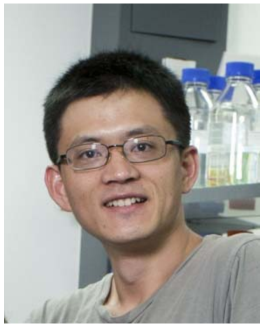
Mixed model association methods: advantages and pitfalls

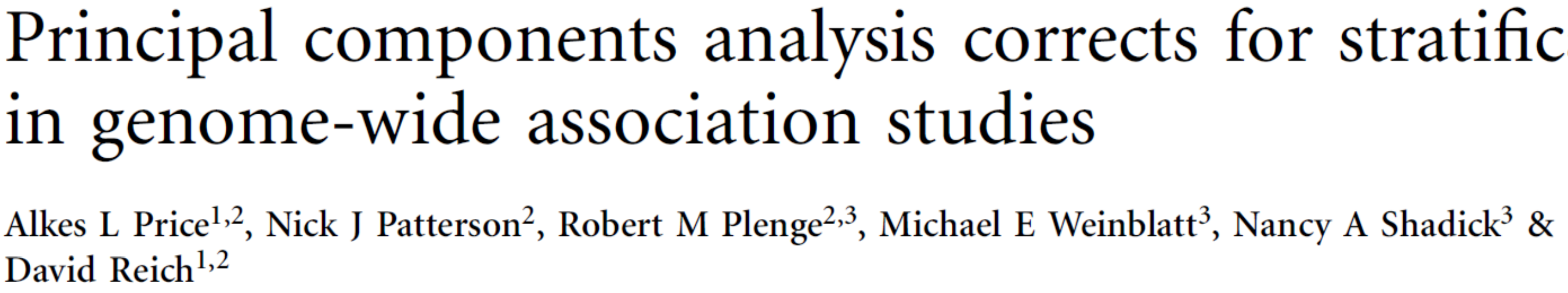
To download slides of this talk: google “Alkes HSPH”

Alkes L. Price
Harvard School of Public Health
October 23, 2013



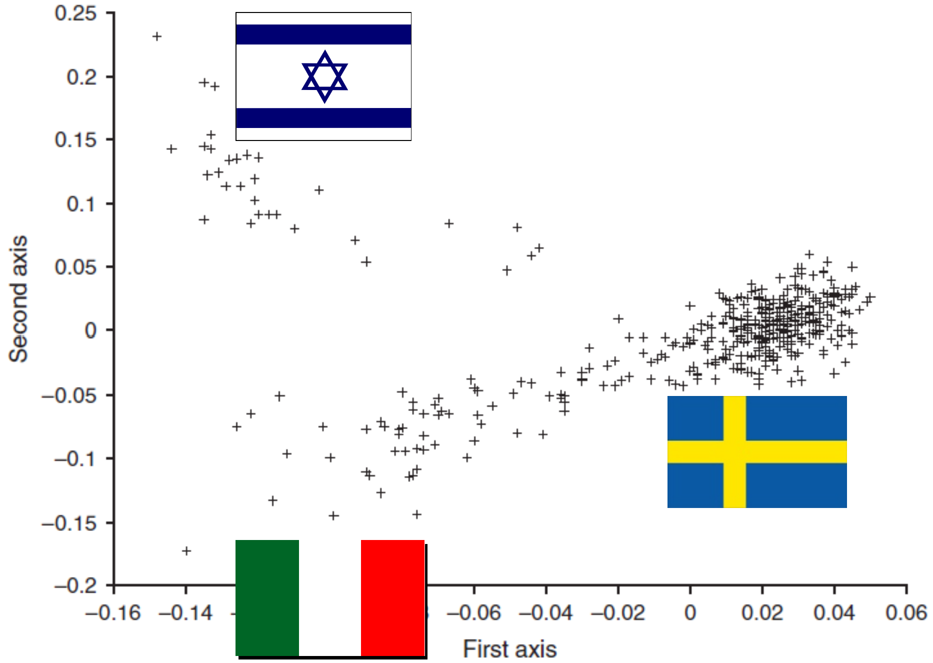
(Yang*, Zaitlen* et al. under revision)

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

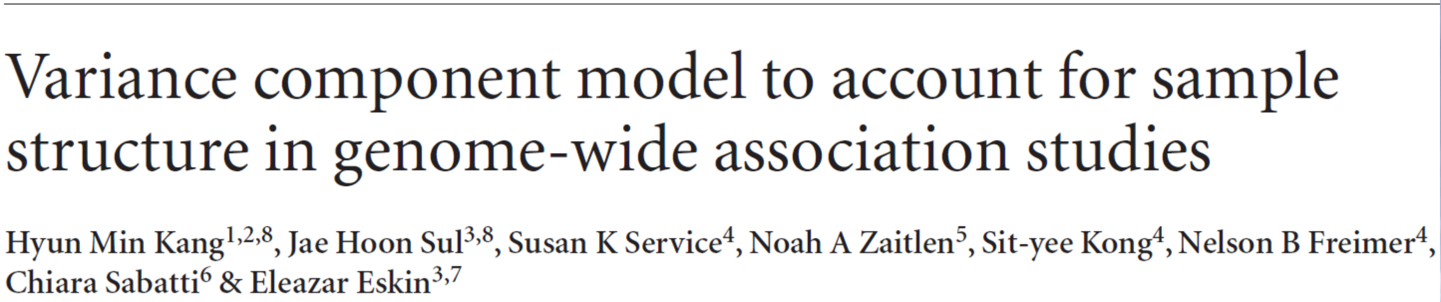
… does not correct for cryptic relatedness

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

reviewed in Price et al. 2010 Nat Rev Genet

Mixed model association saves the day

	λ_{GC}	PCA	MLMA
• 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
also see Sul & Eskin 2013 Nat Rev Genet, Price et al. 2013 Nat Rev Genet

Mixed model = Fixed effects + Random effects

$$Y = \underbrace{X\beta}_{\text{fixed effects}} + \underbrace{u}_{\text{random effects}} + \varepsilon$$

Y = phenotypes

X = candidate SNP genotypes (+ covariates)

β = effect size of candidate SNP (+ covariates)

Var(u) = σ_g^2A, where A is the genetic relationship matrix (GRM)

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

Var(ε) = σ_e^2I

V = Var($u + \varepsilon$) = σ_g^2A + σ_e^2I

Kang et al. 2010 Nat Genet; reviewed in Price et al. 2010 Nat Rev Genet
also see Sul & Eskin 2013 Nat Rev Genet, Price et al. 2013 Nat Rev Genet

Mixed model association: the χ^2 statistic

$$Y = \underbrace{X\beta}_{\text{fixed effects}} + \underbrace{u}_{\text{random effects}} + \varepsilon$$

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

MLMA with no covariates:

Score test: $\boxed{\chi^2 = (X^TV^{-1}Y)^2 / X^TV^{-1}X}$ (Chen & Abecasis 2007 AJHG)

generalizes ATT $\chi^2 = (X^TY)^2 / X^TX$ (Armitage 1955 Biometrics)

Kang et al. 2010 Nat Genet, reviewed in Price et al. 2010 Nat Rev Genet. Also see Lippert et al. 2011 Nat Methods, Listgarten et al. 2012 Nat Methods, Segura et al. 2012 Nat Genet, Zhou & Stephens 2012 Nat Genet, Svishcheva et al. 2012 Nat Genet (Multivariate case: see Korte et al. 2012 Nat Genet, **ASHG 2013 Poster 1790W Zhou**)

Correcting for confounding is important …

CONFOUNDING

Correcting for confounding is important … but so is maximizing power!

CONFOUNDING

POWER

Take-Home Messages

1. Excluding candidate marker from GRM (MLMe) increases power, but including candidate marker in GRM (MLMi) decreases power.

2. Using a small subset of markers in GRM can increase power, but can compromise correction for population stratification.

3. Mixed model methods can suffer a loss in power in ascertained case-control studies .

Take-Home Messages

1. **Excluding candidate marker from GRM (MLMe) increases power, but including candidate marker in GRM (MLMi) decreases power.**

2. Using a small subset of markers in GRM can increase power, but can compromise correction for population stratification.

3. Mixed model methods can suffer a loss in power in ascertained case-control studies .

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 <u>for each candidate SNP</u>
Time cost O(MN^3) where M = #SNPs, N = #samples

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

also see Lippert et al. 2011 Nat Methods, Segura et al. 2012 Nat Genet,
Zhou & Stephens 2012 Nat Genet, Svishcheva et al. 2012 Nat Genet

BUT including candidate SNP in GRM (MLMi) can lead to a decrease in power

Linear Regression (LR):
e.g. Armitage trend test

MLMi:
Mixed model association with candidate marker included in GRM

MLMe:
Mixed model association with candidate marker excluded from GRM

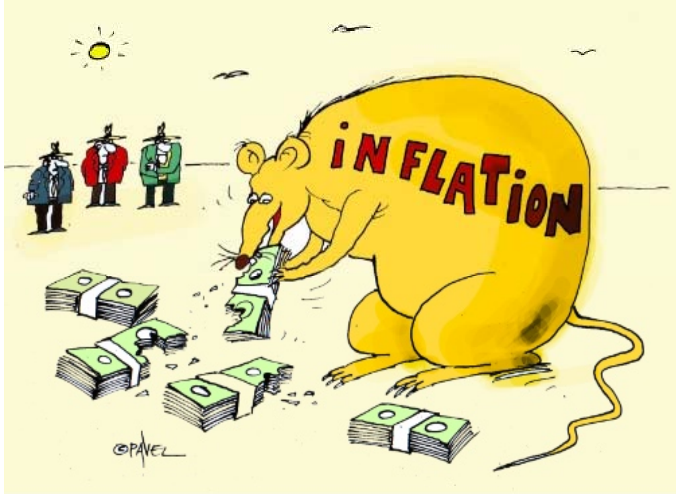
We derive average χ^2 statistics of LR, MLMi, MLMe as a function of N = #samples, M = effective # independent markers, and h_g^2 = heritability explained by genotyped SNPs

Linear Regression: Average $\chi^2 > 1$

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

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

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



$\cancel{\Leftrightarrow}$

CONFOUNDING

New method to determine whether inflation in average χ^2 is due to polygenicity or confounding: **ASHG 2013 poster 1799W Bulik-Sullivan**

MLMi is deflated and decreases power

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

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

MLMi:	All markers	Null markers
(if $N < M$, $r^2 \approx h_g^2 N/M$)	1	$\boxed{\dfrac{1-r^2h_g^2}{h_g^2 N/M+1-r^2h_g^2}}$

MLMi average $\chi^2 < 1$ at null markers: power loss / miscalibration!
Including a candidate SNP in the null model inflates the null likelihood and deflates χ^2 statistics (Listgarten et al. 2012 Nat Methods)

also see Sawcer et al. 2011 Nature, Lippert et al. 2011 Nat Methods

MLMe is well-calibrated and increases power!

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

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

[all derivations validated by extensive simulations]

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

	LR	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

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

New method for MLMe association in O(*MN*) time (!):
<u>ASHG 2013 Poster 1780F Loh</u> (Friday 11:30am-12:30pm).

Take-Home Messages

1. Excluding candidate marker from GRM (MLMe) increases power, but including candidate marker in GRM (MLMi) decreases power.

2. **Using a small subset of markers in GRM can increase power, but can compromise correction for population stratification.**

3. Mixed model methods can suffer a loss in power in ascertained case-control studies .

Using a small subset of markers in GRM can substantially increase power

- MLMe increases power by implicitly conditioning on markers in the GRM. Mathematically equivalent to association analysis on residual phenotype that is residualized for BLUP prediction.

  (Henderson 1975 Biometrics; Svischeva et al. 2012 Nat Genet)

- BLUP prediction using all SNPs is Best Linear Unbiased Pred., but other schemes (e.g. subset of top SNPs) may perform better.

  (de los Campos et al. 2010 Nat Rev Genet; Erbe et al. 2012 J Dairy Sci)

- FaST-LMM-Select approach: build GRM using a subset of $M_T \leq M$ top associated SNPs (excluding candidate SNP: MLMe)

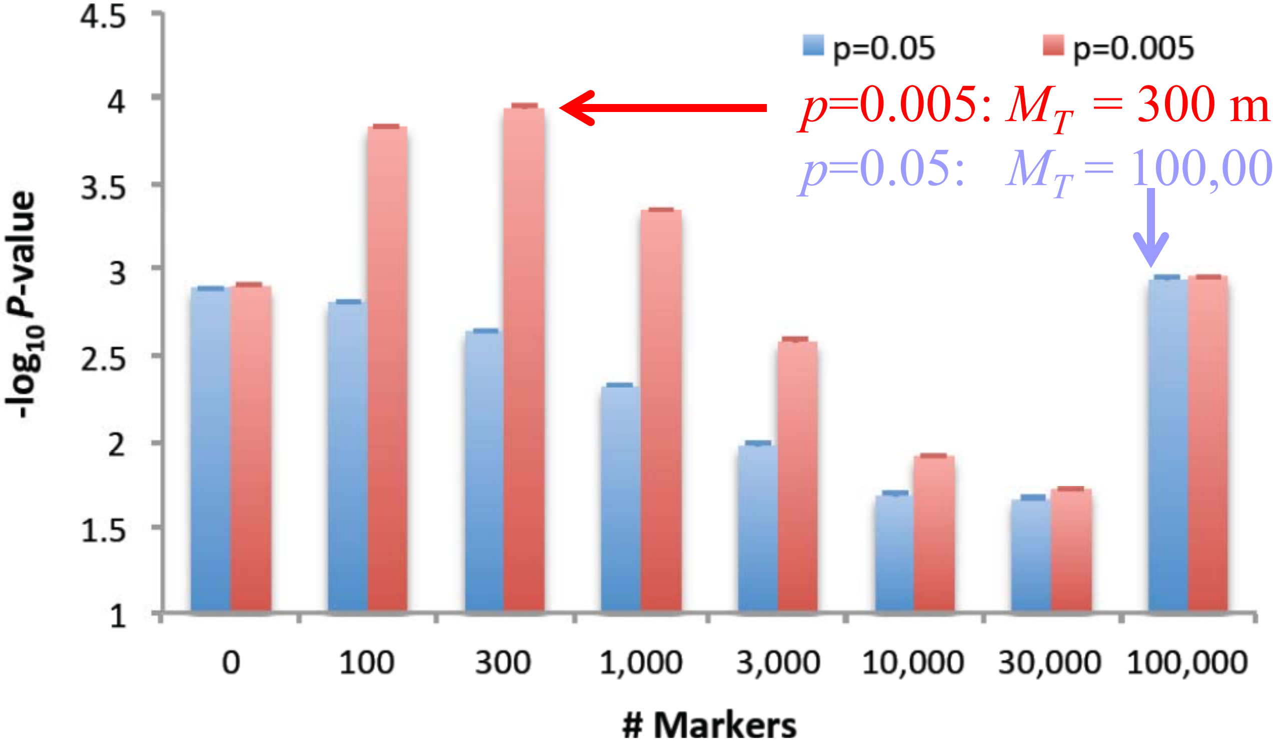
  (Listgarten et al. 2012 Nat Methods; Lippert et al. 2013 Sci Rep; also see Listgarten et al. 2013 Nat Genet)

Using a small subset of markers in GRM can substantially increase power

Quant. trait simulation, N = 10,000 samples, M = 100,000 markers

Mp causal markers explain 50% of trait variance, p = 0.005 or 0.05

M_T top associated markers in GRM, assess power at causal markers



p=0.005: M_T = 300 maximizes power

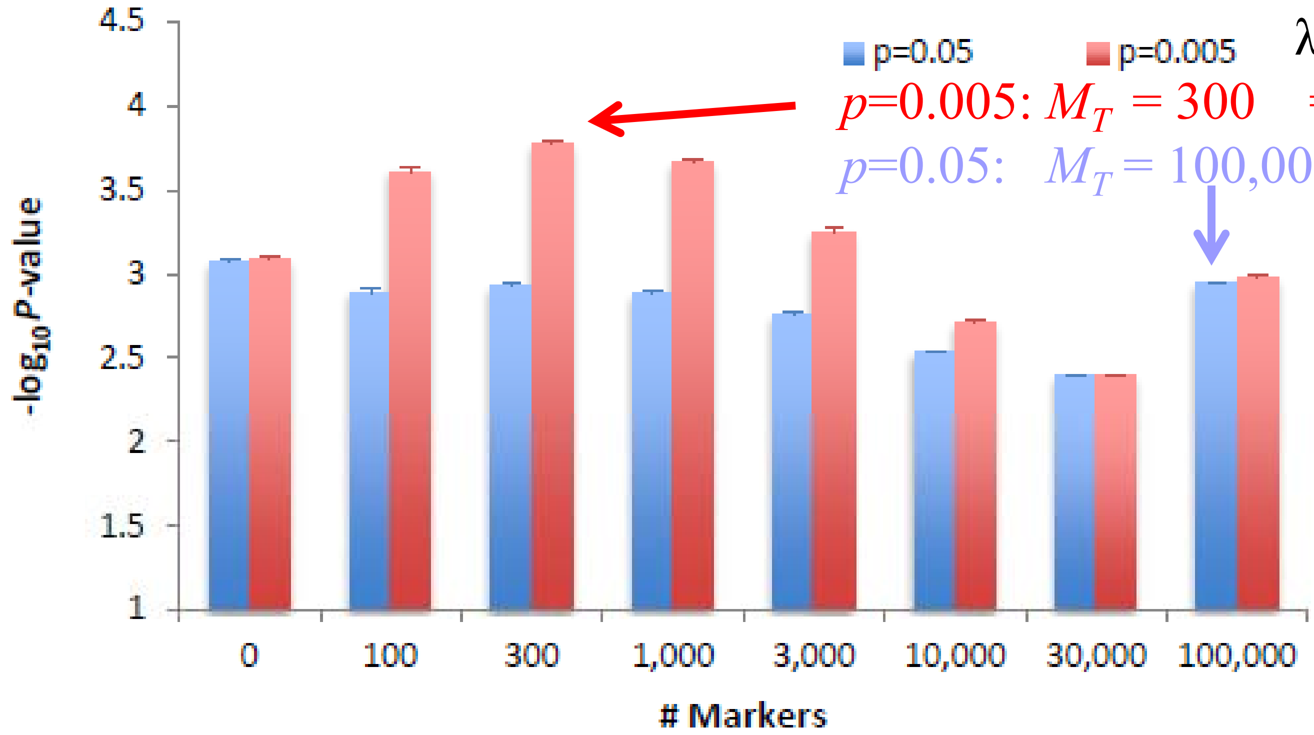
p=0.05: M_T = 100,000 maximizes power

Using a small subset of markers in GRM can compromise correction for stratification

Quant. trait simulation, N = 10,000 samples, M = 100,000 markers

Mp causal markers explain 50% of trait variance, p = 0.005 or 0.05

Add stratification (trait diff. 0.25 between 2 pop. with F_{ST}=0.005)



λ_{GC} (null markers)

p=0.005: M_T = 300 => $\boxed{\lambda_{GC} = 1.26}$

p=0.05: M_T = 100,000 => λ_{GC} = 1.01

Using a subset of markers in GRM: recommendations

- The method that maximizes power may be different from the method that optimizes correction for stratification.
- In data sets where stratification is not a major concern, methods that select a subset of markers in the GRM to maximize power should be used, as they can substantially increase power.
- In data sets with stratification, our current recommendation among published methods is to include all markers in the GRM, but more work is needed.

See ASHG 2013 poster 1775W Heckerman (Wed 10:30am-11:30am) and ASHG 2013 poster 1785T Tucker (Thu 10:30am-11:30am) for ongoing work, including FaST-LMM-Select + PCs approach.

Take-Home Messages

1. Excluding candidate marker from GRM (MLMe) increases power, but including candidate marker in GRM (MLMi) decreases power.

2. Using a small subset of markers in GRM can increase power, but can compromise correction for population stratification.

3. Mixed model methods can suffer a loss in power in ascertained case-control studies.

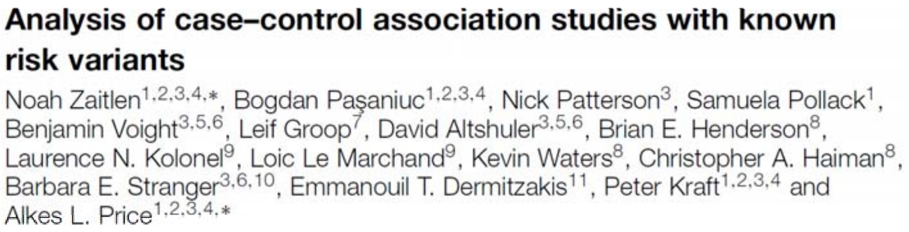
Covariates with case-control ascertainment: standard methods suffer a loss in power

Including known covariates can reduce power to detect genetic effects in case-control studies

Matti Pirinen[1], Peter Donnelly[1,2] & Chris C A Spencer[1]

(Pirinen et al. 2012 Nat Genet)

Our solution: LTSCORE (use posterior mean residual liabilities):



Analysis of case–control association studies with known risk variants

Noah Zaitlen[1,2,3,4,*], Bogdan Paşaniuc[1,2,3,4], Nick Patterson[3], Samuela Pollack[1], Benjamin Voight[3,5,6], Leif Groop[7], David Altshuler[3,5,6], Brian E. Henderson[8], Laurence N. Kolonel[9], Loic Le Marchand[9], Kevin Waters[8], Christopher A. Haiman[8], Barbara E. Stranger[3,6,10], Emmanouil T. Dermitzakis[11], Peter Kraft[1,2,3,4] and Alkes L. Price[1,2,3,4,*]

Informed Conditioning on Clinical Covariates Increases Power in Case-Control Association Studies

Noah Zaitlen[1,2,3,4*], Sara Lindström[1,4], Bogdan Pasaniuc[1,2,3,4], Marilyn Cornelis[5], Giulio Genovese[6], Samuela Pollack[1,2,3,4], Anne Barton[7], Heike Bickeböller[8], Donald W. Bowden[9], Steve Eyre[7], Barry I. Freedman[10], David J. Friedman[6], John K. Field[11], Leif Groop[12], Aage Haugen[13], Joachim Heinrich[14], Brian E. Henderson[15], Pamela J. Hicks[16], Lynne J. Hocking[17], Laurence N. Kolonel[18], Maria Teresa Landi[19], Carl D. Langefeld[20], Loic Le Marchand[18], Michael Meister[21,22], Ann W. Morgan[23], Olaide Y. Raji[11], Angela Risch[22,24], Albert Rosenberger[8], David Scherf[24], Sophia Steer[25], Martin Walshaw[26], Kevin M. Waters[15], Anthony G. Wilson[27], Paul Wordsworth[28], Shanbeh Zienolddiny[13], Eric Tchetgen Tchetgen[1,2], Christopher Haiman[15], David J. Hunter[1,3,4,5], Robert M. Plenge[3,29], Jane Worthington[7], David C. Christiani[1,30], Debra A. Schaumberg[1,31,32], Daniel I. Chasman[32], David Altshuler[3,33,34], Benjamin Voight[3,33,34], Peter Kraft[1,2,3,4], Nick Patterson[3], Alkes L. Price[1,2,3,4*]

(Zaitlen et al. 2012 Bioinformatics;
Zaitlen et al. 2012 PLoS Genet)

reviewed in Mefford & Witte 2012 PLoS Genet
also see Clayton 2012 Genet Epidemol

Mixed models with case-control ascertainment: simulations confirm loss in power

Various values of *N*, *M*, disease prevalence
Effect sizes: 10/*N* of variance on observed scale
$-\log_{10}P$-values ± s.e., based on 100 simulations

# samples (*N*)	#markers (*M*)	Disease prevalence (*F*)	Linear regression	MLMe
10,000	10,000	0.001	3.06 ± 0.15	**2.22 ± 0.12**
10,000	10,000	0.01	3.04 ± 0.16	2.64 ± 0.14
10,000	10,000	0.1	3.04 ± 0.17	3.06 ± 0.17
10,000	100,000	0.001	2.96 ± 0.16	**2.78 ± 0.16**
10,000	100,000	0.01	2.66 ± 0.14	2.54 ± 0.13
10,000	100,000	0.1	3.24 ± 0.16	3.26 ± 0.16

Note: variance component parameters are misestimated by MLMe, even after correction for case-control ascertainment (Lee et al. 2011 Am J Hum Genet).
Also see **ASHG 2013 poster 1000T Golan** **(Thu 11:30am-12:30pm)**

Mixed models with case-control ascertainment: recommendations

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

- For high-prevalence diseases (1%-10%), MLMe is ok.

- Otherwise, for low-prevalence diseases with large N, use other methods (e.g. PCA) to avoid loss in power.

See ASHG 2013 poster 1773T Hayeck (Thu 10:30am-11:30am) and ASHG 2013 poster 953F Weissbrod (Fri 10:30am-11:30am) for ongoing work on mixed models with case-control ascertainment and results on WTCCC phenotypes.

Take-Home Messages

1. **Excluding candidate marker from GRM (MLMe) increases power, but including candidate marker in GRM (MLMi) decreases power.**
   More work is needed!

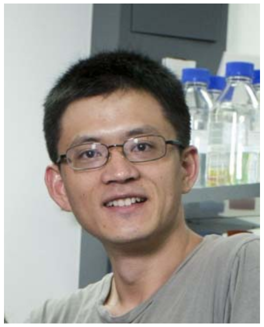
2. **Using a small subset of markers in GRM can increase power, but can compromise correction for population stratification.**
   More work is needed!

3. **Mixed model methods can suffer a loss in power in ascertained case-control studies.**
   More work is needed!

… and more work is needed on studies of rare variants using MLMA.

(Mathieson & McVean 2012 Nat Genet; Listgarten et al. 2013 Nat Genet)

Acknowledgements



Jian Yang, U. of Queensland
Noah Zaitlen, UCSF
Mike Goddard, U. of Melbourne
Peter Visscher, U. of Queensland
(Yang*, Zaitlen* et al. under revision)

With additional thanks to:
N. Patterson, D. Heckerman, J. Listgarten, C. Lippert, E. Eskin, B. Vilhjalmsson, P. Loh, G. Tucker, T. Hayeck, T. Frayling, A. McRae, L. Ronnegart, O. Weissbrod, GIANT consortium, A. Gusev, S. Pollack.

To download slides of this talk: google "Alkes HSPH"