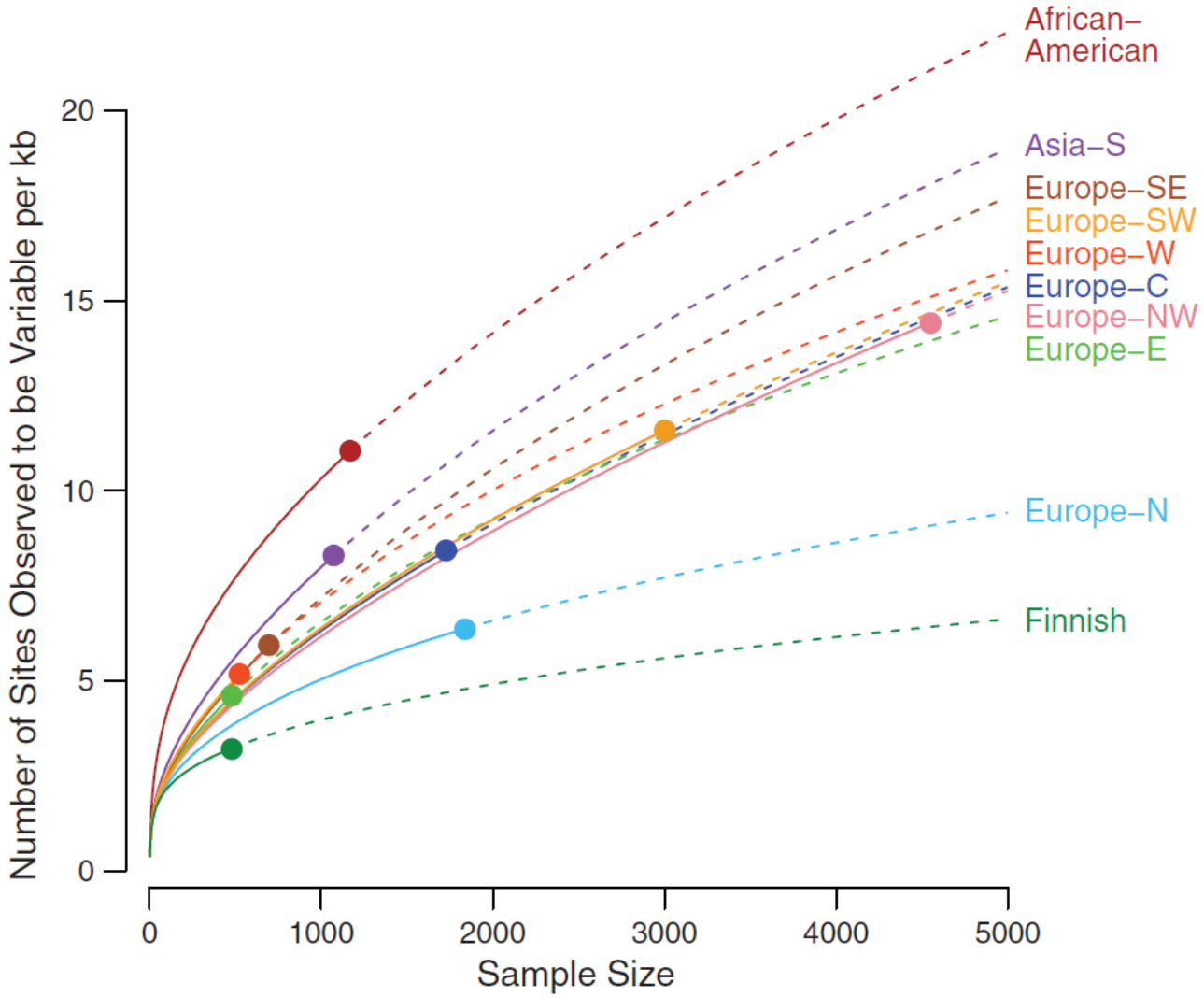
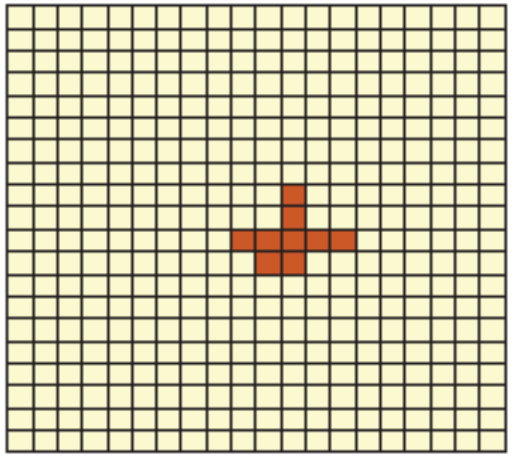


Fig. 3. Number of variants per kilobase of sequence with sample sizes increasing to 5000 people

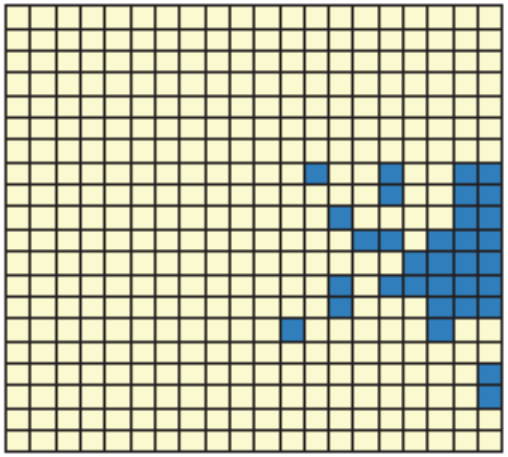
Nelson et al. 2012 Science; also see Ralph & Coop 2013 PLoS Biol

Rare variants are geographically localized (because they are more recent)

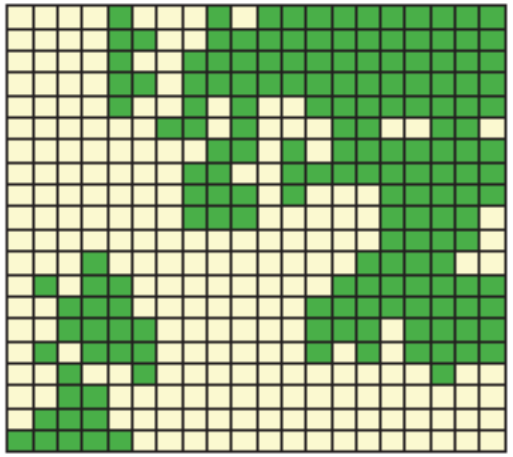
Examples from simulations of rare, low-frequency, common variants:



Rare variant



Low-frequency variant

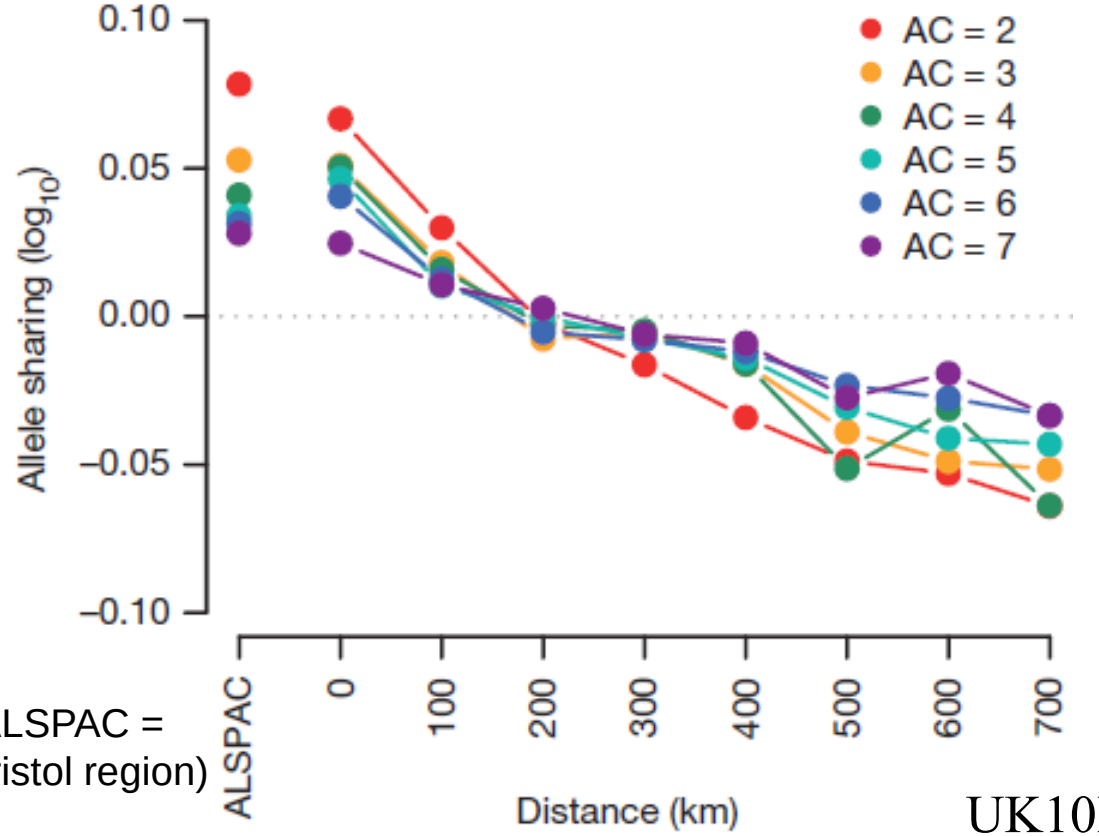


Common variant

Mathieson & McVean 2012 Nat Genet

Rare variants are geographically localized within the UK (because they are more recent)

Allele sharing ratio ($\log_{10}$): how much more likely are 2 individuals at a given geographic distance to share a derived allele compared to expectation for a homogeneous population?



(ALSPAC = Bristol region)

UK10K Consortium 2015 Nature

also see Genome of Netherlands Consortium 2014 Nat Genet

Outline

1. Properties of rare and low-frequency variants

2. Rare variant association tests: methods

3. Rare variant association tests: results

4. Rare variant heritability

Single-variant tests: no new methods needed

A mutation in *APP* protects against Alzheimer's disease and age-related cognitive decline

Jonsson et al. 2012 Nature

Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion

Huyghe et al. 2013 Nat Genet

Exome-wide association study identifies a *TM6SF2* variant that confers susceptibility to nonalcoholic fatty liver disease

Kozlitina et al. 2014 Nat Genet

Rare variant in scavenger receptor BI raises HDL cholesterol and increases risk of coronary heart disease

Zanoni et al. 2016 Science

Mendelian disease: not our main focus

Exome sequencing identifies the cause of a mendelian disorder

Ng et al. 2010a Nat Genet

Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome

Ng et al. 2010b Nat Genet

Whole-exome sequencing identifies recessive *WDR62* mutations in severe brain malformations

Bilguvar et al. 2010 Nature

Exome sequencing as a tool for Mendelian disease gene discovery

reviewed in Bamshad et al. 2011 Nat Rev Genet

Rare variant association tests

Burden tests:

Fixed threshold: Li & Leal 2008 Am J Hum Genet

Weighted test: Madsen & Browning 2009 PLoS Genet

Variable threshold: Price et al. 2010 Am J Hum Genet

Overdispersion tests:

C-alpha: Neale et al. 2011 PLoS Genet

SKAT: Wu et al. 2011 Am J Hum Genet

Combined burden/overdispersion tests:

SKAT-O: Lee et al. 2012 Am J Hum Genet

[gene-based tests, multiple rare coding variants, complex diseases/traits]

reviewed in Lee et al. 2014 Am J Hum Genet

Rare variant association tests

Burden tests:

Fixed threshold: Li & Leal 2008 Am J Hum Genet
Weighted test: Madsen & Browning 2009 PLoS Genet
Variable threshold: Price et al. 2010 Am J Hum Genet

Burden tests assume that all rare variants in a candidate gene have the same direction of effect (e.g. all rare variants increase disease risk).

reviewed in Lee et al. 2014 Am J Hum Genet

Rare variant association tests

Burden tests:

Fixed threshold:	Li & Leal 2008 Am J Hum Genet
Weighted test:	Madsen & Browning 2009 PLoS Genet
Variable threshold:	Price et al. 2010 Am J Hum Genet

Burden tests assume that all rare variants in a candidate gene have the same direction of effect (e.g. all rare variants increase disease risk).

reviewed in Lee et al. 2014 Am J Hum Genet

Fixed threshold test: aggregate all rare variants below a MAF threshold

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

Let $C_{gj} = \sum_{i \in g,\, p_i < T} X_{ij}$ (or "collapse": $C_{gj} = 1$ if at least one $X_{ij} > 0$, or 0 otherwise)

= sum of genotypes of SNPs in gene g in individual j, restricting to SNPs i with MAF $p_i <$ fixed threshold T.

Test association between counts C_{gj} and phenotypes Y_j.

Li & Leal 2008 Am J Hum Genet
also see Morris & Zeggini 2010 Genet Epidemiol

Fixed threshold test: aggregate all rare variants below a MAF threshold

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

Let $C_{gj} = \sum_{i \in g,\, p_i < T} X_{ij}$ (or "collapse": $C_{gj} = 1$ if at least one $X_{ij} > 0$, or 0 otherwise)

= sum of genotypes of SNPs in gene g in individual j, restricting to SNPs i with MAF $p_i <$ fixed threshold T.

Test association between counts C_{gj} and phenotypes Y_j.

To evaluate statistical significance: 3 possibilities:

i. $N\rho(C_g, Y)^2$ is χ^2(1 dof), generalizing Armitage Trend Test.

[Note: $N\rho(C_g, Y)^2$ = square of weighted sum of z-scores $z_i = \sqrt{N}\,\rho(X_i, Y)$]

Li & Leal 2008 Am J Hum Genet

also see Morris & Zeggini 2010 Genet Epidemiol

Fixed threshold test: aggregate all rare variants below a MAF threshold

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

Let $C_{gj} = \sum_{i \in g,\, p_i < T} X_{ij}$ (or "collapse": $C_{gj} = 1$ if at least one $X_{ij} > 0$, or 0 otherwise)

= sum of genotypes of SNPs in gene g in individual j, restricting to SNPs i with MAF $p_i <$ fixed threshold T.

Test association between counts C_{gj} and phenotypes Y_j.

To evaluate statistical significance: 3 possibilities:

i. $N\rho(C_g, Y)^2$ is χ^2(1 dof), generalizing Armitage Trend Test.

[Note: $N\rho(C_g, Y)^2$ = square of weighted sum of z-scores $z_i = \sqrt{N}\rho(X_i, Y)$]

ii. Permutation test: compute test statistics with permuted phenotypes.

Li & Leal 2008 Am J Hum Genet

also see Morris & Zeggini 2010 Genet Epidemiol

Fixed threshold test: aggregate all rare variants below a MAF threshold

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

Let $C_{gj} = \sum_{i \in g,\, p_i < T} X_{ij}$ (or "collapse": $C_{gj} = 1$ if at least one $X_{ij} > 0$, or 0 otherwise)

= sum of genotypes of SNPs in gene g in individual j, restricting to SNPs i with MAF $p_i <$ fixed threshold T.

Test association between counts C_{gj} and phenotypes Y_j.

To evaluate statistical significance: 3 possibilities:

i. $N\rho(C_g, Y)^2$ is χ^2(1 dof), generalizing Armitage Trend Test.

[Note: $N\rho(C_g, Y)^2$ = square of weighted sum of z-scores $z_i = \sqrt{N}\,\rho(X_i, Y)$]

ii. Permutation test: compute test statistics with permuted phenotypes.

iii. Combine (collapsed) rare + common variants via Hotelling's T^2 test

Li & Leal 2008 Am J Hum Genet

also see Morris & Zeggini 2010 Genet Epidemiol

Rare variant association tests

Burden tests:

Fixed threshold:	Li & Leal 2008 Am J Hum Genet
Weighted test:	Madsen & Browning 2009 PLoS Genet
Variable threshold:	Price et al. 2010 Am J Hum Genet

Burden tests assume that all rare variants in a candidate gene have the same direction of effect (e.g. all rare variants increase disease risk).

reviewed in Lee et al. 2014 Am J Hum Genet

Weighted test: weight based on MAF

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

Let $C_{gj} = \sum_{i \in g} w_i X_{ij}$, where $w_i = 1/\sqrt{p_i(1-p_i)}$

= weighted sum of genotypes of SNPs in gene g in individual j.

[Note: this weighting schemes assumes effect sizes $\sim 1/\sqrt{p_i(1-p_i)}$]

Test association between counts C_{gj} and phenotypes Y_j.

Madsen & Browning 2009 PLoS Genet

Weighted test: weight based on MAF

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

Let $C_{gj} = \sum_{i \in g} w_i X_{ij}$, where $w_i = 1/\sqrt{p_i(1-p_i)}$

= weighted sum of genotypes of SNPs in gene g in individual j.

[Note: this weighting schemes assumes effect sizes $\sim 1/\sqrt{p_i(1-p_i)}$]

Test association between counts C_{gj} and phenotypes Y_j.

<u>To evaluate statistical significance</u>:

i. $N\rho(C_g, Y)^2$ is χ^2(1 dof), generalizing Armitage Trend Test.

[Note: $N\rho(C_g, Y)^2$ = square of weighted sum of z-scores $z_i = \sqrt{N}\,\rho(X_i, Y)$]

Madsen & Browning 2009 PLoS Genet

Weighted test: weight based on MAF

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

Let $C_{gj} = \sum_{i \in g} w_i X_{ij}$, where $w_i = 1/\sqrt{p_i(1-p_i)}$

= weighted sum of genotypes of SNPs in gene g in individual j.

[Note: this weighting schemes assumes effect sizes $\sim 1/\sqrt{p_i(1-p_i)}$]

Test association between counts C_{gj} and phenotypes Y_j.

<u>To evaluate statistical significance</u>:

i. $N\rho(C_g, Y)^2$ is χ^2(1 dof), generalizing Armitage Trend Test.

[Note: $N\rho(C_g, Y)^2$ = square of weighted sum of z-scores $z_i = \sqrt{N}\,\rho(X_i, Y)$]

ii. Permutation test: compute test statistics with permuted phenotypes.

Madsen & Browning 2009 PLoS Genet

Weighted test: weight based on MAF

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

Let $C_{gj} = \sum_{i \in g} w_i X_{ij}$, where $w_i = 1/\sqrt{p_i(1-p_i)}$

= weighted sum of genotypes of SNPs in gene g in individual j.

[Note: this weighting schemes assumes effect sizes $\sim 1/\sqrt{p_i(1-p_i)}$]

Test association between counts C_{gj} and phenotypes Y_j.

To evaluate statistical significance:

i. $N\rho(C_g, Y)^2$ is χ^2(1 dof), generalizing Armitage Trend Test.

[Note: $N\rho(C_g, Y)^2$ = square of weighted sum of z-scores $z_i = \sqrt{N}\,\rho(X_i, Y)$]

ii. Permutation test: compute test statistics with permuted phenotypes.

Madsen & Browning 2009 PLoS Genet

Rare variant association tests

Burden tests:

Fixed threshold:	Li & Leal 2008 Am J Hum Genet
Weighted test:	Madsen & Browning 2009 PLoS Genet
Variable threshold:	Price et al. 2010 Am J Hum Genet

Burden tests assume that all rare variants in a candidate gene have the same direction of effect (e.g. all rare variants increase disease risk).

reviewed in Lee et al. 2014 Am J Hum Genet

Variable threshold test: aggregate rare variants below *varying* MAF thresholds

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

For each possible MAF threshold T (e.g. $0.00 < T \leq 0.05$)

Let $C_{gj}(T) = \sum_{i \in g, p_i < T} X_{ij}$

= sum of genotypes of SNPs in gene g in individual j, restricting to SNPs i with MAF $p_i <$ threshold T.

Test association between counts $C_{gj}(T)$ and phenotypes Y_j.

Price et al. 2010 Am J Hum Genet

Variable threshold test: aggregate rare variants below *varying* MAF thresholds

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

For each possible MAF threshold T (e.g. $0.00 < T \leq 0.05$)

Let $C_{gj}(T) = \sum_{i \in g, p_i < T} X_{ij}$

= sum of genotypes of SNPs in gene g in individual j, restricting to SNPs i with MAF $p_i <$ threshold T.

Test association between counts $C_{gj}(T)$ and phenotypes Y_j.

<u>To evaluate statistical significance</u>:

Let $z(T) = \sqrt{N}\,\rho(C_g(T), Y)$ be a z-score for threshold T.

Let $z_{\max}$ be the maximum $z(T)$ across all thresholds T.

Permutation test: compute $z_{\max}$ with permuted phenotypes.

Price et al. 2010 Am J Hum Genet

Rare variant association tests

Burden tests:

Fixed threshold: Li & Leal 2008 Am J Hum Genet
Weighted test: Madsen & Browning 2009 PLoS Genet
Variable threshold: Price et al. 2010 Am J Hum Genet

Overdispersion tests:

C-alpha: Neale et al. 2011 PLoS Genet
SKAT: Wu et al. 2011 Am J Hum Genet

Overdispersion tests assume that rare variants in a candidate gene can have varying direction of effect (e.g. increase or decrease disease risk).

reviewed in Lee et al. 2014 Am J Hum Genet

Rare variant association tests

Burden tests:

Fixed threshold:	Li & Leal 2008 Am J Hum Genet
Weighted test:	Madsen & Browning 2009 PLoS Genet
Variable threshold:	Price et al. 2010 Am J Hum Genet

Overdispersion tests:

C-alpha:	Neale et al. 2011 PLoS Genet
SKAT:	Wu et al. 2011 Am J Hum Genet

Overdispersion tests assume that rare variants in a candidate gene can have varying direction of effect (e.g. increase or decrease disease risk).

reviewed in Lee et al. 2014 Am J Hum Genet

C-alpha test: are rare variant case/control counts overdispersed vs. binomial distribution

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

Let $n_i = \sum_j X_{ij}$ = total allele count of SNP i,

$n_{i,case} = \sum_{j \in case} X_{ij}$ = allele count of SNP i in disease cases,

π_{case} = proportion of individuals who are disease cases.

We expect $n_{i,case} \sim \text{Binomial}(n_i, \pi_{case})$. For rare variants (e.g. MAF<0.01):

Test overdispersion $T = \sum_i \left((n_{i,case} - n_i \pi_{case})^2 - n_i \pi_{case} (1 - \pi_{case}) \right)$.

Neale et al. 2011 PLoS Genet

C-alpha test: are rare variant case/control counts overdispersed vs. binomial distribution

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

Let $n_i = \sum_j X_{ij}$ = total allele count of SNP i,

$n_{i,case} = \sum_{j \in case} X_{ij}$ = allele count of SNP i in disease cases,

π_{case} = proportion of individuals who are disease cases.

We expect $n_{i,case} \sim \text{Binomial}(n_i, \pi_{case})$. For rare variants (e.g. MAF<0.01):

Test overdispersion $T = \sum_i \left((n_{i,case} - n_i \pi_{case})^2 - n_i \pi_{case}(1 - \pi_{case}) \right)$.

<u>To evaluate statistical significance</u>: 2 possibilities:

i. Assume T normally distributed, test $T > 0$ using one-tailed test.
(Caveat: theoretical variance does not account for LD between variants.)

Neale et al. 2011 PLoS Genet

C-alpha test: are rare variant case/control counts overdispersed vs. binomial distribution

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

Let $n_i = \sum_j X_{ij}$ = total allele count of SNP i,

$n_{i,case} = \sum_{j \in case} X_{ij}$ = allele count of SNP i in disease cases,

π_{case} = proportion of individuals who are disease cases.

We expect $n_{i,case} \sim \text{Binomial}(n_i, \pi_{case})$. For rare variants (e.g. MAF<0.01):

Test overdispersion $T = \sum_i \left((n_{i,case} - n_i \pi_{case})^2 - n_i \pi_{case} (1 - \pi_{case}) \right)$.

<u>To evaluate statistical significance</u>: 2 possibilities:

i. Assume T normally distributed, test $T > 0$ using one-tailed test.
(Caveat: theoretical variance does not account for LD between variants.)

ii. Permutation test: compute test statistics with permuted phenotypes.

Neale et al. 2011 PLoS Genet

Rare variant association tests

Burden tests:

Fixed threshold:	Li & Leal 2008 Am J Hum Genet
Weighted test:	Madsen & Browning 2009 PLoS Genet
Variable threshold:	Price et al. 2010 Am J Hum Genet

Overdispersion tests:

C-alpha:	Neale et al. 2011 PLoS Genet
SKAT:	Wu et al. 2011 Am J Hum Genet

Overdispersion tests assume that rare variants in a candidate gene can have varying direction of effect (e.g. increase or decrease disease risk).

reviewed in Lee et al. 2014 Am J Hum Genet

SKAT test: generalize C-alpha to model quantitative traits, LD between variants, etc.

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

Let $z_i = \sqrt{N}\rho(X_i, Y)$ be a z-score for SNP i.

Test statistic: $\sum_i w_i^2 z_i^2$, where weights $w_i \sim \text{Beta}(p_i; a_1, a_2)$

- recommended Beta parameters: $a_1 = 1$, $a_2 = 25$.
- Beta parameters $a_1 = 0.5$, $a_2 = 0.5$ correspond to $1/\sqrt{p_i(1-p_i)}$.
- Constant weights $w_i = 1$ correspond to C-alpha test statistic.

Wu et al. 2011 Am J Hum Genet

SKAT test: generalize C-alpha to model quantitative traits, LD between variants, etc.

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

Let $z_i = \sqrt{N}\rho(X_i, Y)$ be a z-score for SNP i.

Test statistic: $\sum_i w_i^2 z_i^2$, where weights $w_i \sim \text{Beta}(p_i; a_1, a_2)$

- recommended Beta parameters: $a_1 = 1$, $a_2 = 25$.
- Beta parameters $a_1 = 0.5$, $a_2 = 0.5$ correspond to $1/\sqrt{p_i(1-p_i)}$.
- Constant weights $w_i = 1$ correspond to C-alpha test statistic.

<u>To evaluate statistical significance</u>: 2 possibilities:

i. Test statistic follows mixture-χ^2 distribution; compute analytically.
(Note: mixture weights account for linkage disequilibrium between variants.)

Wu et al. 2011 Am J Hum Genet

SKAT test: generalize C-alpha to model quantitative traits, LD between variants, etc.

Let X_{ij} = minor allele count (0, 1, 2) of SNP i in individual j.

Let $z_i = \sqrt{N}\rho(X_i, Y)$ be a z-score for SNP i.

Test statistic: $\sum_i w_i^2 z_i^2$, where weights $w_i \sim \text{Beta}(p_i; a_1, a_2)$

- recommended Beta parameters: $a_1 = 1$, $a_2 = 25$.
- Beta parameters $a_1 = 0.5$, $a_2 = 0.5$ correspond to $1/\sqrt{p_i(1-p_i)}$.
- Constant weights $w_i = 1$ correspond to C-alpha test statistic.

<u>To evaluate statistical significance</u>: 2 possibilities:

i. Test statistic follows mixture-χ^2 distribution; compute analytically.
(Note: mixture weights account for linkage disequilibrium between variants.)

ii. Permutation test: compute test statistics with permuted phenotypes.

Wu et al. 2011 Am J Hum Genet

Burden or overdispersion tests: which are better?

burden!

overdispersion!!

reviewed in Lee et al. 2014 Am J Hum Genet

Burden or overdispersion tests: which are better? It depends.

The Empirical Power of Rare Variant Association Methods: Results from Sanger Sequencing in 1,998 Individuals

Abstract

The role of rare genetic variation in the etiology of complex disease remains unclear. However, the development of next-generation sequencing technologies offers the experimental opportunity to address this question. Several novel statistical methodologies have been recently proposed to assess the contribution of rare variation to complex disease etiology. Nevertheless, no empirical estimates comparing their relative power are available. We therefore assessed the parameters that influence their statistical power in 1,998 individuals Sanger-sequenced at seven genes by modeling different distributions of effect, proportions of causal variants, and direction of the associations (deleterious, protective, or both) in simulated continuous trait and case/control phenotypes. Our results demonstrate that the power of recently proposed statistical methods depend strongly on the underlying hypotheses concerning the relationship of phenotypes with each of these three factors. No method demonstrates consistently acceptable power despite this large sample size, and the performance of each method depends upon the underlying assumption of the relationship between rare variants and complex traits. Sensitivity analyses are therefore recommended to compare the stability of the results arising from different methods, and promising results should be replicated using the same method in an independent sample. These findings provide guidance in the analysis and interpretation of the role of rare base-pair variation in the etiology of complex traits and diseases.

Ladoucer et al. 2012 PLoS Genet

Rare variant association tests

Burden tests:

Fixed threshold: Li & Leal 2008 Am J Hum Genet
Weighted test: Madsen & Browning 2009 PLoS Genet
Variable threshold: Price et al. 2010 Am J Hum Genet

Overdispersion tests:

C-alpha: Neale et al. 2011 PLoS Genet
SKAT: Wu et al. 2011 Am J Hum Genet

Combined burden/overdispersion tests:

SKAT-O: Lee et al. 2012 Am J Hum Genet

Combined burden/overdispersion tests (omnibus tests) provide flexibility to identify rare variants associations either with same direction of effect or with varying direction of effect.

reviewed in Lee et al. 2014 Am J Hum Genet

Rare variant association tests

Burden tests:

Fixed threshold:	Li & Leal 2008 Am J Hum Genet
Weighted test:	Madsen & Browning 2009 PLoS Genet
Variable threshold:	Price et al. 2010 Am J Hum Genet

Overdispersion tests:

C-alpha:	Neale et al. 2011 PLoS Genet
SKAT:	Wu et al. 2011 Am J Hum Genet

Combined burden/overdispersion tests:

SKAT-O:	Lee et al. 2012 Am J Hum Genet

Combined burden/overdispersion tests (omnibus tests) provide flexibility to identify rare variants associations either with same direction of effect or with varying direction of effect.

reviewed in Lee et al. 2014 Am J Hum Genet

SKAT-O test: combine burden and SKAT tests

Let Q_{burden} = test statistic for burden test (e.g. $N\rho(C_g, Y)^2$; see above)

Let Q_{SKAT} = test statistic for SKAT test (e.g. $\sum_i w_i^2 z_i^2$; see above)

SKAT-O test statistic: $rQ_{\text{burden}} + (1 - r)Q_{\text{SKAT}}$ (where $0 \le r \le 1$)
with corresponding P-value P_r for each value of r.

- $r = 1$: burden test
- $r = 0$: SKAT test
- r can be interpreted as the correlation between SNP effect sizes β_i.

<u>To evaluate statistical significance</u>:

Compute minimum P-value: $P_{\text{min}} = \min_r P_r$

Statistical significance of P_{min} can be calculated analytically (one-dimensional numerical integration)

Lee et al. 2012 Am J Hum Genet
also see Uricchio et al. 2016 Genome Res

Rare variant association statistics can be computed using summary statistics

Asking for more

Because of the usefulness of genome-wide association study (GWAS) data for mapping regulatory variation in the human genome, the journal now asks authors to report the co-location of trait-associated variants with gene regulatory elements identified by epigenetic, functional and conservation criteria. We also ask that authors publish or database the genotype frequencies or association *P* values for all SNPs investigated, whether or not they reached genome-wide significance.

—Nat Genet editorial, July 2012

Definition: Summary statistics consist of:

- GWAS association z-scores for each typed or imputed SNP
- Sample sizes on which z-scores were computed (may vary by SNP)

Note: Many applications also require LD information computed from a population reference panel, e.g. 1000 Genomes (2015 Nature).

reviewed in Pasaniuc & Price 2016 Nat Rev Genet

Rare variant association statistics can be computed using summary statistics

Burden test: $N\rho(C_g, Y)^2$ = square of weighted sum of z-scores $z_i = \sqrt{N}\rho(X_i, Y)$

SKAT: test statistic $= \sum_i w_i^2 z_i^2$

Lee et al. 2013 Am J Hum Genet, Hu et al. 2013 Am J Hum Genet, Liu et al. 2014 Nat Genet

Rare variant association statistics can be computed using summary statistics

Burden test: $N\rho(C_g, Y)^2$ = square of weighted sum of z-scores $z_i = \sqrt{N}\rho(X_i, Y)$

SKAT: test statistic $= \sum_i w_i^2 z_i^2$

Caveat: for both burden test and SKAT, in-sample LD is required to obtain correct null distributions and avoid false-positive associations (cannot use LD from population reference panel)

Lee et al. 2013 Am J Hum Genet, Hu et al. 2013 Am J Hum Genet, Liu et al. 2014 Nat Genet

Outline

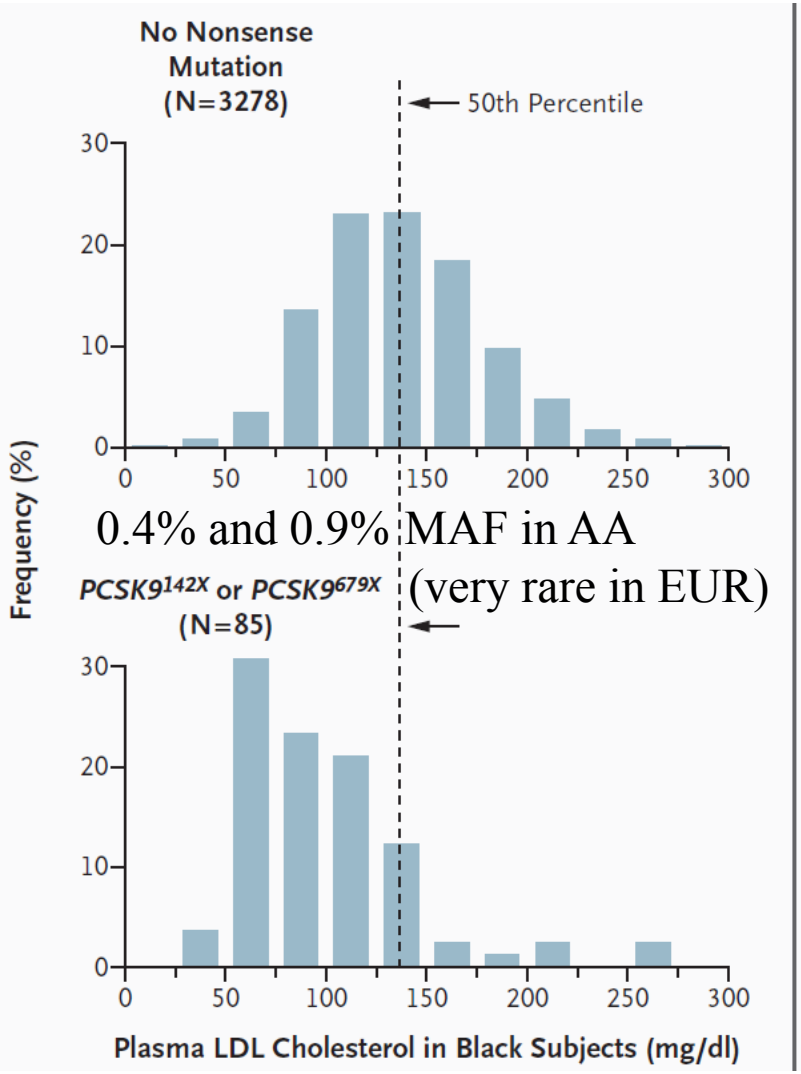
1. Properties of rare and low-frequency variants

2. Rare variant association tests: methods

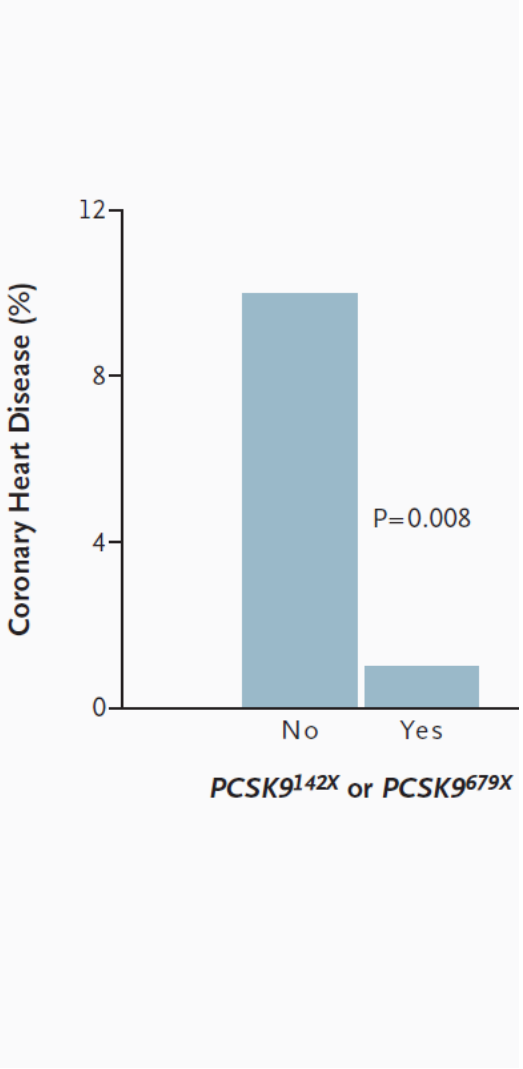
3. Rare variant association tests: results

4. Rare variant heritability

Nonsense/missense mutations in PCSK9 reduce LDL levels and CHD risk



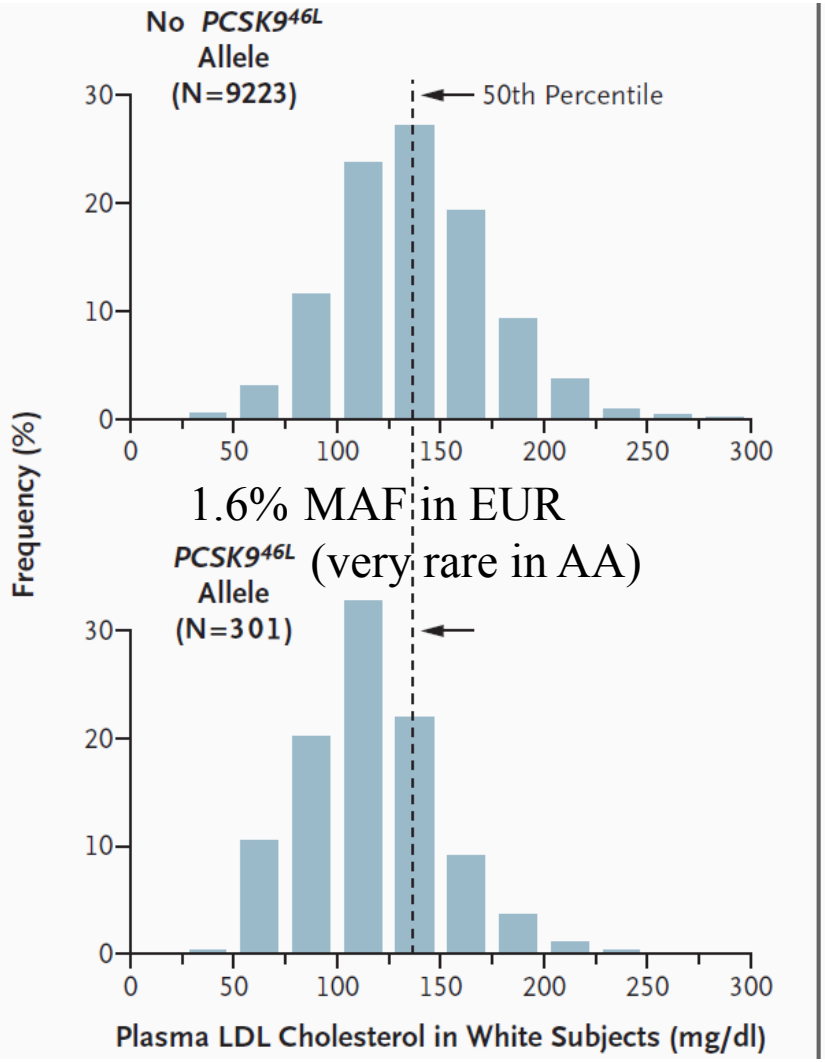
0.4% and 0.9% MAF in AA (very rare in EUR)



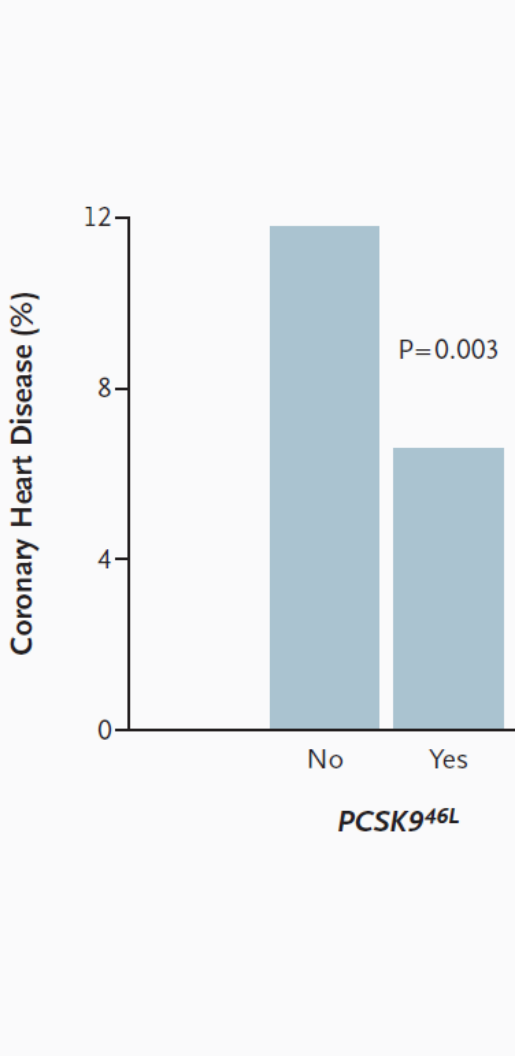
3,278 African Americans

Cohen et al. 2006 New Engl J Med

Nonsense/missense mutations in PCSK9 reduce LDL levels and CHD risk



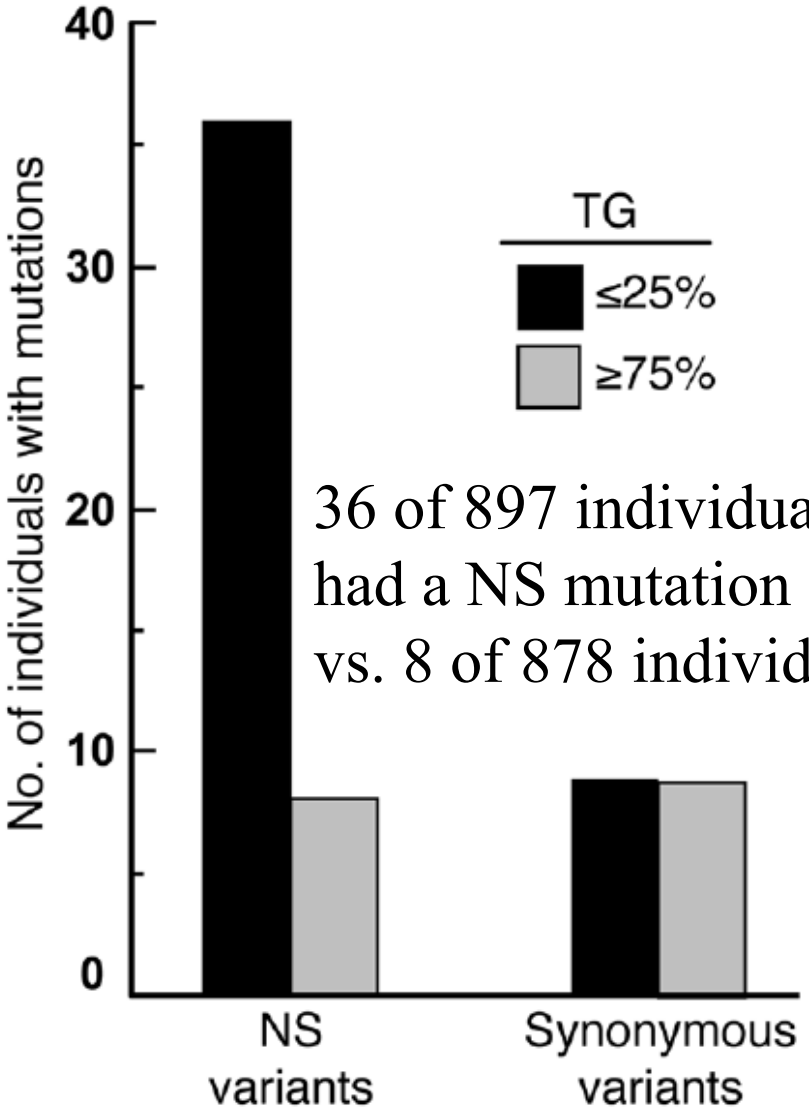
1.6% MAF in EUR (very rare in AA)



9,223 European Americans

Cohen et al. 2006 New Engl J Med

Nonsynonymous mutations in ANGPTL3, ANGPTL4, ANGPTL5 reduce TG levels



36 of 897 individuals in lowest TG quartile had a NS mutation (23 unique NS mutations), vs. 8 of 878 individuals in highest TG quartile

Romeo et al. 2007 Nat Genet
Romeo et al. 2009 J Clin Invest

Additional examples involving IFIH1 and T1D, 21 known monogenic-obesity genes and obesity

Rare Variants of *IFIH1*, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes

Sergey Nejentsev,[1,2]* Neil Walker,[1] David Riches,[3] Michael Egholm,[3] John A. Todd[1]

Nejentsev et al. 2009 Science

Medical Sequencing at the Extremes of Human Body Mass

Nadav Ahituv, Nihan Kavaslar, Wendy Schackwitz, Anna Ustaszewska, Joel Martin, Sybil Hébert, Heather Doelle, Baran Ersoy, Gregory Kryukov, Steffen Schmidt, Nir Yosef, Eytan Ruppin, Roded Sharan, Christian Vaisse, Shamil Sunyaev, Robert Dent, Jonathan Cohen, Ruth McPherson, and Len A. Pennacchio

Ahituv et al. 2007 Am J Hum Genet

Choice of statistical test can affect power

		T1	T5	WE	VT
Romeo 2009	Triglyceride level	0.013	0.00007	0.0020	0.00038*
Nejentsev 2009	Type 1 diabetes	0.001	0.0000002	0.0000004	0.0000008
Ahituv 2007	Obesity	0.032	0.053	0.010	0.010

T1 = Fixed threshold burden test (MAF=1%)
T5 = Fixed threshold burden test (MAF=5%)
WE = Weighted burden test
VT = Variable threshold burden test
*P-value decreases to 0.000095 using **SKAT** (Wu et al. 2011 Am J Hum Genet)

Price et al. 2010 Am J Hum Genet

Incorporating functional data can increase power

		T1	T5	WE	VT	VTP
Romeo 2009	Triglyceride level	0.013	0.00007	0.0020	0.00038	0.00002
Nejentsev 2009	Type 1 diabetes	0.001	0.0000002	0.0000004	0.0000008	0.0000002
Ahituv 2007	Obesity	0.032	0.053	0.010	0.010	0.0017

T1 = Fixed threshold burden test (MAF=1%)
T5 = Fixed threshold burden test (MAF=5%)
WE = Weighted burden test
VT = Variable threshold burden test
VTP = Variable threshold burden test, incorporating PolyPhen-2 weights (posterior prob. of being functional; Adzhubei et al. 2010 Nat Methods)

(to be continued, Tue of Week 7)

Price et al. 2010 Am J Hum Genet

NHLBI Exome Sequencing Project: 0 associations

Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes

Jacob A. Tennessen,[1*] Abigail W. Bigham,[2*†] Timothy D. O'Connor,[1*] Wenqing Fu,[1] Eimear E. Kenny,[3] Simon Gravel,[3] Sean McGee,[1] Ron Do,[4,5] Xiaoming Liu,[6] Goo Jun,[7] Hyun Min Kang,[7] Daniel Jordan,[8] Suzanne M. Leal,[9] Stacey Gabriel,[4] Mark J. Rieder,[1] Goncalo Abecasis,[7] David Altshuler,[4] Deborah A. Nickerson,[1] Eric Boerwinkle,[6,10] Shamil Sunyaev,[4,8] Carlos D. Bustamante,[3] Michael J. Bamshad,[1,2‡] Joshua M. Akey,[1‡] Broad GO, Seattle GO, on behalf of the NHLBI Exome Sequencing Project

As a first step toward understanding how rare variants contribute to risk for complex diseases, we sequenced 15,585 human protein-coding genes to an average median depth of 111× in 2440 individuals of European (n = 1351) and African (n = 1088) ancestry. We identified over 500,000 single-nucleotide variants (SNVs), the majority of which were rare (86% with a minor allele frequency less than 0.5%), previously unknown (82%), and population-specific (82%). On average, 2.3% of the 13,595 SNVs each person carried were predicted to affect protein function of ~313 genes per genome, and ~95.7% of SNVs predicted to be functionally important were rare. This excess of rare functional variants is due to the combined effects of explosive, recent accelerated population growth and weak purifying selection. Furthermore, we show that large sample sizes will be required to associate rare variants with complex traits.

Tennesen et al. 2012 Science

Previous studies predicted low power at NHLBI Exome Sequencing Project sample sizes

Power of deep, all-exon resequencing for discovery of human trait genes

Gregory V. Kryukov[a], Alexander Shpunt[a,b], John A. Stamatoyannopoulos[c], and Shamil R. Sunyaev[a,1]

Table 1. Estimated power of gene mapping by complete resequencing

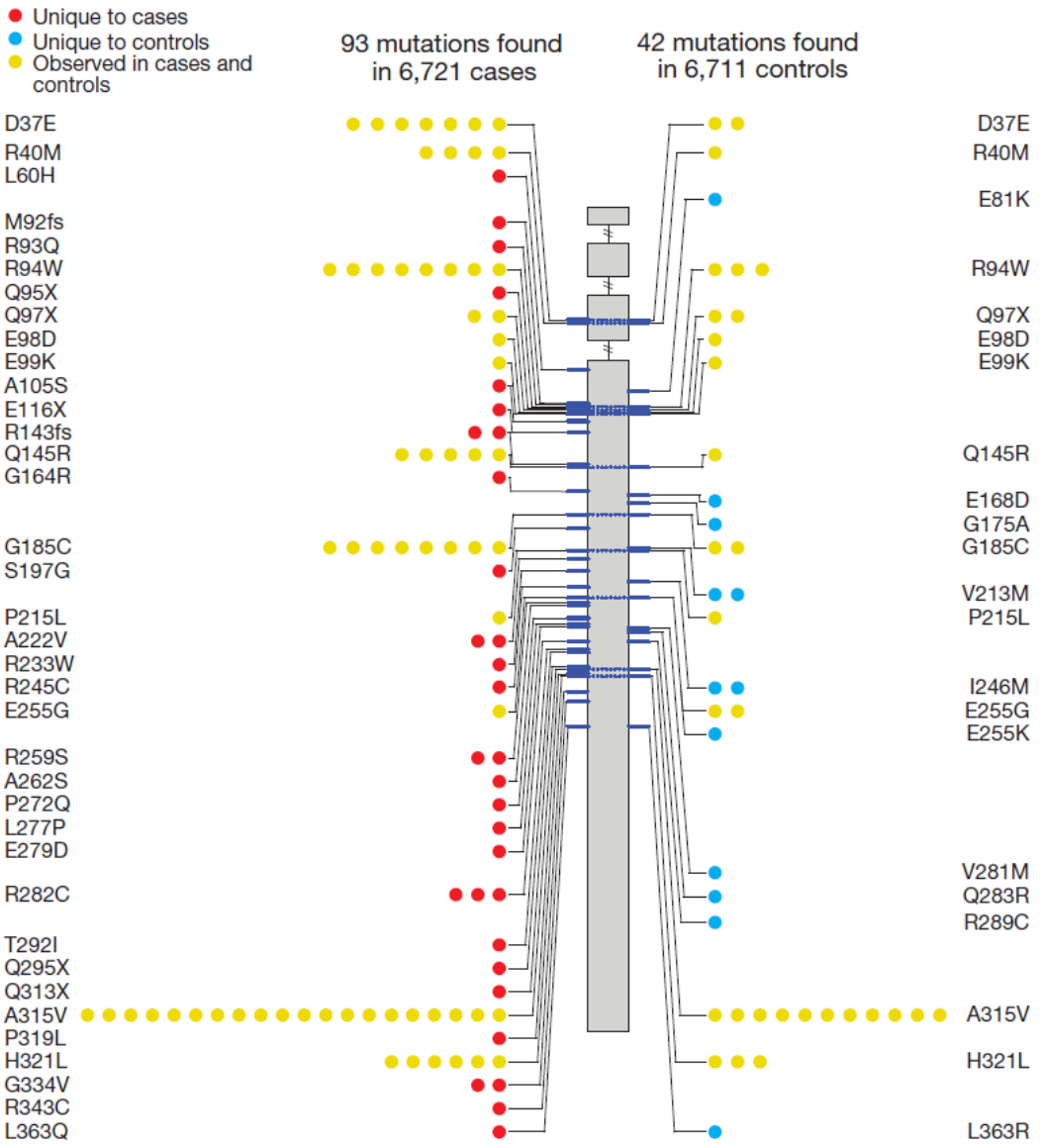
Effect of functional mutations (in fractions of standard deviation)	No. of sequenced individuals	No. of phenotyped individuals				
		12,500	25,000	50,000	100,000	200,000
0.25σ	5,000	0.11	0.18	0.24		
	10,000		0.24	0.31	0.40	
	20,000			0.38	0.51	0.59
0.5σ	5,000	0.36	0.47	0.57		
	10,000		0.56	0.69	0.77	
	20,000			0.76	0.84	0.88

(NHLBI Exome Sequencing Project: N=6,515, distributed across many diseases and traits)

Kryukov et al. 2009 PNAS

also see Kiezun et al. 2012 Nat Genet

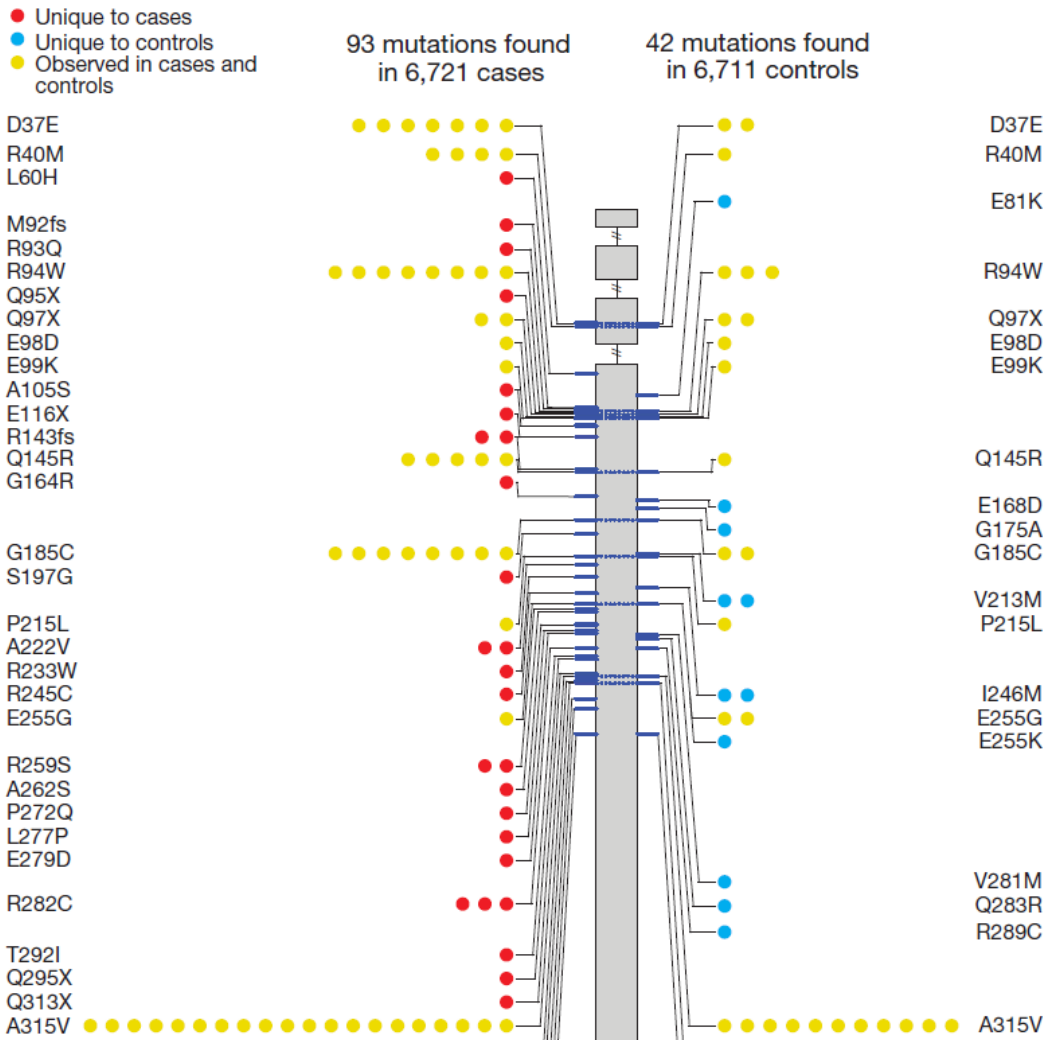
Nonsynonymous mutations in APOA5, LDLR increase risk of myocardial infarction (MI)



APOA5 gene,
MAF < 1% burden test:
93/6,721 cases
42/6,711 controls
$P = 5 \times 10^{-7}$

Do et al. 2015 Nature

Rare variant associations often occur at loci with common variant GWAS associations

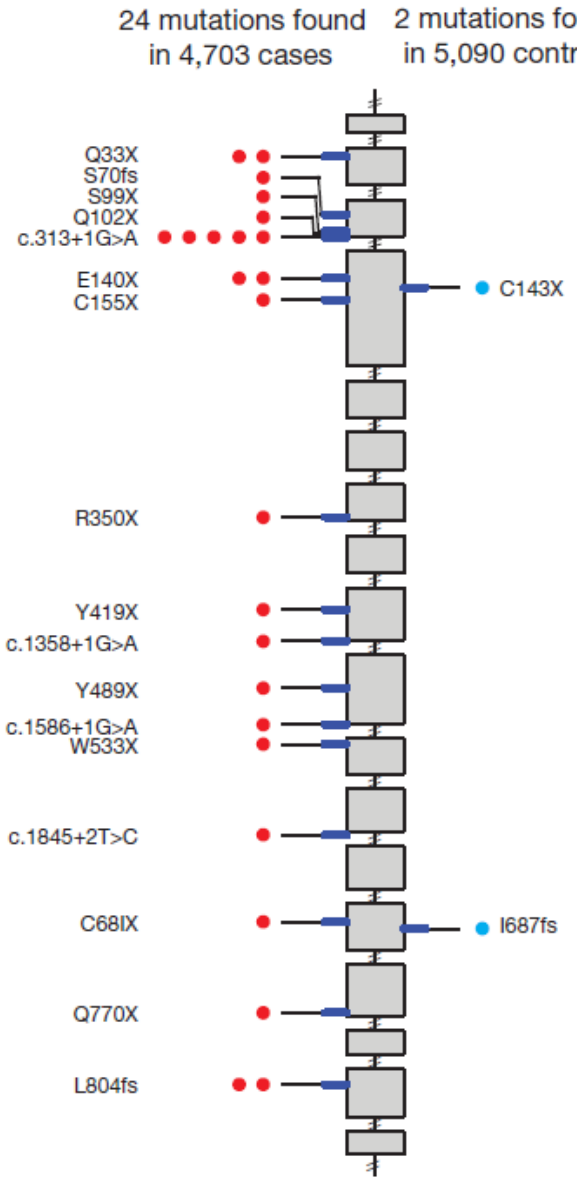


APOA5 gene,
MAF < 1% burden test:
93/6,721 cases
42/6,711 controls
$P = 5 \times 10^{-7}$

GWAS associations at APOA5 locus previously reported for lipid levels and MI (Teslovich et al. 2010 Nature)

Do et al. 2015 Nature; also see Nejentsev et al. 2009 Science, Johansen et al. 2010 Nat Genet, Auer et al. 2014 Nat Genet, Flannick et al. 2014 Nat Genet, Surakka et al. 2015 Nat Genet, Fuchsberger et al. 2016 Nature

Nonsynonymous mutations in APOA5, LDLR increase risk of myocardial infarction (MI)



LDLR gene, MAF < 1% burden test:
285/4,703 cases vs. 208/5,090 controls
$P = 4 \times 10^{-6}$

Restrict to mutations predicted to be damaging by PolyPhen-2:
148/4,703 cases vs. 67/5,090 controls
$P = 1 \times 10^{-11}$

Restrict to nonsense + splice-site + indel frameshift mutations:
24/4,703 cases vs. 2/5,090 controls
$P = 9 \times 10^{-5}$

Note: mutations increase LDL cholesterol

Do et al. 2015 Nature

Nonsynonymous mutations in TBK1 increase risk of amyotrophic lateral sclerosis (ALS)

TBK1 gene,
Fixed threshold burden test (MAF < 5% in test data
and MAF < 0.5% in ExAC; Lek et al. 2016 Nature),
Restrict to mutations predicted to be damaging by PolyPhen-2:
46/4,161 cases vs. 17/6,681 controls
$P = 3.6 \times 10^{-11}$

Cirulli et al. 2015 Science

Disruptive mutations in 2,546 candidate genes confer polygenic burden of schizophrenia risk

Set of 2,546 candidate genes (from previous studies)
Fixed threshold polygenic burden test (MAF < 0.1%)
Restrict to disruptive (nonsense + splice-site + frameshift) mutations:
1,547 mutations in 2,536 cases vs. 1,383 mutations in 2,543 controls
P = 0.0001

Purcell et al. 2014 Nature

Disruptive mutations in 2,546 candidate genes confer polygenic burden of schizophrenia risk

Set of 2,546 candidate genes (from previous studies)
Fixed threshold polygenic burden test (MAF < 0.1%)
Restrict to disruptive (nonsense + splice-site + frameshift) mutations:
1,547 mutations in 2,536 cases vs. 1,383 mutations in 2,543 controls
P = 0.0001

Subset of 28 ARC complex genes: P = 0.0014
Subset of 65 PSD-95 complex genes: P = 0.0009
Subset of 26 voltage-gated calcium ion channel genes: P = 0.0019*
(*: Restrict to singleton SNPs)

Purcell et al. 2014 Nature

Ultra-rare disruptive/damaging mutations confer polygenic burden of schizophrenia risk

Coding regions of all genes

Fixed threshold polygenic burden test: restrict to ultra-rare SNPs (singleton in test data + absent from ExAC; Lek et al. 2016 Nature)

Restrict to disruptive (nonsense + splice-site + frameshift) mutations + mutations predicted to be damaging by PolyPhen-2 and other methods:

+0.25 mutations/individual in 4,877 cases vs. in 6,203 controls

$P = 1.5 \times 10^{-10}$ (OR=1.07; 4.23 mutations/individual overall)

Genovese et al. 2016 Nat Neurosci

Ultra-rare disruptive/damaging mutations in constrained genes increase schizophrenia risk

Coding regions of all genes
Fixed threshold polygenic burden test: restrict to ultra-rare SNPs (singleton in test data + absent from ExAC; Lek et al. 2016 Nature)
Restrict to disruptive (nonsense + splice-site + frameshift) mutations + mutations predicted to be damaging by PolyPhen-2 and other methods:
+0.25 mutations/individual in 4,877 cases vs. in 6,203 controls
$P = 1.5 \times 10^{-10}$ (OR=1.07; 4.23 mutations/individual overall)

Restrict to 1,001 missense-constrained genes (Samocha et al. 2014 Nat Genet)
(genes that contain fewer missense mutations than expected):
larger OR=1.28 ($P = 3.2 \times 10^{-8}$ vs. OR=1.07 overall)

Restrict to 3,488 loss-of-function-intolerant genes (Lek et al. 2016 Nature)
(genes that contain fewer disruptive LoF mutations than expected):
larger OR=1.17 ($P = 1.7 \times 10^{-8}$ vs. OR=1.07 overall)

Genovese et al. 2016 Nat Neurosci

Ultra-rare disruptive/damaging mutations in specifically expressed genes increase SCZ risk

Coding regions of all genes
Fixed threshold polygenic burden test: restrict to ultra-rare SNPs (singleton in test data + absent from ExAC; Lek et al. 2016 Nature)
Restrict to disruptive (nonsense + splice-site + frameshift) mutations + mutations predicted to be damaging by PolyPhen-2 and other methods:
+0.25 mutations/individual in 4,877 cases vs. in 6,203 controls
$P = 1.5 \times 10^{-10}$ (OR=1.07; 4.23 mutations/individual overall)

2,647 genes specifically expressed in brain tissue vs. other tissues (Fagerberg et al. 2014 Mol Cell Proteomics):
larger OR=1.17 ($P = 1.2 \times 10^{-4}$ vs. OR=1.07 overall)

3,388 genes specifically expressed in neurons (Cahoy et al. 2008 J Neurosci):
larger OR=1.17 ($P = 1.9 \times 10^{-7}$ vs. OR=1.07 overall)

Additional enrichments in synaptic gene sets.

Genovese et al. 2016 Nat Neurosci

Ultra-rare disruptive/damaging mutations in constrained genes reduce educational attainment

3,488 loss-of-function-intolerant genes (Lek et al. 2016 Nature)
Fixed threshold polygenic burden test: restrict to ultra-rare SNPs (singleton in test data + absent from ExAC; Lek et al. 2016 Nature)
Restrict to disruptive (nonsense + splice-site + frameshift) mutations:
Years Of Education 3.1 months lower per mutation ($P = 3.3 \times 10^{-8}$)

1,614 missense-constrained genes (Samocha et al. 2014 Nat Genet)
Fixed threshold polygenic burden test: restrict to ultra-rare SNPs (singleton in test data + absent from ExAC; Lek et al. 2016 Nature)
Restrict to damaging mutations (PolyPhen-2 and other methods):
Years Of Education 2.9 months lower per mutation ($P = 1.3 \times 10^{-6}$)

Even larger effects when restricting to top brain-expressed genes (GTEx Consortium 2015 Science).

Ganna et al. 2016 Nat Neurosci

Can rare variant association tests be applied to noncoding variants?

Burden tests:

Fixed threshold: Li & Leal 2008 Am J Hum Genet
Weighted test: Madsen & Browning 2009 PLoS Genet
Variable threshold: Price et al. 2010 Am J Hum Genet

Overdispersion tests:

C-alpha: Neale et al. 2011 PLoS Genet
SKAT: Wu et al. 2011 Am J Hum Genet

Combined burden/overdispersion tests:

SKAT-O: Lee et al. 2012 Am J Hum Genet

[gene-based tests, multiple rare **coding** variants, complex diseases/traits]
What about **noncoding** variants? **(To be continued, Tue of Week 7)**

reviewed in Lee et al. 2014 Am J Hum Genet
also see Zuk et al. 2014 PNAS

Outline

1. Properties of rare and low-frequency variants

2. Rare variant association tests: methods

3. Rare variant association tests: results

4. Rare variant heritability

Negative selection ➔ rare variant heritability

(from Thu of Week 4)

The role of negative selection determines whether the genetic architecture of common diseases is driven by rare or common alleles.

Most Rare Missense Alleles Are Deleterious in Humans: Implications for Complex Disease and Association Studies

Gregory V. Kryukov, Len A. Pennacchio, and Shamil R. Sunyaev

Evaluating empirical bounds on complex disease genetic architecture

Vineeta Agarwala[1–3,9], Jason Flannick[2,4,5,9], Shamil Sunyaev[1–3,6], GoT2D Consortium[7] & David Altshuler[2,4,5,8]

Kryukov et al. 2007 Am J Hum Genet, Agarwala et al. 2013 Nat Genet; also see Pritchard 2001 Am J Hum Genet, Eyre-Walker 2010 PNAS, Zuk et al. 2014 PNAS

Functional SNPs are more likely to be rare

Whole-exome sequencing (NHLBI Exome Sequencing Project) of 1,351 European Americans (EA) + 1,088 African Americans (AA):

95.7% of functional* coding SNPs are rare (MAF<0.5%) vs.
84.1% of non-functional* coding SNPs are rare (MAF<0.5%)

*: as predicted by consensus of PolyPhen-2 and 6 other methods
"Functional" ≈ "Damaging"

Odds Ratio = 4.2 (95% CI: 4.0-4.3, $P < 10^{-15}$)

Functional, damaging SNPs have their minor allele frequencies pushed down by negative selection.

Tennessen et al. 2012 Science
also see Zuk et al. 2014 PNAS

Rare SNPs are more likely to be functional

Whole-exome sequencing (NHLBI Exome Sequencing Project) of 1,351 European Americans (EA) + 1,088 African Americans (AA):

18.8% of rare (MAF<0.5%) coding SNPs are functional* vs.
5.2% of non-rare (MAF>0.5%) coding SNPs are functional*

*: as predicted by consensus of PolyPhen-2 and 6 other methods
"Functional" ≈ "Damaging"

Odds Ratio = 4.2 (95% CI: 4.0-4.3, $P < 10^{-15}$)

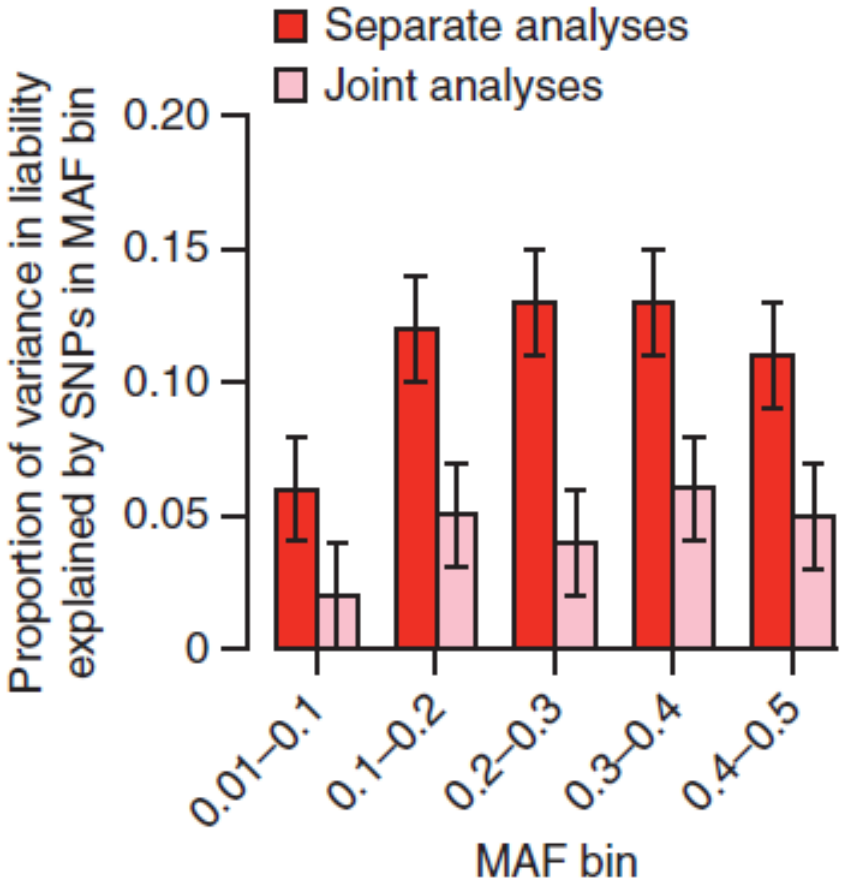
How much do rare SNPs contribute to disease/trait heritability?

Tennessen et al. 2012 Science
also see Zuk et al. 2014 PNAS

Partitioning h_g^2 by minor allele frequency (MAF) in a schizophrenia data set (N=21,258)

$V = h_{g,1}^2 A_1 + \ldots + h_{g,5}^2 A_5 + (1 - h_g^2)I$

($A_1, \ldots, A_5$ computed from 5 disjoint MAF bins)



h_g^2 in this data set is primarily coming from common SNPs.

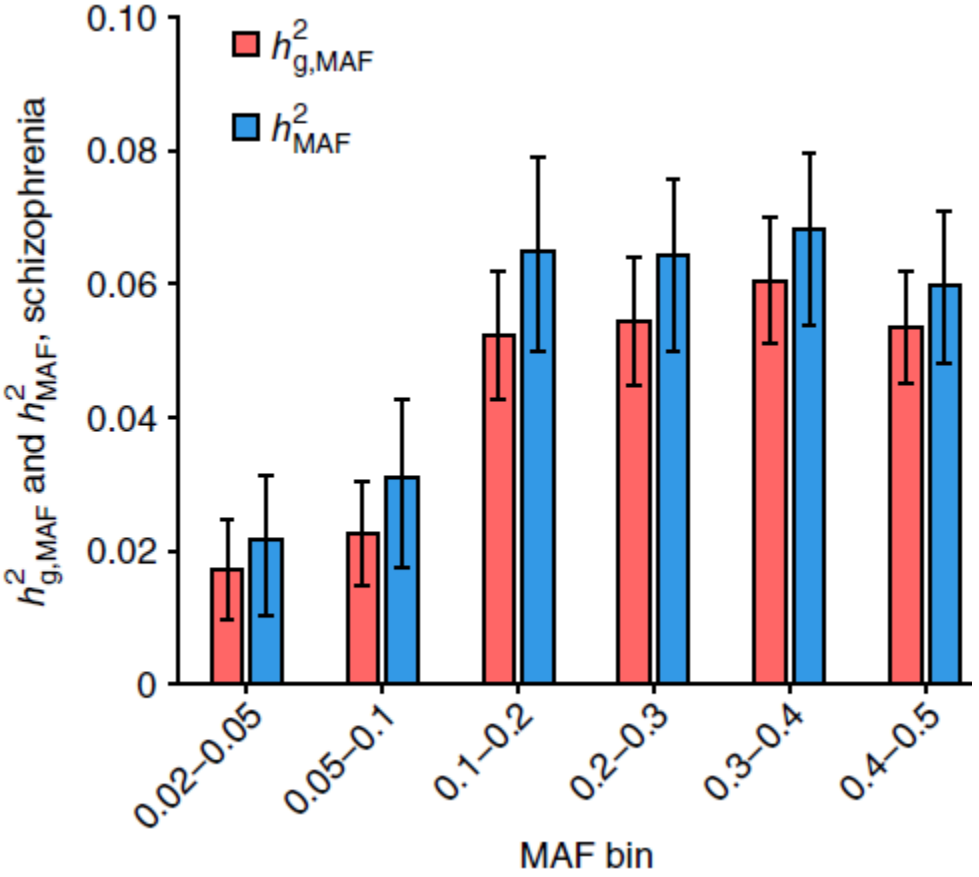
(from Tue of Week 5)

Lee et al. 2012 Nat Genet

Partitioning h_g^2 by minor allele frequency (MAF) in a schizophrenia data set (N=49,806)

$$V = h_{g,1}^2 A_1 + \ldots + h_{g,6}^2 A_6 + (1 - h_g^2)I$$

($A_1, \ldots, A_6$ computed from 6 disjoint MAF bins)



h_g^2 in this data set is primarily coming from common SNPs.

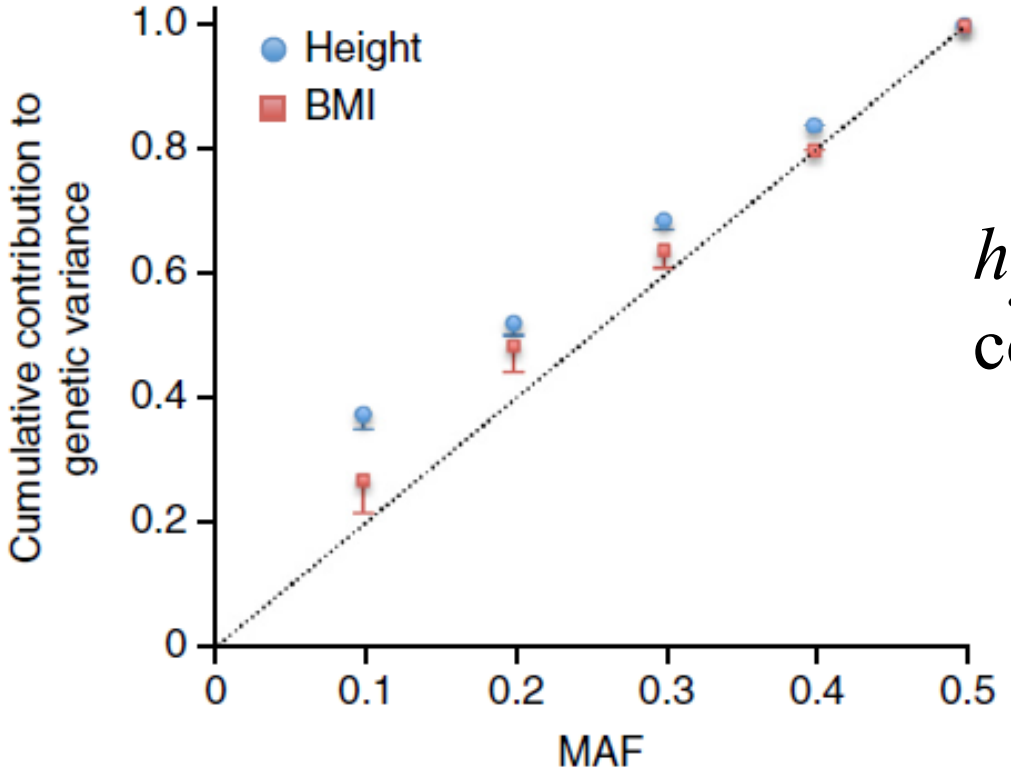
Loh, Bhatia et al. 2015 Nat Genet

Partitioning h_g^2 by minor allele frequency (MAF) in height and BMI data sets (N=44,126)

$$V = h_{g,1}^2 A_1 + \ldots + h_{g,7}^2 A_7 + (1 - h_g^2)I$$

($A_1, \ldots, A_7$ computed from 7 disjoint MAF bins*)

*actually 7 MAF bins x 4 regional LD bins, to deal with LD-dependent architectures



h_g^2 in this data set is primarily coming from common SNPs.

Yang et al. 2015 Nat Genet

Functional (damaging) SNPs are likely to have larger disease effect sizes; how much larger?

Eyre-Walker model:

Absolute effect size $|\beta| = Cs^{\tau}(1 + \varepsilon)$, where

C = a constant

s = selection coefficient (e.g. s = 0.0001-0.01 for damaging SNPs)

τ = strength of coupling between s and effect size (e.g. $0 \leq \tau \leq 1$)

ε = normally distributed with mean 0 and some variance

Eyre-Walker 2010 PNAS

Functional (damaging) SNPs are likely to have larger disease effect sizes; how much larger?

Eyre-Walker model:

Absolute effect size $|\beta| = Cs^{\tau}(1 + \varepsilon)$, where

C = a constant

s = selection coefficient (e.g. s = 0.0001-0.01 for damaging SNPs)

τ = strength of coupling between s and effect size (e.g. $0 \leq \tau \leq 1$)

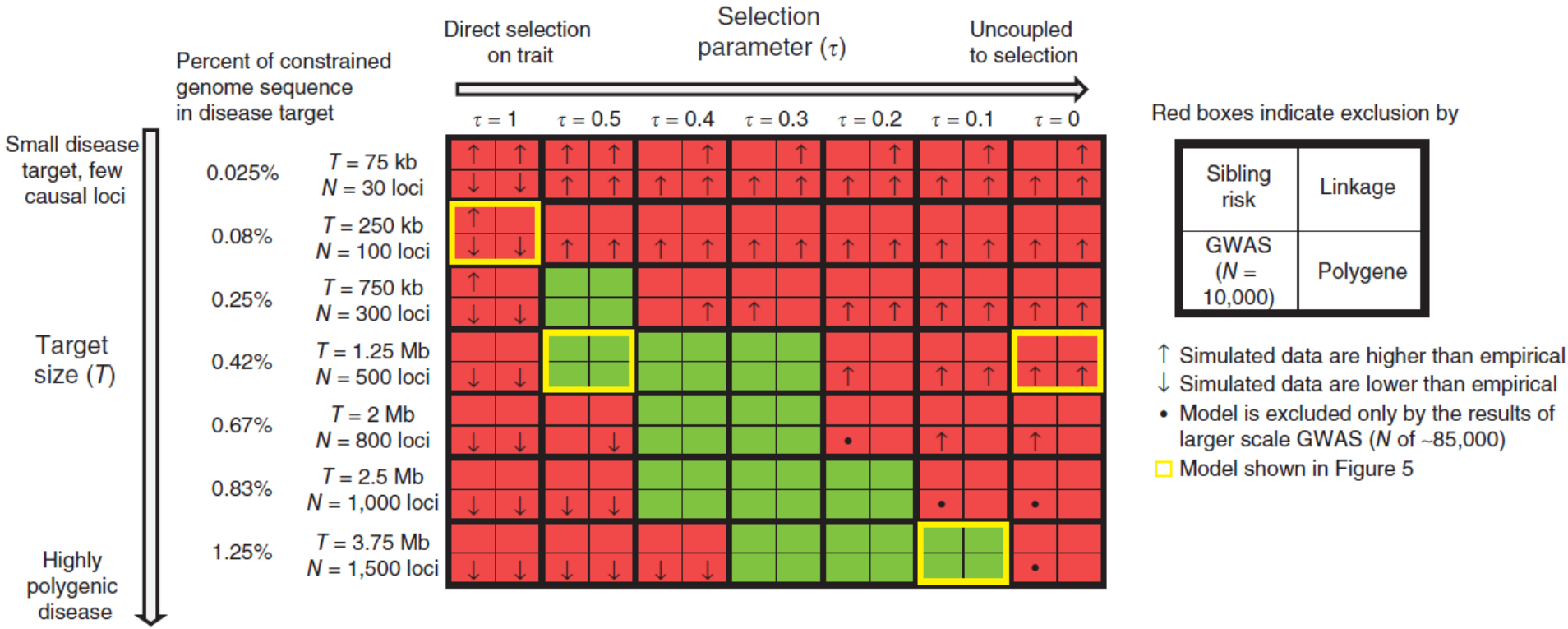
ε = normally distributed with mean 0 and some variance

$\tau = 1$: strong coupling; disease-associated SNPs are primarily rare

$\tau = 0$: no coupling; disease-associated SNPs are primarily common

Eyre-Walker 2010 PNAS

Estimating the selection coupling parameter τ from type 2 diabetes GWAS results



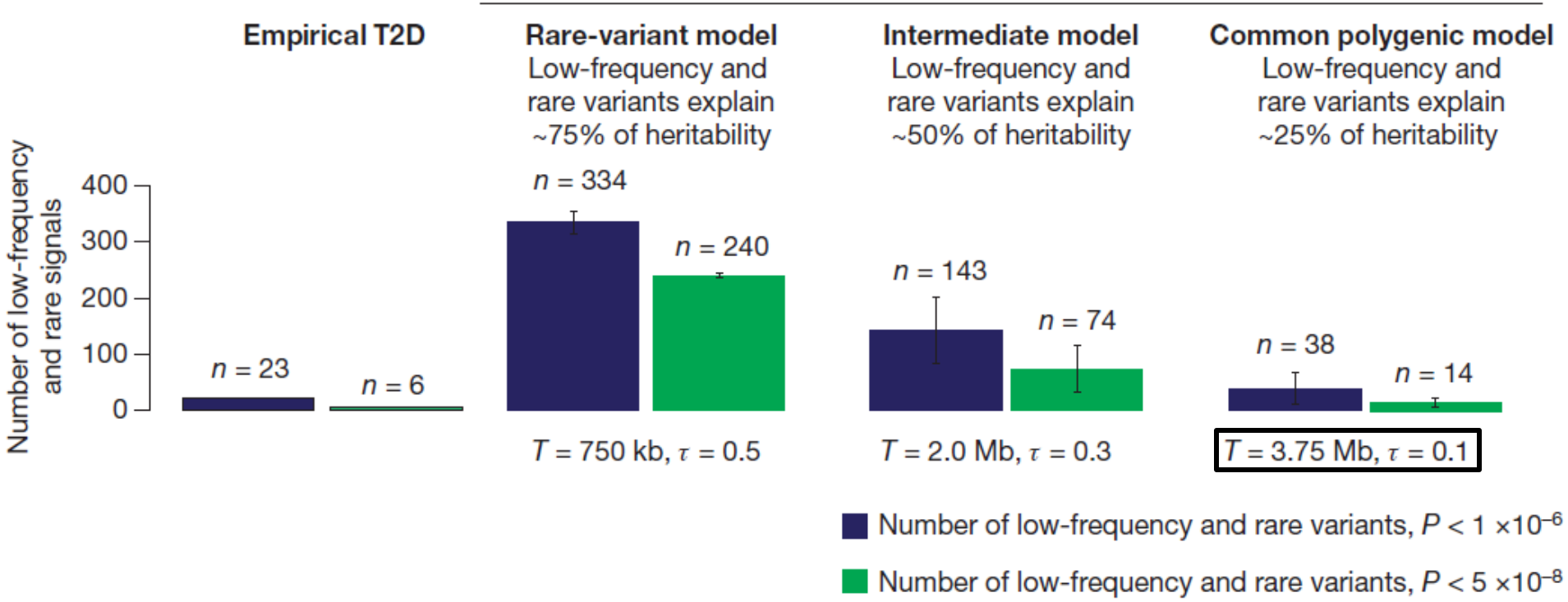
$\tau = 1$: implausible: too few GWAS hits vs. observed

$\tau = 0$: implausible: too many GWAS hits vs. observed

$0 < \tau < 1$: plausible, depending on disease polygenicity

Agarwala et al. 2013 Nat Genet

Estimating the selection coupling parameter τ from type 2 diabetes GWAS+imputation results



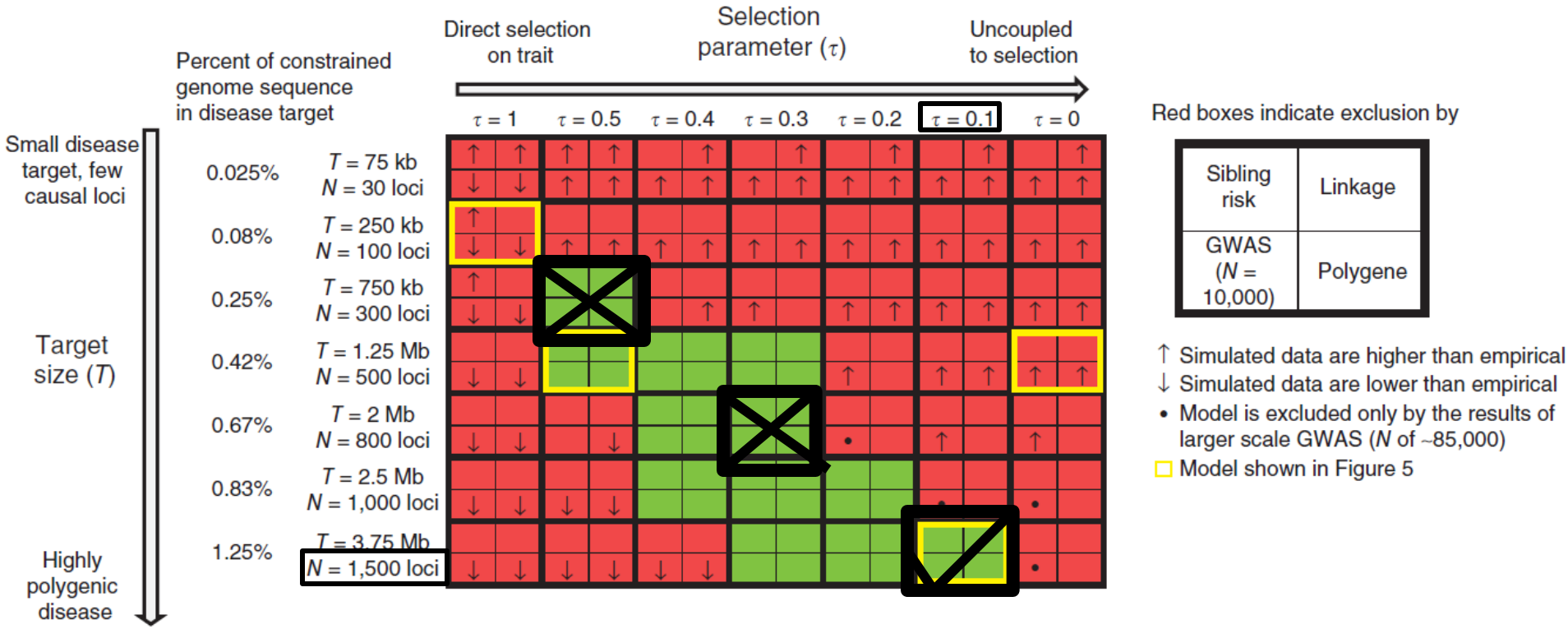
$\tau = 0.5$: implausible: too many rare associations vs. observed

$\tau = 0.3$: implausible: too many rare associations vs. observed

$\tau = 0.1$: most plausible

Fuchsberger et al. 2016 Nature

Estimating the selection coupling parameter τ from type 2 diabetes GWAS+imputation results



$\tau = 0.5$: implausible: too many rare associations vs. observed

$\tau = 0.3$: implausible: too many rare associations vs. observed

$\tau = 0.1$: most plausible

Agarwala et al. 2013 Nat Genet

Fuchsberger et al. 2016 Nature

Estimating the selection coupling parameter τ from prostate cancer $h_{g,rare}^2$ at GWAS loci

$h_{g,rare}^2$ estimates from sequencing data at 63 known GWAS loci:

Ancestry	Sample size	h_g^2 index SNPs (s.e.)	$h_{g,\text{rare}}^2$ (s.e.)	*P* value	$h_{g,\text{common}}^2$ (s.e.)	*P* value
African	4,006	0.06 (0.01)	0.12 (0.05)	2.29×10^{-3}	0.17 (0.03)	7.08×10^{-13}
European	1,753	0.10 (0.01)	0.00 (0.06)	5.00×10^{-1}	0.27 (0.06)	5.83×10^{-11}
Japanese	1,770	0.08 (0.01)	0.05 (0.07)	2.68×10^{-1}	0.13 (0.04)	3.09×10^{-5}
Latino	1,708	0.06 (0.01)	0.00 (0.06)	5.00×10^{-1}	0.14 (0.05)	2.38×10^{-5}

rare: $0.1\% \leq \text{MAF} < 1\%$ common: $\text{MAF} \geq 1\%$

Mancuso et al. 2015 Nat Genet

Estimating the selection coupling parameter τ from prostate cancer $h_{g,rare}^2$ at GWAS loci

$h_{g,rare}^2$ estimates from sequencing data at 63 known GWAS loci:

Ancestry	Sample size	h_g^2 index SNPs (s.e.)	$h_{g,\text{rare}}^2$ (s.e.)	*P* value	$h_{g,\text{common}}^2$ (s.e.)	*P* value
African	4,006	0.06 (0.01)	0.12 (0.05)	2.29×10^{-3}	0.17 (0.03)	7.08×10^{-13}
European	1,753	0.10 (0.01)	0.00 (0.06)	5.00×10^{-1}	0.27 (0.06)	5.83×10^{-11}
Japanese	1,770	0.08 (0.01)	0.05 (0.07)	2.68×10^{-1}	0.13 (0.04)	3.09×10^{-5}
Latino	1,708	0.06 (0.01)	0.00 (0.06)	5.00×10^{-1}	0.14 (0.05)	2.38×10^{-5}

➔

Ancestry	Sample size	Mean τ	95% confidence interval	$h_{g,\text{rare}}^2$
African	4,006	0.48	0.19, 0.78	0.12 (0.05)
European	1,753	0.28	–0.08, 0.90	0.00 (0.06)
Japanese	1,770	0.38	–0.07, 0.92	0.05 (0.07)
Latino	1,708	0.39	–0.08, 1.05	0.00 (0.06)
Meta-analysis	9,237	0.42	0.22, 0.62	0.05 (0.03)

τ parameter estimated via simulations

Mancuso et al. 2015 Nat Genet

Conclusions

- The human genome harbors a large number of rare variants, most of which have arisen in the past 5,000 years.
- Burden tests and overdispersion tests each have the potential to identify rare variant associations.
- Rare variant association studies require large sample sizes (just like GWAS), but some associations have been identified. Polygenic analyses of gene sets/pathways can increase power.
- The contribution of rare variants to heritability depends on the role of negative selection, and varies across diseases/traits.