Genome-wide complex trait analysis and extensions

Matthew Keller
Teresa de Candia

University of Colorado at Boulder

Outline

- Overview of GCTA (Keller)
  - how it works
  - what it tells us
- Practical – using GCTA to get "SNP heritability" for three traits
- Issues and extensions of GCTA (de Candia)
  - Assumptions
  - SNP data quality control
  - Additional topics

Outline

- Overview of GCTA (Keller)
  - how it works
  - what it tells us
- Practical – using GCTA to get "SNP heritability" for three traits
- Issues and extensions of GCTA (de Candia)
  - Assumptions
  - SNP data quality control
  - Additional topics

Missing heritability

- The sum of R^2 of significantly associated SNPs typically < 5%. Why?
- One possibility: large number of small-effect SNPs (the 'infinitesimal model'; Fisher, 1918) that failed to reach genome-wide significance (many type-II errors)
- GCTA[1] designed to test this

[1]Yang et al, 2010, Nature Genetics

GCTA

- Determine extent to which genetic similarity at SNPs is related to phenotypic similarity
- By treating genetic effects as random effects, a mixed linear model derives unbiased estimate of V_A captured by measured (common) SNPs
  - Need to remove 'close' relatives, like 2nd-cousins, to minimize any confounding of shared environment with pi-hat
  - Need to control for 'ethnic PCs' to minimize confounding of ethnicity (and cultural factors) with pi-hat

Similarity bw GCTA & twin studies

$$\theta_{ij} = Z_i Z_j$$

product of centered scores (here, z-scores)

$$E[\theta_{ij}] = COV(Z_i, Z_j)$$

$$E[\theta_{ij} \mid \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is an estimate of h^2)

Similarity bw GCTA & twin studies

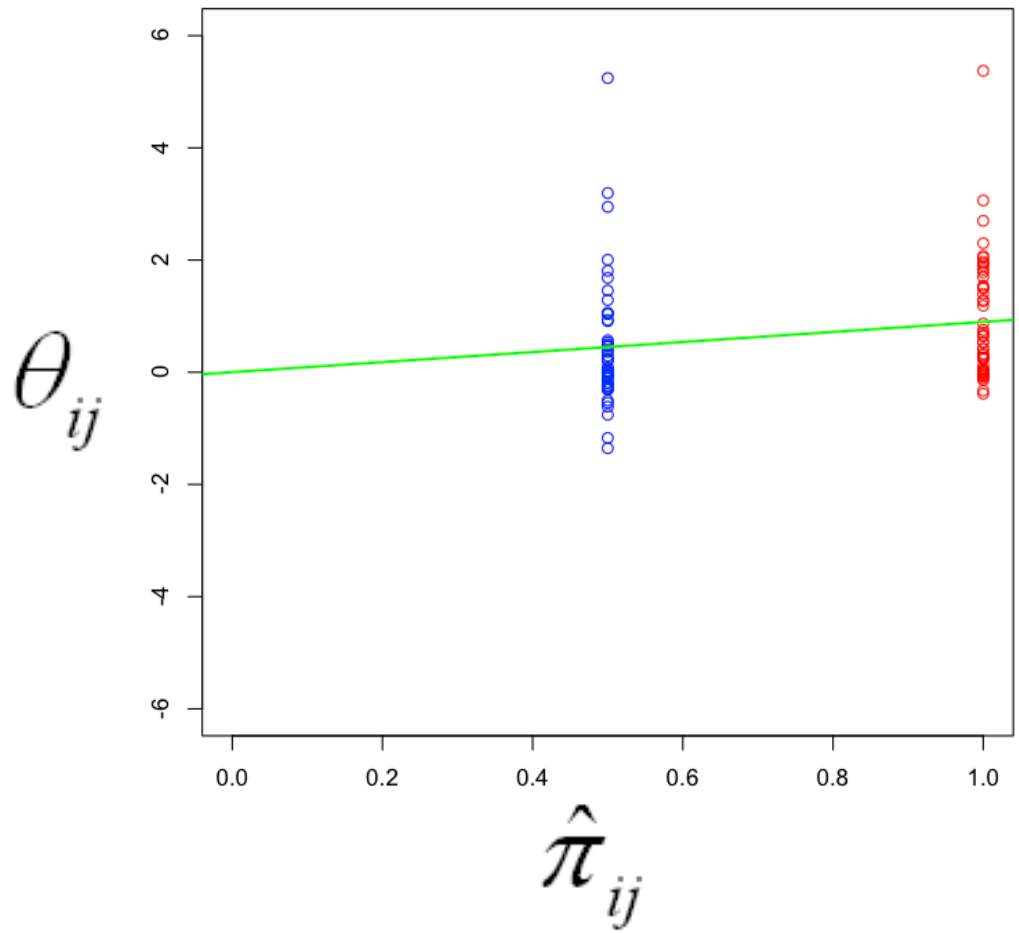
$$\theta_{ij} = Z_i Z_j$$

$$E[\theta_{ij}] = COV(Z_i, Z_j)$$

$$E[\theta_{ij} \mid \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is an estimate of h^2)



Similarity bw GCTA & twin studies

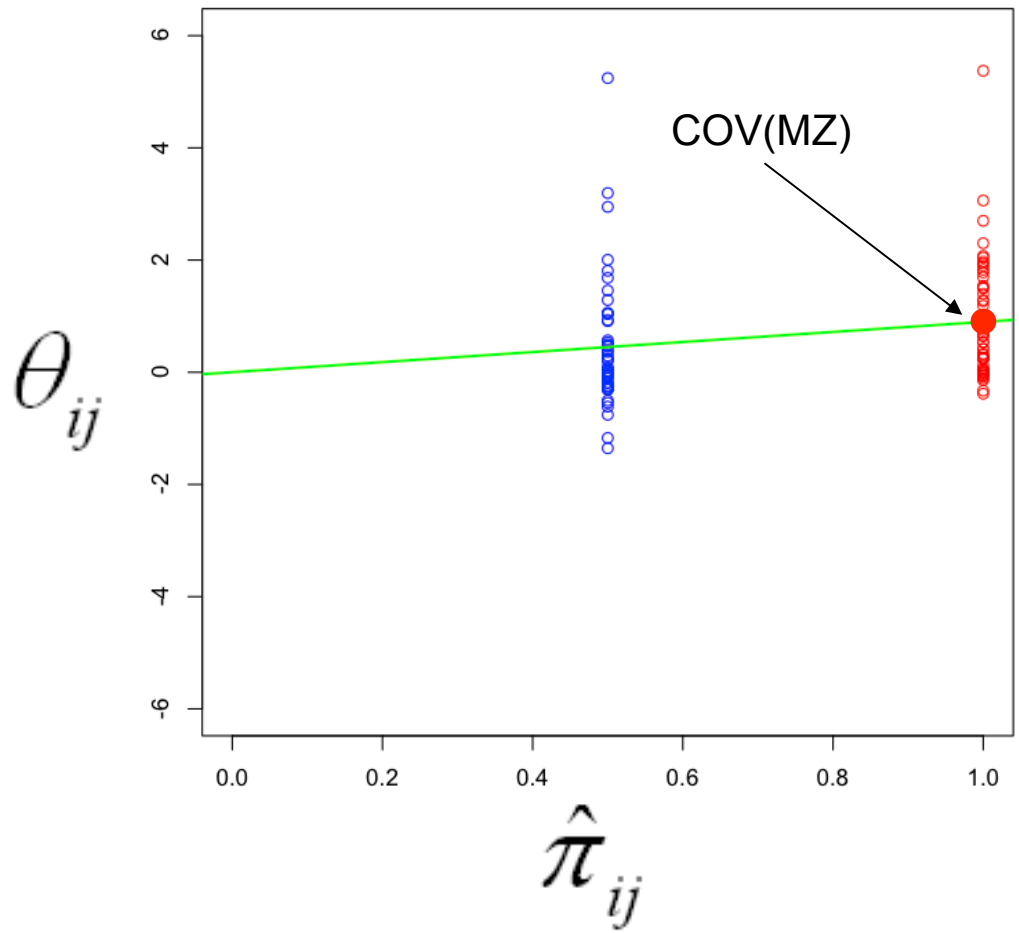
$$\theta_{ij} = Z_i Z_j$$

$$E[\theta_{ij}] = COV(Z_i, Z_j)$$

$$E[\theta_{ij} \mid \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is an estimate of h^2)



Similarity bw GCTA & twin studies

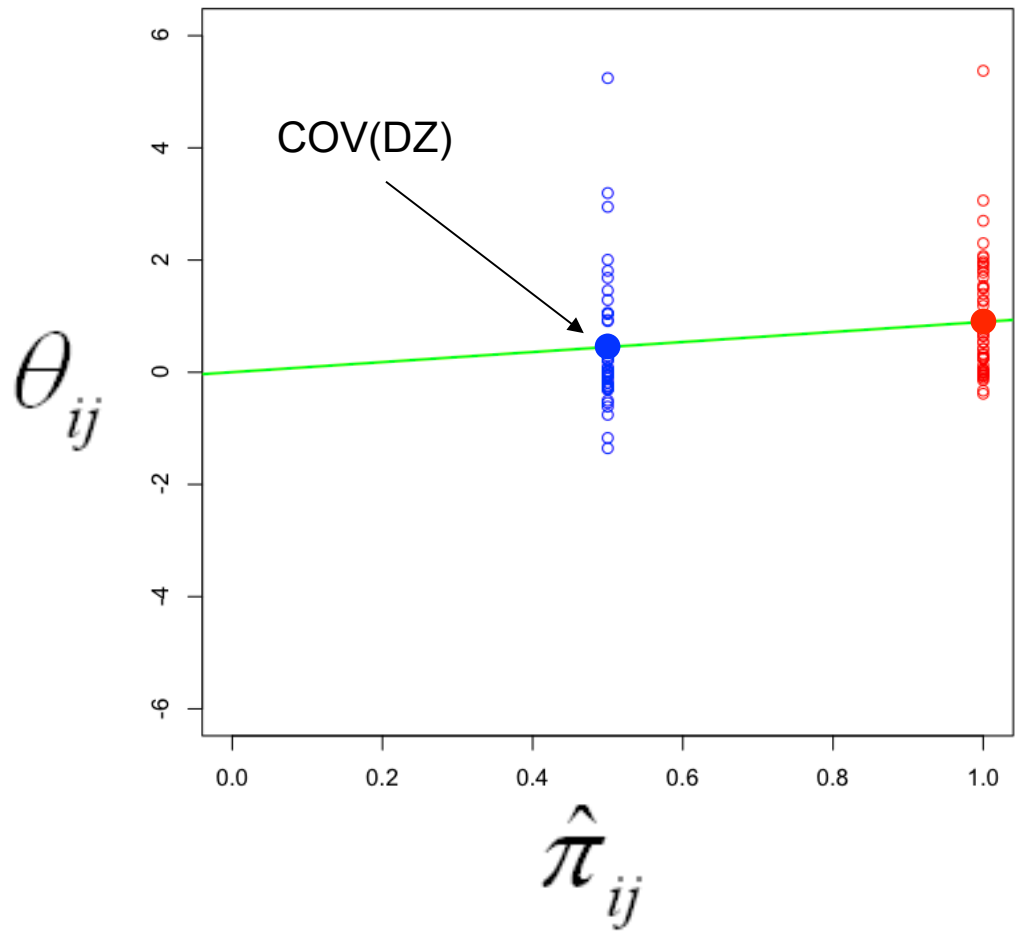
$$\theta_{ij} = Z_i Z_j$$

$$E[\theta_{ij}] = COV(Z_i, Z_j)$$

$$E[\theta_{ij} \mid \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is an estimate of h^2)



Similarity bw GCTA & twin studies

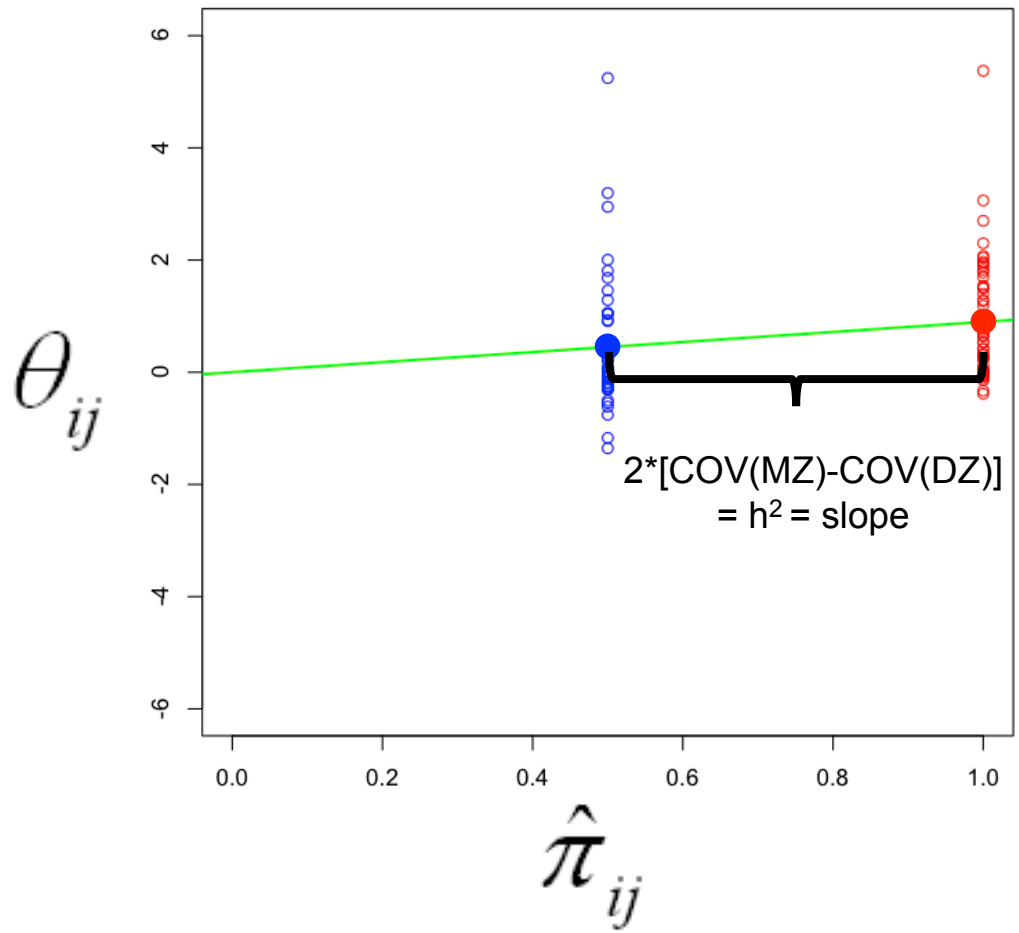
$$\theta_{ij} = Z_i Z_j$$

$$E[\theta_{ij}] = COV(Z_i, Z_j)$$

$$E[\theta_{ij} \mid \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is an estimate of h^2)



Similarity bw GCTA & twin studies

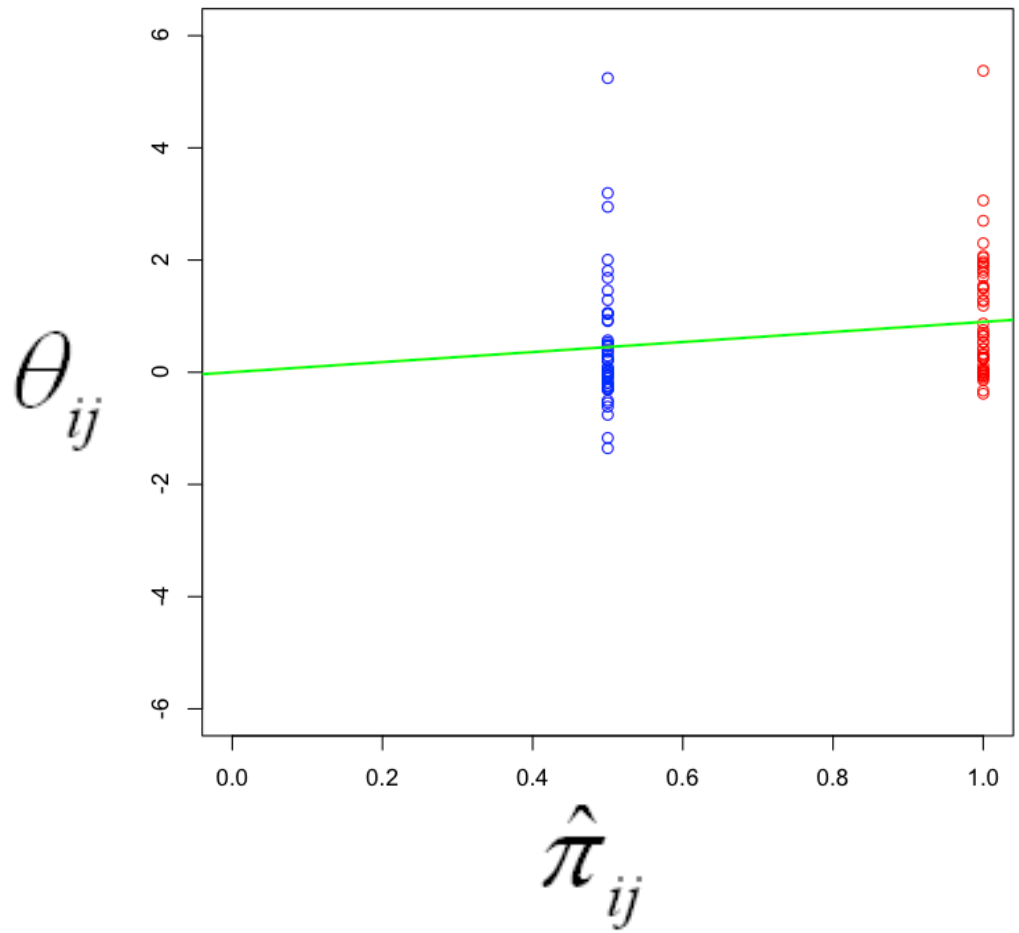
$$\theta_{ij} = Z_i Z_j$$

$$E[\theta_{ij}] = COV(Z_i, Z_j)$$

$$E[\theta_{ij} \mid \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is an estimate of h^2)



Similarity bw GCTA & twin studies

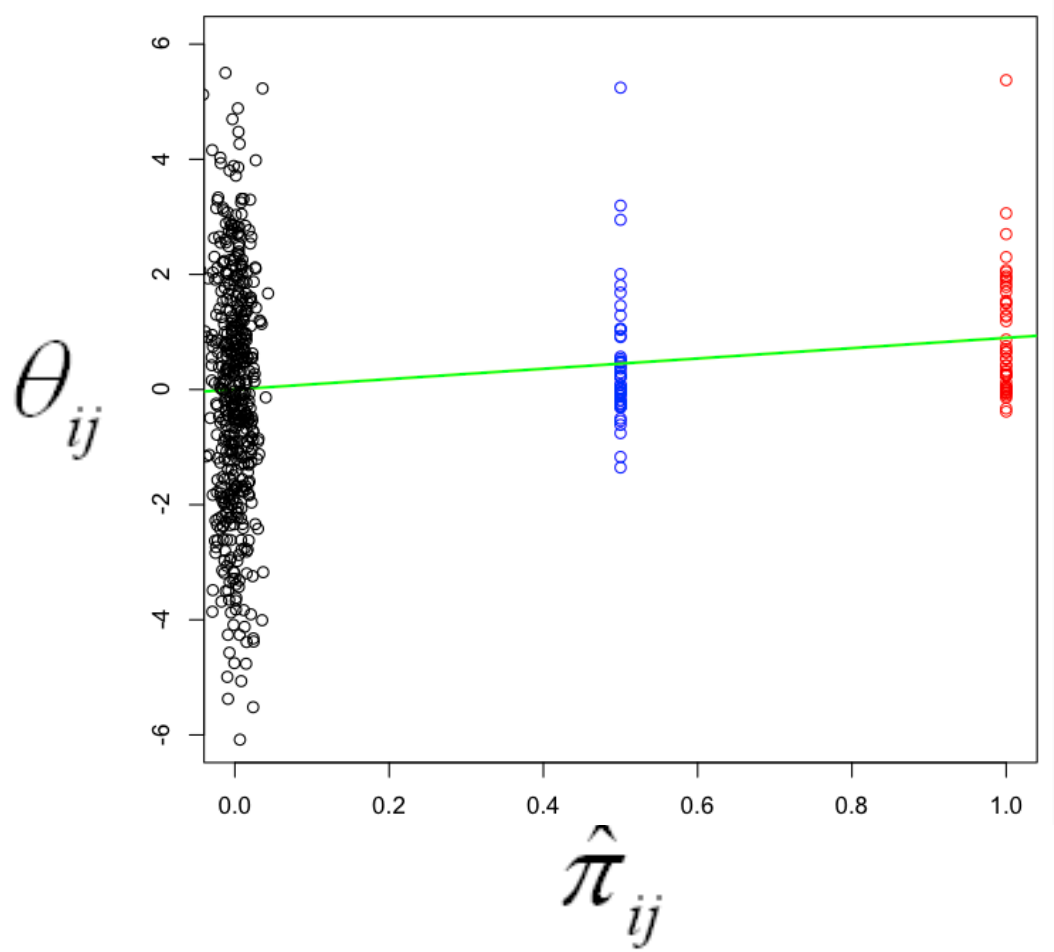
$$\theta_{ij} = Z_i Z_j$$

$$E[\theta_{ij}] = COV(Z_i, Z_j)$$

$$E[\theta_{ij} \mid \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is an estimate of h^2)



Similarity bw GCTA & twin studies

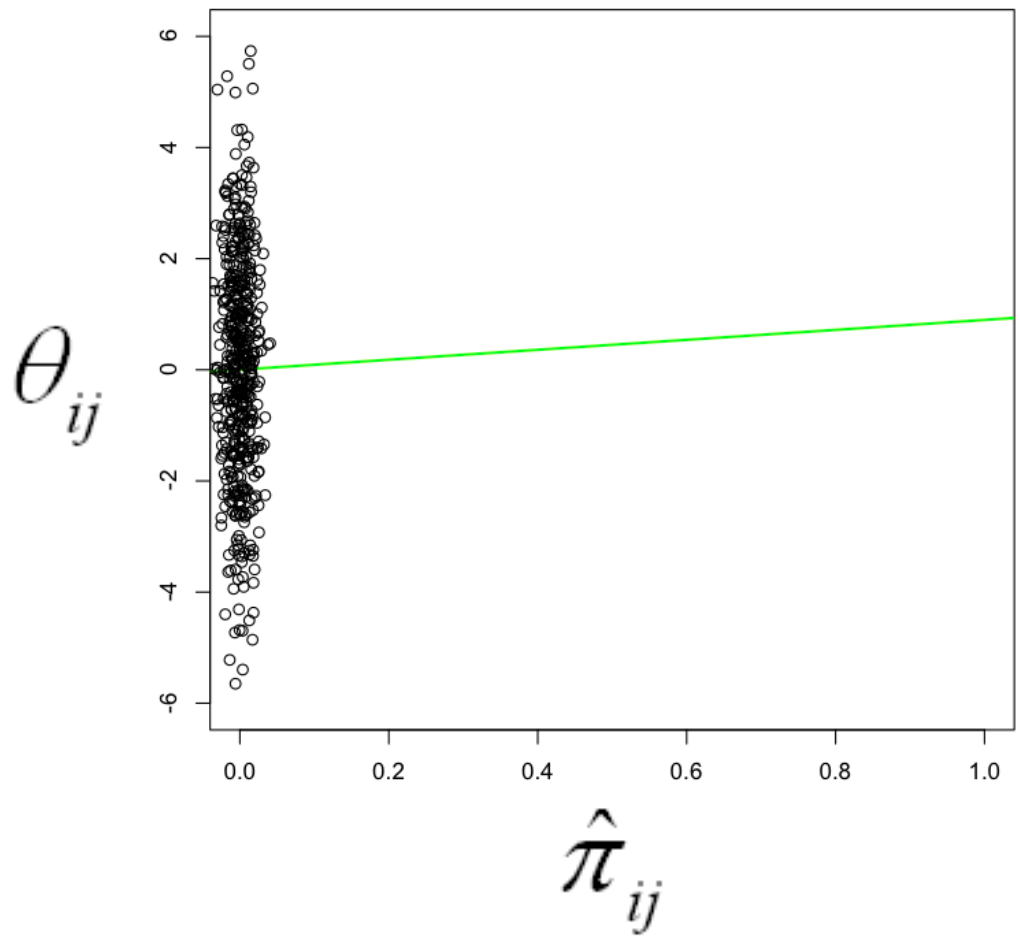
$$\theta_{ij} = Z_i Z_j$$

$$E[\theta_{ij}] = COV(Z_i, Z_j)$$

$$E[\theta_{ij} \mid \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is an estimate of h^2)



Genetic Relationship Matrix (GRM)

- Rather than $n/2$ twin pairs, we fit this model to $n(n+1)/2$ genetic pairwise relationships
- Each element of this GRM matrix is:

$$\hat{\pi}_{ij} = \frac{1}{N}\sum_k \frac{(x_{ki} - 2p_k)(x_{kj} - 2p_k)}{2p_k(1-p_k)}$$

where x_{ki} is the kth SNP (k=1...N) of the ith person, taking the value of 0, 1, or 2 if it is AA, Aa, aa.

- GCTA estimates V_A using REML rather than least squares regression

GRM example, 3 individuals, 1 SNP

$$\hat{\pi}_{ij} = \frac{1}{N}\sum_k \frac{(x_{ki} - 2p_k)(x_{kj} - 2p_k)}{2p_k(1-p_k)} \quad *$$

For one SNP (N=1), say that P(a) = .45, P(A)=.55.
Person 1 is AA, 2 is aa, and 3 Aa:

AA 0	aa 2	Aa 1	x_{ki}	
1.63	-2.0	-0.18	0	AA
-2.0	2.44	0.22	2	aa
-0.18	0.22	0.02	1	Aa

*Note: GCTA uses a slightly modified formula when i=j

GRM example, 3 individuals, 1 SNP

$$\hat{\pi}_{ij} = \frac{1}{N}\sum_k \frac{(x_{ki} - 2p_k)(x_{kj} - 2p_k)}{2p_k(1-p_k)}$$

For one SNP (N=1), say that P(a) = .45, P(A)=.55.
Person 1 is AA, 2 is aa, and 3 Aa:

$$\hat{\pi}_{1,1} = \frac{(0-.9)^2}{2\times .45\times .55}$$

AA	aa	Aa	x_{ki}	
0	2	1		
1.63	-2.0	-0.18	0	AA
-2.0	2.44	0.22	2	aa
-0.18	0.22	0.02	1	Aa

GRM example, 3 individuals, 1 SNP

$$\hat{\pi}_{ij} = \frac{1}{N}\sum_k \frac{(x_{ki} - 2p_k)(x_{kj} - 2p_k)}{2p_k(1-p_k)}$$

For one SNP (N=1), say that P(a) = .45, P(A)=.55.
Person 1 is AA, 2 is aa, and 3 Aa:

$$\hat{\pi}_{1,2} = \frac{(0-.9)(2-.9)}{2\times .45\times .55}$$

AA	aa	Aa	x_{ki}	
0	2	1		
1.63	-2.0	-0.18	0	AA
-2.0	2.44	0.22	2	aa
-0.18	0.22	0.02	1	Aa

GRM example, 3 individuals, 1 SNP

$$\hat{\pi}_{ij} = \frac{1}{N}\sum_k \frac{(x_{ki} - 2p_k)(x_{kj} - 2p_k)}{2p_k(1-p_k)}$$

For one SNP (N=1), say that P(a) = .45, P(A)=.55.
Person 1 is AA, 2 is aa, and 3 Aa:

AA	aa	Aa	x_{ki}	
0	2	1		
1.63	-2.0	-0.18	0	AA
-2.0	2.44	0.22	2	aa
-0.18	0.22	0.02	1	Aa

$$\hat{\pi}_{1,3} = \frac{(0-.9)(1-.9)}{2\times .45\times .55}$$

GRM, 3 individuals, 1000 SNPs

$$\hat{\pi}_{ij} = \frac{1}{N} \sum_k \frac{(x_{ki} - 2p_k)(x_{kj} - 2p_k)}{2p_k(1 - p_k)}$$

1.04	-.005	.012
-.005	.995	-.01
.012	-.01	1.21

Mixed linear effect model in GCTA

$$\widehat{\mathrm{var}}(y) = A\hat{\sigma}_g^2 + I\hat{\sigma}_e^2$$

Mixed linear effect model in GCTA

est. add. genetic variance (scalar)

est. environmental variance (scalar)

$$\widehat{\mathrm{var}}(y) = A\hat{\sigma}_g^2 + I\hat{\sigma}_e^2$$

implied n x n var-covar matrix of y's

n x n matrix of pi-hats (GRM)

n x n Identity matrix

Mixed linear effect model in GCTA

Goal of REML is to change $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ in order to get the observed and implied var-covar matrices to be as similar as possible.

$$\mathrm{var}(y) \longleftrightarrow \widehat{\mathrm{var}}(y)$$

observed n x n var-covar matrix of y's

implied n x n var-covar matrix of y's

Mixed linear effect model in GCTA

$$\mathrm{var}(y) \Leftrightarrow A\hat{\sigma}_g^2 + I\hat{\sigma}_e^2$$

$$\begin{bmatrix} 1.57 & -.97 & 2.23 \\ -.97 & .59 & -1.36 \\ 2.23 & -1.36 & 3.16 \end{bmatrix} \Leftrightarrow \begin{bmatrix} 1.04 & -.005 & .012 \\ -.005 & .995 & -.01 \\ .012 & -.01 & 1.21 \end{bmatrix} \hat{\sigma}_g^2 + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \hat{\sigma}_e^2$$

$$\begin{bmatrix} -1.25 \\ 0.77 \\ -1.77 \end{bmatrix} * \begin{bmatrix} -1.25 & .77 & -1.77 \end{bmatrix}$$

(i.e., outer product of the y vector of centered scores for each individual)

What GCTA tells us

- Estimate of V_A captured by common SNPs
- Gives idea of the aggregate importance of common causal variants (bc rare ones poorly tagged by common SNPs)
- Upper bound of how much V_A GWAS can detect
- By not using relatives who also share environmental effects:
  - (a) V_A estimate is 'uncontaminated' by V_C
  - (b) does not rely on assumption that r(MZ) > r(DZ) for purely genetic reasons
  - Allows investigation into several heretofore difficult/ impossible-to-study questions

Outline

- Overview of GCTA (Keller)
  - how it works
  - what it tells us
- Practical – using GCTA to get "SNP heritability" for three traits
- Issues and extensions of GCTA (de Candia)
  - Assumptions
  - SNP data quality control
  - Additional topics

GCTA software

- Several options:
  - Data management (similar to PLINK)
  - Estimation of GRM from genome wide SNPs
  - Estimation variance explained via REML from GRM
  - PCA, Estimation LD structure, Simulation….

Input Files

- Binary PLINK files
  - Fam file `(.fam)`
  - Bim file `(.bim)`
  - Bed file `(.bed)`

Data management

- **Inclusion criteria**
  - --keep mylist.txt, --remove mylist.txt
  - --extract mysnps.txt, --exclude mysnps.txt
  - --chr 6, --autosome
- **Using phenotypes files**
  - --pheno
- **Using covariate files**
  - --covar, --qcovar

Genetic Relationship Matrix (GRM)

- GRM:

```
gcta -bfile simd
--make-grm --out simd.gcta
```

- Generates:
  - `simd.gcta.grm.gz`
  - `simd.gcta.grm.id`

Genetic Relationship Matrix (GRM)

snpdat.gcta.grm.id

```
10  01
10  02
17  01
28  01
33  01
33  02
37  50
38  01
45  50
46  01
```

snpdat.gcta.grm.gz

```
1  1  273588  0.99629
2  1  273566  0.47804
2  2  273600  0.99192
3  1  269152  0.00656
3  2  269164  0.00215
3  3  269192  0.99075
4  1  273582  0.00004
```

```
gcta --bfile simd --make-grm --out simd.gcta --thread-num 2
```

Estimate SNP h^2 in GCTA

- Estimate proportion of phenotypic variance explained by genome wide SNPs for trait1

```
gcta --grm simd.gcta -pheno
simd.pheno --mpheno XXX --reml -out
simd.results
```

"XXX" will be 1 for phenotype data in 3rd column, 2 for phenotype data in 4th, and 3 for phenotype data in 5th. Run it on all 3 phenotypes

Practical - overview

- SNP and trait data are from simulated 20 Mb of SNP data (about 3000 SNPs) on 2000 people
- QC already done (`simd.<bim/bed/fam>`)
- Use “GCTA.Practical.R” to do all this
- First use GCTA to get GRM. Look at the pi-hat distribution
- Then use REML in GCTA to get SNP h^2 estimate for your 3 phenotypes
- Then use least squares regression to get same
- HELP: http://www.complextraitgenomics.com/software/gcta/

Practical - results

- Different h^2 between traits is due to MAF of causal variants (CVs):
  Trait 1: h^2=.60, CV MAF .10-.50
  Trait 2: h^2=.60, CV MAF .01-.05
  Trait 3: h^2=.60, CV MAF .0005-.002

- GCTA works by taking advantage of LD between SNPs and CVs

Outline

- Overview of GCTA (Keller)
  - how it works
  - what it tells us
- Practical – using GCTA to get "SNP heritability" for three traits
- Issues and extensions of GCTA (de Candia)
  - LD
  - quality control
  - additional topics

LD Caveats

- Datasets with fewer SNPs will give lower genetic variance estimates
- Lower MAF CVs will give lower h^2 estimates because poorly tagged by common SNPs
- Regions with higher LD get overrepresented, lower LD underrepresented (Speed et al. 2012)

LD Caveats

- Datasets with fewer SNPs will give lower genetic variance estimates
- **Lower MAF CVs will give lower h^2 estimates because poorly tagged by common SNPs**
- Regions with higher LD get overrepresented, lower LD underrepresented (Speed et al. 2012)

D as a measure of LD

- D compares the observed frequency of a haplotype (e.g., A_1B_1) to the expected when alleles are in LE
  - $D = x_{11} - p_1q_1$, where
    - x_{11} is frequency of A_1B_1
    - p_1 and q_1 are frequencies of A_1 and B_1, respectively
    - Note: requires phased data; iterative procedures can estimate it with unphased data assuming random mating
  - Its range depends on frequency of alleles

Normalized measures of LD

- D` = D / D_{max} , where
  - D_{max} is the theoretical maximum D between two alleles
  - D` varies between -1 and 1

- r = D / sqrt($p_1p_2q_1q_2$)
  - Measure that we're interested in since h^2 that we can infer is function of variances in CVs tagged by measured SNPs

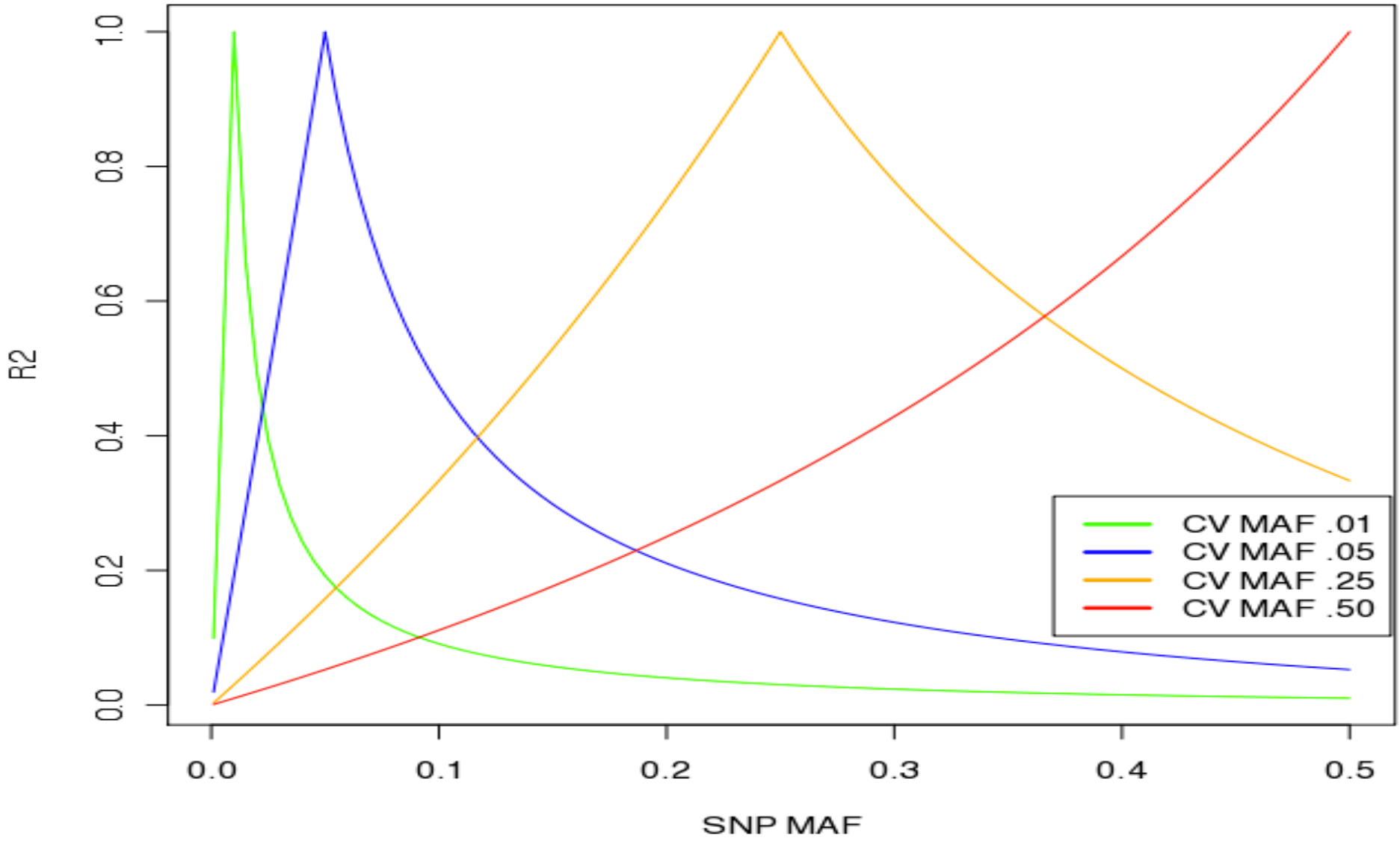
- D` and r are not the same
  - When D` = 1, p_1 = .2, q_1 = .2, r = 1.00
  - When D` = 1, p_1 = .2, q_1 = .5, r = 0.50

h^2 estimates lower for traits influenced by rarer CVs

- SNPs pick up most variance in CVs when they are the same frequency as the CVs
  - GWAS doesn't include lowest MAF SNPs (especially if well cleaned) so lowest MAF CVs unlikely to be tagged perfectly in GWA data

R^2 as function of SNP MAFs

for different CVs



h^2 estimates lower for traits influenced by rarer CVs

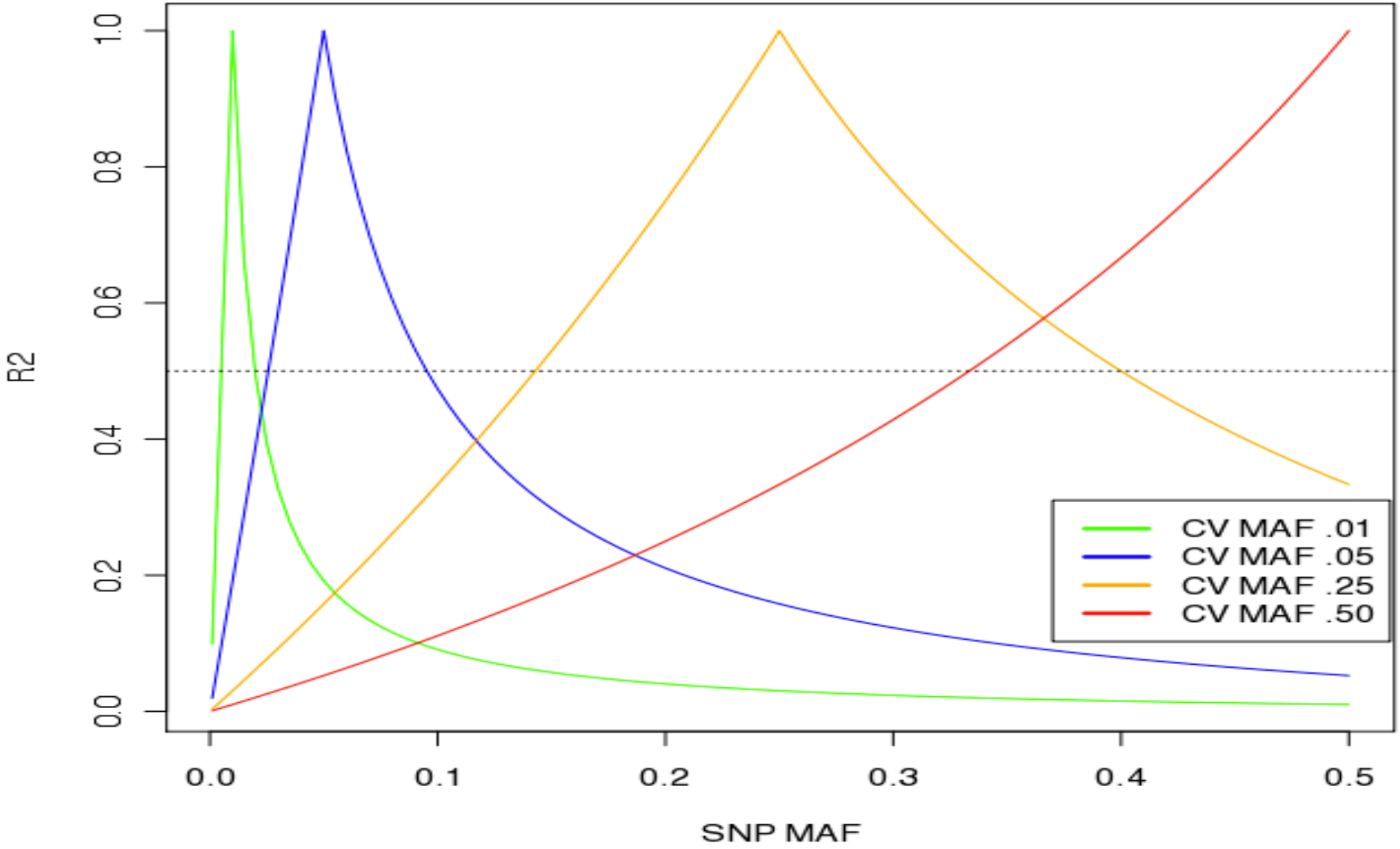
- SNPs pick up most variance in CVs when they are the same frequency as the CVs

  GWAS doesn't include lowest MAF SNPs (especially if well cleaned) so lowest MAF CVs unlikely to be tagged perfectly in GWA data

- The more common the CV, the larger the range of SNPs that will detect it

R^2 as function of SNP MAFs

for different CVs



QC Procedures

1. Reduce standard errors by including covariates and reducing error variance in genotypes

2. Reduce bias in variance estimates by eliminating possible confounds

y = fixed + random g + random e

var(y) = var(g) + var(e)

We assume cov(fixed,g)=0 and cov(g,e)=0

- Be especially careful with case control data

QC Procedures

1. Reduce standard errors by including covariates and reducing error variance in genotypes

2. Reduce bias in variance estimates by eliminating possible confounds

y = fixed + random g + random e

var(y) = var(g) + var(e)

We assume cov(fixed,g)=0 and cov(g,e)=0

- Be especially careful with case control data

QC Procedures to reduce st. error

- Clean data for
  - Subjects missing > ~.02
  - SNPs missing > ~.05
  - HWE p < 10e-6
  - MAF < ~.01
  - Plate effects:
    - Remove plates with extreme average inbreeding coefficients or high average missingness

QC Procedures

1. Reduce standard errors by including covariates and reducing error variance in genotypes

2. Reduce bias in variance estimates by eliminating possible confounds

y = fixed + random g + random e

var(y) = var(g) + var(e)

We assume cov(fixed,g)=0 and cov(g,e)=0

- Be especially careful with case control data

QC Procedures to reduce bias in h^2

- Remove close relatives (e.g., --grm-cutoff 0.05)
  - Correlation between pi-hats and shared environment can inflate h^2 estimates
- Control for stratification (usually 5 or 10 PCs)
  - Different prevalence rates (or ascertainments) between populations can show up as h^2
- Control for plates and other possible technical artifacts
  - With case-control data, be very careful if cases & controls are not randomly placed on plates (can create upward bias in h^2)

Additional Topics - bivariate

- Bivariate analyses can be used to look at genetic overlap between traits and datasets

```
gcta --grm snpdat.gcta --pheno pheno1 --reml-bivar 1 2 --qcovar snpdat.eigenvec --
   covar snpdat.covars --reml-bivar-lrt-rg 0 --out results.pheno1
```

– especially useful for examining overlap between rare traits that are very unlikely to co-occur within families

– $r_g < 1$ between datasets can be due to artifactual differences, or genetic/phenotypic differences between populations

Additional Topics - binning

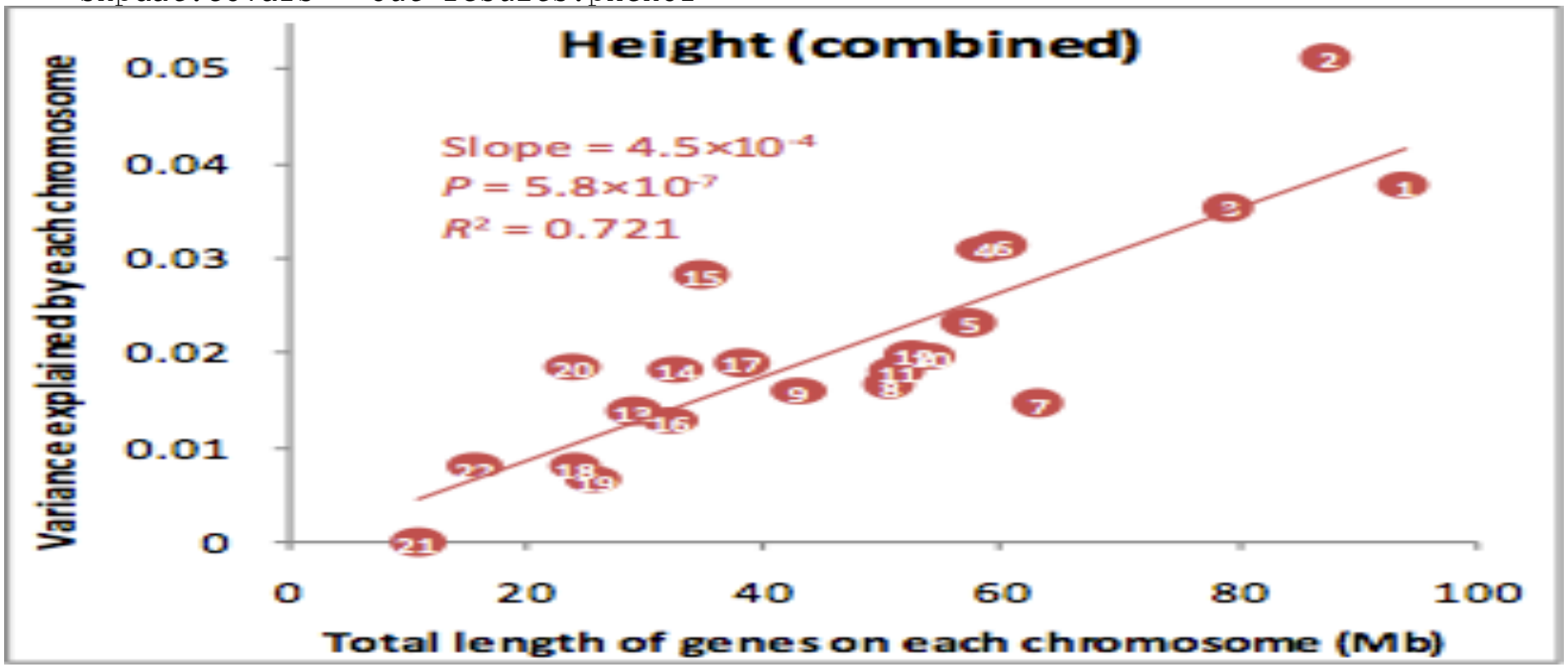
- Bins (i.e., --mgrm) are nice for looking at relevance of functional classes (exonic vs intronic, CNS vs other genes, etc.) and polygeneity

```
gcta --mgrm snpdat.gcta.txt --pheno pheno1 --qcovar snpdat.eigenvec --covar
   snpdat.covars --out results.pheno1
```

Additional Topics - binning

- Bins (i.e., --mgrm) are nice for looking at relevance of functional classes (exonic vs intronic, CNS vs other genes, etc.) and polygeneity

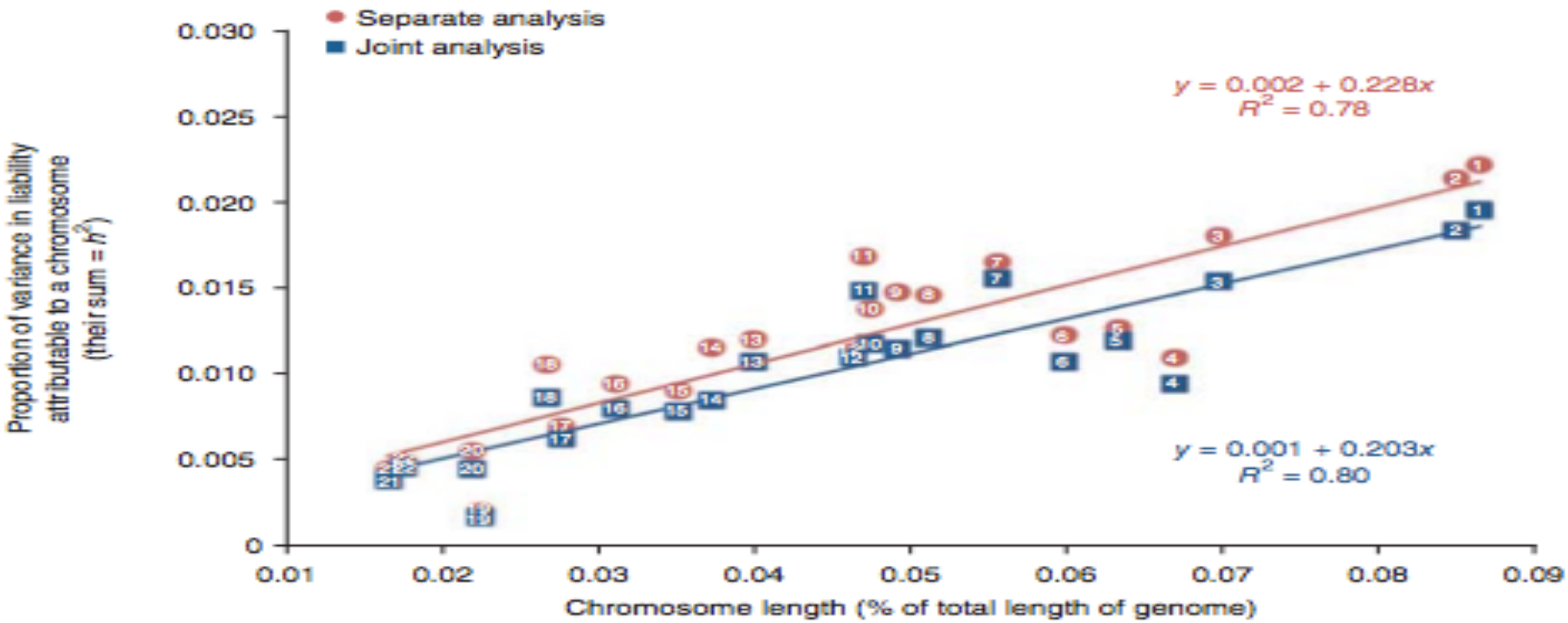
```
gcta --mgrm snpdat.gcta.txt --pheno pheno1 --qcovar snpdat.eigenvec --covar
    snpdat.covars --out results.pheno1
```



Additional Topics - binning

- Bins (i.e., --mgrm) are nice for looking at relevance of functional classes (exonic vs intronic, CNS vs other genes, etc.) and polygeneity

```
gcta --mgrm snpdat.gcta.txt --pheno pheno1 --qcovar snpdat.eigenvec --covar
    snpdat.covars --out results.pheno1
```



Additional Topics - GWAS

- GWAS by including random genetic effects along with fixed effects (SNP 'covariate').
  - SNPs being tested can be included as fixed effects
  - Pi-hats shouldn’t be calculated based on SNPs in LD with SNPs of interest
- Can control for all factors that can inflate h^2 estimates in GCTA - stratification, QC, etc.
- Can increase power by reducing phenotypic variance by the estimated h^2

Acknowledgments

Peter Visscher

Mike Goddard

Jian Yang

Lee Hong

Naomi Wray