Estimation of relationships from markers in population samples

Key concepts

- Relationships between gametes or individuals can be estimated from markers in the absence of a clear ‘base population’
- Estimated relationships can be interpreted as the expected genetic covariance between pairs of individuals
- Methods to estimate relatedness from SNP data differ in how they weight rare vs common variants

Recap: Estimate relationships from markers

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

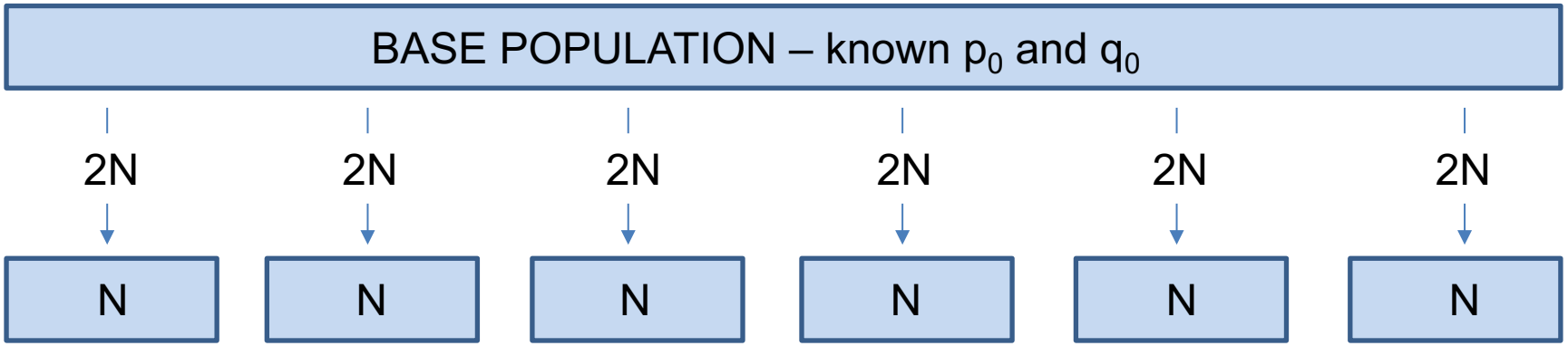
Well defined, recent base (reference)

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

Estimate relationships from markers

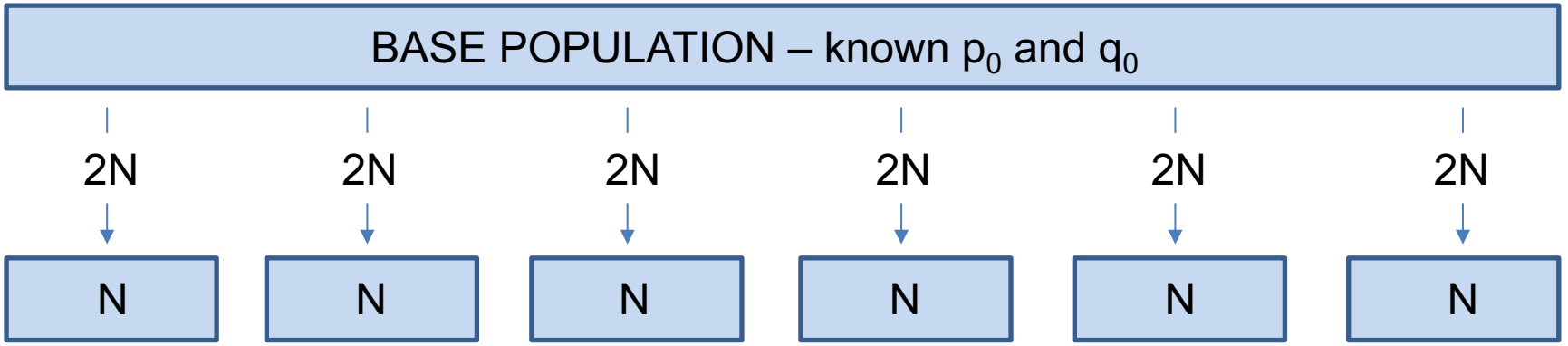
2. Less well defined, less recent base

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



Estimate relationships from markers

2. Less well defined, less recent base



mean of allele frequencies across samples = p_0

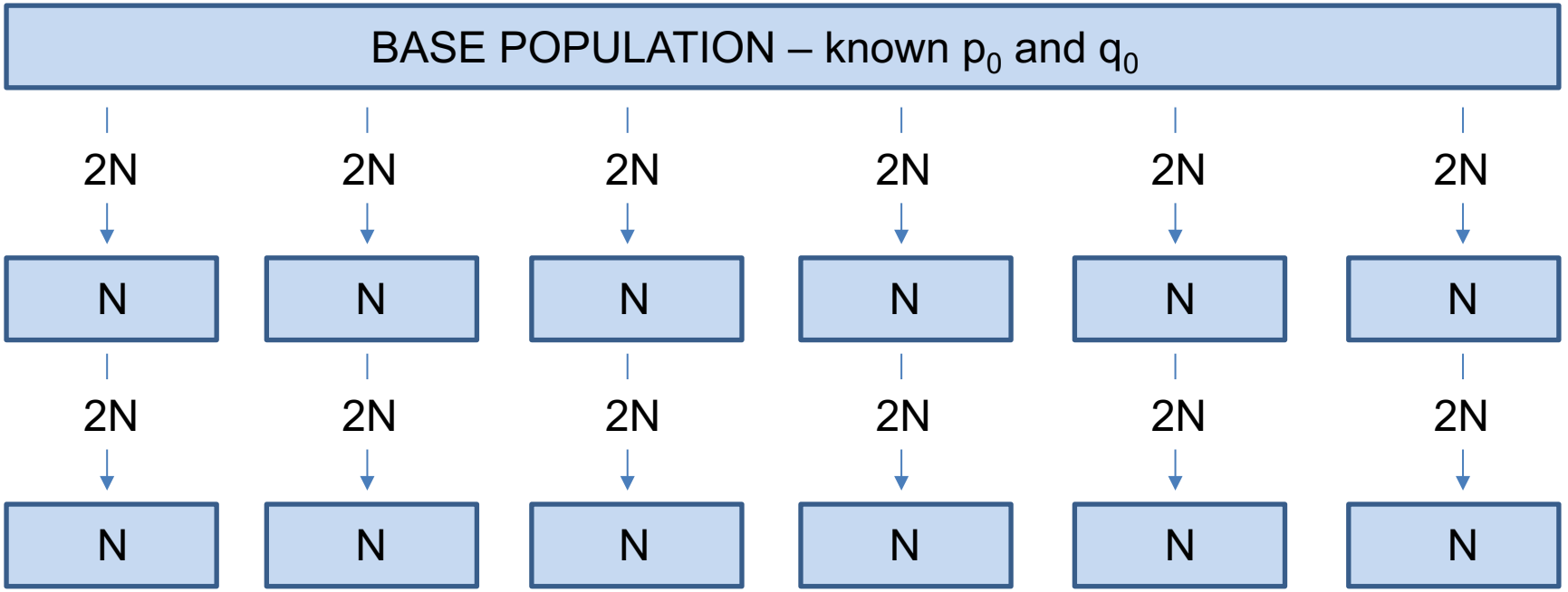
$\sigma_p^2 = p_0 q_0 / 2N$

= change in allele frequency due to sampling

samples differ in allele frequency but population mean is unchanged

Estimate relationships from markers

2. Less well defined, less recent base

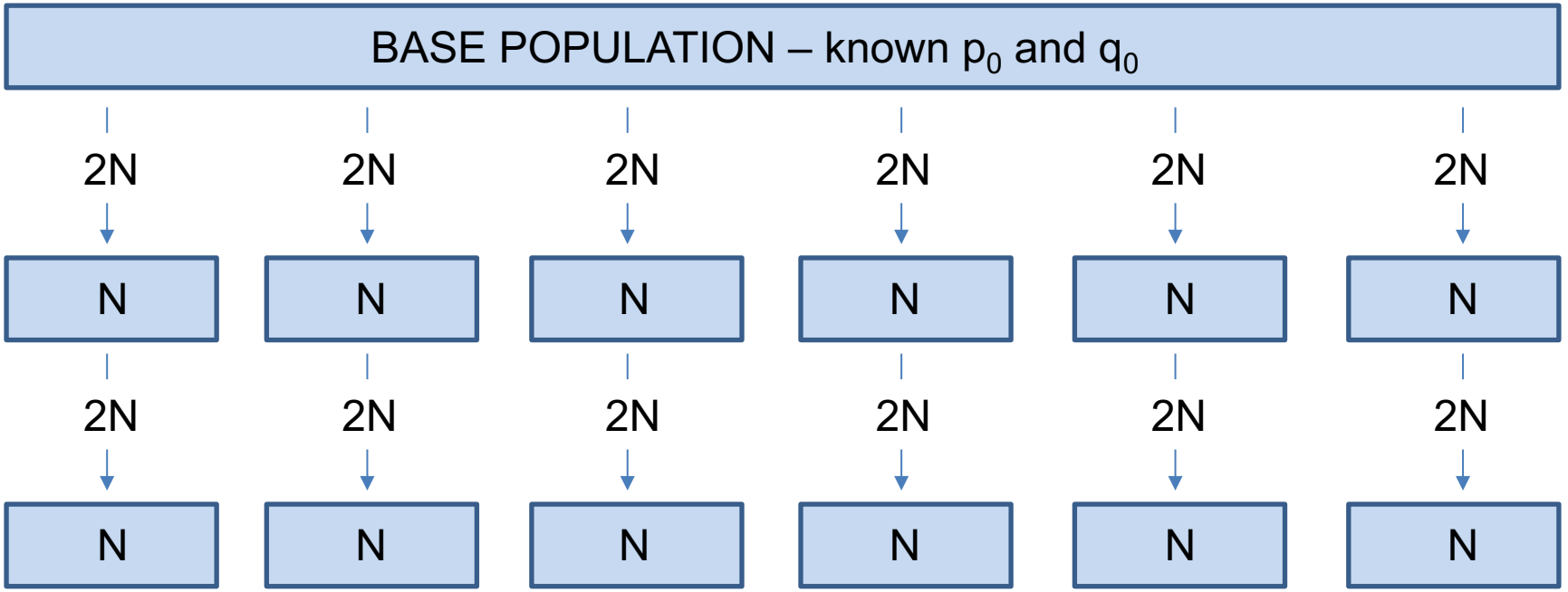


more generations mean repeated sampling.

what is the variance in allele frequency?

Estimate relationships from markers

2. Less well defined, less recent base

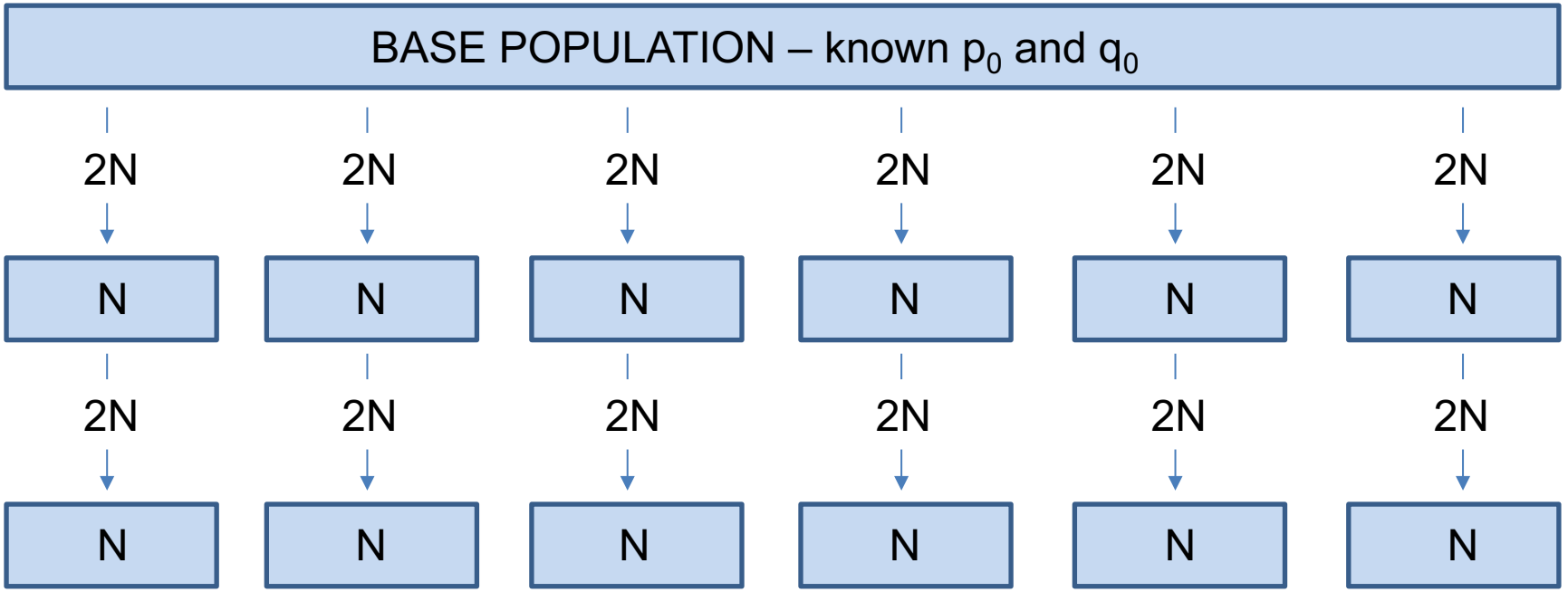


Drift (inbreeding) occurs as a result of sampling

Coefficient of inbreeding (*F*) is probability that two alleles at a locus in an individual are IBD.

Estimate relationships from markers

2. Less well defined, less recent base



generation 1: $F_1 = \frac{1}{2N}$ probability same gametes uniting

generation 2: $F_2 = \frac{1}{2N} + (1 - \frac{1}{2N})F_1$ = increment + remainder

Estimate relationships from markers

2. Less well defined, less recent base

$F_2 = \frac{1}{2N} + (1 - \frac{1}{2N})F_1$ = increment + remainder

increment = $\Delta F = \frac{1}{2N}$

$F_t = \Delta F + (1 - \Delta F)F_{t-1}$

$F_t = 1 - (1 - \Delta F)^t$, inbreeding coefficient relative to base population

back to variance of allele frequency…

$\sigma^2_{\Delta p} = p_0 q_0 / 2N = p_0 q_0 \Delta F$

$\sigma^2_p = p_0 q_0 F_t$

Estimate relationships from markers

2. Less well defined, less recent base

$$\sigma_p^2 = p_0 q_0 F_t$$

The probability of sampling different genotypes is then:

A and A	A and B	B and B
$q^2 + pqF$	$2pq(1-F)$	$p^2 + pqF$

We want to estimate relationship between gametes and ask: *what fraction of the gametes are IBD (G)?*

[Falconer 1960]

Estimate relationships from markers

2. Less well defined, less recent base

What fraction of the gametes are IBD (G)?

SNP genotypes	A and A	A and B	B and B
probability	$q^2 + pqF$	$2pq(1-F)$	$p^2 + pqF$

x_i = SNP dosage for minor allele B, (0, 1), mean of minor allele x_i is p_0

The covariance for sampled gametes x_j and x_k is:

$$\mathrm{E}\left[\left(x_j - p\right)\left(x_k - p\right)\right] = \mathrm{E}\left[x_j x_k\right] - \mathrm{E}\left[x_j\right]\mathrm{E}\left[x_k\right] = \mathrm{E}\left[x_j x_k\right] - p^2$$

$$= \left[\left(p^2 + pqF\right)\times 1\times 1\right] + \left[2pq(1-F)\times 0\times 1\right] + \left[\left(q^2 + qpF\right)\times 0\times 0\right] - p^2$$

$$= pqF$$

Estimate relationships from markers

2. Less well defined, less recent base

What fraction of the gametes are IBD (G)?

SNP genotypes	A and A	A and B	B and B
probability	$q^2 + pqF$	$2pq(1-F)$	$p^2 + pqF$

x_i = SNP dosage for minor allele B, (0, 1), mean of minor allele x_i is p_0

The covariance for sampled gametes x_j and x_k is:

$$\mathrm{E}\left[\left(x_j - p\right)\left(x_k - p\right)\right] = pqF$$

An estimate of relationship (correlation) between two randomly sampled gametes is then:

$$\hat{F}_{jk} = \frac{\left(x_j - p\right)\left(x_k - p\right)}{p(1-p)}$$

Estimate relationships from markers

2. Less well defined, less recent base

For *M* SNPs with allele frequencies p_i the estimate of $\hat{F}_{jk}$ can differ depending upon how we want to weight the contribution of each locus.

$$\hat{G}_{jk} = (\frac{1}{M}) \sum_{i=1}^{M} \frac{(x_{ij} - p_i)(x_{ik} - p_i)}{p_i(1 - p_i)}$$

$$z_{ij} = x_{ij} - p_i$$

$$w_{ij} = (x_{ij} - p_i)/\sqrt{p_i(1 - p_i)} = z_{ij}/\sigma_{x_i}$$

$$\hat{G}_{jk} = (\frac{1}{M}) \sum_{i=1}^{M} \frac{(z_{ij} z_{ik})}{p_i(1 - p_i)} = (\frac{1}{M}) \sum_{i=1}^{M} (w_{ij} w_{ik})$$

Estimate relationships from markers

2. Less well defined, less recent base

$$\hat{G}_{jk} = \left(\frac{1}{M}\right) \sum_{i=1}^{M} \frac{\left(z_{ij} z_{ik}\right)}{p_i(1-p_i)} = \left(\frac{1}{M}\right) \sum_{i=1}^{M} \left(w_{ij} w_{ik}\right)$$

score for pairs of gametes from population in H-W

p = 0.1, q = 0.9, A = 0, B = 1

	0 (0.9)	1 (0.1)
0 (0.9)	0.11	-1
1 (0.1)	-1	9

[(1 – 0.1)(1 – 0.1)] / 0.1(0.9) = 9

[(0 – 0.1)(0 – 0.1)] / 0.1(0.9) = 0.11

[(0 – 0.1)(1 – 0.1)] / 0.1(0.9) = -1

Estimate relationships from markers

2. Less well defined, less recent base

score for pairs of gametes from population in H-W

p = 0.1, q = 0.9, A = 0, B = 1

	0 (0.9)	1 (0.1)
0 (0.9)	0.11	-1
1 (0.1)	-1	9

[(1 – 0.1)(1 – 0.1)] / 0.1(0.9) = 9

[(0 – 0.1)(0 – 0.1)] / 0.1(0.9) = 0.11

[(0 – 0.1)(1 – 0.1)] / 0.1(0.9) = -1

mean G for the SNP:

$= q^2\ G_{1,1} + 2pqG_{1,0} + p^2G_{0,0}$

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

Estimate relationships from markers

2. Less well defined, less recent base

score for pairs of gametes from population in H-W

p = 0.1, q = 0.9

	0 (0.9)	1 (0.1)
0 (0.9)	0.11	-1
1 (0.1)	-1	9

p = 0.4, q = 0.6

	0 (0.6)	1 (0.4)
0 (0.6)	0.67	-1
1 (0.4)	-1	1.5

more weight is given to rare alleles as 1/p(1-p) is larger for small p

Estimate relationships from markers

$\hat{G}_{jk}$ = mean of *s* across loci

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.

Fewer possibilities sampling B if rare:

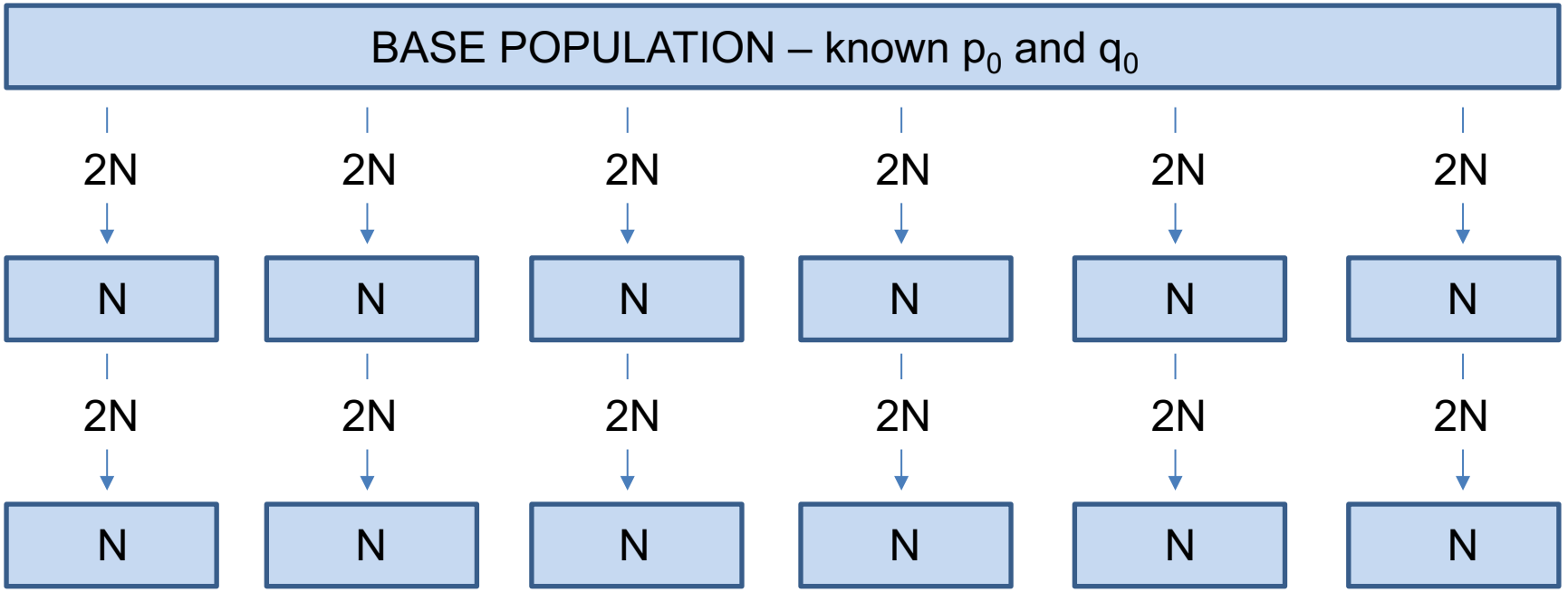
AAAAAAAAAAAAAABB ..so more likely to be IBD

Gives minimum error variance of relationship under some conditions.

Estimate relationships from markers

2. Less well defined, less recent base

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



Repeated sampling leads to drift from base population p_0

Estimate relationships from markers

2. Less well defined, less recent base

score for pairs of gametes from population in H-W but after allele frequency drift to p = 0.2, q = 0.8

	0 (0.9)	1 (0.1)
0 (0.9)	0.11	-1
1 (0.1)	-1	9

[(1 – 0.1)(1 – 0.1)] / 0.1(0.9) = 9

[(0 – 0.1)(0 – 0.1)] / 0.1(0.9) = 0.11

[(0 – 0.1)(1 – 0.1)] / 0.1(0.9) = -1

mean G for the SNP:

$= q^2\ G_{1,1} + 2pqG_{1,0} + p^2G_{0,0}$

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

mean G has shifted relative to the base population

Estimate relationships from markers

2. Current population as the base

Mean $\hat{G}_{jk}$ may be larger than 0 due to drift when estimated with respect to a base population

In a random sample from population we don't know allele frequency in the base.

Use the current population as the base and take sample allele frequencies from the current generation

Estimate relationships from markers

2. Current population as the base

Use sample allele frequencies from the current generation

Problem: some G<0

not probabilities but can still interpret as covariances

If **g** = genetic value, V(**g**) = **G** V_A, where G is calculated using allele frequencies in current population.

Estimate relationships from markers

2. Current population as the base

Weighting 1:

$$\hat{G}_{jk} = \left(\frac{1}{M}\right) \sum_{i=1}^{M} \frac{\left(z_{ij} z_{ik}\right)}{p_i(1-p_i)} = \left(\frac{1}{M}\right) \sum_{i=1}^{M} \left(w_{ij} w_{ik}\right)$$

but there are other weighting schemes one could use…

Estimate relationships from markers

2. Current population as the base

Weighting 1:

$$\hat{G}_{jk} = (\frac{1}{M}) \sum_{i=1}^{M} \frac{(z_{ij} z_{ik})}{p_i(1-p_i)} = (\frac{1}{M}) \sum_{i=1}^{M} (w_{ij} w_{ik})$$

Weighting 2:

$$\hat{G}_{jk} = (\frac{1}{\sum_{i=1}^{M} p_i(1-p_i)}) \sum_{i=1}^{M} (z_{ij} z_{ik})$$

this gives equal weighting to all SNPs

[VanRaden 2008]

Estimate relationships from markers

Score for pairs of gametes from population in H-W

p = 0.2, q = 0.8

z as deviation from E(x) = 0.2

x	p	z	0 (0.8) -0.2	1 (0.2) 0.8
0	(0.8)	-0.2	0.04	-0.16
1	(0.2)	0.8	-0.16	0.64

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

Estimate relationships from markers

p = 0.2, q = 0.8

weighting 1 = $w_{ij}w_{ik}$			weighting 2 = $z_{ij}\ z_{ik}$			
	0 (0.8)	1 (0.2)			0 (0.8)	1 (0.2)
				z	-0.2	0.8
0 (0.8)	0.25	-1	0 (0.8)	-0.2	0.04	-0.16
1 (0.2)	-1	4	1 (0.2)	0.8	-0.16	0.64

scores are proportional to each other, but the proportion will not be constant for different SNPs

Estimate relationships from markers

2. Current population as the base

Weighting 1:

$$\hat{G}_{jk} = (\frac{1}{M}) \sum_{i=1}^{M} \frac{(z_{ij} z_{ik})}{p_i(1-p_i)} = (\frac{1}{M}) \sum_{i=1}^{M} (w_{ij} w_{ik})$$

Weighting 2:

$$\hat{G}_{jk} = (\frac{1}{\sum_{i=1}^{M} p_i(1-p_i)}) \sum_{i=1}^{M} (z_{ij} z_{ik})$$

General case:

$$\hat{G}_{jk} = (\frac{1}{\sum_{i=1}^{M} [p_i(1-p_i)]^{1+S}}) \sum_{i=1}^{M} (z_{ij} z_{ik}) [p_i(1-p_i)]^{S}$$

Estimate relationships from markers

2. Current population as the base

General case:

$$\hat{G}_{jk} = \left(\frac{1}{\sum_{i=1}^{m}[p_i(1-p_i)]^{1+S}}\right)\sum_{i=1}^{m}\left(z_{ij}z_{ik}\right)[p_i(1-p_i)]^{S}$$

when $S = -1$ then this is weighting scheme 1

when $S = 0$ then this becomes weighting scheme 2

$S = -1$ gives more weight to rare variants better captures IBD from IBS

Estimate relationships from markers

2. Current population as the base

Diploid models are straightforward

General case:

$$\hat{G}_{jk} = \left(\frac{1}{\sum_{i=1}^{m}[2p_i(1-p_i)]^{1+S}}\right)\sum_{i=1}^{m}\left(z_{ij}z_{ik}\right)[2p_i(1-p_i)]^{S}$$

$x_i = 0, 1, 2$ for genotypes AA, AB, and BB

$z_{ij} = \left(x_{ij} - 2p_i\right)$

[Speed et al. 2012]

Estimate relationships from markers

2. Current population as the base

Weighting 1:

$$\hat{G}_{jk} = \left(\frac{1}{M}\right)\sum_{i=1}^{M}\frac{\left(z_{ij}z_{ik}\right)}{2p_i(1-2p_i)} = \left(\frac{1}{M}\right)\sum_{i=1}^{M}\left(w_{ij}w_{ik}\right)$$

where $w_{ij} = \left(x_{ij} - 2p_i\right)/\sqrt{2p_i(1-p_i)}$

[Yang et al. 2010,2011]

Weighting 2:

$$\hat{G}_{jk} = \left(\frac{1}{\sum_{i=1}^{M} 2p_i(1-p_i)}\right)\sum_{i=1}^{M}\left(z_{ij}z_{ik}\right)$$

[VanRaden 2008]

Key concepts

- Relationships between gametes or individuals can be estimated from markers in the absence of a clear ‘base population’
- Estimated relationships can be interpreted as the expected genetic covariance between pairs of individuals
- Methods to estimate relatedness from SNP data differ in how they weight rare vs common variants