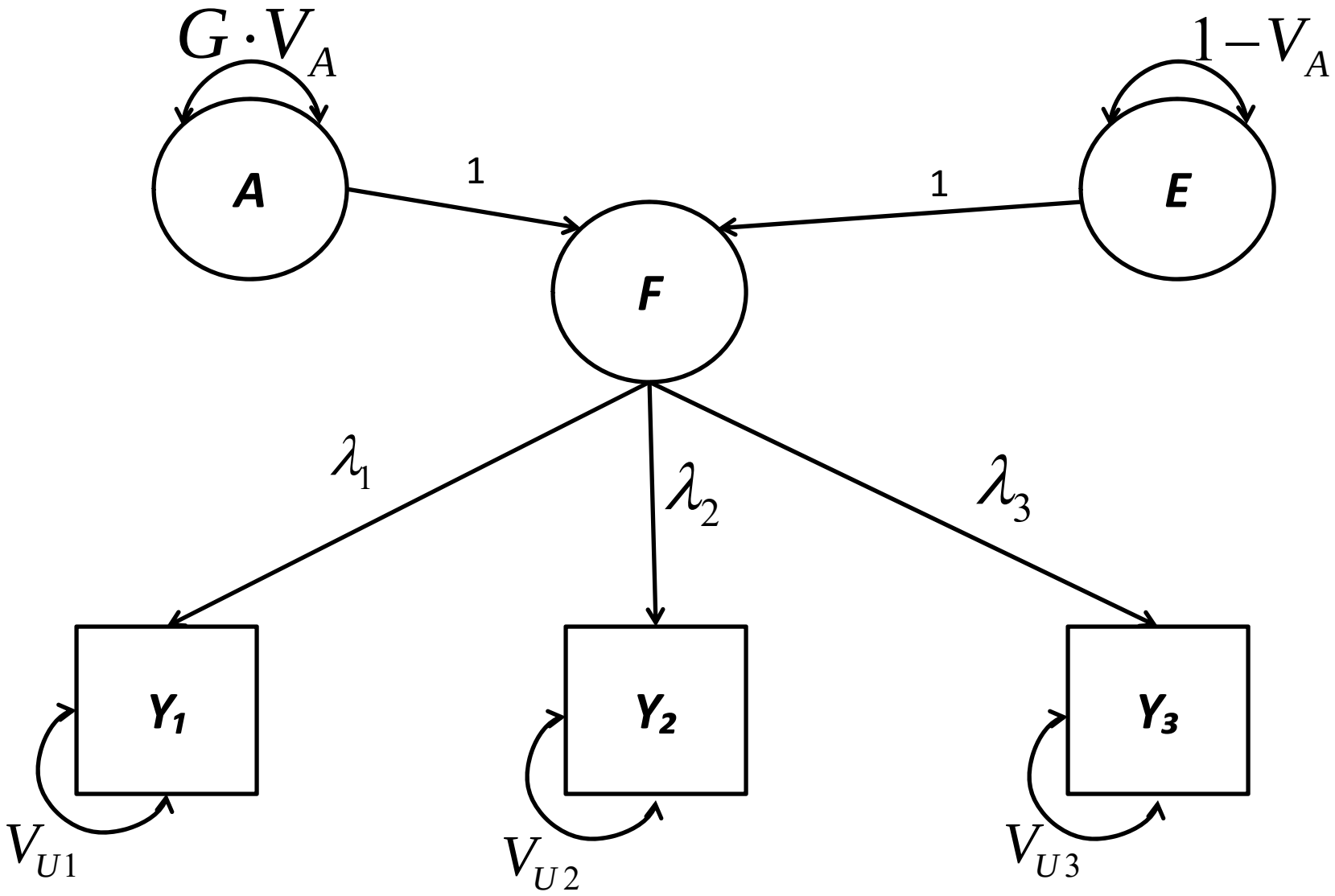
Combining SEM & GREML in *OpenMx*

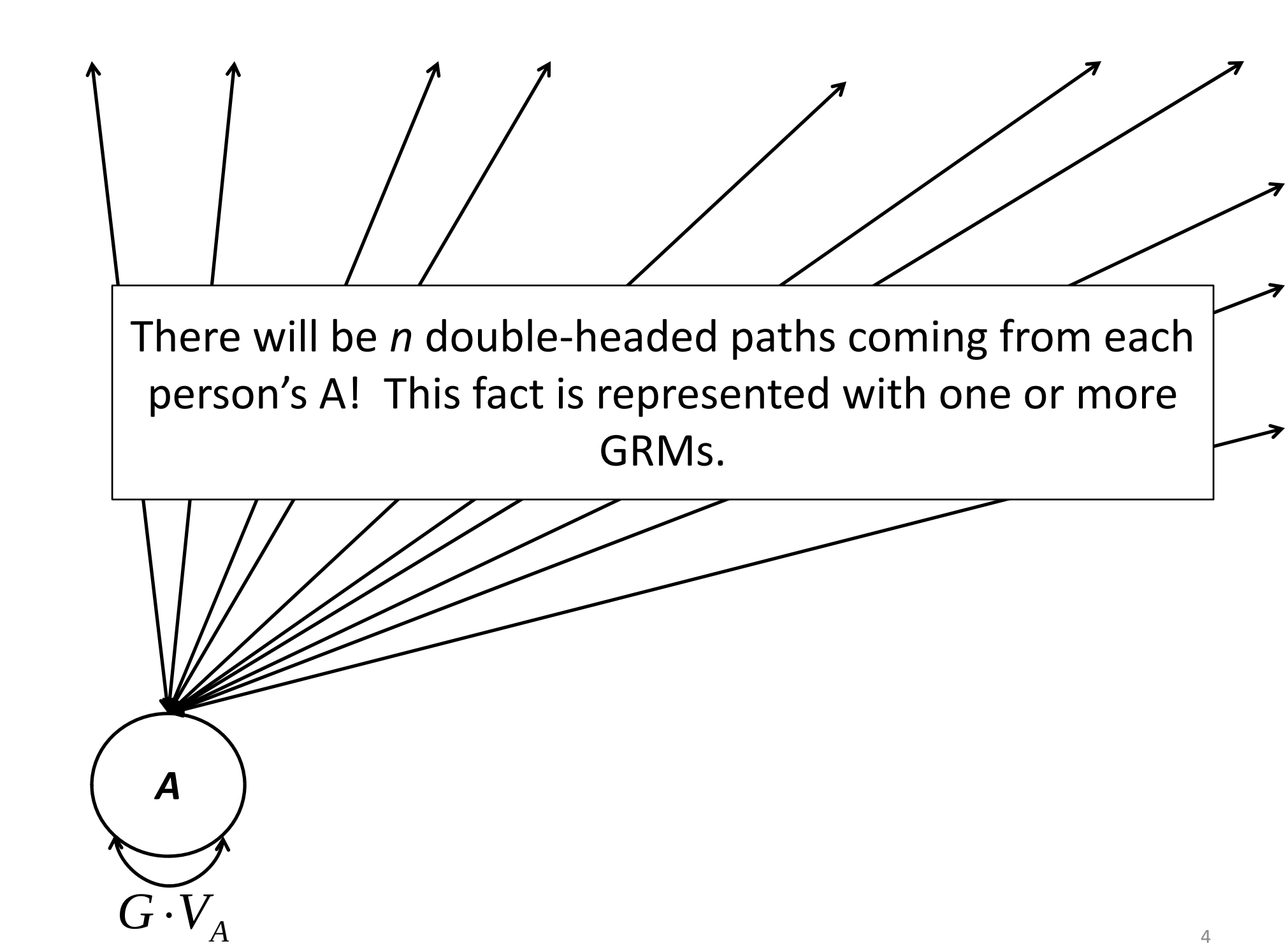
Rob Kirkpatrick

3/11/16

Overview

I. Introduction.
II. mxGREML Design.
III. mxGREML Implementation.
IV. Applications.
V. Miscellany.


$G \cdot V_A$
$1 - V_A$
A
E
F
1
1
λ_1
λ_2
λ_3
Y_1
Y_2
Y_3
V_{U1}
V_{U2}
V_{U3}

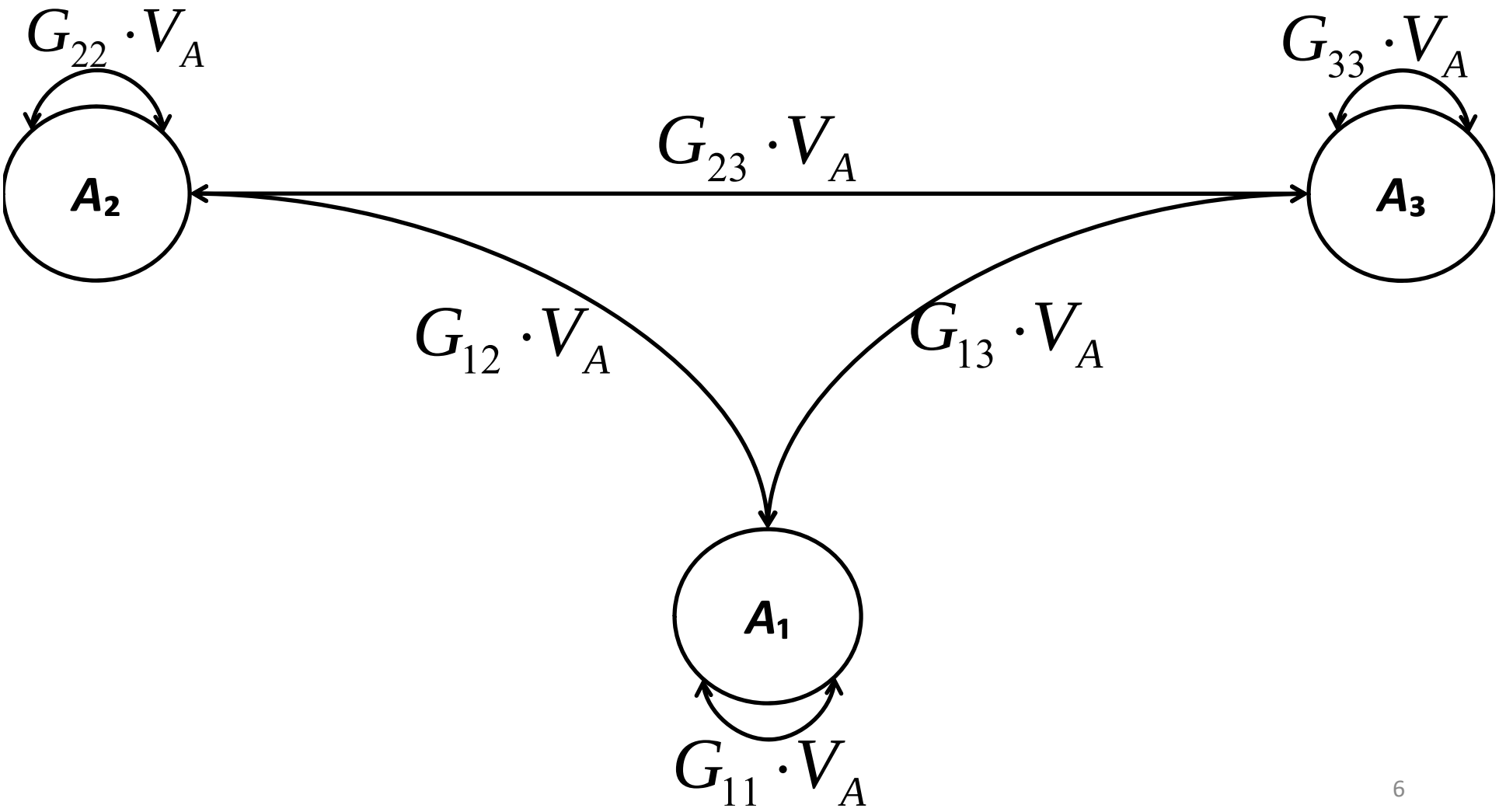

There will be *n* double-headed paths coming from each person's A! This fact is represented with one or more GRMs.
A
$G \cdot V_A$

Consider 3 random participants

Let $\mathbf{G}$ denote the GRM, where

$$\mathbf{G} = \begin{bmatrix} G_{11} & G_{12} & G_{13} \\ G_{21} & G_{22} & G_{23} \\ G_{31} & G_{32} & G_{33} \end{bmatrix}$$

Consider 3 random participants…



Biometric SEM with Unrelateds

- Freed from assumptions of twin/family/adoption study.
- Use observed, genotyped markers to assess relatedness & explain variance or risk.
- In particular—explain variance in *latent* variables…
- "Bin" markers (e.g., chromosome or biological pathway).

Biometric SEM with Unrelateds

- Caveats:
  - Does not *directly* contribute to discovering & identifying causal polymorphisms underlying trait.
  - REML seems not to be appropriate for case-control studies of disease.[1]

[1]Golan, D., et al. (2014). *PNAS, 111*(49), E5272–E5281. doi: 10.1073/pnas.1419064111

Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits

Noah Zaitlen[1]*, Peter Kraft[2,3,4], Nick Patterson[4], Bogdan Pasaniuc[5], Gaurav Bhatia[2,3,4], Samuela Pollack[2,3,4], Alkes L. Price[2,3,4]*

1 Department of Medicine, Lung Biology Center, University of California San Francisco, San Francisco, California, United States of America, **2** Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, United States of America, **3** Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, United States of America, **4** Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **5** Interdepartmental Program in Bioinformatics Pathology and Laboratory Medicine, University of California Los Angeles, Los Angeles, California, United States of America

Abstract

Important knowledge about the determinants of complex human phenotypes can be obtained from the estimation of heritability, the fraction of phenotypic variation in a population that is determined by genetic factors. Here, we make use of extensive phenotype data in Iceland, long-range phased genotypes, and a population-wide genealogical database to examine the heritability of 11 quantitative and 12 dichotomous phenotypes in a sample of 38,167 individuals. Most previous estimates of heritability are derived from family-based approaches such as twin studies, which may be biased upwards by epistatic interactions or shared environment. Our estimates of heritability, based on both closely and distantly related pairs of individuals, are significantly lower than those from previous studies. We examine phenotypic correlations across a range of relationships, from siblings to first cousins, and find that the excess phenotypic correlation in these related individuals is predominantly due to shared environment as opposed to dominance or epistasis. We also develop a new method to jointly estimate narrow-sense heritability and the heritability explained by genotyped SNPs. Unlike existing methods, this approach permits the use of information from both closely and distantly related pairs of individuals, thereby reducing the variance of estimates of heritability explained by genotyped SNPs while preventing upward bias. Our results show that common SNPs explain a larger proportion of the heritability than previously thought, with SNPs present on Illumina 300K genotyping arrays explaining more than half of the heritability for the 23 phenotypes examined in this study. Much of the remaining heritability is likely to be due to rare alleles that are not captured by standard genotyping arrays.

Citation: Zaitlen N, Kraft P, Patterson N, Pasaniuc B, Bhatia G, et al. (2013) Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. PLoS Genet 9(5): e1003520. doi:10.1371/journal.pgen.1003520

Editor: Peter M. Visscher, The University of Queensland, Australia

Received September 27, 2012; **Accepted** April 6, 2013; **Published** May 30, 2013

Funding: This work was funded by NIH grant R03HG005732 (NZ and ALP), NIH fellowship 5T32ES007142-27 (NZ), and the Rose Traveling Fellowship Program in Chronic Disease Epidemiology and Biostatistics (NZ). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: noah.zaitlen@ucsf.edu (NZ); aprice@hsph.harvard.edu (ALP)

II. mxGREML Design

Overview of GREML in *OpenMx*

- All participants' scores on all phenotypes get "stacked" into a single vector, **y**.
- "Definition variables" not allowed/needed.
- Ordinal phenotypes must be treated as continuous…
- User must specify model for **y**.
  - Mean of **y** conditioned on covariates, which are columns of matrix **X**.
  - var(**y**) is covariance matrix, **V**, which user must define.

GREML in *OpenMx* is *flexible*

- Key distinguishing characteristic from other analyses in *OpenMx* (e.g., FIML): phenotype vector **y** is a *single realization* of a random vector that cannot, in general, be partitioned into independent subvectors.

- (Applicable to analyses in disciplines other than genetics.)

GREML in *OpenMx*: assumptions

1. Conditional on covariates **X**, phenotype vector **y** is a single draw from a multivariate-normal distribution having (in general) dense covariance matrix, **V**.
2. The parameters of **V** are of primary interest.
3. Random effects are normally distributed.
4. GLS regression (using $\mathbf{V}^{-1}$) is adequate model for phenotypic mean.

III. mxGREML Implementation

Models in *OpenMx* 2.x

- "Objectives" from version 1 now split between:
  - Expectation
    - Model's specification.
    - For example, MxExpectationNormal specified in terms of covariance, means, and (sometimes) thresholds.
  - Fitfunction
    - Loss function to be minimized.
    - For instance, MxFitFunctionML uses -2 times multivariate-normal loglikelihood.

Overview of mxGREML Feature

0. Condensed matrix slots.

1. GREML expectation.

2. Data-handling helper function.

3. GREML fitfunction.

Large Matrices and Memory Efficiency

- Demo script…
- Main idea—when your *OpenMx* script involves large matrices that contain no free parameters:
  1. Place `options(mxCondenseMatrixSlots=TRUE)` near beginning of script.
  2. Always access slots of MxMatrix objects with $, and never with @.

GREML Expectation

- Compatible with GREML fitfunction and ML fitfunction.
- Requires raw continuous data.
- User tells it:
  - Which algebra/matrix is **V**.
  - Whether & with what arguments to call `mxGREMLDataHandler()` at runtime.
  - Whether & how to resize **V** at runtime due to missing data.

mxGREMLDataHandler()

- User provides:
  - Dataframe or matrix, in “wide” format.
  - Column names of phenotypes (for **y**).
  - Column names of covariates (for **X**).
- Creates **X** & **y**, and automatically trims NAs out of them.
- Can be called by user, or automatically at runtime.
- Can structure data for multiple phenotypes or for clustered/repeated measures.

GREML fitfunction

- Can OPTIONALLY accept analytic first partial derivatives of **V**:
  - User needs to know some calculus…
  - User provides names of matrices/algebras that equal first partial derivatives of **V** w/r/t free parameters.
  - *OpenMx* uses them to calculate derivatives of REML loglikelihood during optimization.
  - Calculating derivatives of REML logL is distributed over multiple processors.

GREML fitfunction

- With custom compute plan using Newton-Raphson:
  - Backend does average-information REML[1].
  - *OpenMx* gives analytic standard errors from average-information matrix at solution.
  - Note: N-R cannot handle MxConstraints.
- Both REML and ML -2log*L* returned from backend; use the ML -2log*L* for model comparison.

[1]Johnson, D. L., & Thompson, R. (1995). *Journal of Dairy Science, 78*, 449-456.

IV. Applications

Two Demo Scripts In My Folder

- For your reference; we won't go over them.
- One is a trivially simple example.
- The other carries out the factor model diagrammed earlier in this presentation.
- Warning! They use some advanced features:
  - Custom compute plan.
  - Newton-Raphson optimizer.
  - Analytic derivatives.

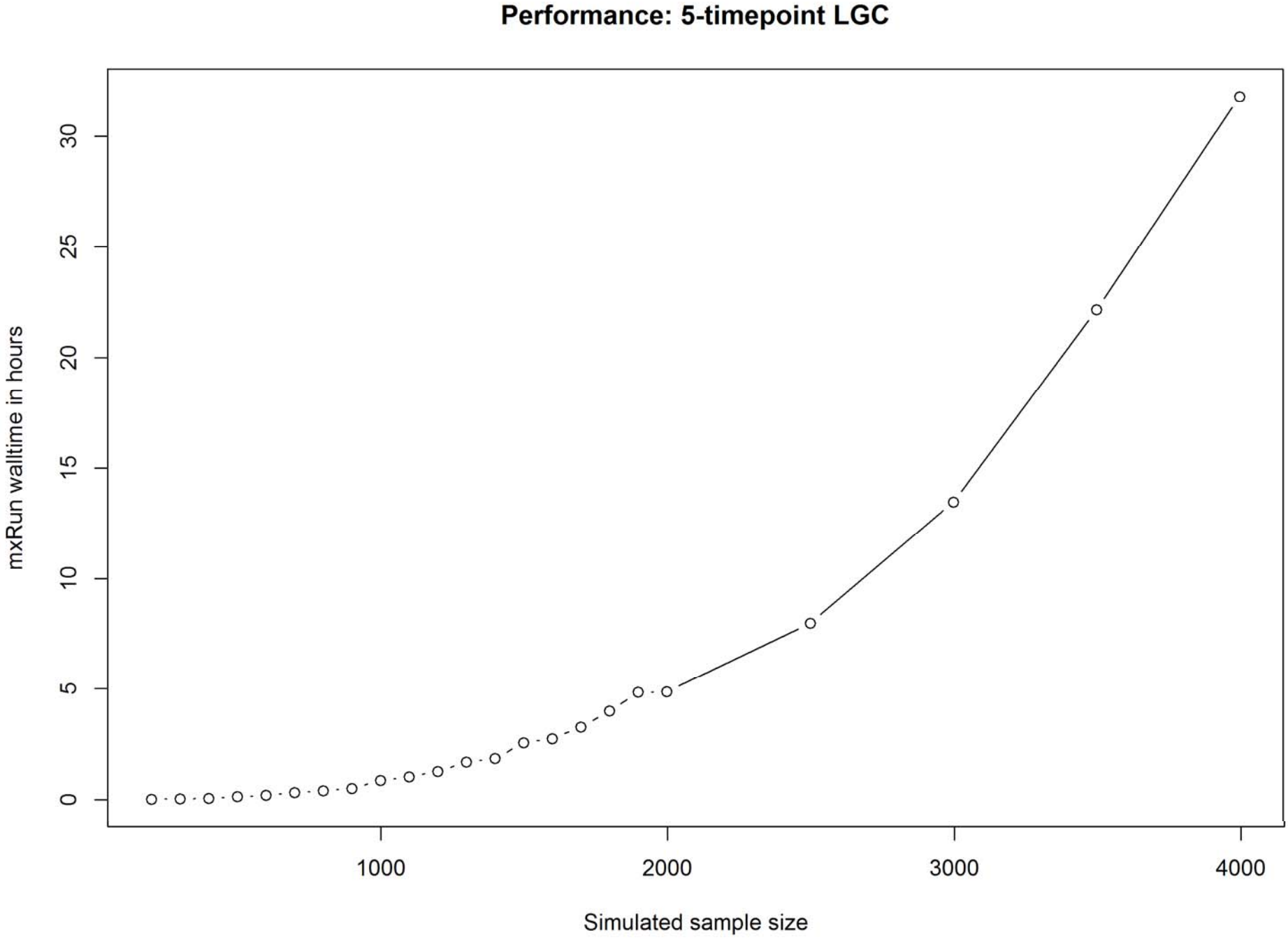
Validation

- Reproduced FIML and *GCTA* results from Eaves et al. (2014) simulation.
- Reproduced *GCTA* results for FSIQ reported in Kirkpatrick et al. (2014).

Eaves, L. J., et al. (2014). *Behavior Genetics, 44*, 445-455.
Kirkpatrick, R. M., et al. (2014). *PLoS ONE 9*(11): e112390.

Potential Applications

- Use GRM(s) & sample of classically unrelated participants to model:
  - Initiation & frequency of drug use.
  - Common vs. independent pathways.
  - Continuous biometric moderation (*a la* Purcell, 2002).
- Latent Growth-Curve analysis…


Performance: 5-timepoint LGC
mxRun walltime in hours
0
5
10
15
20
25
30
1000
2000
3000
4000
Simulated sample size

Potential Applications

- Use GRM(s) & sample of classically unrelated participants to model:
  - Initiation & frequency of drug use.
  - Common vs. independent pathways.
  - Continuous biometric moderation (*a la* Purcell, 2002).
- Latent Growth-Curve analysis.
- Multicategory ordinal phenotypes.

GREML with Binary Phenotypes

ARTICLE DOI 10.1016/j.ajhg.2011.02.002

Estimating Missing Heritability for Disease from Genome-wide Association Studies

Sang Hong Lee,[1] Naomi R. Wray,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1,*]

Genome-wide association studies are designed to discover SNPs that are associated with a complex trait. Employing strict significance thresholds when testing individual SNPs avoids false positives at the expense of increasing false negatives. Recently, we developed a method for quantitative traits that estimates the variation accounted for when fitting all SNPs simultaneously. Here we develop this method further for case-control studies. We use a linear mixed model for analysis of binary traits and transform the estimates to a liability scale by adjusting both for scale and for ascertainment of the case samples. We show by theory and simulation that the method is unbiased. We apply the method to data from the Wellcome Trust Case Control Consortium and show that a substantial proportion of variation in liability for Crohn disease, bipolar disorder, and type I diabetes is tagged by common SNPs.

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 28 no. 19 2012, pages 2540–2542*
doi:10.1093/bioinformatics/bts474

Genetics and population analysis Advance Access publication July 26, 2012

Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood

S.H. Lee[1,*], J. Yang[2], M.E. Goddard[3], P.M. Visscher[1,2] and N.R. Wray[1]

[1]The University of Queensland, Queensland Brain Institute, Brisbane, QLD 4072, [2]The University of Queensland Diamantina Institute, Princess Alexandra Hospital, Brisbane, QLD 4102 and [3]Department of Agriculture and Food Systems, University of Melbourne, VIC 3010, Melbourne, Australia

Associate Editor: Jeffrey Barrett

Multicategory Ordinal GREML

- Under suitable conditions[1], h^2 of ordinal phenotype is squared polyserial correlation between observed phenotype scores and latent genetic liability.
- Analyze ordinal scores as though continuous, and then adjust h^2 on observed scale to h^2 on latent scale via generalization of Dempster-Lerner-Robertson transformation[1].
- Must treat thresholds as though known, even though estimated from data.

[1]Dempster, E. R., & Lerner, I. M. (1950). *Genetics 35*(212), 212-236.

V. Miscellaney

Miscellaneous—stuff I didn't really cover

- **Be careful using GREML with any kind of ascertained sample.**
- Use of >1 GRM (or other such "relatedness matrix").
- GREML with family data.
- Technical aspects of computing GRMs.
- Computational shortcuts available for simple models (e.g., diagonalization).

Possible Improvements

- More-robust N-R optimizer.
- Enabling N-R to find confidence intervals.
- Interface to provide NPSOL & CSOLNP with analytic derivatives of constraints.
- User-specified sparse matrices?

Acknowledgements

- NIH grant DA026119
- Mike Neale (PI)
- Lindon Eaves
- Mike Hunter & Joshua Pritikin
- The rest of the *OpenMx* Development Team