Genome-wide Complex Trait Analysis and extensions

Matthew Keller
Teresa de Candia

University of Colorado at Boulder

Outline

- Issues and extensions of GCTA (de Candia)
  - SNP variance estimates and heritability
  - Estimating multiple genetic variances (e.g., two groups of SNPs)
  - Bivariate models (e.g., two traits)
  - Practical

Outline

- Issues and extensions of GCTA (de Candia)
  - SNP variance estimates and heritability
  - **Estimating multiple genetic variances**
  - Bivariate models
  - Practical

Estimating Multiple Genetic Variances

- Just as we can simultaneously estimate partial or independent effects of several predictors in standard linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + .. + \varepsilon$$

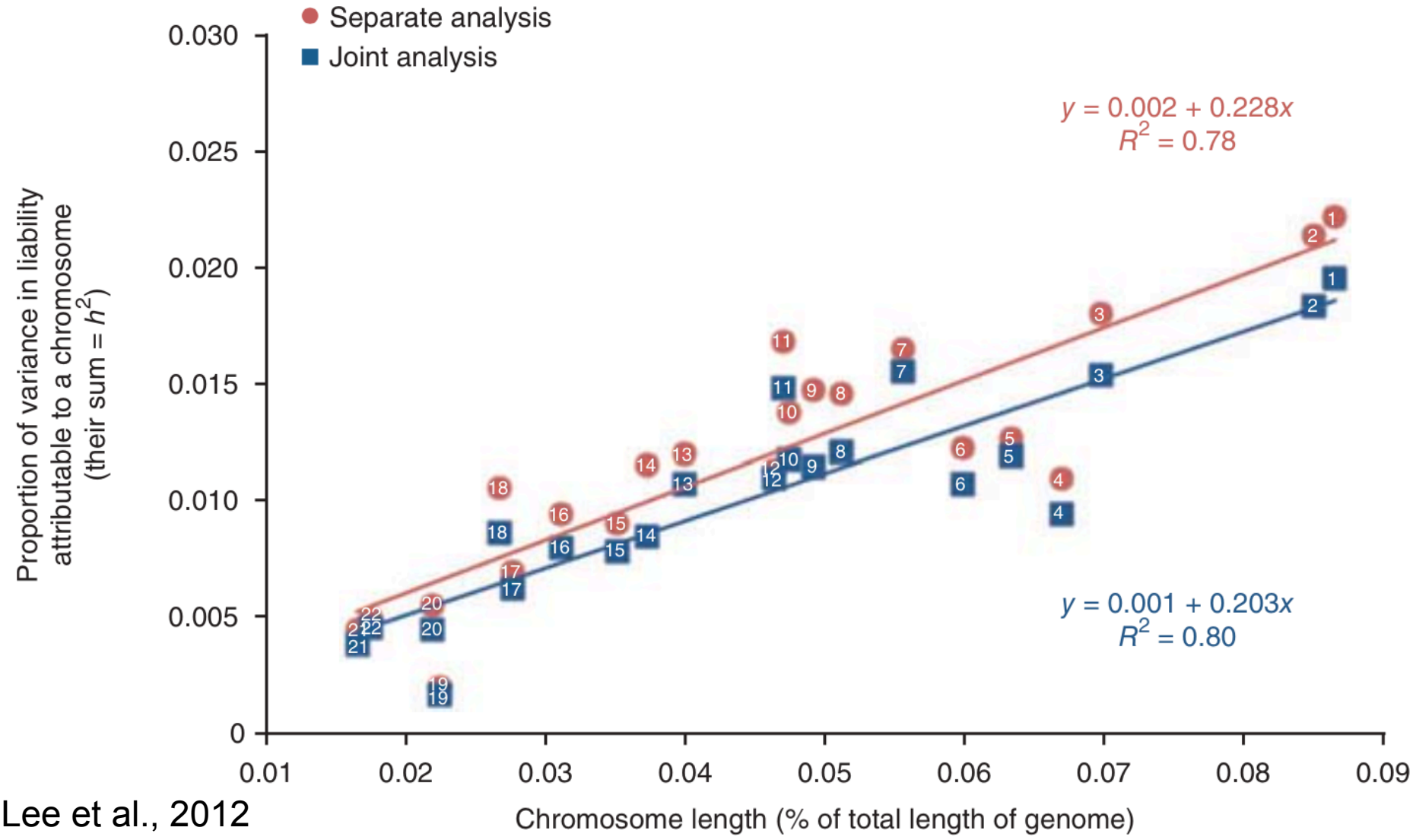
- We can similarly estimate the <u>independent</u> effects of several groups of SNPs:

$$Var(Y) = G_1\sigma^2_{g1} + G_2\sigma^2_{g2} + G_3\sigma^2_{g3} + \ldots + I\sigma^2_e$$

Estimating Multiple Genetic Variances

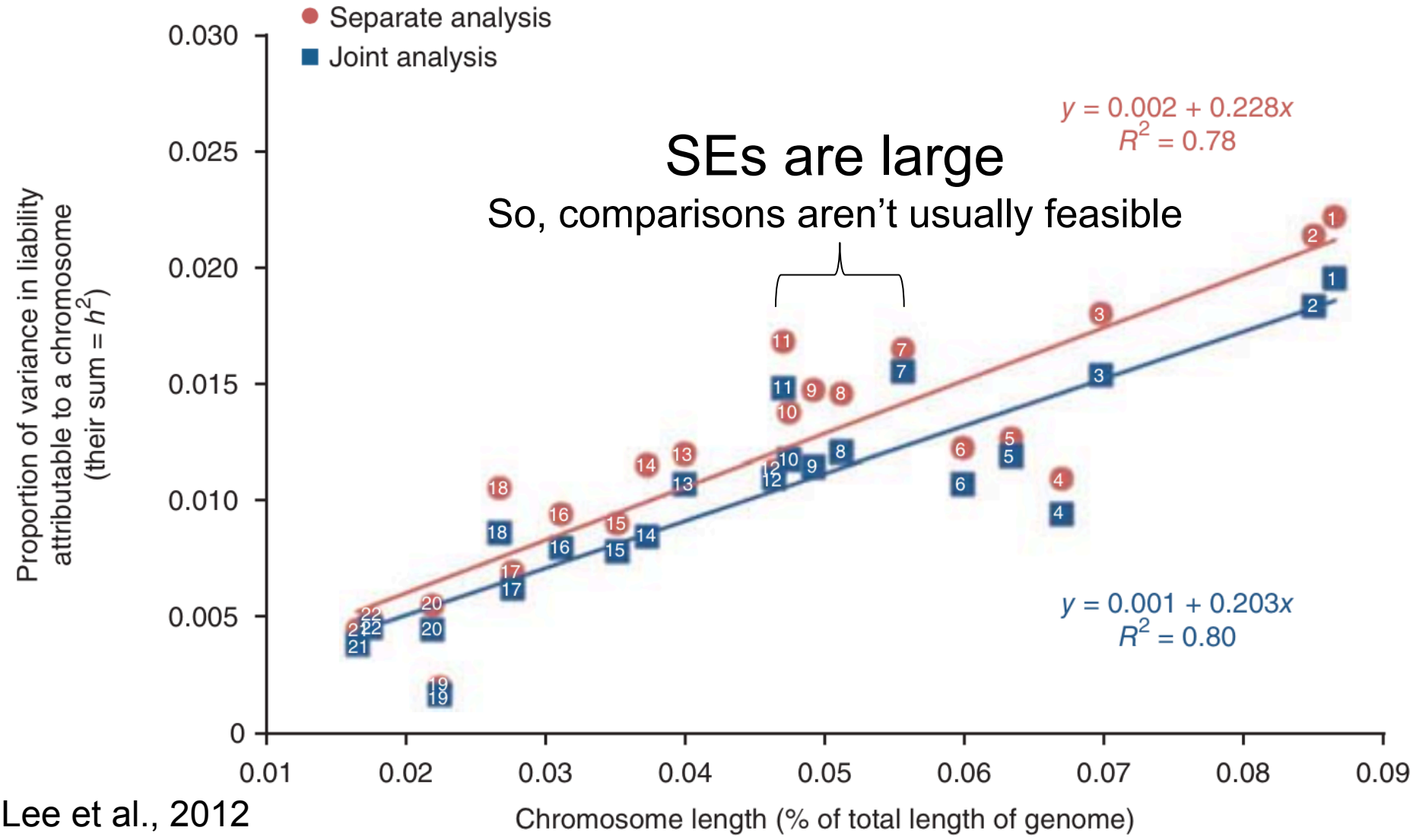
- Advantages:
  - Ability to question specific categories of CVs, by grouping similar SNPs together. Some examples of SNP categories: chromosome, expression, pathway, allele frequency, etc.
  - Increase robustness of model to overestimates due to confounding of GRM due to:
    - Stratification when environmental influences **DO** differ by ethnicity: π-hats confounded w/ environmental effects,
    - **Systematic** plate effects (e.g., case control data): π-hats confounded w/ genotyping artifacts,
    - Cryptic relatedness: π-hats confounded w/ rare or non-additive genetic variance, or shared environmental effects

SNP variances for schizophrenia, by chromosome in ethnically homogeneous sample



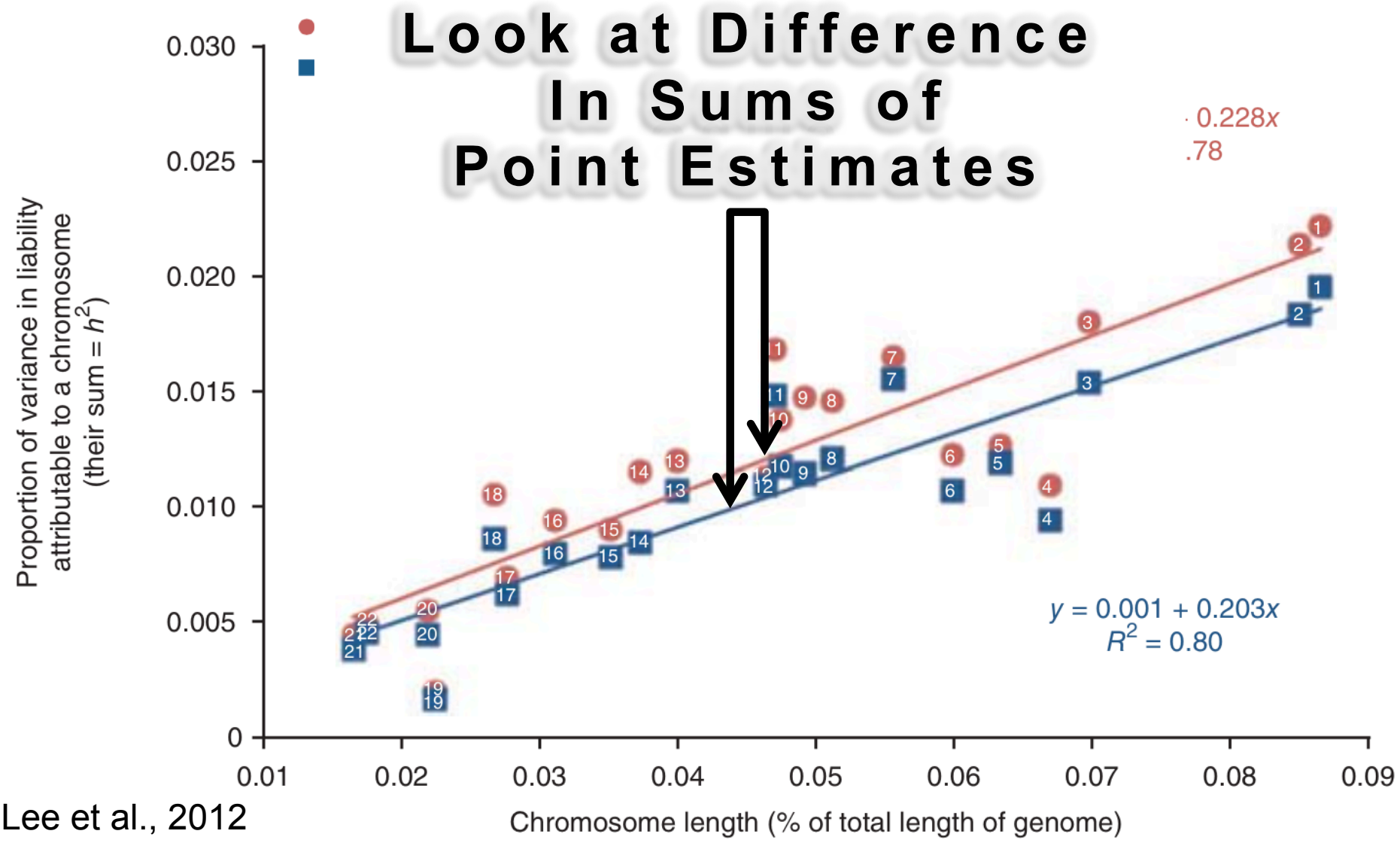
Lee et al., 2012

SNP variances for schizophrenia, by chromosome in ethnically homogeneous sample



Lee et al., 2012

SNP variances for schizophrenia, by chromosome in ethnically homogeneous sample



Lee et al., 2012

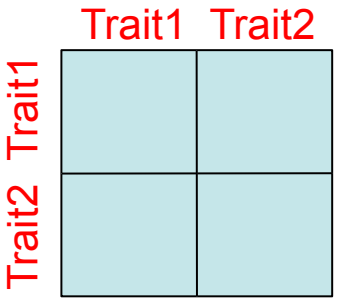
Outline

- Issues and extensions of GCTA (de Candia)
  - SNP variance estimates and heritability
  - Estimating multiple genetic variances
  - **Bivariate models**
  - Practical

Bivariate Models

- Can be used to examine genetic overlap between two separate measures thought to be related, such as:
  - Different phenotypes
  - Same phenotype across different datasets or genotyping procedures
  - Same phenotype across different populations or environments
- Importantly, model estimation does not require individuals to be assessed on both measures
  - Useful for examining rare traits

Bivariate Models



- For 2 measures, using all available pairs of individuals i and j, we use the following 3 different parts of the G & Y matrices:
  - 1 matrix for each of measures 1 and 2
  - 1 matrix for covariances between measures 1 and 2
- This model simultaneously estimates 3 genetic parameters: σ^2_{g1}, σ^2_{g2}, σ_{g12}
- Using these we can calculate a SNP correlation: $r_{SNP} = \sigma_{g12} / (\sigma_{g1}\sigma_{g2})$

Bivariate Models

SNP correlations (r_{SNP}) are only our best estimates of underlying genetic correlations (r_g):

- They will reflect the extent to which more common CVs are shared between traits
- r_{SNP} is not a direct estimate of the correlation of effect sizes of causal alleles. Systematic genotyping artifacts and population structure (distinct populations with MAF and background LD differences) will produce underestimates of r_g
- If we look across different traits, each of which is measured in separate datasets, then r_{SNP} between traits can be biased downward. It is important to make apples to apples comparisons (different traits, same dataset), and/or to use benchmarks (e.g., same trait, different datasets)

Outline

- Issues and extensions of GCTA (de Candia)
  - **SNP variance estimates and heritability**
  - Estimating multiple genetic variances
  - Bivariate models
  - Practical

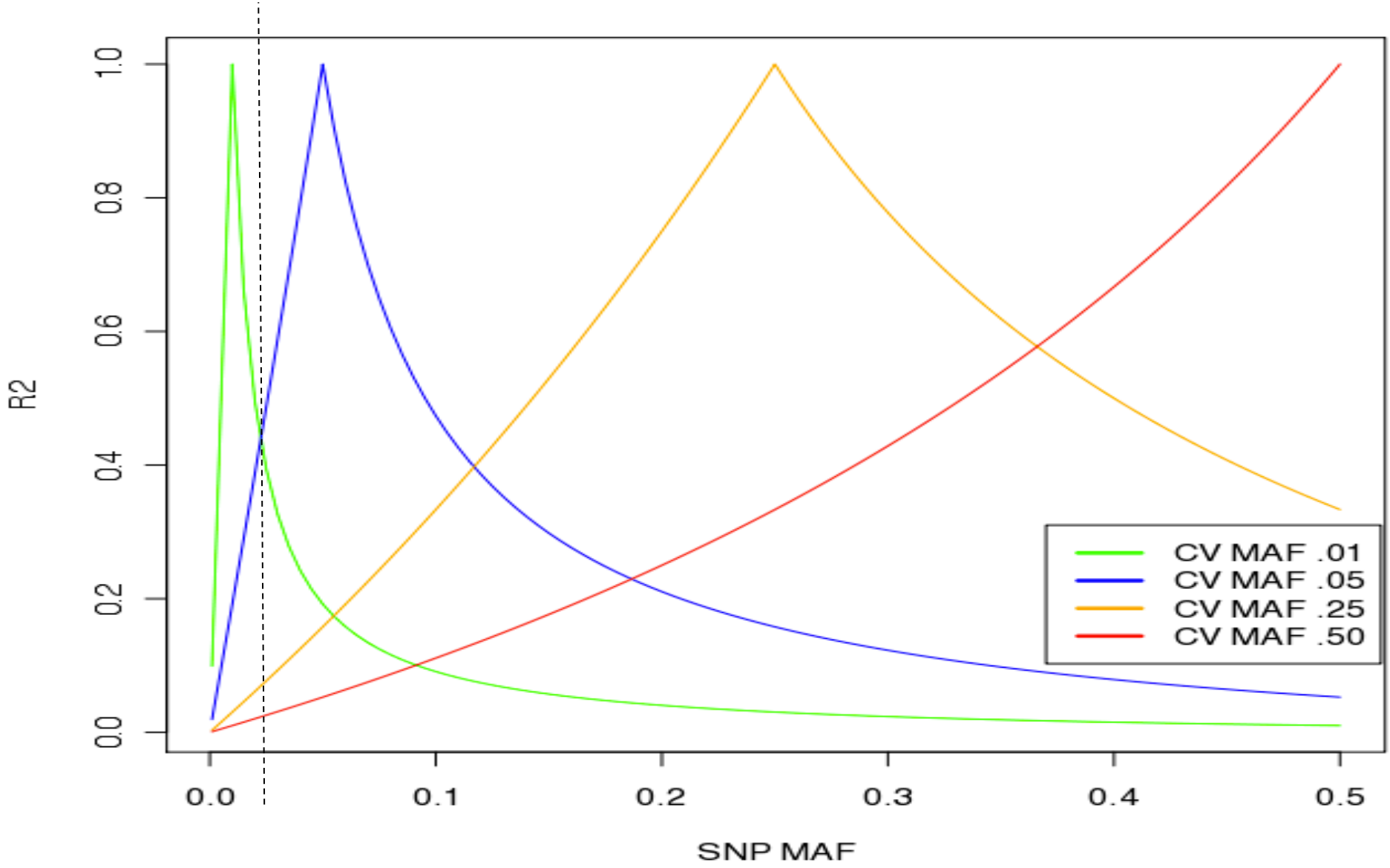
SNP-h^2 < Narrow-sense h^{2*}

1. Estimates rely on LD between SNPs and causal variants (CVs), and are therefore imperfect:

- Datasets with lower SNP density will capture less heritability
- Estimates are biased if background LD around CVs does not mirror that around SNPs. LDAK has been used to correct for this but may over-adjust (Speed et al., 2013)
- If allele frequency spectrum of SNPs is different than that of CVs, estimate will be too low. Rarer CVs are not well represented by SNP panels

*usually, maybe

Max R^2 as function of SNP MAF for several CV allele frequencies



SNP-h^2 < Narrow-sense h^{2*}

2. Noise in SNP calls (thus, π-hats) and phenotypes tends to bias SNP-h^2 downward when:

- **Random** plate effects inflate variance across π-hats compared with πs. This seems to be a problem with ascertained case-control samples as well (Golan et al., 2014)

$$G_{SNP} = cG_{CV}$$

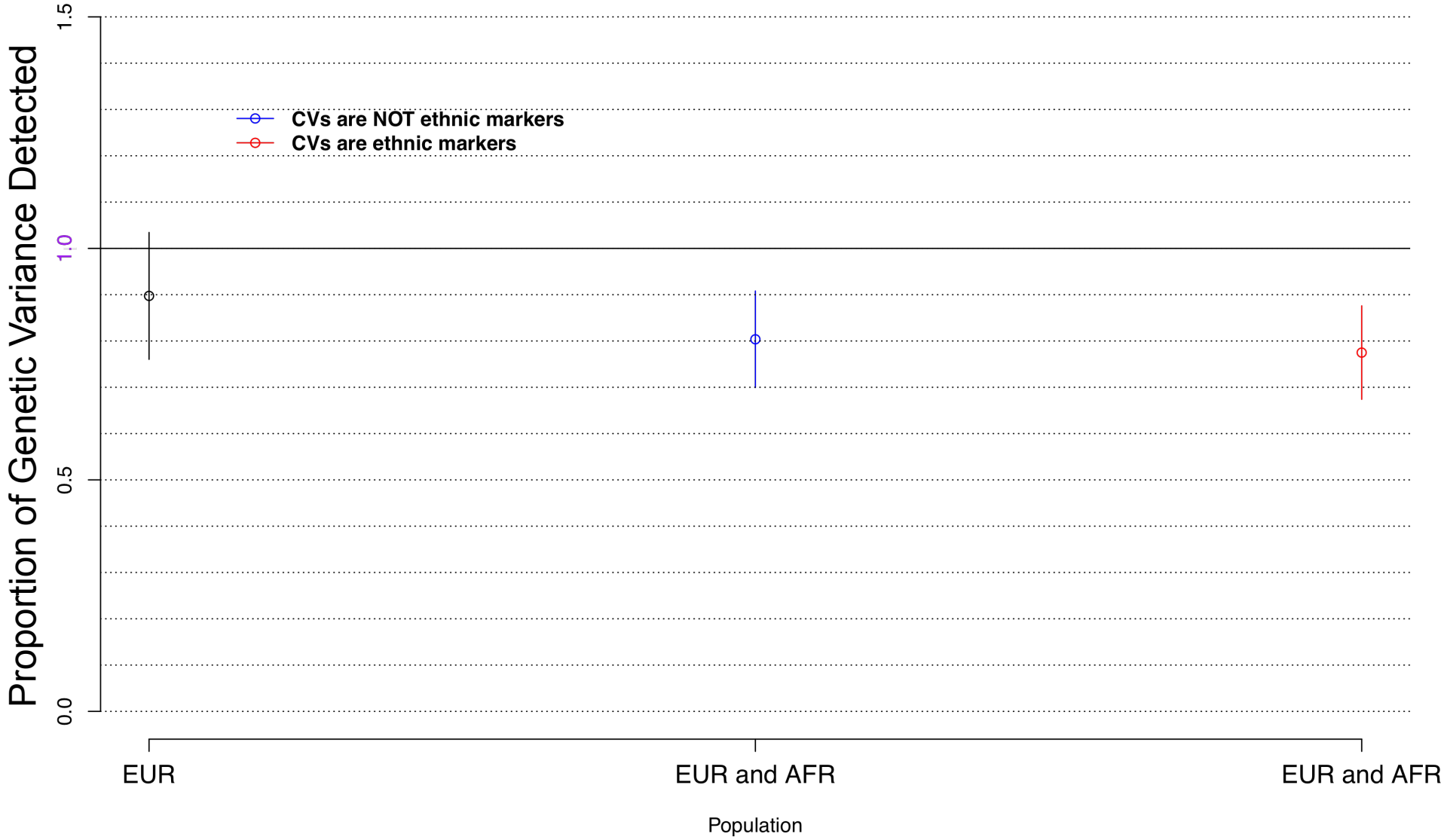
$$Var(Y) = cG_{CV1}\sigma^2_g + I\sigma^2_e$$

h^2 is underestimated when c is a scalar >1 on G_{CV}, inflating var(π-hat)

- Stratification is present, but environmental influences **DO NOT** differ by ethnicity
- Genetic heterogeneity exists, such that two "phenotypes" that are genetically quite different are regarded as the same thing

*usually, maybe

SNP variances for simulated trait, as a function of stratification



SNP-h^2 < Narrow-sense h^{2*}

But, not so shabby. Overall, things seem to work:

- SNPs pick up a substantial proportion of variance for a lot of tested traits, and simulations of phenotypes using real data confirm that methods are relatively robust to most assumptions
- Given published SNP h^2, amount of missing heritability is not surprising, especially if we take into account that estimates from family studies include rare and non-additive heritability as well as shared environmental effects

*usually, maybe

Outline

- Issues and extensions of GCTA (de Candia)
  - SNP variance estimates and heritability
  - Estimating multiple genetic variances
  - Bivariate models
  - **Practical**

Practical Objectives

- Let's estimate some univariate and bivariate models using simulated genotype and phenotype data.
- Suppose we have plink binary files for two studies - dat1 and dat2. (We also have merged these plink files into a single file - dat.)
- Suppose the first dataset (dat1) is of 2k females measured on height, the second (dat2) is of 2k females measured on BMI. Each individual is only present in one dataset.
- Our aim is to first estimate heritability separately for each of the two traits in univariate models, and then to jointly estimate two heritabilities and a genetic correlation in a bivariate model.
- To do this we will be a) calculating GRMs, and b) running REML to estimate model parameters.

GCTA Software

- Can be used for:
  - Data management (similar to PLINK)
  - Calculation of GRM from genome-wide SNPs (this can also be done in PLINK)
  - Model estimation by REML
  - PCA, simulations, etc.

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

Calculating GRM

- GRM:

```
gcta -bfile dat1 --make-grm-gz --
  thread-num 2 --out dat1.gcta
```

- Generates:
  - dat1.gcta.grm.gz
  - dat1.gcta.grm.id

Genetic Relationship Matrix (GRM)

example.grm.id	example.grm.gz
10 01 10 02 17 01 28 01 33 01 33 02 37 50 38 01 45 50 46 01	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

Estimating SNP h^2

- Estimate SNP h^2 for trait1 (and then do the same for trait 2):

  ```
  gcta –grm-gz dat1.gcta –pheno dat.pheno --mpheno XXX
  --reml –out dat1.results
  ```

- Jointly estimate SNP h2 for both measures as well as SNP-correlation:

  ```
  gcta –grm-gz dat.gcta –pheno dat.pheno –reml-bivar
  XXX XXX –out dat.results
  ```

  "XXX" will be 1 for phenotype data in 3rd column, 2 for phenotype data in 4th column. Exactly two columns must be specified for bivariate model

- Both traits are in phenotype file dat.pheno. Height is in column 3 and BMi is in column 2.
- Extension of results files is ".hsq"

Phenotype File

	example.pheno			
Dataset 1	fdfs1	fdfs1	0.99629	NA
	absd2	absd2	-0.47804	NA
	edgg3	edgg3	0.49192	NA
	rkls4	rkls4	0.00656	NA
Dataset 2	eedf1	eedf1	NA	0.00215
	aaaa2	aaaa2	NA	-0.99075
	bbbb3	bbbb3	NA	0.00004

Help and Questions

- Go to your desktop's "Home" directory, create a directory called "GCTA", and copy everything from Matt's subdirectory "Boulder2015" into it
- Use "GCTA_2015.Practical.R" to do all this
- GCTA website: http://www.complextraitgenomics.com/software/gcta/
- What is SNP h^2 for measure1? Measure2? And what are the SEs on those point estimates?
- How genetically correlated are they?
- What SNP r would you expect across datasets assessed on the same exact measure?

Results

- Pretty different SNP h^2 between traits:
  SNP h^2 estimated for height: ~.78

  (simulated h^2 was .8)

  SNP h^2 for BMI: ~.36

  (simulated h^2 was .4)
- Relatively high SNP r between traits: ~.64
- Are these simulated SNP h^2 estimates realistic?

From Wikipedia "Apples and Oranges"

- "At least two tongue-in-cheek scientific studies have been conducted on the subject, each of which concluded that apples can be compared with oranges fairly easily and on a low budget and the two fruits are quite similar. …[One] study … concluded: '…the comparing apples and oranges defense should no longer be considered valid. This is a somewhat startling revelation. It can be anticipated to have a dramatic effect on the strategies used in arguments and discussions in the future.'"
- "In many languages, oranges are, implicitly or explicitly, referred to as a type of apple".
- "Oranges, like apples, grow on trees."
- Additionally, one figure with the subtitle "Not all apples are alike", at the very least, possibly calls into question the use of the phrase "apples-with-apples comparison".