Pedigrees or markers: which are better in estimating relatedness and inbreeding coefficient?

Jinliang Wang

Institute of Zoology, Zoological Society of London
Outline

• Introduction
  ❖ Relatedness ($r$) and inbreeding coefficient ($F$)
  ❖ Estimators: Pedigrees ($r_p, F_p$) & markers ($r_M, F_M$)
  ❖ Which are better for genomic relatedness ($r_G$) and inbreeding coefficient ($F_G$)?
  ❖ How much better?
  ❖ Under which conditions?

• Methods
  ❖ Simulations of $F_G, F_p, F_M, r_G, r_p, r_M$, inbreeding depression (ID) of viability
  ❖ Power of $F_P, F_M$ in estimating ID measured by proportion of replicates in which ID is significant

• Results & discussions

• Conclusions
Introduction

- **Inbreeding coefficient \((F)\)**
  - Correlation coefficient (Wright, 1921): range \([-1, 1]\)
  - Probability (PIBD, Malecot 1948): range \([0, 1]\)
  - Both have a (arbitrary) reference \((F=0, r=0)\)
  - Correlation reference: ancestral \((\bar{F}>0)\), current \((\bar{F}=0)\), descendant \((\bar{F}<0)\)
  - IBD reference: ancestral \((\bar{F}>0)\).
  - Equivalent in most cases, but different in others

- **Genomic inbreeding coefficient \((F_G)\)**
  - Realized \(F\) at a single locus: 0 or 1
  - \(F_G\): Average \(F\) across loci in an individual genome
  - \(F_G\): Expected \(F\) of a locus taken at random from an individual genome
  - Variable among individuals with the same pedigree
  - Variation depends on genome size and pedigree (Hill & Weir 2011)
Introduction – Cont 1

• Pedigrees
  ✤ Estimate $F$ by path analysis (Wright 1921)
  ✤ Expected value across loci and across individuals
  ✤ $F_P \geq 0$ (reference: ancestral population, founders)
  ✤ Problems: reference, complete sampling, $F_P \neq F_G$
• Genome $r \neq$ Pedigree $r$

- For $E[r_g] = r_p = 0.5$, $Var[r_g]$ is

$$Var(r_g) = \frac{1}{16L} - \frac{22}{64L^2}$$

- For human, $L \approx 36$ Morgan, and $SD(r_g) = 0.04$ (Visscher 2006)

- Empirical distribution of genomic $r_g$, with mean=0.498, SD=0.04

- SD increases with a decreasing genome

---

**Fig. 1** Empirical distribution of genome-wide coefficients of additive relationships from 4,401 pairs of fullsibs (Visscher et al. 2006)
Introduction – Cont 3

• Markers
  - Estimate $F$ from marker genotypes
  - Average value across the particular marker loci
  - $F_M$ can be positive & negative (Ref: current gen.)
  - Average $F_M$ across individuals close to 0
  - Problems: high sampling variance among loci
Introduction – Cont 4

• Pedigrees or markers?
  
  ❖ Which best estimate $F_G$ and $r_G$?
  
  ❖ In pre-genomic era, pedigrees are better, and markers should better be used to validate, amend and construct pedigrees rather than to replace them (e.g. Pemberton 2008)
  
  ❖ Can genomic markers become better? How much & when?
Methods

• Simulations

- $F_G$ and $r_G$ from many ($L_{IBD}=10^4$) IBD loci, each with $2N$ alleles in $N$ founders
- $F_M$ and $r_M$ estimated from $L_{SNP}$ biallelic loci (SNPs) scattered in the genome
- Initial SNP allele frequency drawn from a uniform distribution
- Inbreeding depression of viability determined by many QTL, each with $s=0.01$ and $h=0.2$, lethal equivalent=2.0
- Founders non-inbred and unrelated, all loci in HD and linkage equilibrium
- $F_P$ and $r_P$ calculated from simulated pedigrees
- Population size $N$ (half males, half females) variable
- Genome size (map length) $L_G$ variable
- Number of generations $G$ variable
- 1000 replicates for a parameter combination
Methods – *Cont 1*

- **Accuracy**
  - Absolute values of $F_G$, $F_P$ and $F_M$ incomparable
  - Relative values are relevant in most applications (e.g. ID, $h^2$)
  - Correlation between $F_P$ (or $F_M$) and $F_G$ to measure accuracy
  - Proportion of replicates in which Inbreeding depression is detected (significant) by $F_P$ (or $F_M$) to measure power
Results & Discussion – Genome size

Other parameters: $N=64$, $G=8$, $L_{SNP}=10^4$, and $L_{IBD}=10^4$
Results & Discussion – Number of SNPs

Other parameters: \( N=64, \ G=8, \ L=32M, \) and \( L_{IBD}=10^4 \)
Other parameters: $L_{\text{SNP}}=10^4$, $G=8$, $L=32\text{M}$, and $L_{\text{IBD}}=10^4$
Results & Discussion – Generations

Correlation coefficient

- $r_P$, IBD=Pedigree Ref (variable)
- $r_M$, IBD=Pedigree Ref (variable)
- $F_P$, IBD=Pedigree Ref (variable)
- $F_M$, IBD=Pedigree Ref (variable)

Other parameters: $L_{SNP} = 10^4$, $N = 64$, $L = 32M$, and $L_{IBD} = 10^4$
Results & Discussion – Generations

Correlation coefficient

Generations (G)

Other parameters: $L_{SNP}=10^4$, $N=64$, $L=32M$, and $L_{IBD}=10^4$
Results & Discussion – Detecting ID

Parameters: $L_{SNP}=10^4$ (variable in B), $N=64$ (variable in D), $L=32M$ (variable in A), $G=8$ (variable in C), and $L_{IBD}=10^4$, $h=0.2$, $s=0.01$, and $B=2.0$. 
Conclusions

• Genomic markers can yield better $F$ and $r$ estimates than pedigrees
• Many SNPs ($\sim 10^4$) required
• Small genome, small $N$, and any factors (e.g. non-random mating) resulting in LD will increase the power of SNPs
• Pedigree replaceable by dense SNPs? Too early!
Conclusions – Cont 1

Detailed IBD coefficients between 4 genes of 2 individuals

Complicated IBD coefficients between multiple individuals
203 IBD states, 66 condensed IBD states for 3 individuals!

Complicated IBD coefficients for multiple loci between 2 individuals

Power of markers decline (close to) exponentially with a decreasing relatedness (Donnelly 1983)
Conclusions – Cont 2

Probability of no detectable relationship

Donnelly 1983, TPB, 23: 34-63
Thank you for your attention!

Questions?