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^2_g) is the proportion of phenotypic variation that can be explained by a regression of phenotypes on markers

Relationship between h^2 and h^2_g

- Genomic values (g) are only an approximation of 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'

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

Estimating genetic relatedness via pedigrees

- **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 realtionship 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**: *nxm* 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 (**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
```

- **WW′** is an *nxn* **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 $\mathbf{ZZ}'$
  - To be comparable we must make $\mathbf{ZZ}'$ independant of the number of markers
  - Dividing by the sum of variances at each locus $2\sum_1^i p_i(1-pi)$ gives us **G** a **realized** relationship matrix that has the similar properties to **A**

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

```
##              [,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

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

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

Estimating genetic parameters with $\mathbf{G}$ via the "Animal" model

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

$$\lambda = \frac{\sigma_e^2}{\sigma_g^2} = \frac{1 - h^2}{h^2} \tag{2}$$

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
  - R: 'nlme', 'lme4', 'ASREML-R'

A cautionary tale of genomic heritability

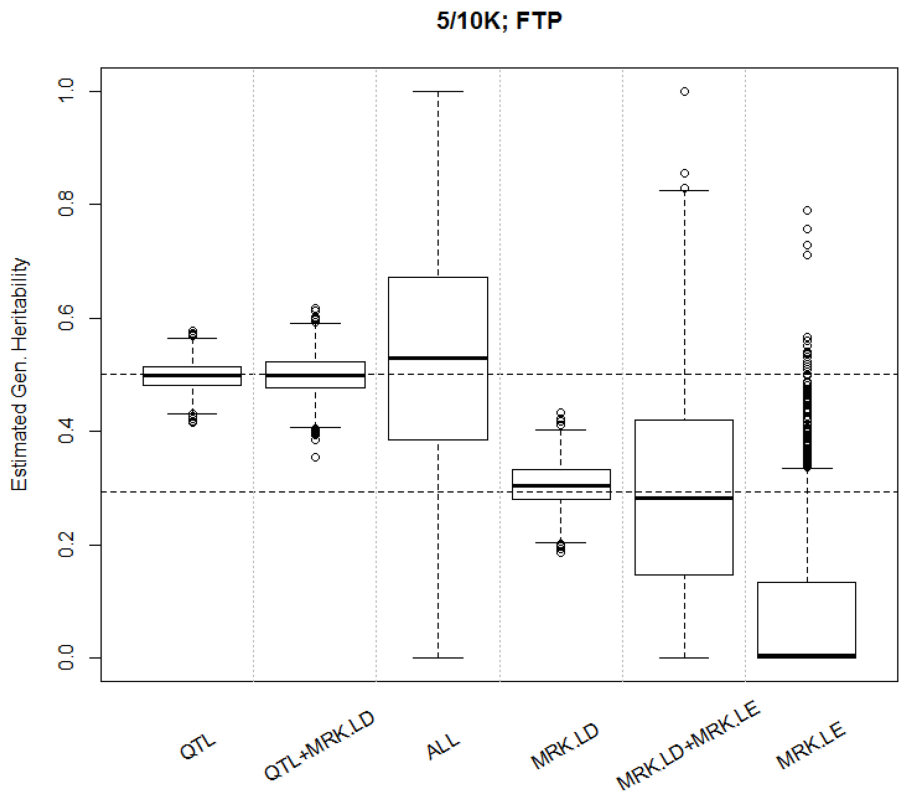
- Is genomic heritability a good approximation of h^2?
- $\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2\sum_1^i p_i(1-pi)}$
  - What information is captured with $\mathbf{Z}$?

A cautionary tale of genomic heritability

- Is genomic heritability a good approximation of h^2?
- $\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2\sum_1^i p_i(1-pi)}$
  - 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^2_p); 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^2_g in related (FHS) and unrelated populations (GEN)

Scenario	Genetic Information Used to Compute Relationships	h^2_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

Figure 2: dIC2

- 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