Estimating relationship from marker genotypes

Mike Goddard
mike.goddard@ecodev.vic.gov.au

Relationships

We use relationship data

- to estimate genetic variance
- to estimate demographic history
- …

Relationships

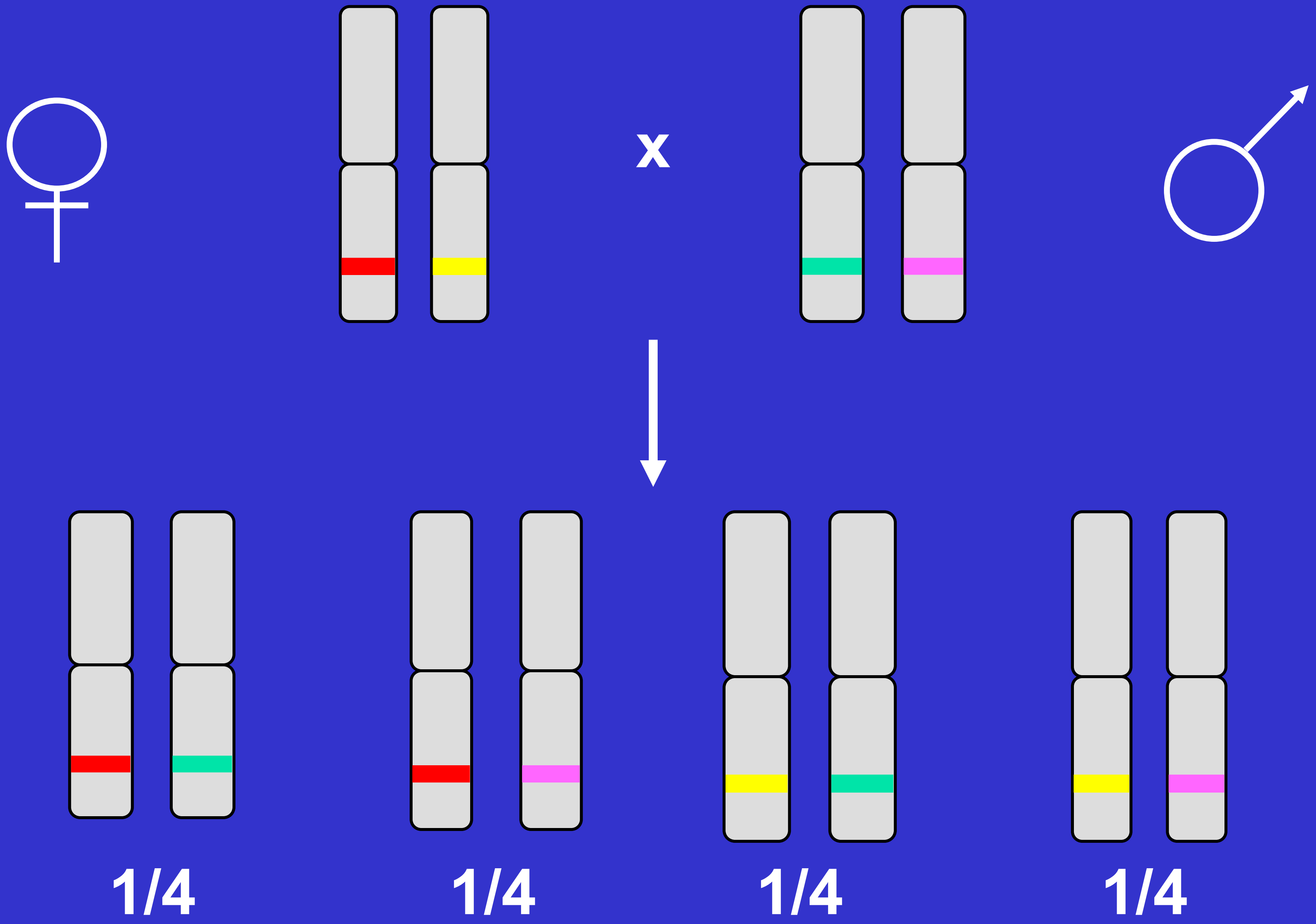
Additive genetic relationship G(i, j)

= proportion of the genome in i and j that is IBD

Pedigree relationship A(i,j) = Prob (IBD)

= E(G(i,j))

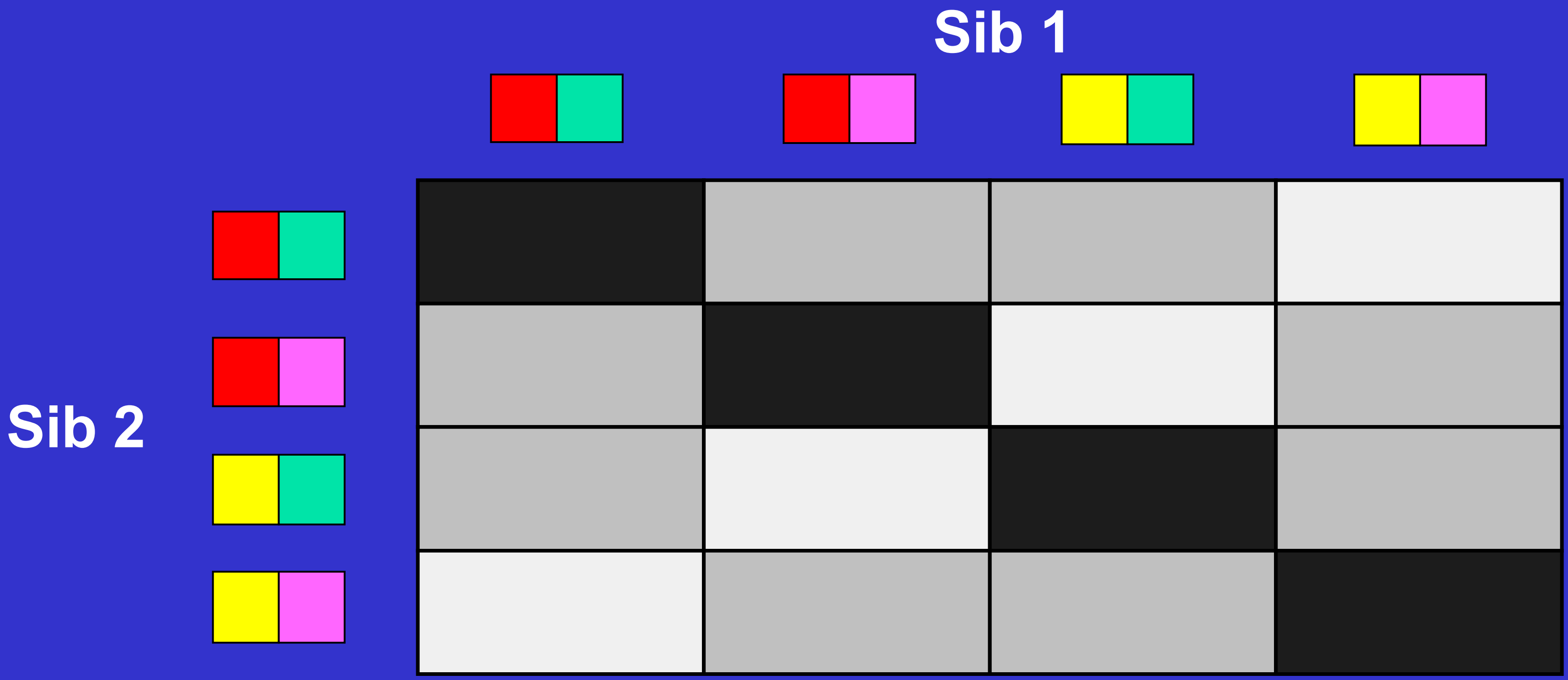
Actual relationship deviates randomly from this expectation



♀ x ♂

1/4 1/4 1/4 1/4

IDENTITY BY DESCENT



■ 4/16 = 1/4 sibs share BOTH parental alleles G = 1

▒ 8/16 = 1/2 sibs share ONE parental allele G = ½

□ 4/16 = 1/4 sibs share NO parental alleles G = 0

Relationships

Summary of single locus case, full sibs

Pairs of sibs share

0 alleles	25% of the time
1 allele	50%
2 alleles	25%

E(G) = A = 0.5 but actual relationship G varies from 0 to 1

Estimate relationship from markers

G is a more accurate description of relationship than A

- G captures unknown pedigree information
- pedigree can be incorrect
- G captures deviations from A

Therefore, can use G in

- Random sample of population (“unrelated individuals”)
- Individuals with same pedigree

Estimate relationship from markers

1. Well defined (recent) base
2. No well defined base

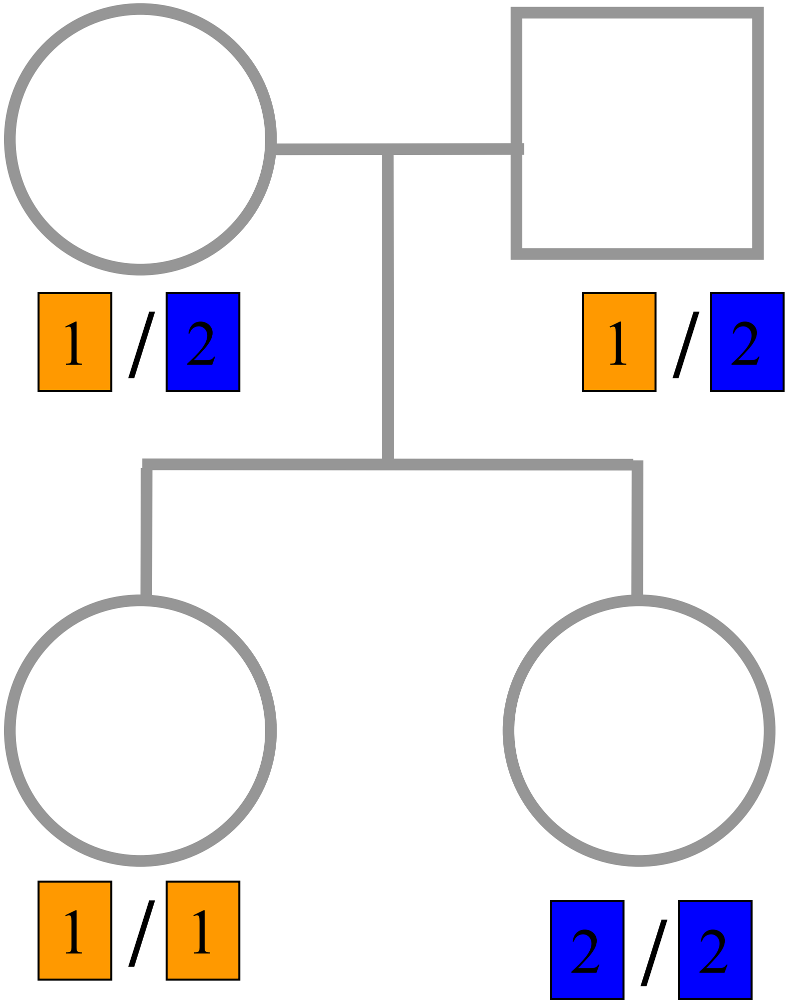
1. Well defined, recent base

Eg Data on families of full-sibs and parents of sibs are the base

Estimating relatedness with markers

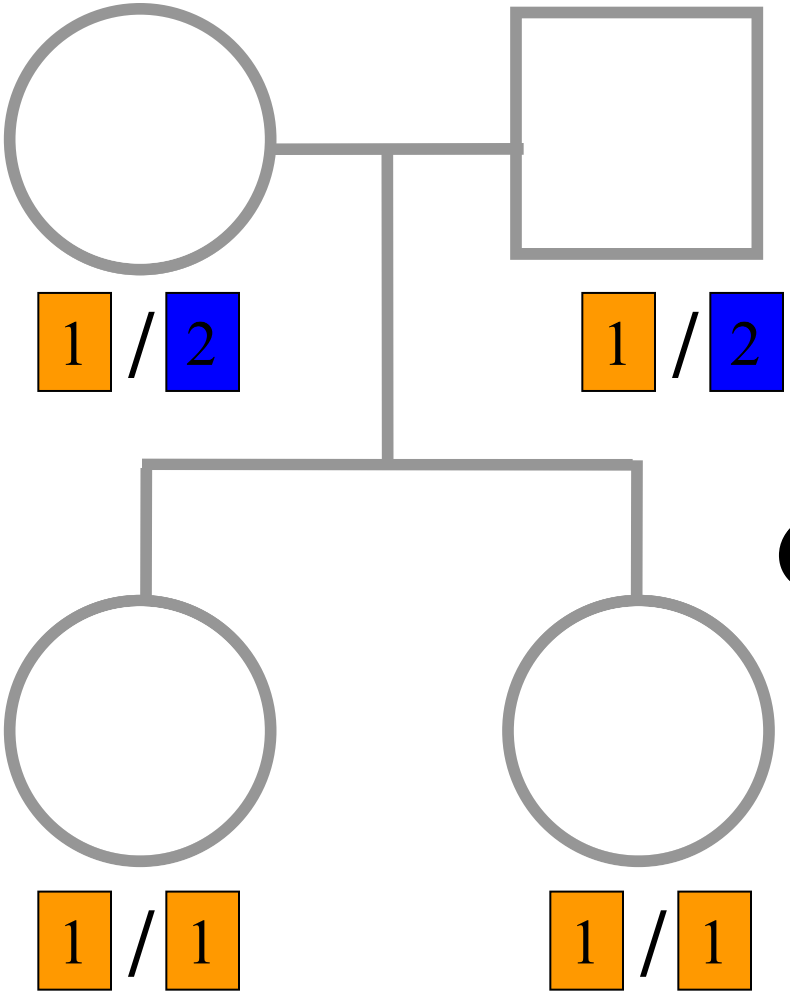
- Using:
  - Observed data (SNP genotypes)
  - Mendelian segregation rules (prior probability of sharing alleles IBD)
  - Marker allele frequencies in the population

IBD can be trivial…

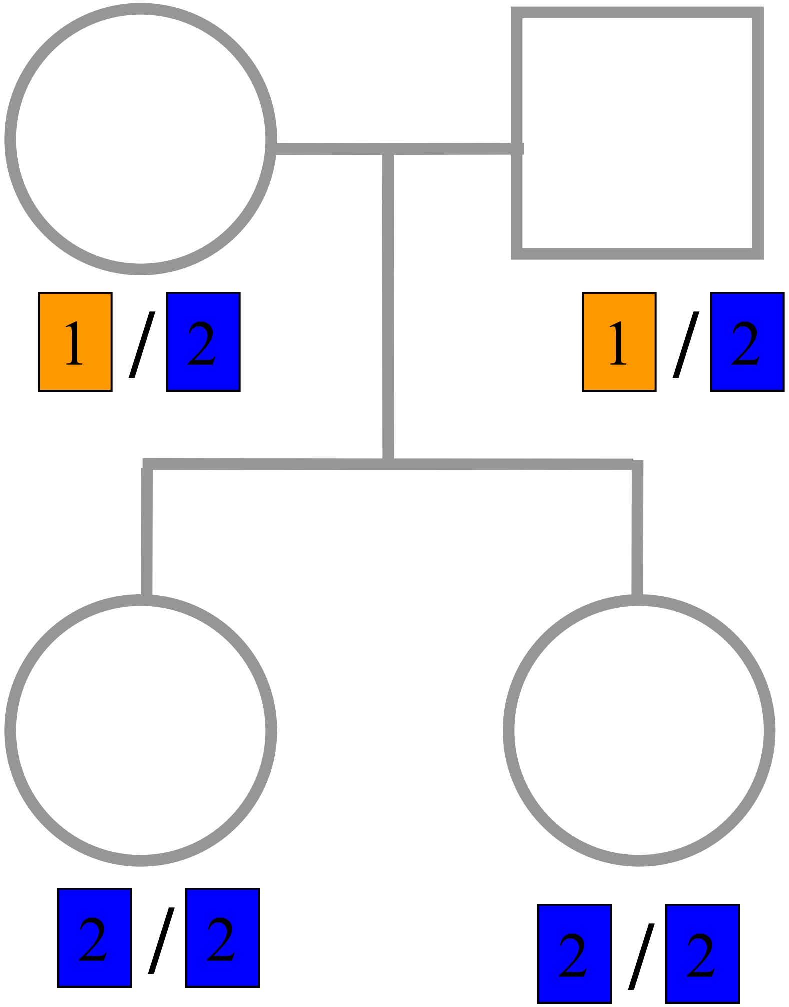


G=0

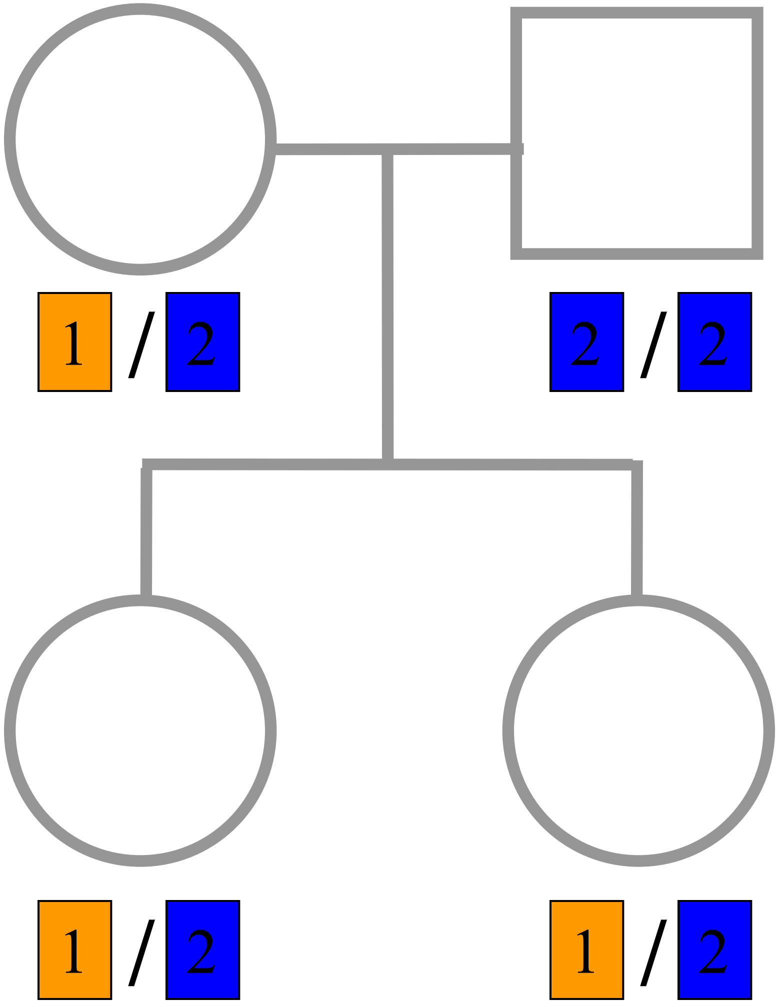
Two Other Simple Cases…



G=1



A little more complicated…



1 / 2

2 / 2

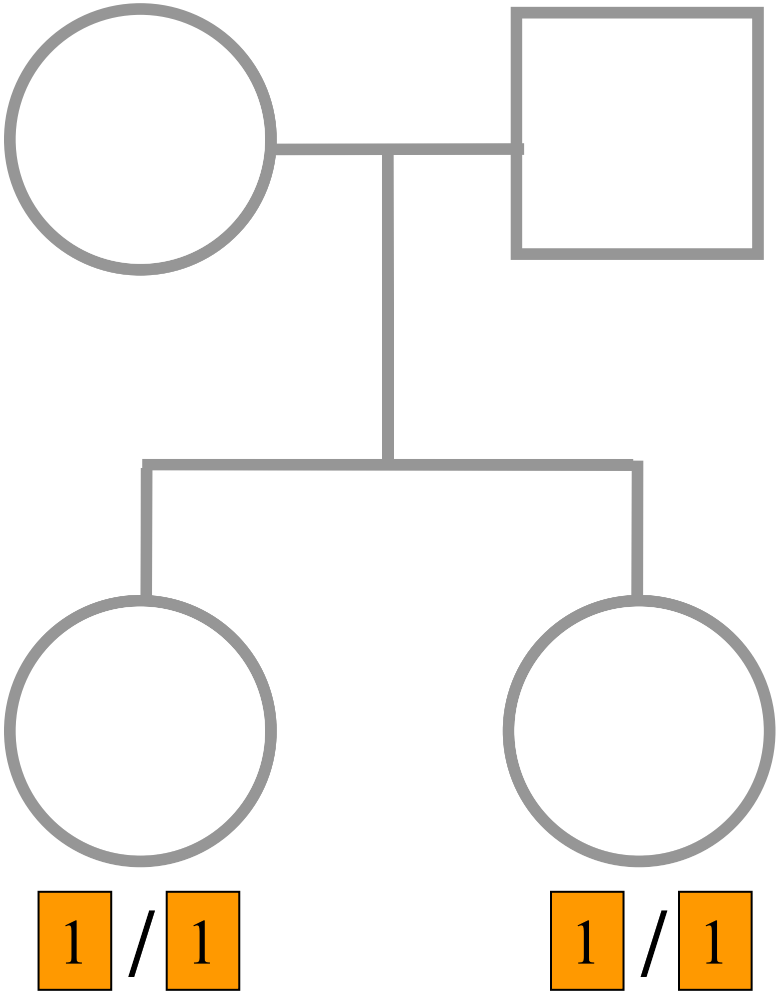
G= ½
(50% chance)

1 / 2

G=1
(50% chance)

1 / 2

And even more complicated…



G=?

Bayes Theorem for IBD Probabilities

posterior

$$P(IBD=i\,|\,Genotypes) = \frac{\mathrm{P(IBD=}\ i,Genotypes)}{P(Genotypes)}$$

prior

$$= \frac{P(IBD=i)P(Genotypes\,|\,IBD=i)}{P(Genotypes)}$$

$$= \frac{P(IBD=i)P(Genotypes\,|\,IBD=i)}{\sum_{j} P(IBD=j)P(Genotypes\,|\,IBD=j)}$$

Prob(data)

E(G) = ½P(IBD=1|Genotypes) + P(IBD=2|Genotypes)

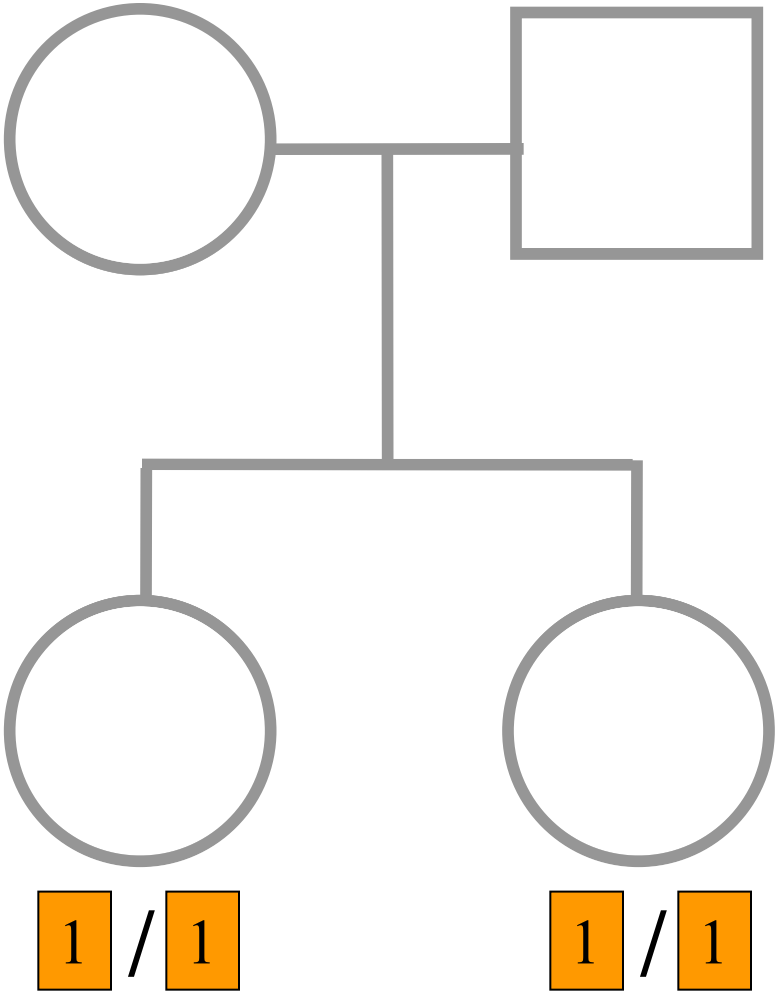
P(Marker Genotype|IBD State)

		IBD		
Sib	**CoSib**	0	1	2
(a,b)	(c,d)	$p_a p_b p_c p_d$	0	0
(a,a)	(b,c)	$p_a^2 p_b p_c$	0	0
(a,a)	(b,b)	$p_a^2 p_b^2$	0	0
(a,b)	(a,c)	$p_a^2 p_b p_c$	$p_a p_b p_c$	0
(a,a)	(a,b)	$p_a^3 p_b$	$p_a^2 p_b$	0
(a,b)	(a,b)	$p_a^2 p_b^2$	$p_a p_b^2 + p_a^2 p_b$	$p_a p_b$
(a,a)	(a,a)	p_a^4	p_a^3	p_a^2
Prior Probability		¼	½	¼

[Assumes Hardy-Weinberg proportions of genotypes in the population]

$p_1 = 0.5$

Worked Example



$$P(Genotypes \mid IBD = 0) = p_1^4 = 1/16$$

$$P(Genotypes \mid IBD = 1) = p_1^3 = 1/8$$

$$P(Genotypes \mid IBD = 2) = p_1^2 = 1/4$$

$$P(Genotypes) = 1/4\, p_1^4 + 1/2\, p_1^3 + 1/4\, p_1^2 = 9/64$$

$$P(IBD = 0 \mid Genotypes) = \frac{1/4\, p_1^4}{P(Genotypes)} = 1/9$$

$$P(IBD = 1 \mid Genotypes) = \frac{1/2\, p_1^3}{P(Genotypes)} = 4/9$$

$$P(IBD = 2 \mid Genotypes) = \frac{1/4\, p_1^2}{P(Genotypes)} = 4/9$$

$\boxed{E(G) = {}^2/_3}$

Estimating IBD from marker data

- Elston-Stewart algorithm

Handles large pedigrees, but small nr of loci, exact IBD distributions (Elston and Stewart, 1971)

- Lander-Green algorithm

Handles small pedigrees, but large nr of loci, exact IBD distributions (Lander and Green, 1987). Software: Merlin

- MCMC methods

Calculates approximate IBD distributions (Heath, 1997). Software: Loki

- Average sharing methods.

Calculates approximate IBD distributions (Fulker et al., 1995; Almasy and Blangero, 1998). Software: SOLAR

Estimate relationship from markers

1. **Well defined, recent base**

Eg Data on families of full-sibs and parents of sibs are the base

a) Calculate Bayesian probability of IBD status at each SNP

→ E(G) at each SNP

average over SNPs

b) Use haplotypes ?

Estimate relationship from markers

2. Less well defined, less recent base

Eg Data on current population, base = ancestors 1000 years ago and allele frequencies in base are known (p and q)

Consider haploid gametes of SNP alleles instead of genotypes

What fraction of the gametes are IBD (G)?

At a single SNP, there are 3 possible data sets and their probabilities are

A and A	A and B	B and B
p^2 +pqG	2pq(1-G)	q^2+pqG

Estimate relationship from markers

SNP genotypes	A and A	A and B	B and B
Probability	$p^2 +pqG$	$2pq(1-G)$	q^2+pqG
score (x)	q/p	-1	p/q

Estimate G(i,j) from the mean value of x over SNPs

This is a relationship between gametes. Calculate G for individuals from the 4 gametic relationships.

See Yang et al (2010) and Powell et al (2010) for the diploid formulae.

Estimate relationship from markers

E.g. Score (x) for pairs of gametes from population in H-W

p(A) = 0.9, q(B) = 0.1

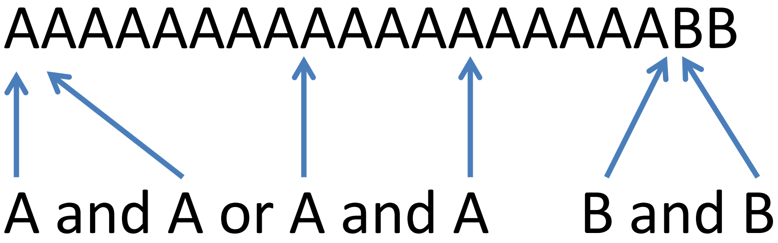
	A (0.9)	B (0.1)
A (0.9)	0.11	-1
B (0.1)	-1	9

Mean G = 0.81 * 0.11 + 0.18 *(-1) + 0.01 *9 = 0

Estimate relationship from markers

E.g. Score (x) for pairs of gametes from population in H-W

p(A) = 0.9, q(B) = 0.1



AAAAAAAAAAAAAAAAAABB

A and A or A and A   B and B

Estimate relationship from markers

E.g. Score (x) for pairs of gametes from same parent

p(A) = 0.9, q(B) = 0.1

Parent	AA	AB	BB
Freq.	0.81	0.18	0.01
	AA (x = 0.11)	AA (0.11)	BB (9)
		AB (-1)	
		BB (9)	

Mean G = 0.81*0.11 + 0.18*(0.25*0.11+0.5*(-1)+0.25*9) + 0.01 *9

= 0.5

Estimate relationship from markers

E.g. Score (x) for pairs of gametes from population in H-W but after allele frequency has drifted to p(A) = 0.8, q(B) = 0.2

	A	B
	(0.8)	(0.2)
A (0.8)	0.11	-1
B (0.2)	-1	9

Mean G = 0.64 * 0.11 + 0.32 *(-1) + 0.04 *9 = 0.11

Estimate relationship from markers

2. No well defined base

Eg random sample from population but don’t know allele frequency in the base.

a) Use the current population as the base

Problem: Some G <0

Cannot interpret as probabilities but still interpret as covariances

If g = genetic value, V(g) = **G** V_A

where G is calculated as above but using allele frequencies in current population.

Estimate relationship from markers

E.g. Score (x) for pairs of gametes from population in H-W but after allele frequency has drifted to p(A) = 0.8, q(B) = 0.2 and using allele frequencies in modern population

	A (0.8)	B (0.2)
A (0.8)	0.25	-1
B (0.2)	-1	4

Mean G = 0.64 * 0.25 + 0.32 *(-1) + 0.04 *4 = 0

Estimate relationship from markers

2. No well defined base

b) Assume SNPs are a random sample of loci as are QTL

$\mathbf{y}$ = mean + $\mathbf{g}$ + $\mathbf{e}$

$\mathbf{y}$ = mean + $\mathbf{Zu}$ + $\mathbf{e}$

Z_{ij} = 0 for AA, 1 for AB or 2 for BB

$\mathbf{u} \sim N(0,\mathbf{I}\sigma_u^2) \rightarrow \mathbf{g} = \mathbf{Zu} \sim N(0,\mathbf{ZZ}'\sigma_u^2)$, $\mathbf{ZZ}'\sigma_u^2 = \mathbf{G}\sigma_g^2$, if $\sigma_g^2 = N\sigma_u^2$

where $N = \Sigma 2pq$ across SNPs

Therefore, $\mathbf{G} = \mathbf{ZZ}'/N$

Estimate relationship from markers

E.g. Score for pairs of gametes from population in H-W

p(A) = 0.8, q(B) = 0.2

		A	B
		(0.8)	(0.2)
	z	0	1
A (0.8)	0	0	0
B (0.2)	1	0	1

Mean G = 0.04 *1 = 0.04

Estimate relationship from markers

E.g. Score for pairs of gametes from population in H-W

p(A) = 0.8, q(B) = 0.2

		A	B
		(0.8)	(0.2)
	z	-0.2	0.8
A (0.8)	-0.2	0.04	-0.16
B (0.2)	0.8	-0.16	0.64

Mean G = 0.64*0.04 + 0.32 * (-0.16) + 0.04 *0.64 = 0

Comparing 2a and 2b

E.g. p(A) = 0.8, q(B) = 0.2

		2b	
		A	B
		(0.8)	(0.2)
	z	-0.2	0.8
A (0.8)	-0.2	0.04	-0.16
B (0.2)	0.8	-0.16	0.64

	2a	
	A	B
A	0.25	-1
B	-1	4

Estimate relationship from markers

2a and 2b compared for gametic relationships

SNP data	A and A	A and B	B and B
score (x)	q/p	-1	p/q
weight (w)	pq	pq	pq

2a) G = mean of x

2b) G = weighted mean of x = Σwx/Σw

This could be described as using the IBS status of SNPs instead of IBD

Estimate relationship from markers

E.g. Score (x i.e. method 2a) for pairs of gametes p(A) = 0.8, q(B) = 0.2 and weighting by pq = 0.16

	A (0.8)	B (0.2)
A (0.8)	0.25*0.16 =0.04	-1*0.16 = -0.16
B (0.2)	-1*0.16 = -0.16	4*0.16 = 0.64

Same as 2b

Estimate relationship from markers

2a) G = mean of x

gives more emphasis to sharing rare alleles

Makes sense because individuals who share rare alleles are more likely to be closely related than individuals who share common alleles.

Gives minimum error variance of relationship under some conditions

Estimate relationship from markers

2. No well defined base

c) Assume SNPs are a random sample of loci as are QTL but effect of SNP decreases as heterozygosity increases

$\mathbf{y}$ = mean + $\mathbf{g}$ + $\mathbf{e}$

$\mathbf{y}$ = mean + $\mathbf{Zu}$ + $\mathbf{e}$

Z_{ij} = 0 for AA, 1 for AB or 2 for BB

$\mathbf{u} \sim N(0,\mathbf{D}\sigma_u^2) \rightarrow \mathbf{g} = \mathbf{Zu} \sim N(0,\mathbf{ZDZ}'\sigma_u^2)$, $\mathbf{ZDZ}'\sigma_u^2 = \mathbf{G}\sigma_g^2$, if $\sigma_g^2 = N\sigma_u^2$

where $N = \Sigma(p_iq_i)$

Therefore, $\mathbf{G} = \mathbf{ZDZ}'/N$

$D_{ii} = 1/(p_iq_i)$

That is, assume the effect of SNPs is proportional to $\surd(p_iq_i)$

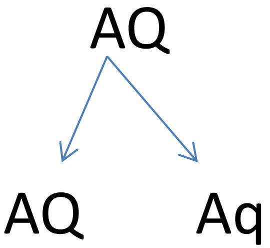
So variance explained by SNPs is not affected by allele frequency

2c = 2a

Estimate relationship from markers

Relationship depends on the markers or QTL

Eg QTL are due to recent mutations



AQ

AQ Aq

Marker is the same but QTL is different

Rare SNP alleles tend to be a recent mutation

Therefore, treat SNPs differently according to MAF

Estimate relationship from markers

Relationship depends on the markers or QTL

Therefore, treat SNPs differently according to MAF

y = mean + g1 + g2 +g3 + g4 +g5 +e

$V(g_i) = (\mathbf{ZZ}'/N)\sigma_i^2$ for SNPs in MAF bin i

Estimate relationship from markers **Summary**

1. In families

2. In the general population

   Express relationship relative to current population

   G can be negative

   G is not a probability

   $V(\mathbf{g}) = \mathbf{G}\sigma_g^2$

   two formulae (2a and 2b)

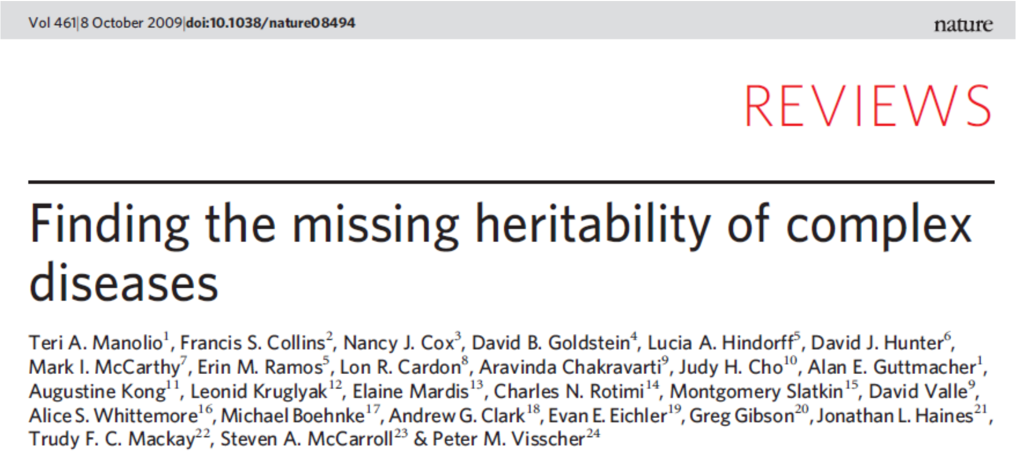
   Same except 2a gives more weight to rare alleles

Application: estimation of SNP-heritability from GWAS data

- Background
  - 2008: GWAS was perceived by many to have failed as an experimental design
  - Missing heritability: discrepancy between pedigree heritability and variance captured by associated SNPs

Disease	Number of loci	Percent of Heritability Measure Explained	Heritability Measure
Age-related macular degeneration	5	50%	Sibling recurrence risk
Crohn's disease	32	20%	Genetic risk (liability)
Systemic lupus erythematosus	6	15%	Sibling recurrence risk
Type 2 diabetes	18	6%	Sibling recurrence risk
HDL cholesterol	7	5.2%	Phenotypic variance
Height	40	5%	Phenotypic variance
Early onset myocardial infarction	9	2.8%	Phenot vari
Fasting glucose	4	1.5%	Phenc varianc

Where is the Dark Matter?

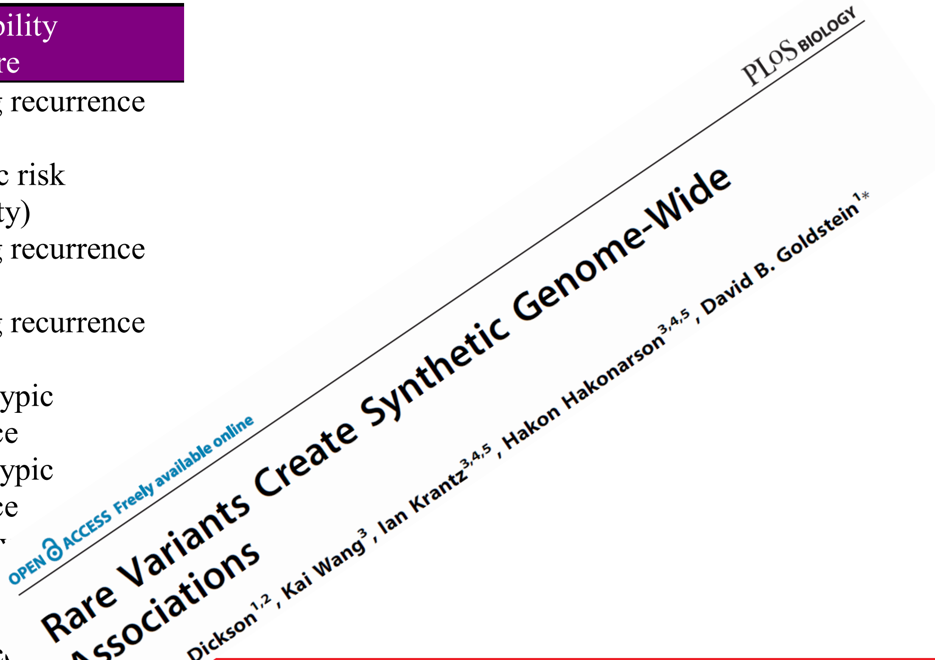


Vol 461|8 October 2009|doi:10.1038/nature08494 nature

REVIEWS

Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

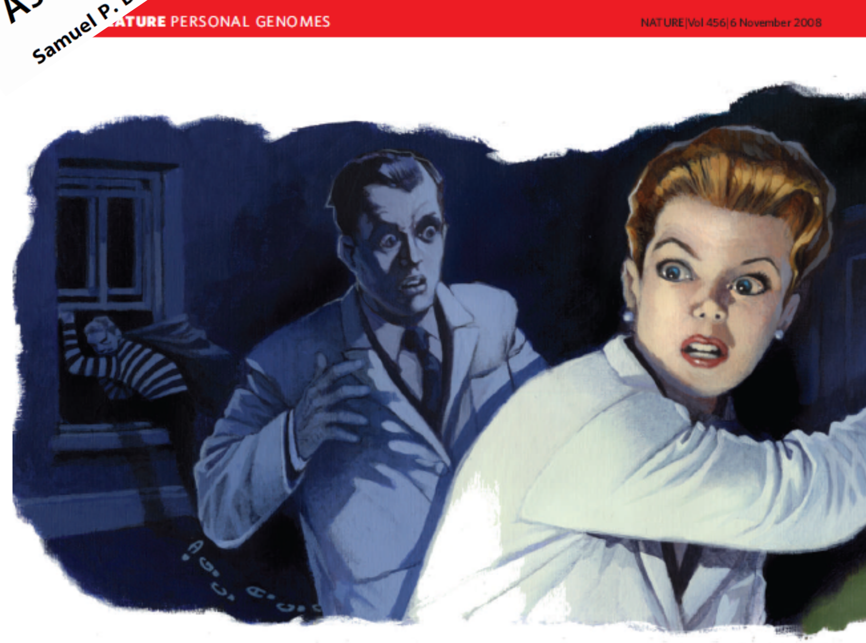


PLoS BIOLOGY

OPEN ACCESS Freely available online

Rare Variants Create Synthetic Genome-Wide Associations

Samuel P. Dickson[1,2], Kai Wang[3], Ian Krantz[3,4,5], Hakon Hakonarson[3,4,5], David B. Goldstein[1*]



NATURE PERSONAL GENOMES NATURE|Vol 456|6 November 2008

The case of the missing heritability

Hypothesis testing vs. Estimation

- GWAS = hypothesis testing
  - Stringent p-value threshold
  - Estimates of effects biased (“Winner’s Curse”)

- Can we estimate the total proportion of variation accounted for by all SNPs?

A model for a single causal variant

	AA	AB	BB
frequency	$(1-p)^2$	$2p(1-p)$	p^2
x	0	1	2
effect	0	b	2b
$z = [x-E(x)]/\sigma_x$	$-2p/\sqrt{\{2p(1-p)\}}$	$(1-p)/ \sqrt{\{2p(1-p)\}}$	$2(1-p)/ \sqrt{\{2p(1-p)\}}$

$y_j = \mu' + x_{ij}b_i + e_j$ x = 0, 1, 2 {standard association model}

$y_j = \mu + z_{ij}u_j + e_j$ $u = b\sigma_x$; $\mu = \mu' + b\sigma_x$

Multiple (m) causal variants

$$y_j = \mu + \Sigma z_{ij}u_j + e_j$$

$$= \mu + g_j + e_j$$

Weighting scheme 2a

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{g} + \mathbf{e}$$

$$= \mu\mathbf{1} + \mathbf{Zu} + \mathbf{e}$$

Equivalence

Let u be a random variable, u ~ $N(0, \sigma_u^2)$

Then $\sigma_g^2 = m\sigma_u^2$ and

$$\begin{aligned} \text{var}(\mathbf{y}) &= \mathbf{ZZ'}\,\sigma_u^2 + \mathbf{I}\sigma_e^2 \\ &= \mathbf{ZZ'}\,(\sigma_g^2/m) + \mathbf{I}\sigma_e^2 \\ &= \mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_e^2 \end{aligned}$$

Model with individual genome-wide additive values using relationships (**G**) at the causal variants is equivalent to a model fitting all causal variants

We can estimate genetic variance just as if we would do using pedigree relationships

But we don’t have the causal variants

If we estimate **G** from SNPs:

- lose information due to imperfect LD between SNPs and causal variants
- how much we lose depends on
  - density of SNPs
  - allele frequency spectrum of SNPs vs. causal variants
- estimate of variance → missing heritability

Let **A** be the estimate of **G** from N SNPs:

$A_{jk} = (1/N) \Sigma \{ x_{ij} – 2p_i)(x_{ik} – 2p_i) / \{2p_i(1-p_i)\}$

$= (1/N) \Sigma z_{ij}z_{ik}$

Methods

- Estimate realised relationship matrix from SNPs
- Estimate additive genetic variance

$$y_i = g_i + e_i$$

$$\mathrm{var}(\mathbf{y}) = \mathbf{V} = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_e^2$$

$$A_{ijk} = \frac{\mathrm{cov}(x_{ij}a_i, x_{ik}a_i)}{\sqrt{\mathrm{var}(x_{ij}a_i)\,\mathrm{var}(x_{ik}a_i)}} = \frac{\mathrm{cov}(x_{ij}, x_{ik})}{2p_i(1-p_i)}$$

Base population = current population

$$A_{jk} = \frac{1}{N}\sum_i A_{ijk} = \begin{cases} \frac{1}{N}\sum_i \frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)}, \; j \neq k \\ 1+\frac{1}{N}\sum_i \frac{x_{ij}^2-(1+2p_i)x_{ij}+2p_i^2}{2p_i(1-p_i)}, \; j = k \end{cases}$$

Statistical analysis

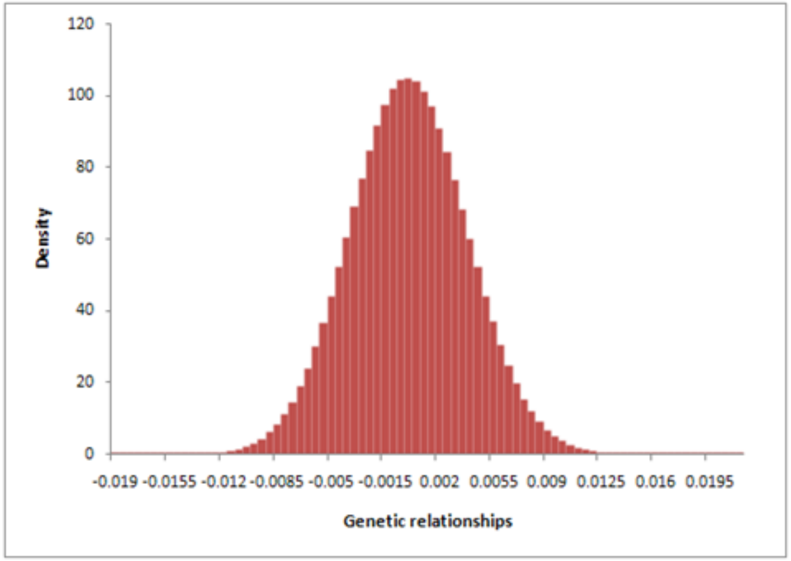
$$\mathrm{var}(\mathbf{y}) = \mathbf{V} = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_e^2$$

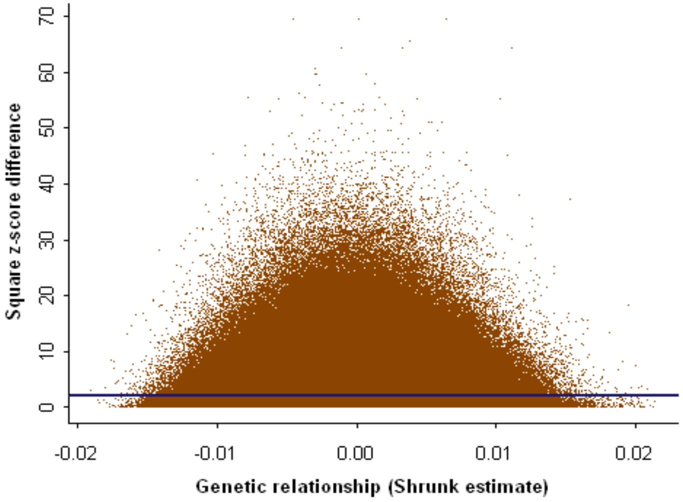
y standardised ~N(0,1)

No fixed effects other than mean

A estimated from SNPs

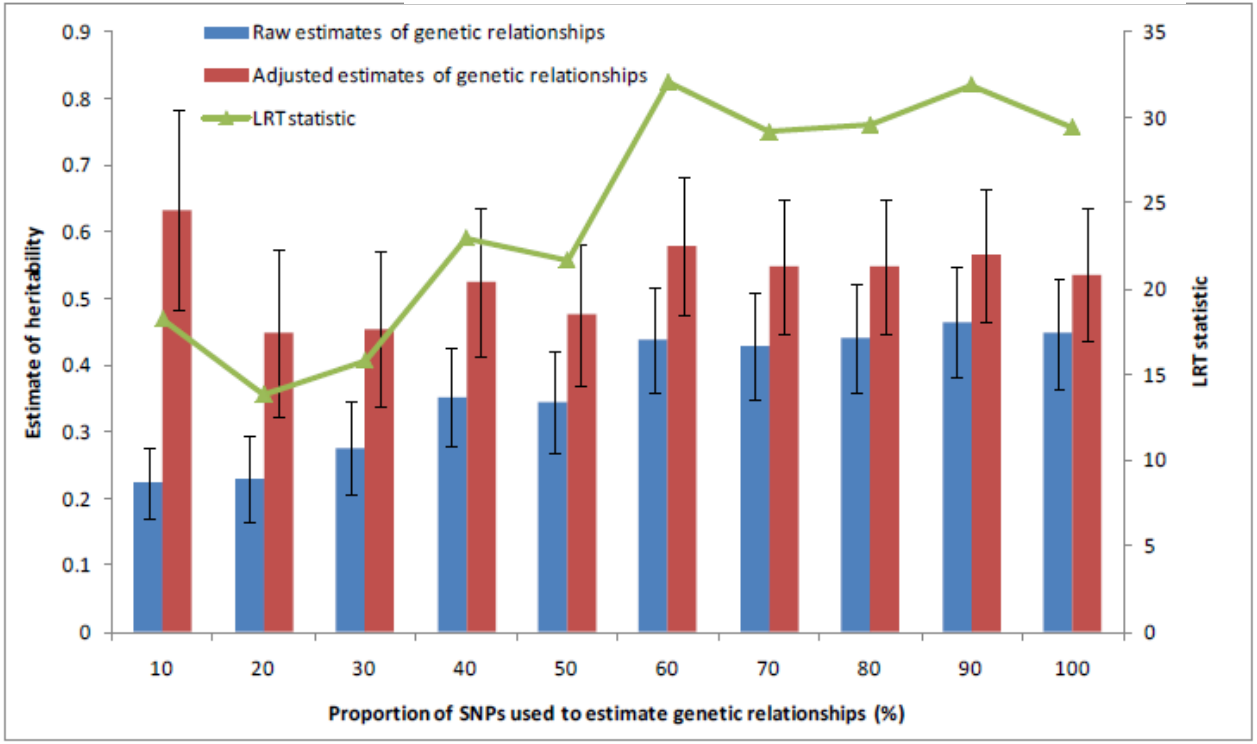
Residual maximum likelihood (REML)





Results

h^2 ~ 0.5 (SE 0.1)



Yang et al. 2010, Nature Genetics, ~ 2000 citations

Checking for population structure

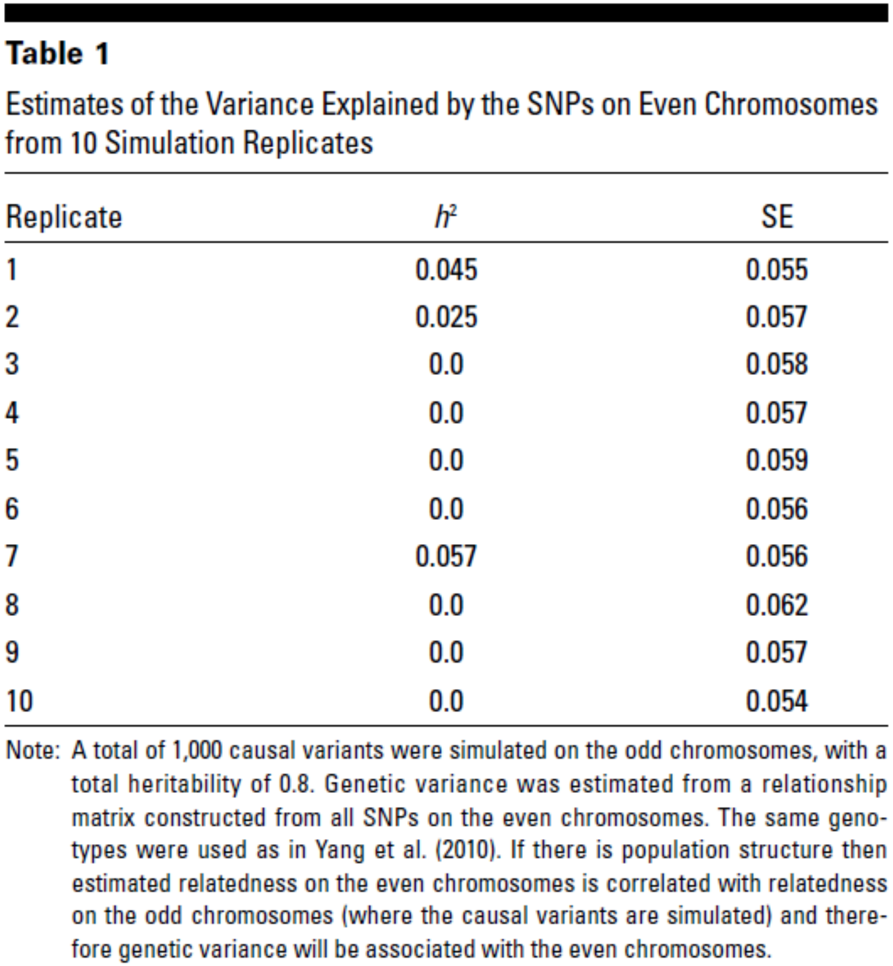


Table 1

Estimates of the Variance Explained by the SNPs on Even Chromosomes from 10 Simulation Replicates

Replicate	h^2	SE
1	0.045	0.055
2	0.025	0.057
3	0.0	0.058
4	0.0	0.057
5	0.0	0.059
6	0.0	0.056
7	0.057	0.056
8	0.0	0.062
9	0.0	0.057
10	0.0	0.054

Note: A total of 1,000 causal variants were simulated on the odd chromosomes, with a total heritability of 0.8. Genetic variance was estimated from a relationship matrix constructed from all SNPs on the even chromosomes. The same genotypes were used as in Yang et al. (2010). If there is population structure then estimated relatedness on the even chromosomes is correlated with relatedness on the odd chromosomes (where the causal variants are simulated) and therefore genetic variance will be associated with the even chromosomes.

Visscher et al. 2010, Twin Research and Human Genetics

Conclusions

- Genetic variance associated with all SNPs can be estimated from GWAS data
  - use SNPs to estimate **G**
  - use phenotypes on “unrelated” individuals and **G** to estimate genetic variance
- Empirical results: most additive genetic variation for height is captured by common SNPs
  - little ‘missing’ heritability
  - GWAS works fine