Kinship/relatedness

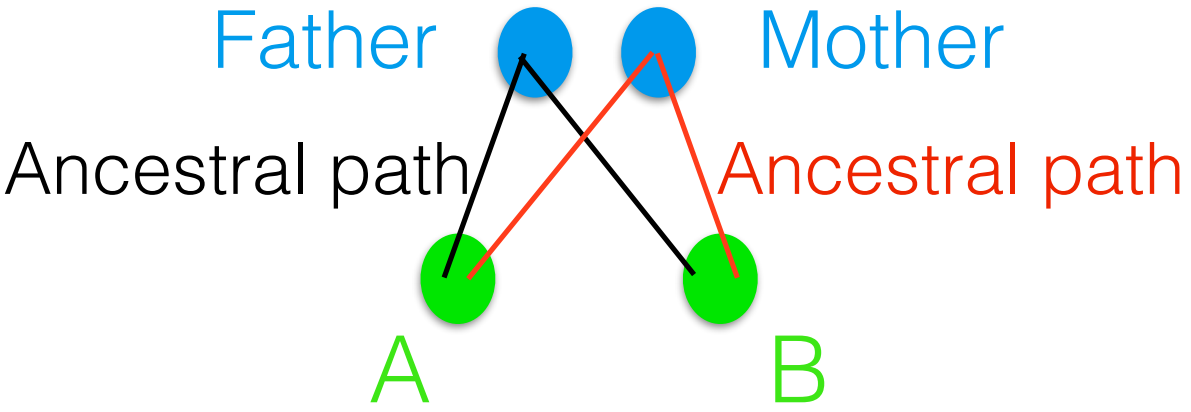
David Balding
Professor of Statistical Genetics
University of Melbourne, and
University College London

2 Feb 2016

1. Ways to measure relatedness
2. Pedigree-based kinship coefficients
3. The statistics of IBD
4. SNP-based measures of genomic similarity
5. Prediction and kinship

Relatedness: what is it? how do we measure it?

- Basic unit is simple: all relationships are made up of parent-child links.
- An ancestral path is a sequence of distinct parent-child steps to each of two individuals starting from a shared ancestor.
- Informally we describe our relationships in terms of the shortest ancestral path(s):
  - siblings are linked by 2 paths of length 2 (both paths have one step up and one step down);
  - half-siblings are linked by 1 path of length 2;
  - half-second cousins are linked by one path of length 6 (three steps up followed by three steps down).



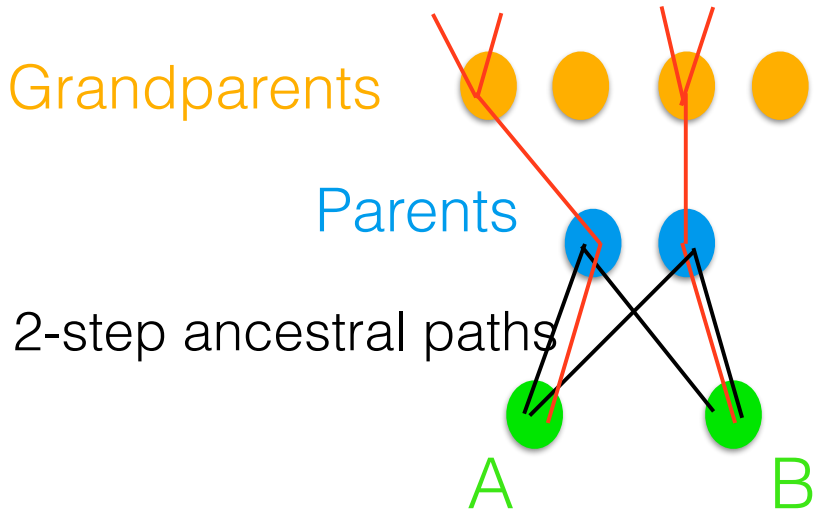
The two 2-step ancestral paths linking two outbred siblings are shown in red and black.

Relatedness: what is it? how do we measure it?

Reality is more complex: we are all linked by very many ancestral paths.

- even pairs of sibs have differing levels of relatedness (see figure);
- there is no such thing as "unrelated", that term just means that the relationship does not include any short ancestral paths;
- long ancestral paths are neglected in many applications,
  - but how to define "long"?

Some of the many longer ancestral paths



In addition to the two 2-step ancestral paths, there are many longer ancestral paths corresponding to the possible ancestries of alleles not shared IBD from a parental allele.

Relatedness: what is it? how do we measure it?

Relatedness is often summarised as a single-number "kinship coefficient",[1] which has become a fundamental concept in quantitative genetics:

- **Heritability** can be estimated as the amount of observed phenotypic variation that can be "explained by" kinships (similar to "variance explained" in a regression model).
- A similar statistical model underlies phenotype **prediction**.

The kinship coefficient is so fundamental to thinking about genetics, that the fact that it is not well defined has been overlooked.

In this module we will take a critical look at different attempts to measure/define relatedness. We closely follow:

> *Speed D, Balding D, "Relatedness in the post-genomic era: is it still useful?"* Nat Rev Genet *Jan 2015*

[1] Alternatively the relatedness coefficient = 2 x kinship.

Pedigree-based kinship coefficients

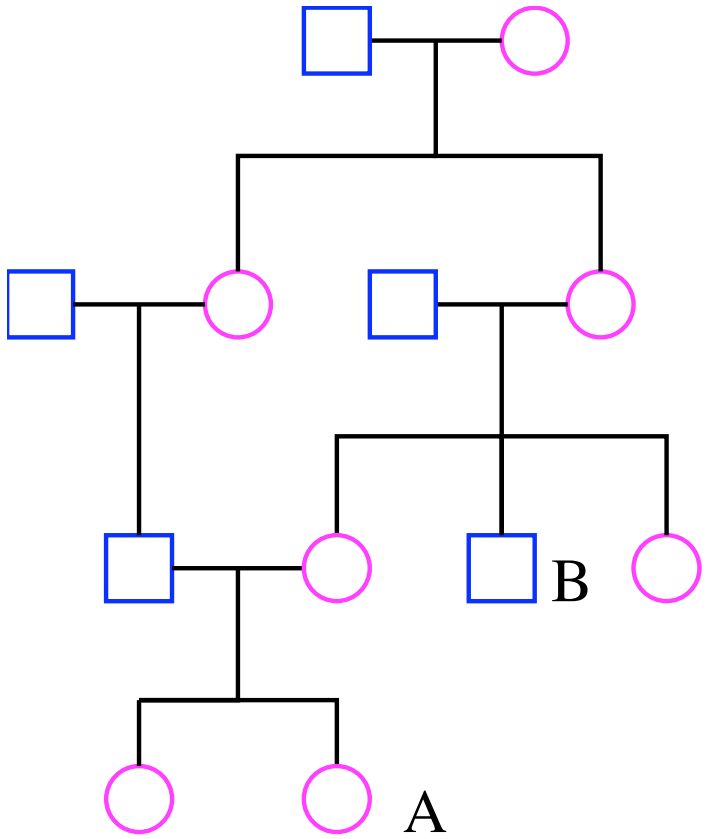
Based on known relationships in a specified pedigree.

Most important is coancestry $\theta(A, B)$, the probability that a random allele from A is Identical by Descent (IBD) with one from B assuming Mendelian probabilities:

$$\theta(A, B) = \sum_X \frac{1 + f_X}{2^{g_X}}.$$

Sum is over most recent common ancestors X of A and B *within the pedigree*;

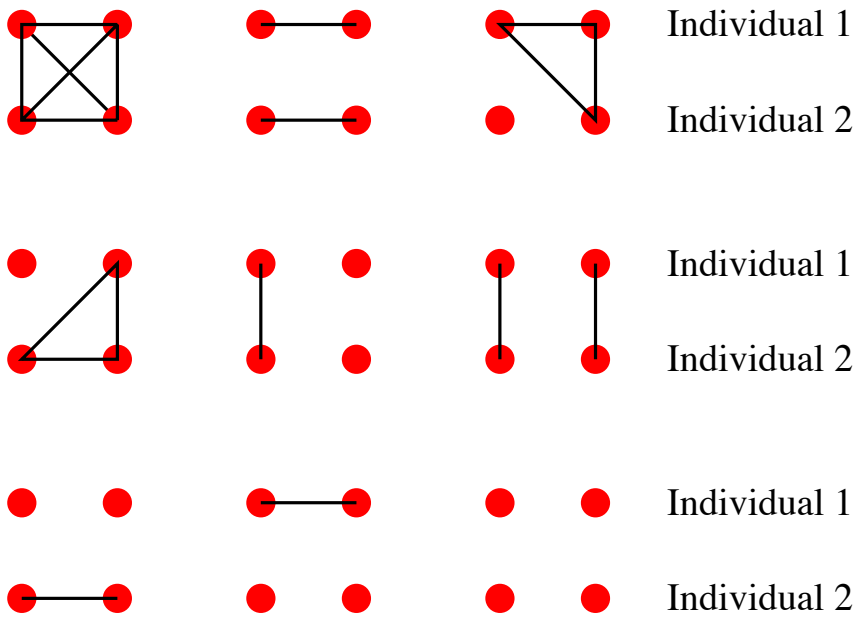
- $f_X = \theta(M(X), F(X))$ inbreeding coefficient of X = coancestry of parents of X;
- g_X is path length from A to B via X.



Additive kinship coefficient based on pedigrees

- 16 possible IBD states among 4 alleles of 2 diploid individuals;
- reduces to 9 ignoring within-individual ordering.
- Also ignoring inbreeding: 3 IBD states (IBD = 0, 1, 2).
- Also ignoring dominance: 1 additive kinship (coancestry) coefficient, $\theta = \mathbb{E}[\text{IBD}]/4 = \mathbb{P}[\text{IBD}{=}1]/4 + \mathbb{P}[\text{IBD}{=}2]/2$.

9 IBD states:



circles = alleles, arcs = IBD.

The θ for you and me is the expected fraction shared IBD in a haploid genome chosen at random from each of us.

For two outbred individuals we write (k_0, k_1, k_2) for the probabilities that they have exactly 0, 1, and 2 alleles IBD. Then $\theta = k_1/4 + k_2/2$.

Relating kinship to phenotypic correlation

Relative pair	(k_0, k_1, k_2)	θ
MZ twins	(0,0,1)	1/2
parent-child	(0,1,0)	1/4
siblings	(0.25,0.5,0.25)	1/4
uncle-niece	(0.5,0.5,0)	1/8
half-sib	(0.5,0.5,0)	1/8
grandparent-grandchild	(0.5,0.5,0)	1/8

Phenotypic covariance among relatives: Individuals i and j have relationship vector (k_0, k_1, k_2) and phenotypes Y_i and Y_j. Then, ignoring epistatic effects, we might assume the following model:

$$\text{Cov}[Y_i, Y_j] = 2\theta\sigma_a^2 + k_2\sigma_d^2 + \gamma\sigma_c^2$$

where $\gamma = 1$ if i and j have the same environment (e.g. same household in childhood), otherwise $\gamma = 0$.

Estimating components of variance

Relative pair	phenotypic covariance
MZ twins	$\sigma_a^2 + \sigma_d^2 + \sigma_c^2$
Parent-child	$\sigma_a^2/2$
Siblings	$\sigma_a^2/2 + \sigma_d^2/4 + \sigma_c^2$
Uncle-niece	$\sigma_a^2/4$

- By computing the phenotypic variance-covariance matrix for many individuals of varying relationships, for example in multiple extended pedigrees, we can estimate σ_a^2, σ_d^2 and σ_c^2. By subtracting these estimates from σ^2 (estimated from unrelated individuals) can estimate σ_e^2.
- Researchers can fit different models depending on their assumptions about sources of variation:
  - an ACE model includes shared environmental effects (C) but not dominance (D) or epistatic (I) effects;
  - an ADE model includes D but not C or I effects.

Problem 1:
θ depends on the pedigree you happen to have available

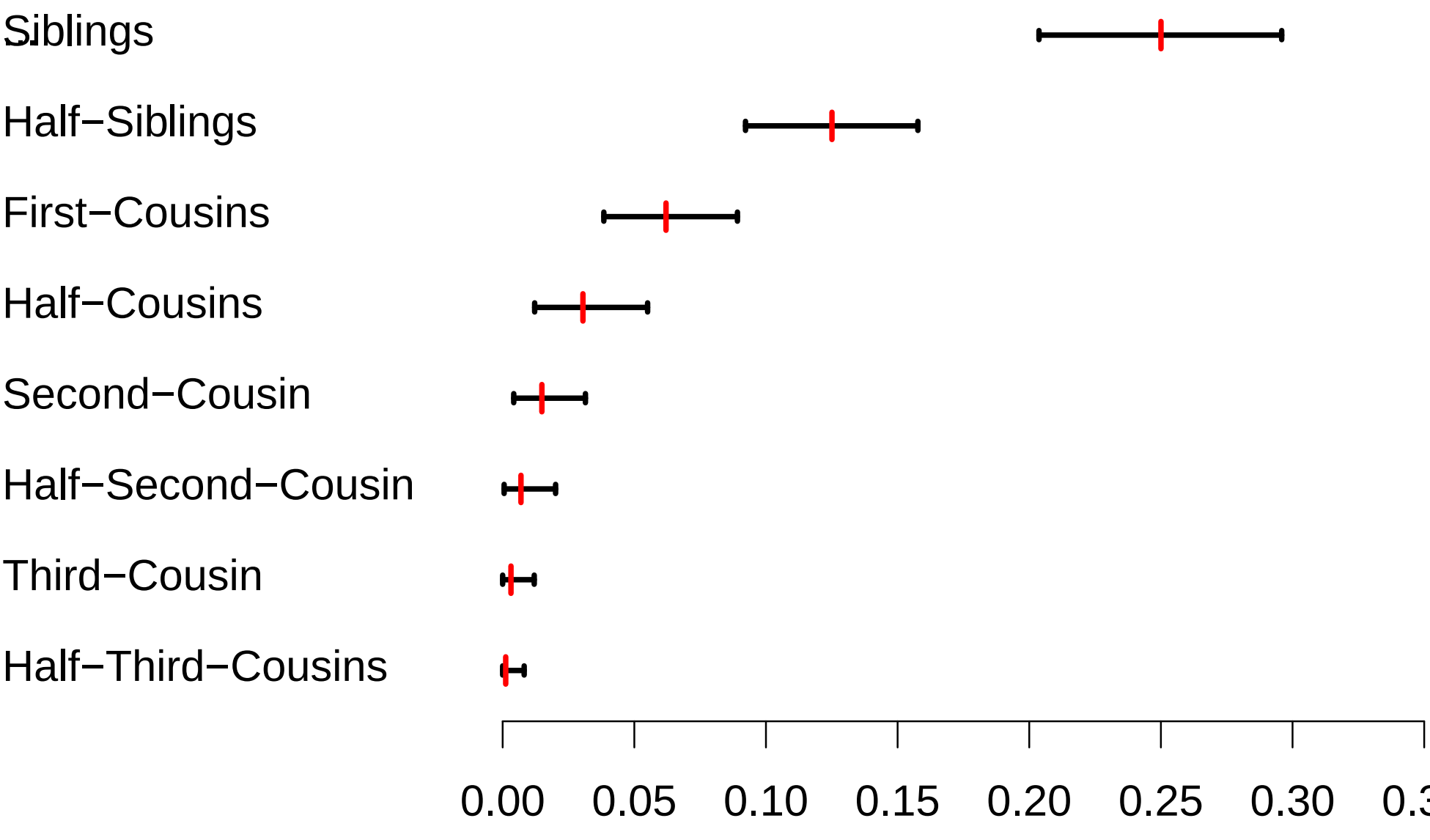
- For diploids, **there is no such thing as a complete pedigree**.
- As more ancestors are added, θ among original pedigree members can only increase and eventually converges to one;
  - so if a complete pedigree were possible, it would be useless.
- **There is also no "ideal" pedigree** in any other sense.
- Similarly for inbreeding (θ between parents): an inbreeding coefficient depends on the available pedigree, and always increases with increasing pedigree information.

Didn't matter much in the past because we could only make use of close relatedness, but with genome-wide date now we can "see" relatives separated by 10 or more meioses.

Problem 2:
θ only captures *expected*, and not *realised*, genome-sharing

- θ for half-sibs is 0.125, but 95% CI is (0.092,0.158).
- Just 6 parent-child transmissions can result in no DNA remaining from the founder.
  - Two children may share no DNA from their common great-grandparent.
    - So they are pedigree-related but not genetically related.
- Conversely, $\theta = 0$ for many pairs of individuals, yet the levels of genome-sharing among "unrelateds" can vary substantially; this has been exploited e.g. for prediction or to estimate SNP heritability.

Genome sharing from recent shared ancestors

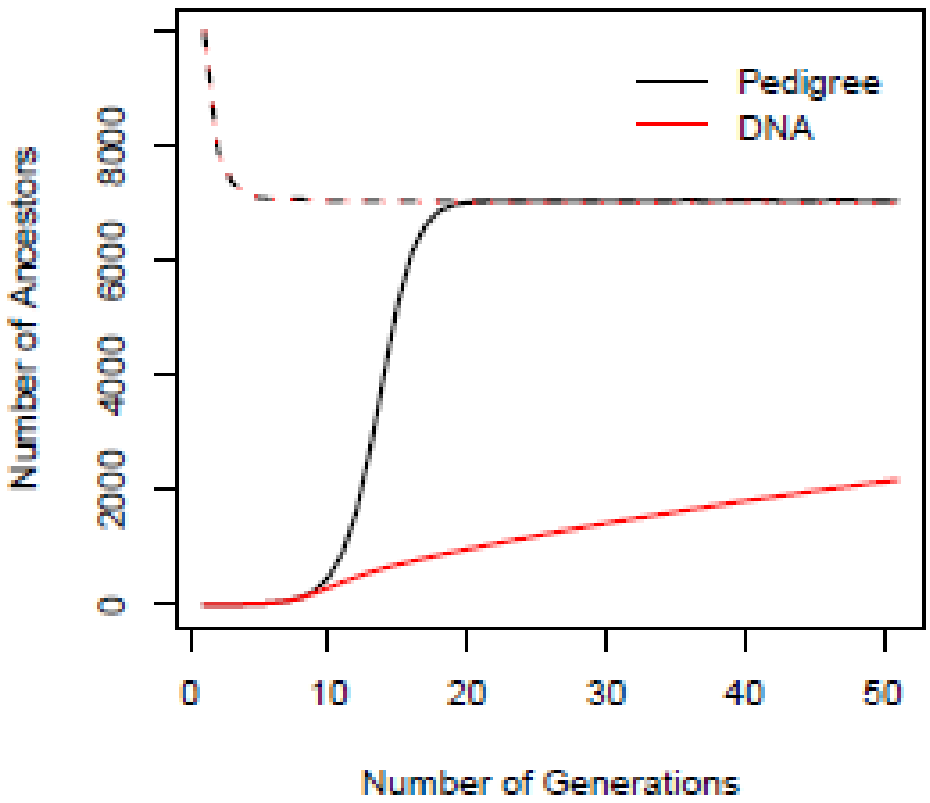


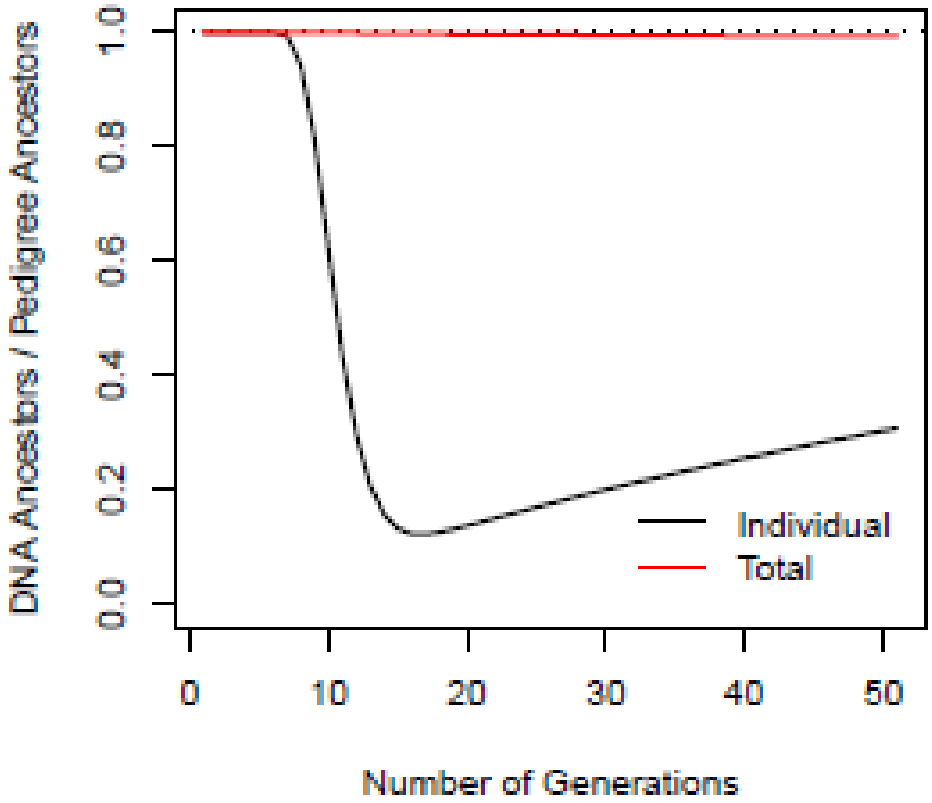
Statistics of IBD from recent shared ancestors (update of Donnelly 1983)

Relationship	# G	# A	$\theta(A,B)$ $\mathbb{E}$[IBD]/4	95% CI	$\mathbb{P}$[IBD>0]	$\mathbb{E}$[#sr]	$\mathbb{E}$[rl] (Mb)
Sibling	1	2	0.250	(0.204,0.296)	1.000	85.3	31.3
1/2-sib	1	1	0.125	(0.092,0.158)	1.000	42.6	"
Cousin	2	2	0.063	(0.039,0.089)	1.000	37.1	18.0
1/2-cuz	2	1	0.031	(0.012,0.055)	1.000	18.5	"
2nd-cuz	3	2	0.016	(0.004,0.031)	1.000	13.2	12.6
1/2-2nd-cuz	3	1	0.008	(0.001,0.020)	0.995	6.6	"
3rd-cuz	4	2	0.004	(0.000,0.012)	0.970	4.3	9.7
1/2-3rd-cuz	4	1	0.002	(0.000,0.008)	0.832	2.2	"
	5	2	0.001	(0.000,0.005)	0.675	0.7	7.9
	7	2	$(1/2)^{14}$	(0.000,0.001)	0.098	0.1	5.5
	9	2	$(1/2)^{18}$		0.009	0.0	4.4

G: generations; A: ancestors; sr = shared regions; rl = region length

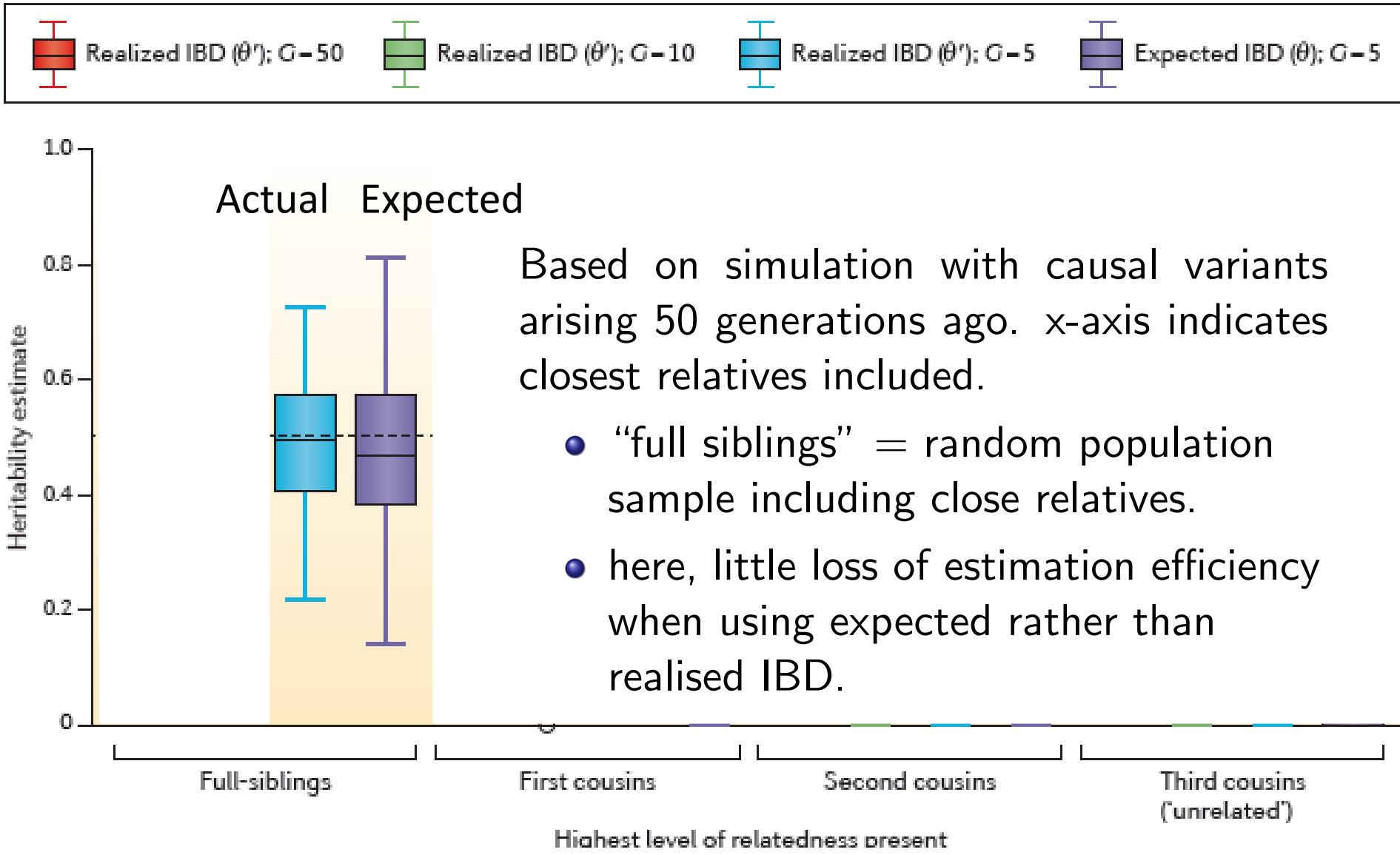
Pedigree ancestors vs DNA ancestors (simple simulation)





The gap between solid red and black lines (left panel; expressed as a fraction on right) corresponds to ancestors in your pedigree (individuals from whom you are descended) from whom you inherited no DNA.

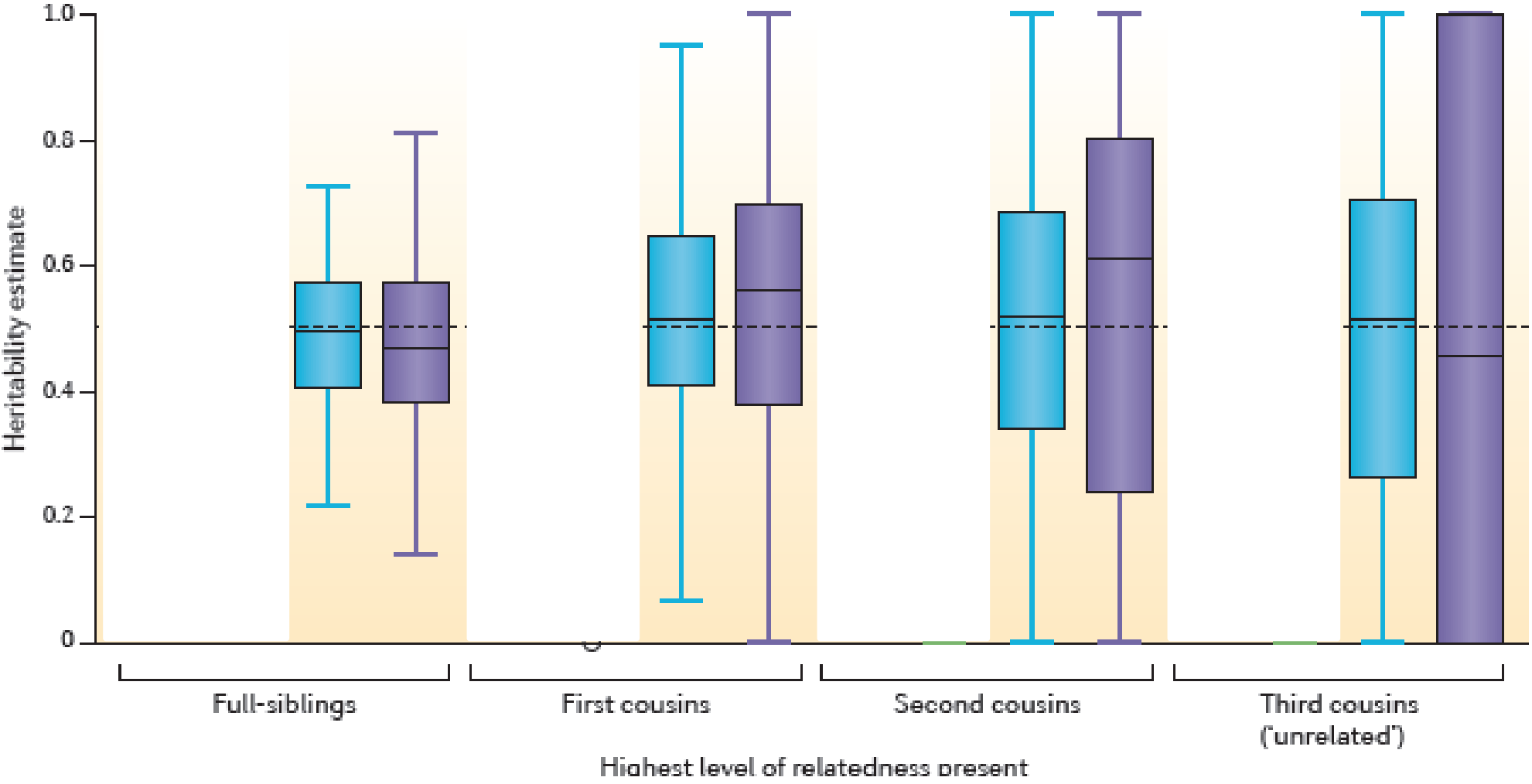
Effect on h^2 estimation of realised versus expected IBD



Based on simulation with causal variants arising 50 generations ago. x-axis indicates closest relatives included.

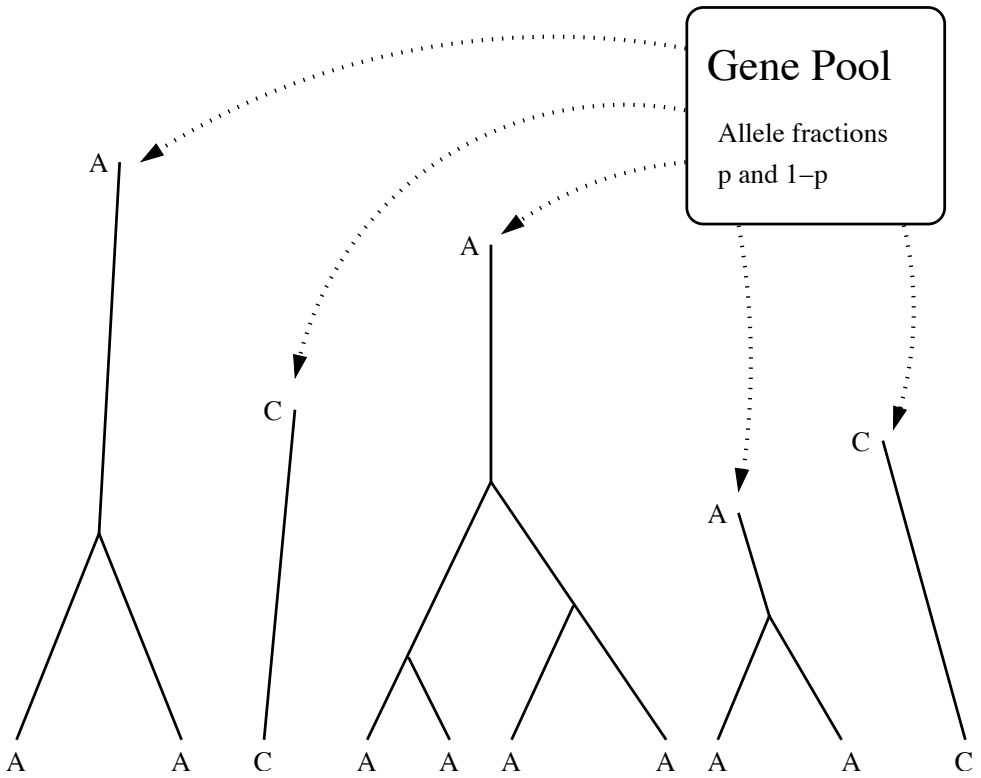
- "full siblings" = random population sample including close relatives.
- here, little loss of estimation efficiency when using expected rather than realised IBD.

Effect on h^2 estimation of realised versus expected IBD



Substantial loss of information for estimation when close relatives excluded.

Kinships based on unobserved pedigrees



- Many population genetics models define kinship in terms of excess allele sharing, measured as a correlation (no reference to a pedigree).
- The correlation coefficient = pedigree θ **if** individuals come from a finite pedigree with unrelated founders, and **if** allele probabilities in founders are known.

Pop gen textbooks and practice put much weight on this theory

- but the underpinning assumptions don't hold;
- negative estimates are frequent yet θ is positive by definition.

IBD genome segments

Homologous segments from two haploid genomes are (recombination-sense) IBD if there has been no recombination within the segment since their MRCA (mutation is ignored).

With sequence data, it is now common to think of relatedness in terms of numbers and sizes of IBD segments.

Advantages:

- No need for an explicit pedigree and no founder population.

Problems:

- Recombinations cannot always be inferred.
- Easy to identify if shared segment is large, but most shared segments are short, even for close relatives.
- Limited use as a measure relatedness: two haploid genomes are entirely IBD, relatedness is reflected in distribution of IBD fragment lengths, which is hard to infer.

IBD genome segments

Homologous segments from two haploid genomes are (recombination-sense) IBD if there has been no recombination within the segment since their MRCA (mutation is ignored).

With sequence data, it is now common to think of relatedness in terms of numbers and sizes of IBD segments.

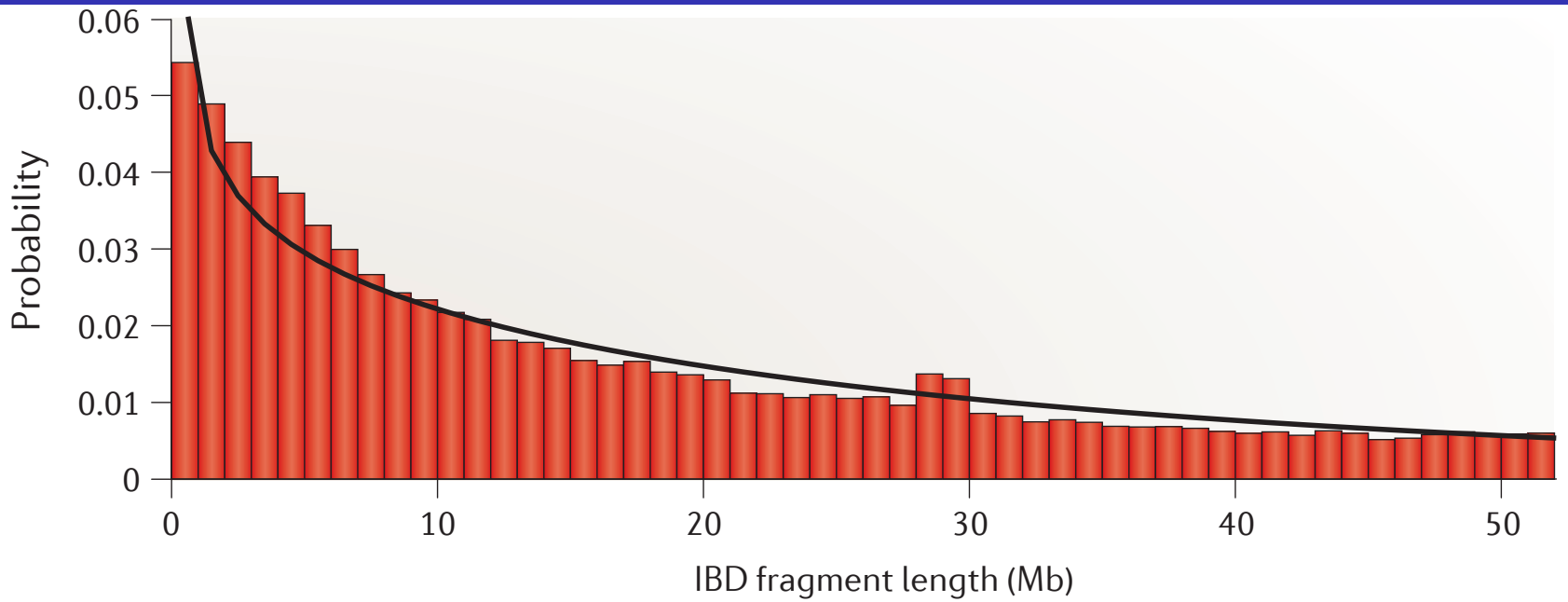
Advantages:

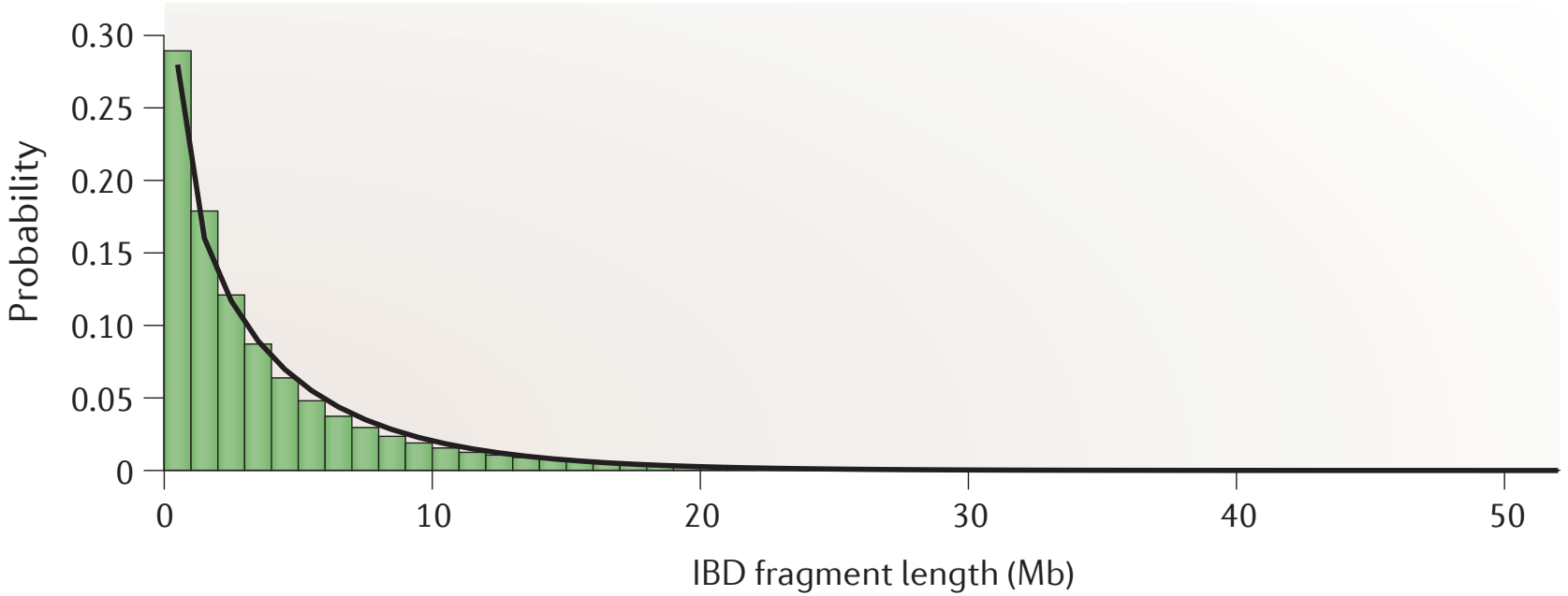
- No need for an explicit pedigree and no founder population.

Problems:

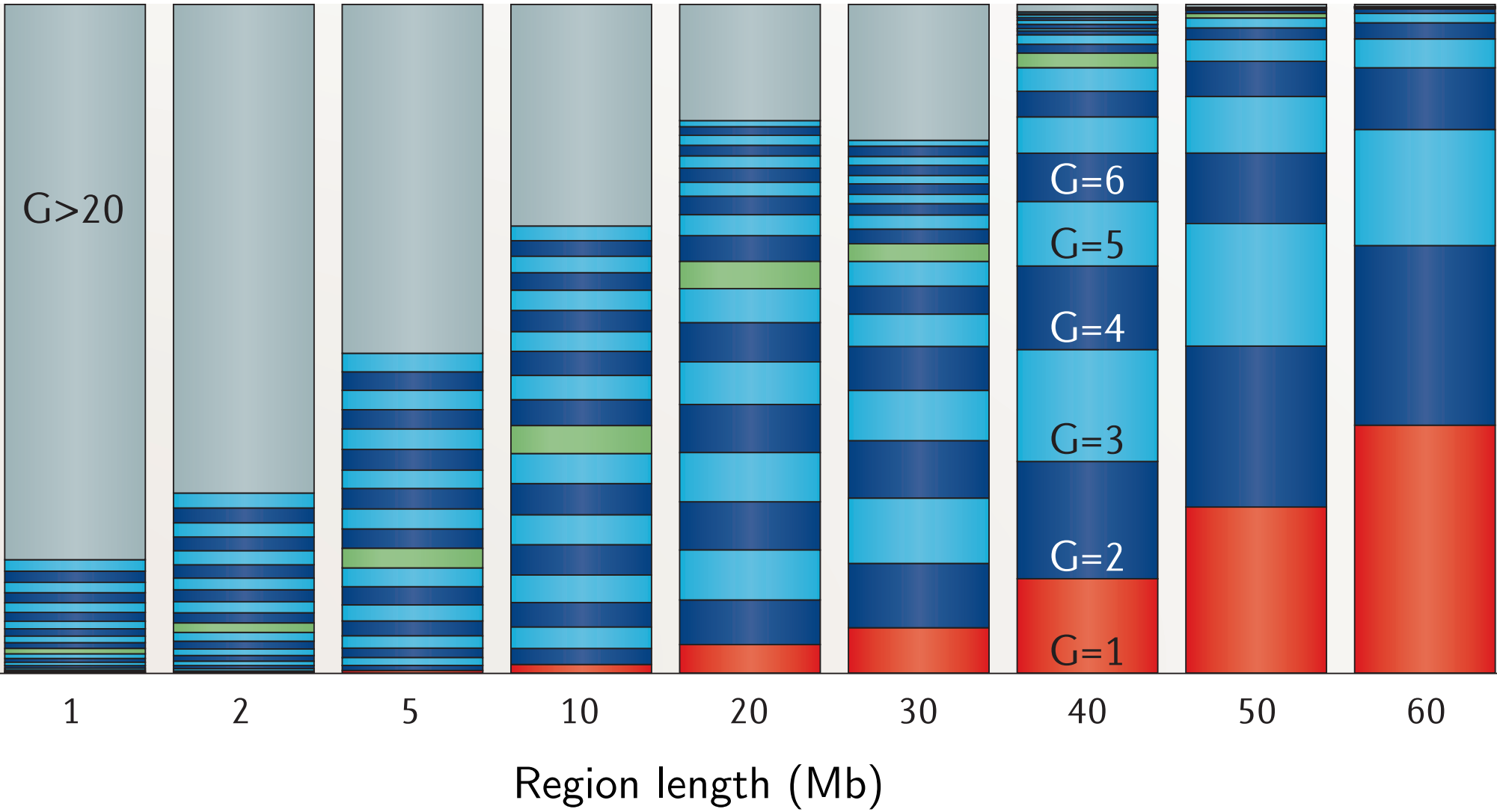
- Recombinations cannot always be inferred.
- Easy to identify if shared segment is large, but most shared segments are short, even for close relatives.
- Limited use as a measure relatedness: two haploid genomes are entirely IBD, relatedness is reflected in distribution of IBD fragment lengths, which is hard to infer.

Fragment lengths IBD from 1 and 10 generations ago





Distribution of TMRCA given IBD fragment length



Consumer genetics and IBD

- Large consumer genetics companies have $\sim 10^6$ customers genotyped at $\sim 10^6$ SNPs.
- They are interested to identify IBD segments in order to infer (remote) pedigree relationships.
  - The relationship is usually expressed in terms of the shortest ancestral path (e.g. 3rd cousin, two paths each of length 8) but these are hard to distinguish from many other relationships e.g. involving multiple ancestral paths.
- Why should a customer prefer a poorly-inferred pedigree relationship to a direct measure of genome similarity?

Summary so far

- Classical measures of relatedness had serious flaws, but were good enough for many applications in the pre-genome era.
  - With genome-wide data now available, we need new concepts definitions and measures (not estimates!).
- Many researchers still regard pedigree kinships as "gold standard",
  - but they are unsatisfactory as a definition;
  - they were only a convenient proxy when we didn't have genome data.
- Only actual genome similarity matters for most purposes.

So how do we measure genome similarity?

SNP-based kinships

There are many ways to measure genetic similarity of two individuals from genome-wide genetic markers (SNPs),

- which one is the best?

One difficulty in humans is that we are all closely related:

- Any two haploid human genomes share over 99.9% sequence identity due to shared ancestry.
- This isn't evident for SNPs because they are highly polymorphic, but
  - measures of similarity can depend sensitively on the Minor Allele Fraction (MAF) spectrum.
  - more low-MAF sites $\Rightarrow$ greater similarity.
  - depends on SNP chip **and** QC.

SNP-based kinships

Two approaches:

1. **Average haplotype sharing.**
2. **Genome-wide average of a single-SNP measure.**

We briefly discuss approach 1 here, approach 2 on following slides.

Average haplotype sharing:

- Identify genome segments that are IBD between two individuals.
- Measure kinship by the number of shared fragments, or their total length.
- Useful in some settings, but small (e.g. $<$ 1Mb) shared fragments are informative yet hard to exploit:
  - Because any two human genomes are $>$ 99.9% IBD, an arbitrary decision must be made to ignore small IBD fragments.
  - This decision can have a big impact on the resulting measure of kinship.

Single-SNP approach 1: Average allele-sharing

- Code SNP genotypes as 0,1 and 2, and so the genotypes of the two individuals can be represented as a pair, such as (0,1): individual A has genotype 0 while B has genotype 1.
- Pairs of genotypes are assigned a score $= \mathbb{P}$(allele drawn at random from A $=$ allele drawn at random from B):

$$\begin{array}{rcc} (0,0) \text{ or } (2,2) & \rightarrow & 1 \\ (0,1), (1,1) \text{ and } (1,2) & \rightarrow & 1/2 \\ (0,2) & \rightarrow & 0 \end{array}$$

Note similarity with the definition of coancestry (θ), but instead of the probability that the two alleles are descended from a common ancestor within the pedigree (which can be computed without genotypes) we use the probability that the alleles are observed to be the same (sometimes called Identity By State, IBS).

Single-SNP approach 1: Average allele-sharing

- Using above definition, the kinship of an individual with itself is $(1 + h)/2$, where h is the fraction of heterozygous sites.
- This is similar in form to the pedigree-kinship of an individual with itself which is $(1 + f)/2$, where f is the individual's inbreeding coefficient (coancestry of its parents).
- Disagreement about how to code heterozygotes: PLINK is highly influential and it codes (1,1) as 1, rather than 0.5.
  - Now, the kinship can be represented in a simple formula

$$1 - \frac{1}{2m} \sum_{j=1}^{m} |G_{Aj} - G_{Bj}|$$

    where $G_{Aj} \in \{0, 1, 2\}$ is the genotype of A at the jth locus.
  - The kinship of an individual with itself is always 1.
- Not clear which coding is better, and often not clear which coding has been used in a calculation.

Single-SNP approach 2: Average allelic correlation

Average allele sharing has the advantage of not requiring MAF values, but disadvantages:

- Matching common alleles score the same as matching rare alleles;
- The result is very sensitive to the MAF spectrum of the SNPs.

The coancestry θ can be represented as a correlation coefficient, which suggests the following expression for the kinship of A and B:

$$\frac{1}{m}\sum_{j=1}^{m}\frac{(G_{Aj}-2p_j)(G_{Bj}-2p_j)}{2p_j(1-p_j)}$$

a genome-wide average of single-SNP sample-size-1 correlation estimates.

- This expression upweights the sharing of rare shared alleles (which provides more evidence for a recent common ancestor).
- Not clear what MAF values to use (the p_j) and these have a big impact on the results.
- Usually sample MAFs are used, which implies that many negative kinship values will be observed.

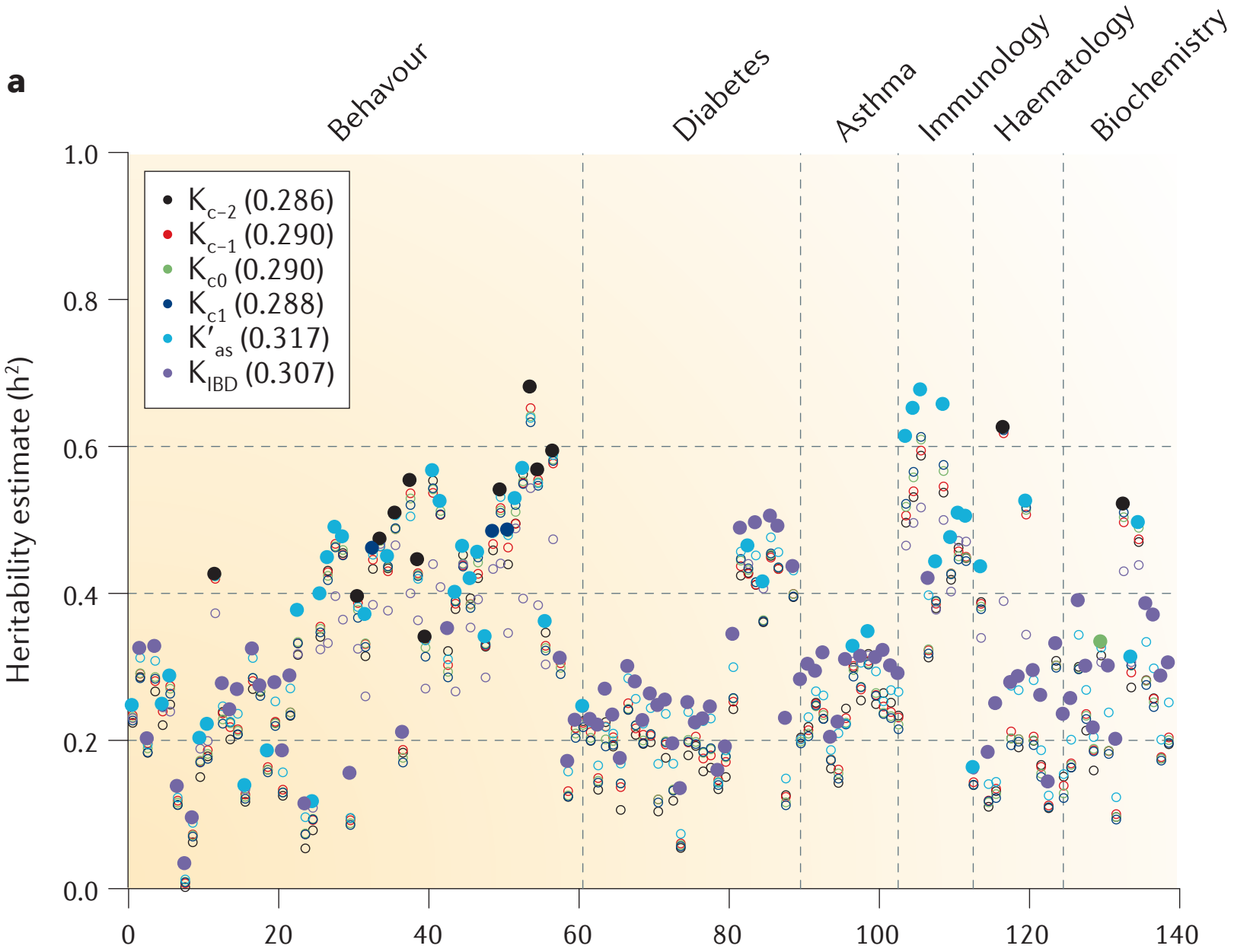
Single-SNP approach 3: A more general formula

The kinship formula introduced above (Single-SNP approach 2) is the case $\alpha = -1$ of a more general formula:

$$K_\alpha = \frac{1}{m}\sum_{j=1}^{m}(G_{Aj} - 2p_j)(G_{Bj} - 2p_j) \times [2p_j(1-p_j)]^\alpha$$

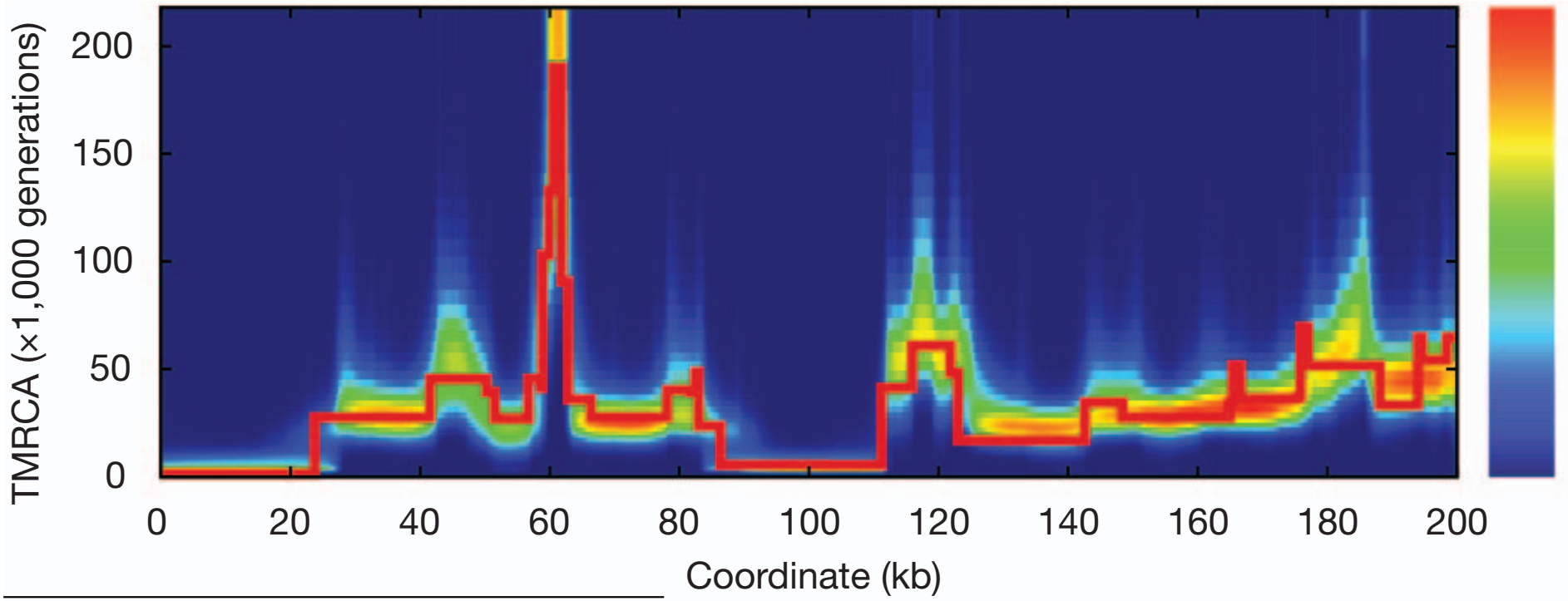
- Animal/plant breeders tend to use $\alpha = 0$, human geneticists $\alpha = -1$
- For many applications, the value of α encodes an assumption about the relationship between the MAF of an allele and its effect size.
  - $\alpha = 0$ implies the same effect size distribution for each SNP, irrespective of MAF.
  - $\alpha = -1$ implies that each SNP is expected to contribute the same to total heritability, which implies that effect size is inversely proportional to MAF.
  - Other values of α imply different MAF/effect size relationships

$\hat{h}^2$ for139 mouse traits using various kinship matrices



A better way to measure relatedness?

- Relatedness is a property of ALL the shared ancestors of two individuals from whom they both inherited DNA;
- Better to use (genome-wide average) Time since the MRCA.
- TMRCA estimated from markers/sequence + demographic model.
- Estimates used for inferring historical demographic parameters.[2]



[2]Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature* 475, 493-496 (2011).

Historically prediction of phenotype was understood in terms of exploiting relatedness summarised by kinship coefficients:

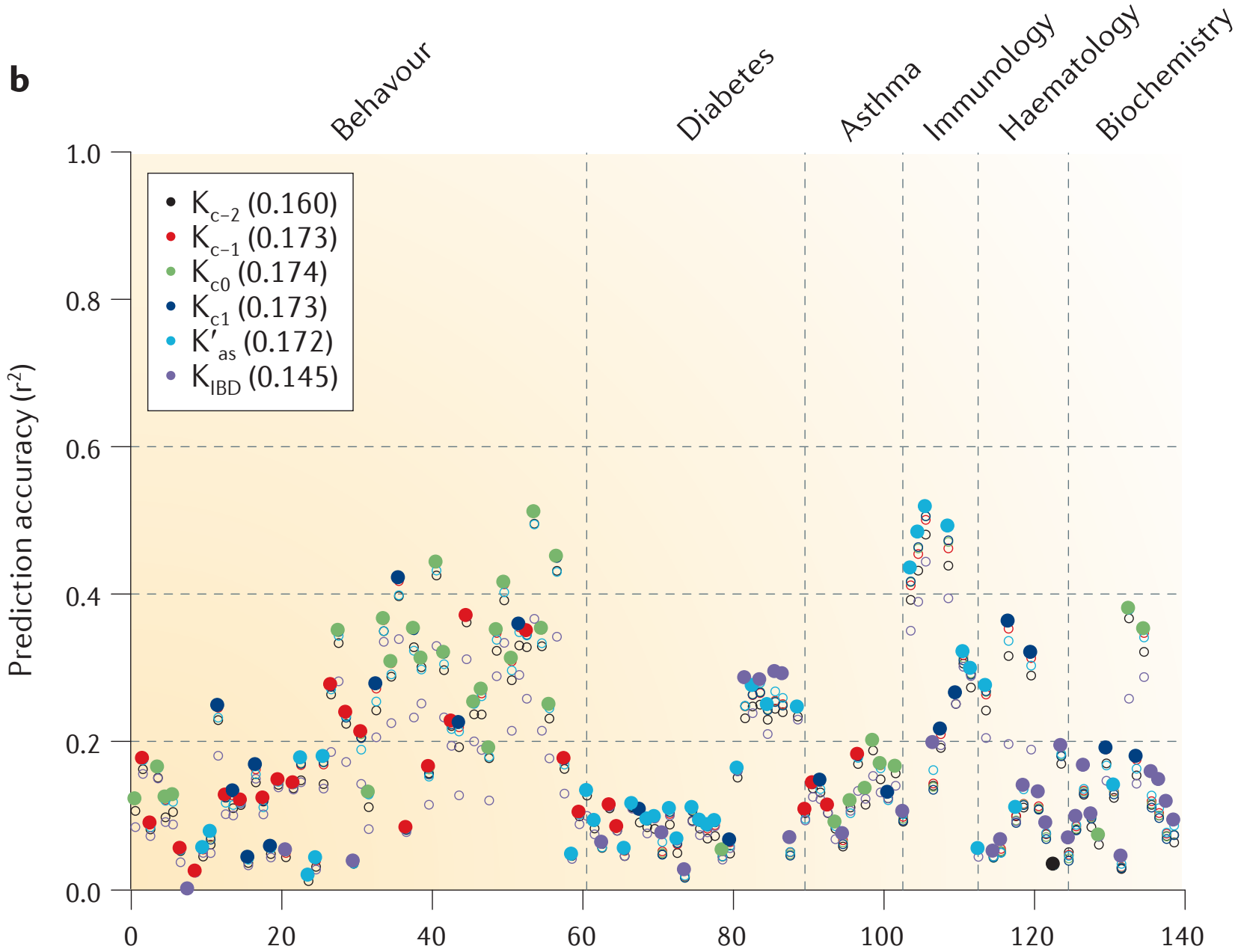
- mathematically the standard formulation involved a matrix of kinship coefficients, usually understood to be uniquely defined.

Now we have many different kinship coefficients:

- we are free to tailor the kinship coefficient to match the genetic architecture of the trait
- can use multiple different kinship coefficients
  - for example corresponding to different genome regions
  - or for pedigree relationships and SNPs (after adjusting for pedigree)

Exciting new possibilities, but the traditional notion of kinship coefficients is no longer useful - more tomorrow.

Prediction of 139 mouse traits, various kinship matrices



Model likelihood, heritability and prediction for 7 human disease traits, kinship K_α for $\alpha \in \{-2, -1, 0, 1\}$

	Log likelihood				Heritability ($\hat{h}^2$)				Prediction accuracy (r^2)			
	K_{c-2}	K_{c-1}	K_{c0}	K_{c1}	K_{c-2}	K_{c-1}	K_{c0}	K_{c1}	K_{c-2}	K_{c-1}	K_{c0}	K_{c1}
order	−97	0*	−12	−32	1.00*	0.98	0.92	0.81	0.040	0.074*	0.073	0.069
rtery disease	−24	−3	0*	−1	0.33	0.41*	0.17	0.06	0.000	0.017	0.020*	0.02
ease	−178	−5	0*	−3	1.00	1.00	1.00	1.00	0.057	0.096	0.098*	0.095
ion	−32	−3	0*	−1	0.57*	0.48	0.21	0.08	0.005	0.024	0.026*	0.026
d arthritis	−125	0*	−15	−72	0.77	1.00*	0.99	0.17	0.016	0.043	0.042	0.043*
etes	−65	0*	−7	−16	0.85*	0.82	0.41	0.16	0.031	0.060	0.060*	0.056
etes	−28	0*	0	−3	0.64*	0.52	0.22	0.08	0.009	0.026*	0.025	0.024
	−78	−2*	−5	−18	0.74	0.74*	0.56	0.34	0.022	0.048	0.049*	0.047

- Log likelihoods, computed under the mixed model, are reported relative to the maximum observed over the four α values.
- $\hat{h}^2$ values correspond to the observed scale (not directly interpretable but OK for comparisons here).
- The GSMs marked by asterisks indicate those that maximize the model likelihood, $\hat{h}^2$ and r^2.