

Genomic Heritability

Malachy Campbell

10/16/2018

Learning Objectives

- ▶ Understanding heritability
- ▶ Quantify relatedness
 - ▶ Identity by descent (IBD) and Identity by state (IBS)
 - ▶ Pedigree-based approach (brief): A matrix
 - ▶ Marker-based approach: G matrix
- ▶ Estimating genetic variances via the “Animal” model (brief)
- ▶ Cautionary tale of genomic heritability

What is heritability?

What is heritability?

- ▶ Proportion of total phenotypic variance in a population explained by total genetic (H^2) or additive genetic effects (h^2)

$$Y = G + E$$

- ▶ In the broad sense:

$$V_Y = V_A + V_D + V_I + V_E$$

$$H^2 = \frac{V_A + V_D + V_I}{V_A + V_D + V_I + V_E}$$

- ▶ In the narrow sense:

$$h^2 = \frac{V_A}{V_A + V_D + V_I + V_E}$$

Genetic values (a)

- ▶ Genetic values (a) are a linear function of allele content at **QTL**

$$y = a + e = \alpha' z_i + e$$

- ▶ Thus $a = \alpha' z_i$, where α' is a vector of additive effects at QTL and z' is a vector of genotypes at each QTL
- ▶ h^2 is the proportion of phenotypic variation that can be explained by a regression of phenotypes on QTL

Genomic values (g)

- ▶ Genomic values (g) are a linear function of allele content at **markers**
- ▶ Thus $g = \beta' x_i$, where β' is a vector of additive effects at markers and x' is a vector of genotypes at each marker
- ▶ Genomic heritability (h_g^2) is the proportion of phenotypic variation that can be explained by a regression of phenotypes on markers

Relationship between h^2 and h_g^2

- ▶ Inferred genetic values (g) are only an approximation of true genetic values (a)

$$\alpha' z_i = \beta' x_i + e$$

$$a = g + \bar{g}$$

- ▶ The true genetic value (a) is the genomic value (g) plus some genetic effects (\bar{g}) that cannot be captured by markers
 - ▶ \bar{g} is the 'missing heritability'
 - ▶ The true a is unknown, but is best approximated via **large pedigrees**

'missing heritability'

- ▶ Dependant on how well markers capture QTL (e.g. LD between markers and QTL)
- ▶ Genomic heritability is $h_g^2 = \rho_{zx}^2 h^2$, where h^2 is the 'true' heritability, ρ_{zx} is the correlation between QTL and marker (Gianola et al, 2015)
 - ▶ Thus, 'missing heritability' $h_g^2 = (1 - \rho_{zx}^2) h^2$

Estimating h^2 from genetic relatedness via pedigrees

- ▶ Rationale: Phenotypic similarity is due to QTL that are shared between related individuals
 - ▶ Utilize geneological data to calculate the expected genetic resemblance between relatives
- ▶ **Coefficient of ancestry (kinship coefficient):** What is the **probability** that two alleles at the same locus sampled at random from two individuals are derived from a common ancestor (e.g. are Identical By Descent, IBD)
 - ▶ Two randomly sampled alleles at a locus are IBD if they have the same ancestral origin.



Estimating genetic relatedness via pedigrees: expected relatedness

- ▶ The expected relatedness between individuals is twice the kinship coefficient
 - ▶ This expected relatedness matrix (A-matrix) represents the expected additive genetic relationships between individuals in a population

Marker-based approach

- ▶ **Goal:** Derive a relationship matrix, like **A**, that represents the **realized** genetic similarities between individuals using genetic markers
 - ▶ Genomic relationship matrix (**G**)
- 1. Determine the proportion of chromosome segments shared via the **identical by state (IBS)** matching of marker alleles.
- 2. Scale markers to more closely reflect **IBD** relationships

Tiny GRM example

- **M**: $n \times m$ matrix of markers ($aa = 0$, $Aa = 1$, $AA = 2$)

```
##      [,1] [,2] [,3] [,4]
## [1,]    0    1    0    2
## [2,]    2    1    1    1
## [3,]    2    0    0    0
```

- **W**: subtract 1 to rescale **M** to -1, 0, 1

```
##      [,1] [,2] [,3] [,4]
## [1,]   -1    0   -1    1
## [2,]    1    0    0    0
## [3,]    1   -1   -1   -1
```

Tiny GRM example: **identical by state (IBS)** matching of marker alleles

- ▶ Determine the homozygous identity in state matching by taking cross product (\mathbf{WW}')
 - ▶ Diagonal = # of homozygous loci in each individual
 - ▶ Off-diagonal = (# of loci with matching homozygous genotypes) - (# number non-matching homozygous loci)

##		[,1]	[,2]	[,3]
##	[1,]	3	-1	-1
##	[2,]	-1	1	1
##	[3,]	-1	1	4

- ▶ \mathbf{WW}' is an $n \times n$ **IBS** similarity matrix

Marker-based approach: Rescale to reflect **identical by state (IBD)**

- ▶ In the previous example common and rare alleles have the same weight on genomic relatedness between individuals
- ▶ If two individuals share a rare allele there should be a greater chance that they are closely related
 - ▶ Therefore for each marker i in \mathbf{M} , center the marker scores by the mean marker score ($2\hat{p}_i$) where p_i is the minor allele frequency (MAF) of marker i

Marker-based approach: Rescale to reflect **identical by state (IBD)**

- ▶ Suppose the four markers have a MAFs of 0.01, 0.15, 0.25, and 0.5. Subtracting $(2\hat{p}_i)$ from **M** will give us **Z**
- ▶ **ZZ'**

##		[,1]	[,2]	[,3]
##	[1,]	1.7404	0.2004	-0.9996
##	[2,]	0.2004	4.6604	3.4604
##	[3,]	-0.9996	3.4604	5.2604

- ▶ **WW'**

##		[,1]	[,2]	[,3]
##	[1,]	3	-1	-1
##	[2,]	-1	1	1
##	[3,]	-1	1	4

Marker-based approach: scale **G** so that it is analogous to **A**

- ▶ As the number of markers in **Z** increases, so does the elements of **ZZ'**
 - ▶ To be comparable we must make **ZZ'** independent of the number of markers
 - ▶ Dividing by the sum of variances at each locus $2 \sum_1^i p_i(1 - p_i)$ gives us **G** a **realized** relationship matrix that has the similar properties to **A**

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2 \sum_1^i p_i(1 - p_i)}$$

##		[,1]	[,2]	[,3]
##	[1,]	2.7312576	0.7096886	-1.495912
##	[2,]	0.7096886	3.5661854	1.527222
##	[3,]	-1.4959123	1.5272221	2.905201

Genomic relationship matrix **G**

- ▶ Elements of **G** are twice the realized kinship coefficients
- ▶ Diagonal elements indicate the degree of inbreeding for individuals ($E(\Theta_{ii}) = 1 + F_i$, where F_i is the inbreeding coefficient)
 - ▶ 1 for non-inbred individuals

##		[,1]	[,2]	[,3]
##	[1,]	2.7312576	0.7096886	-1.495912
##	[2,]	0.7096886	3.5661854	1.527222
##	[3,]	-1.4959123	1.5272221	2.905201

Estimating genetic parameters with **G** via the “Animal” model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

- ▶ \mathbf{y} ($n \times 1$) vector of observations
- ▶ $\boldsymbol{\beta}$ ($p \times 1$) vector of fixed effects (year, location, etc.)
- ▶ \mathbf{u} ($q \times 1$) vector of genetic values
 - ▶ q is all the individuals in **G**, $q > n$
 - ▶ $\mathbf{u} \sim N(0, \mathbf{G}\sigma_g^2)$, covariance in genetic values follows from genetic covariance between individuals
- ▶ \mathbf{e} residual effects
 - ▶ $\mathbf{e} \sim N(0, \mathbf{I}_n\sigma_e^2)$

Estimating genetic parameters with **G** via the “Animal” model

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \quad (1)$$

$$\lambda = \frac{\sigma_e^2}{\sigma_g^2} = \frac{1 - h^2}{h^2} \quad (2)$$

Estimating genetic parameters with **G** via the “Animal” model

- ▶ Use maximum likelihood (ML) or restricted maximum likelihood to solve model
 - ▶ Goal is to find a set of parameters that maximizes the likelihood of the data
 - ▶ ML: Estimates all parameters together; assumes no error in estimating fixed effects
 - ▶ REML: Allows for loss of degrees of freedom for estimating fixed effects
- ▶ Little difference between ML or REML if # of fixed effects is small
- ▶ No need to solve it by hand!
 - ▶ standalone: ASREML, GCTA, BLUPF90, Wombat
 - ▶ R: ‘nlme’, ‘lme4’, ‘ASREML-R’, ‘rrBLUP’, ‘bWGR’, ‘BGLR’

REML algorithms

- ▶ Some terminology. . .
 - ▶ Score function: derivative of the log-likelihood; expected value = 0
 - ▶ Information: variance of score function
- ▶ Newton-Raphson: Based on second-order Taylor series approximation of LL
 - ▶ $\theta \cong \tilde{\theta} - \left[H(\tilde{\Theta}) \right]^{-1} s(\tilde{\theta})$, where $s(\tilde{\theta})$ is the score function and $\left[H(\tilde{\Theta}) \right]^{-1}$ is the Hessian matrix
 - ▶ At each iteration take a current estimate for parameter and use it to update score function and Hessian matrix

REML algorithms

- ▶ Fisher's scoring: replaces the Hessian matrix in NR with the information matrix $\theta \cong \tilde{\theta} - \left[I(\tilde{\Theta}) \right]^{-1} s(\tilde{\theta})$, where $s(\tilde{\theta})$ is the score function and $\left[H(\tilde{\Theta}) \right]^{-1}$
- ▶ AI-REML (Average-information REML): Use an average of the observed and expected information matrices rather than expected information matrix in the Fisher's score algorithm; more computationally efficient
 - ▶ Gilmour et al (1995)

REML algorithms

- ▶ EM-REML (expectation-maximization REML): If values of random effects are known, variances can be directly estimated from them via $\sigma_u^2 = \frac{E[u_i' u_i]}{n_i}$ and $\sigma_e^2 = \frac{E[e_i' e_i]}{n}$, where n is the number of elements for u or e
 - ▶ Iterative approach where each iteration consists of two steps (1) calculate conditional expected values for parameters of distribution, (2) reestimation of parameters by maximization of the expected log-likelihood of the data; repeat until convergence is achieved
 - ▶ May take many iterations to converge

Alternative methods (Bayesian Gibbs sampling approach)

- ▶ Gibbs sampling generates random drawings from marginal posterior distributions through iterative sampling from the conditional posterior distributions
 - ▶ From the animal model ($\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$) we have a vector of parameters ($\boldsymbol{\beta}$, \mathbf{u} , σ_u^2 , and σ_e^2) and a conditional distribution that generates the data ($\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma_e^2 \sim N(\mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{R}\sigma_e^2)$)
- Process: **1.** Choose priors ($\boldsymbol{\beta} \rightarrow \text{flat } P(\boldsymbol{\beta}) \sim \text{constant}$; $u|G, \sigma_u^2 \sim N(0, \mathbf{G}\sigma_u^2)$; scaled inverted chi-square distributions for variance components **2.** For each parameter calculate the full conditional posterior distribution assuming all other parameters are known, sample from this distribution **3.** Rerun (2) using updated values, repeat until convergence

A cautionary tale of genomic heritability

- ▶ Is genomic heritability a good approximation of h^2 ?
- ▶ $\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2 \sum_1^I p_i(1-p_i)}$
 - ▶ What information is captured with \mathbf{Z} ?

A cautionary tale of genomic heritability

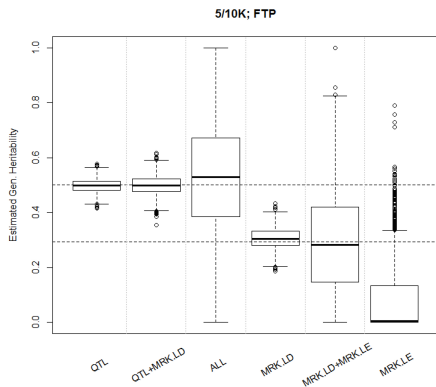
- ▶ Is genomic heritability a good approximation of h^2 ?

- ▶ $\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2 \sum_1^i p_i(1-p_i)}$

- ▶ What information is captured with \mathbf{Z} ?
 - ▶ QTLs are typed by markers
 - ▶ Markers in LD with QTL
 - ▶ Markers in LE with QTL

A cautionary tale of genomic heritability: de los Campos, Sorrensen, Gianola (2015)

- ▶ Six simulated scenarios for construction of **G**: 1. Only QTL; 2. QTL and markers in LD with QTL; 3. All loci (causal and non-causal); 4. Only markers in LD with QTL; 5. Only markers in LE with QTL (non-informative for σ_p^2); 6. All markers (LD and LE with QTL)
 - ▶ h^2 fixed at 0.5



A cautionary tale of genomic heritability: de los Campos, Sorrensen, Gianola (2015)

- ▶ Comparing h^2 and h_g^2 in related (FHS) and unrelated populations (GEN)

Scenario	Genetic Information Used to Compute Relationships	$\hat{h}_G^{(1)}$		R^2 (TST) ⁽²⁾	
		FHS	GEN	FHS	GEN
RAND	Causal Loci	0.775	0.773	0.545	0.517
		(0.009)	(0.010)	(0.040)	(0.031)
	Markers	0.774	0.737	0.263	0.071
		(0.018)	(0.040)	(0.048)	(0.023)
	Pedigree	0.764	—	0.223	—
		(0.020)	—	(0.047)	—
Low-MAF	Causal Loci	0.777	0.775	0.551	0.536
		(0.007)	(0.008)	(0.026)	(0.026)
	Markers	0.748	0.573	0.240	0.049
		(0.018)	(0.058)	(0.029)	(0.019)
	Pedigree	0.755	—	0.224	—
		(0.023)	—	(0.033)	—

FHS = Framingham Heart Study; GEN = GENEVA; RAND: in this scenario causal and marker loci were drawn from the same distribution; Low-MAF: in this scenario marker loci were drawn at random and causal loci were drawn over-sampling loci with low minor allele frequency; TST = Testing data set.

⁽¹⁾: average (over 30 MC replicates) estimated posterior mean of the ratio of genomic variance over the sum of genomic and residual variance;

⁽²⁾: average prediction R^2 (phenotypes) over 30 training ($N=5,300$)/testing ($N=500$) partitions.

doi:10.1371/journal.pgen.1003608.t002

- ▶ Proportion of allele sharing at markers and QTL are greater for related individuals compared to unrelated individuals
 - ▶ Cosegregation of markers and QTL are due to recent relationships

Questions?

Questions?

How do we measure heritability?

- ▶ Prior to genetic markers what did we need?

How do we measure heritability?

- ▶ What do we need?
 - ▶ Need a population with related individuals
 - ▶ Need a means to assess genetic relatedness between individuals
 - ▶ Need methods to partition variance into genetic and non-genetic effects