

1.1.1 GENETIC TERMINOLOGY

Chromosome— long string of double-strand DNA

Cell nucleus — has 46 chromosomes
(22 pairs of autosomes, and 2 sex chromosomes, X,Y)

Locus— position on a chromosome, or DNA at that position, or the piece of DNA coding for a trait.

Allele— type of the DNA at a particular locus on particular chromosome

Genotype— (unordered) pair of alleles at a particular locus in a particular individual.

Homozygote— a genotype with two like alleles.

Heterozygote — a genotype with two unlike alleles.

Phenotype— observable characteristics of an individual

1.1.2 EXAMPLE: ABO blood types

The ABO locus is on chromosome 9

The (main) alleles at the locus are A, B, and O

The 6 genotypes are AA, AO, BB, BO, AB and OO

Homozygotes are *AA, BB, OO*.

Heterozygotes are *AO, BO* and *AB*.

The 4 phenotypes are blood types A, B, AB and O

O allele is recessive to A and to B

A and B are each dominant to O

A and B are codominant

What is a gene??

– the chunk of DNA coding for a functional protein.

Not a locus. Not an allele.

1.1.3 MENDEL'S LAWS (1866)

1. At any given locus, each individual has two genes, one maternal and the other paternal. Each individual segregates a randomly chosen one of its two genes to each offspring, independently to each offspring, independently of gene segregated by the spouse, independently of gene segregated from parent.

2. Independently for different loci.

(Not true; segregation of genes at loci on the same chromosome are dependent)

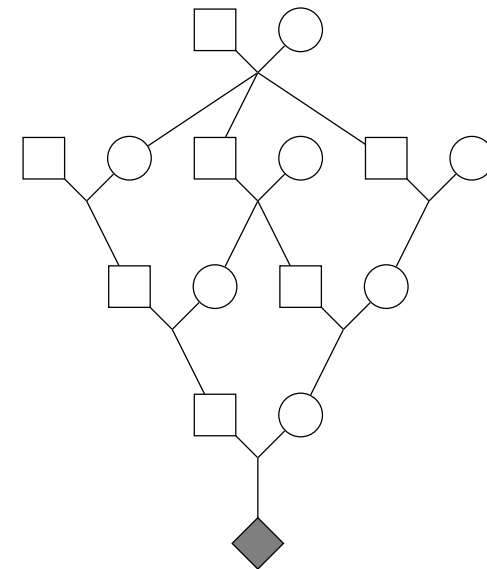
Mendel's first law says all meioses are independent.

Meiosis is the biological process of offspring gamete formation.

The total number of meioses, m , is just the total count of parent-offspring transmissions occurring in the data. For every individual with both parents specified there are two meioses, the one from his/her father (the paternal meiosis) and the one from his/her mother (the maternal meiosis). This will be clearer when we start to talk more about pedigree data.

1.2.1 GRAPHICAL REPRESENTATION OF PEDIGREES

- Three graphical representations.
 - The parent-offspring links.
 - The sibship representation.
 - The marriage-node graph.
- Founders and non-founders (no half-founders)
- Gender: male, female, and unknown. (square, circle, diamond)
- Shading or labelling of individuals



1.2.2 SPECIFICATION OF PEDIGREES

- Unique individual identifiers (“names”)
- Parent-offspring trios. (default: ind, dad, mom)
- Specification of founders. (parent “names” =0)
- Gender: male, female, and unknown. (1, 2, 0) or (M, F, U)
- Phenotypic, covariate, and marker data.
- “Chronological” (partial) ordering of pedigrees.

name	dad	mom	sex	other data
101	0	0	1	——-
102	0	0	2	——-
201	101	102	2	——-
204	101	102	1	——-
206	101	102	1	——-
fred	0	0	1	——-
203	0	0	2	
joe	fred	201	1	——-
jane	204	203	2	
dave	204	203	1	
hugh	joe	jane	1	
etc				

1.3.1 A SAMPLE OF GENES

Consider a single genetic locus, with two codominant alleles A and B . Suppose each independent gene has allelic type A with probability q . We say q is the (population) allele frequency of allele A .

For a random sample of n genes from the population: the number of A alleles is $T \sim Bin(n, q)$. The MLE of q is T/n , which is unbiased since $E(T/n) = nq/n = q$.

The variance of the MLE is $q(1 - q)/n$ which is the smallest possible variance for any unbiased estimator.

Since $\Pr(T = t) \propto q^t(1 - q)^{n-t}$ the log-likelihood is

$$\ell = t \log(q) + (n - t) \log(1 - q)$$

So differentiating the log-likelihood

$$\frac{\partial \ell}{\partial q} = \frac{t}{q} - \frac{n-t}{1-q} = \frac{n}{q(1-q)} \left(\frac{t}{n} - q \right)$$

So the MLE of q is t/n . Also

$$\begin{aligned} \frac{\partial^2 \ell(q; t)}{\partial q^2} &= -\frac{t}{q^2} - \frac{n-t}{(1-q)^2} \\ \mathbb{E} \left(-\frac{\partial^2 \ell(q; T)}{\partial q^2} \right) &= \frac{n}{q} + \frac{n}{(1-q)} = \frac{n}{q(1-q)} \end{aligned}$$

So the Fisher information is $n/q(1-q)$ and the (large-sample) variance of the MLE is $q(1-q)/n$. In this example, it is the variance for any sample size. For large n , MLE's are approx unbiased, and have approx the smallest possible variance.

1.3.2 A SAMPLE OF INDIVIDUALS

Suppose we sample n individuals, and that n_1 have genotype AA , n_2 have genotype AB and n_3 have genotype BB . ($n_1 + n_2 + n_3 = n$). Then we have $(2n_1 + n_2)$ genes of allelic type A , in a sample of size $2n$. We can estimate q by $(2n_1 + n_2)/2n$, but properties of the estimator depend on the genotype frequencies.

$$\ell = \log L = n_1 \log(\Pr(AA)) + n_2 \log(\Pr(AB)) + n_3 \log(\Pr(BB))$$

Two extreme cases:

(i) Complete positive dependence;
there are no AB individuals in the population ($n_2 = 0$). The two homologous genes in an individual are of the same allelic type. The estimator is n_1/n and in effect we have a sample of n genes.

(ii) Hardy-Weinberg equilibrium (HWE);
There is independence of the allelic types of the two homologous genes within an individual. So $\Pr(AA) = q^2$, $\Pr(AB) = 2q(1-q)$ and $\Pr(BB) = (1-q)^2$

(iii) See next page for a model of intermediate dependence.

1.3.3 POPULATION STRUCTURE

Suppose populations i , each in HWE, with q_{ij} the freq of allele A_j in population i , and α_i the proportion of population i . So $\Pr(A_j) = q_{\cdot j} = \sum_i \alpha_i q_{ij}$

$$\begin{aligned} \Pr(A_j A_j) - (\Pr(A_j))^2 &= \sum_i \alpha_i q_{ij}^2 - q_{\cdot j}^2 \\ &= \sum_i \alpha_i (q_{ij} - q_{\cdot j})^2 \geq 0 \end{aligned}$$

$$\begin{aligned} \Pr(A_j A_l) - 2\Pr(A_j)\Pr(A_l) &= 2(\sum_i \alpha_i q_{ij} q_{il} - q_{\cdot j} q_{\cdot l}) \\ &= 2\sum_i \alpha_i (q_{ij} - q_{\cdot j})(q_{il} - q_{\cdot l}) \end{aligned}$$

Thus, population subdivision results in homozygote excess relative to HWE. This excess is known as the Wahlund variance.

In total, we therefore have heterozygote deficiency, but NOT necessarily for each heterozygote.

For two alleles, let $q_{i1} = q_i$, $q_{i2} = 1 - q_i$, $q = q_{\cdot}$.

If $\sigma_f^2 = \sum_i \alpha_i (q_i - q)^2$, then the three genotype freqs are $q^2 + \sigma_f^2$, $2q(1 - q) - 2\sigma_f^2$ and $(1 - q)^2 + \sigma_f^2$.

1.3.4 ESTIMATION in the case of HWE

$$\begin{aligned} \ell &= n_1 \log(q^2) + n_2 \log(2q(1 - q)) + n_3 \log((1 - q)^2) \\ &= (2n_1 + n_2) \log(q) + (n_2 + 2n_3) \log(1 - q) \end{aligned}$$

The MLE of \mathbf{q} is $(2n_1 + n_2)/2n$. If $T = 2n_1 + n_2$, $T \sim \text{Bin}(2n, q)$. $\text{var}(T/2n) = q(1 - q)/2n$ — back to binomial sampling.

Note: One generation of random mating establishes HWE, since, by definition, the two genes in an individual are copies of independently sampled parental genes.

1.3.5 CASE OF A RECESSIVE ALLELE A

$t = n_1$ of type AA, and $n - t$ not of type AA.

Assuming HWE, $\Pr(AA) = q^2$, so

$$\ell = t \log(q^2) + (n - t) \log(1 - q^2)$$

Differentiating

$$\begin{aligned} \frac{\partial \ell}{\partial q} &= \frac{2t}{q} - \frac{(n - t)2q}{1 - q^2} \\ &= \frac{2}{q(1 - q^2)}(t - nq^2) \end{aligned}$$

So the MLE of q is $\sqrt{t/n}$. Why should this be expected?

Variance and information:

Now $T \sim \text{Bin}(n, q^2)$, but how can we find the variance of this MLE?

$$\begin{aligned} \frac{\partial^2 \ell}{\partial q^2} &= -\frac{2t}{q^2} - \frac{2(n - t)}{(1 - q^2)} - \frac{(n - t)4q^2}{(1 - q^2)^2} \\ E\left(-\frac{\partial^2 \ell}{\partial q^2}\right) &= 2n + 2n + \frac{4q^2 n}{(1 - q^2)} = \frac{4n}{1 - q^2} \end{aligned}$$

The variance of the MLE of q is approx. $(1 - q^2)/4n$.

Note this is larger than $q(1 - q)/2n$.

We have to make assumptions (HWE)

Variance of the estimator is larger.

We can measure the information lost.

1.3.6 ESTIMATING FROM DATA ON RELATIVES

For simplicity we consider just mother-baby pairs and assume HWE. See next page for tables of conditional and joint probabilities.

$$\begin{aligned}
 \ell &= \sum_{(i,j)} n_{ij} \log \Pr(g_i, g_j) \\
 &= n_{00} \log(q^3) + n_{01} \log(q^2(1-q)) + n_{10} \log(q^2(1-q)) \\
 &\quad + n_{11} \log(q(1-q)) + n_{12} \log(q(1-q)^2) \\
 &\quad + n_{21} \log(q(1-q)^2) + n_{22} \log((1-q)^3) \\
 &= (3n_{00} + 2(n_{01} + n_{10}) + n_{11} + n_{12} + n_{21}) \log q + \\
 &\quad (3n_{22} + 2(n_{21} + n_{12}) + n_{11} + n_{10} + n_{01}) \log(1-q) \\
 &= m_A \log q + m_B \log(1-q)
 \end{aligned}$$

The MLE of q is $m_A/(m_A+m_B)$, where $(m_A+m_B) = 3n - n_{11}$ and $m_A = (3n_{00} + 2(n_{01} + n_{10}) + n_{11} + n_{12} + n_{21})$.

parent genotype	probability	Pr(child parent)		
		AA	AB	BB
AA	q^2	q	$1-q$	0
AB	$2q(1-q)$	$\frac{1}{2}q$	$\frac{1}{2}$	$\frac{1}{2}(1-q)$
BB	$(1-q)^2$	0	q	$(1-q)$

parent geno.	Pr(parent, child).			Data count		
	AA	AB	BB	AA	AB	BB
AA	q^3	$q^2(1-q)$	0	n_{00}	n_{01}	0
AB	$q^2(1-q)$	$q(1-q)$	$q(1-q)^2$	n_{10}	n_{11}	n_{12}
BB	0	$q(1-q)^2$	$(1-q)^3$	0	n_{21}	n_{22}

1.3.7 ALTERNATIVES TO THE MLE

The MLE is “best”, but there are simpler estimators that are not so bad.

(a) Use only founders (the moms):

estimate q by $(2n_{AA} + n_{AB})/2n$ where n_{AA} is number of AA moms, and n_{AB} is number of AB moms. ($n_{AA} = n_{00} + n_{01}$).

(b) Use everyone, disregarding relationship:

estimate q by $(2m_{AA} + m_{AB})/4n$, where m_{AA} is total number of AA individuals, and m_{AB} is total number of AB individuals. ($m_{AA} = 2n_{00} + n_{01} + n_{10}$).

These are both unbiased estimators, but asymptotically the MLE has smaller variance.

1.4.1 TESTING Hardy-Weinberg PROPORTIONS (HWE)

Consider the following three samples, each of 100 individuals. Each has 120 A alleles, so the MLE of q is 0.6, but different genotypic counts n_c in genotype class c .

n	AA	AB	BB	$\hat{\ell}$	\hat{q}	$\tilde{\ell}$	$2(\hat{\ell} - \tilde{\ell})$
100	36	48	16	-101.33	0.6	-101.33	0
100	30	60	10	-89.79	0.6	-93.01	6.5
100	45	30	25	-106.71	0.6	-113.81	14.2

With probability p_c for class c ,

$$\ell = \text{const} + \sum_c n_c \log(p_c) \quad \text{with} \quad \sum_c p_c = 1$$

With no constraints, MLE of p_c is n_c/n , and maximized value of the log-likelihood is

$$\hat{\ell} = \sum_c n_c \log(n_c/n) = \sum_c n_c \log(n_c) - n \log(n)$$

Assuming HWE,

$$\begin{aligned} \tilde{p}_c &= (\hat{q}^2, 2\hat{q}(1-\hat{q}), (1-\hat{q})^2) = (0.36, 0.48, 0.16) \\ \tilde{\ell} &= \sum_c n_c \log(\tilde{p}_c) \end{aligned}$$

Now, if HWE is true, $2 \log \Lambda = 2(\hat{\ell} - \tilde{\ell})$ is approximately χ_1^2 , and larger otherwise. In our three examples, the values are 0, 6.5 and 14.2. What do we conclude?

1.4.2 TESTING THE ABO BLOOD GROUP MODEL

	factor freq.		phenotype frequencies			
	<i>A</i>	<i>B</i>	<i>A</i>	<i>B</i>	<i>AB</i>	<i>O</i>
Data			0.422	0.206	0.078	0.294
<i>H</i> ₁ theory	<i>p</i>	<i>q</i>	<i>p</i> (1 - <i>q</i>)	(1 - <i>p</i>) <i>q</i>	<i>pq</i>	(1 - <i>p</i>)(1 - <i>q</i>)
<i>H</i> ₁ fitted	0.500	0.284	0.358	0.142	0.142	0.358
<i>H</i> ₂ theory	<i>p</i>	<i>q</i>	<i>p</i> ² + 2 <i>pr</i>	<i>q</i> ² + 2 <i>qr</i>	2 <i>pq</i>	<i>r</i> ²
<i>H</i> ₂ fitted	0.295	0.155	0.411	0.194	0.091	0.303

Bernstein reported *ABO* blood types on a sample of 502 individuals: 42.2% type *A*, 20.6% type *B*, 7.8% type *AB* and 29.4% type *O*. (Did he drop 2 individuals?)

For the general model

$$\begin{aligned}\hat{\ell} &= 502(.422 \log .422 + .206 \log .206 + .078 \log .078 + .294 \log .294) \\ &= -626.71\end{aligned}$$

1.4.3 TESTING GOODNESS OF FIT

*H*₁: *A* and *B* are independently inherited factors
Frequency of individuals having the factor *A* is 0.500 and of *B* is 0.284. Independence of the factors would give an *AB* frequency of 0.500 × 0.284 = 0.142 much larger than the 0.078 observed.

Under *H*₁ the estimated frequencies are as shown in Table, and the log-likelihood is

$$\begin{aligned}\ell_1 &= 502(.422 \log .358 + .206 \log .142 + .078 \log .142 + .294 \log .358) \\ &= -647.50\end{aligned}$$

Twice the log-likelihood difference is 41.58, and would be the value of a χ^2_1 random variable if *H*₁ were true. Clearly, *H*₁ is rejected.

Testing another alternative: H_2

Under H_2 : A and B are the two non-null alleles of a single system. Assuming HWE, if the three alleles A , B and O have frequencies p , q and r ($p+q+r=1$), then the frequencies of the four blood types are p^2+2pr , q^2+2qr , $2pq$ and r^2 .

Bernstein pointed out that the sum of the A and O blood type frequencies is $(p+r)^2$, or one minus the square root of this frequency is $(1-p-r)=q$. Similarly one minus the square root of the sum of the B and O blood type frequencies is p , and the square root of the O blood type frequency is r . The sum of these three numbers should be one. For his data

$$(1 - \sqrt{0.422 + 0.294}) + (1 - \sqrt{0.206 + 0.294}) + \sqrt{0.294} = 0.99$$

which is close to one, suggesting a good fit.

Likelihood ratio test for H_2 :

More formally, we may perform a likelihood ratio test. Finding the MLEs of the parameters p , q and r is not simple; in fact, we shall see later that these MLEs are $\hat{p} = 0.2945$ and $\hat{q} = 0.1547$, with the resulting fitted frequencies given in the table. The log-likelihood is

$$\begin{aligned} \ell_2 = 502(&.422 \log .4114 + .206 \log .1942 + .078 \log .0911 \\ &+ .294 \log .3033) = -627.52 \end{aligned}$$

Twice the log-likelihood difference between this and the general alternative is now only 1.62. Again, this is the value of a χ_1^2 random variable if H_2 is true: H_2 is not rejected.

1.5.1 GENE COUNTING: CASE OF RECESSIVE TRAIT

current q	current $2q/(1+q)$	recessive phenotype $t_1 = 36$ AA	dominant phenotype $t_2 + t_3 = 64$ AB BB		new $q =$ $(2t_1 + t_2)/2n$
0.5	0.667	36	42.67	21.33	0.573
0.573	0.729	36	46.64	17.36	0.593
0.593	0.745	36	47.66	16.34	0.598
0.598	0.749	36	47.91	16.09	0.600
0.600	0.750	36	48.00	16.00	0.600

The three genotypes are AA , AB and BB , with counts say t_i , ($i = 1, 2, 3$). Now, $n_1 = t_1$, but the counts of AB and BB are unobservable since B is dominant to A .

The counting algorithm:

If counts, t_2 and t_3 , were known, then the number of A alleles is $m_1 = 2t_1 + t_2$, and the MLE of q would be $(2t_1 + t_2)/2n$. Further,

$$\Pr(AB \mid AB \text{ or } BB) = \frac{2q(1-q)}{1-q^2} = \frac{2q}{1+q}$$

so

$$E_q(t_2 \mid n_2 = t_2 + t_3 = 64) = 64 \frac{2q}{1+q}.$$

The EM-algorithm implements the sequence of iterates shown. Starting from an arbitrary initial value $q = 0.5$, the proportion $2q/(1+q)$ is computed, and the 64 individuals of dominant phenotype divided into the expected numbers t_2 and t_3 that are AB and BB , respectively (E-step). Then a new value of q is estimated as $(2t_1 + t_2)/2n$ (M-step).

1.5.2 EM ALGORITHM FOR MULTINOMIAL DATA

In latent variable problems, suppose the actual data are \mathbf{Y} , and the ideal data that would make the problem easy are (\mathbf{Y}, \mathbf{X}) . The complete-data log-likelihood is

$$\ell^* = \log \Pr((\mathbf{Y}, \mathbf{X}) = (\mathbf{y}, \mathbf{x})).$$

The actual log-likelihood to be maximized is

$$\ell = \log \Pr(\mathbf{Y} = \mathbf{y}) = \log \left(\sum_{\mathbf{x}} \Pr((\mathbf{Y}, \mathbf{X}) = (\mathbf{y}, \mathbf{x})) \right).$$

E-step (expectation):

At the current estimate θ^* compute ECDLL

$$H_{\mathbf{y}}(\theta; \theta^*) = E_{\theta^*}(\log P_{\theta}(\mathbf{X}, \mathbf{Y}) \mid \mathbf{Y} = \mathbf{y})$$

M-step (maximization):

Maximize $H_{\mathbf{y}}(\theta; \theta^*)$ w.r.t. θ to obtain a new estimate $\tilde{\theta}$.

Theoretical result: $\ell(\tilde{\theta}) \geq \ell(\theta^*)$.

Thus the EM algorithm for finding MLEs alternates E-steps and M-steps. The likelihood is non-decreasing over the process. Where the likelihood surface is unimodal, convergence to the MLE is assured, although it may be slow. Where computable, evaluate the (log)-likelihood to assess convergence.

For multinomial data, let n_c be actual data-counts, and m_{c^*} complete-data counts for idealized data. So $\ell^* = \sum_{c^*} m_{c^*} \log(p_{c^*})$, and finding the ECDLL just means finding

$$E(m_{c^*} | n_c) = n_c \Pr(c^* | c, \theta^*) = n_c \frac{p_{c^*}(\theta^*)}{\sum_{c^* \rightarrow c} p_{c^*}(\theta^*)}$$

1.5.3 The ABO log-likelihood

- $\ell = \sum_{\text{obs counts}} \#(Y_i) \log P(Y_i)$
- $\ell^* = \sum_{\text{all counts}} \#(X_i) \log P(X_i)$

Data $Y = (N_A, N_B, N_{AB}, N_O)$

Complete-data $X = (n_{AA}, n_{AO}, \dots)$

- Do not confuse ℓ and ℓ^*

ℓ^* is just a tool that lets us maximize ℓ .

- Compute $E(\ell^* | Y)$ – in the multinomial case this just involves imputing the “hidden” counts – but only because ℓ^* is a linear function of these counts.

1.5.4 ESTIMATION OF ABO ALLELE FREQUENCIES

For the MLE of ABO blood group allele frequencies, the EM-algorithm is one of the easiest ways to find the MLEs. (see table, next page)

E-step: partition the A phenotypes into expected counts of AA and AO genotypes, and similarly B into BB and BO:

$$\Pr(AO \mid \text{type } A) = \frac{2pr}{p^2 + 2pr} = \frac{2r}{p + 2r}$$

$$\Pr(BO \mid \text{type } B) = \frac{2qr}{q^2 + 2qr} = \frac{2r}{q + 2r}$$

M-step: Then $\tilde{p} = \Pr(AA) + (\Pr(AO) + \Pr(AB))/2$,
and $\tilde{q} = \Pr(BB) + (\Pr(BO) + \Pr(AB))/2$.

Note \tilde{p} does not change monotonely, but $\tilde{\ell}$ does.
($\tilde{\ell}$ is the current value of ℓ , not of ℓ^* .)

current values				phenotype A		phenotype B		...
p	q	$\frac{2r}{p+2r}$	$\frac{2r}{q+2r}$	Pr(A) = 0.422	Pr(B) = 0.206
				AA	AO	BB	BO	...
0.3	0.3	0.73	0.73	0.115	0.307	0.056	0.150	...
0.308	0.170	0.77	0.86	0.096	0.326	0.029	0.177	...
0.298	0.156	0.79	0.87	0.091	0.331	0.026	0.180	...
0.295	0.155	0.79	0.88	0.089	0.333	0.025	0.181	...

...	phen AB	phen O	new values		$\tilde{\ell}$
...	Pr(AB) =	Pr(OO) =	\tilde{p}	\tilde{q}	
...	0.078	0.294			-687.12
...	0.078	0.294	0.308	0.170	-629.00
...	0.078	0.294	0.298	0.156	-627.57
...	0.078	0.294	0.295	0.155	-627.53
...	0.078	0.294	0.295	0.155	-627.52

1.6 HAPLOTYPES AND ALLELIC ASSOCIATION

1.6.1 ESTIMATING PHASE: 2 loci

Consider diallelic loci (e.g. SNPs):

Label alleles 0 and 1.

At two loci there are 4 haplotypes: 00, 01, 10, 11.

There are 10 phased two-locus genotypes.

eg; 10/00, 11/00, 10/01

Observable (unphased) genotype is a pair of pairs:

There are $9 = 3 \times 3$ observable two-locus genotypes

e.g. (00,00), (10, 11),

Only the double-heterozygotes (10,10) is ambiguous:

can be 11/00 or 10/01.

For other genotype pairs we just count.

Suppose the current estimates of haplotype frequencies

are $q_{00}, q_{01}, q_{10}, q_{11}$.

Suppose there are H double-heterozygotes.

Then $E(\#(11/00) | H) = Hq_{11}q_{00}/(q_{11}q_{00} + q_{10}q_{01})$

So EM is easily implemented. For two (or a very few)

loci it works well.

1.6.2 ESTIMATING PHASE AND HAPLOTYPE FREQUENCIES

Consider diallelic loci (e.g. SNPs):
Label alleles 0 and 1.

Then the genotypes are a set of pairs e.g. 00, 10, 11, 10, 10, and haplotypes a string such as 01010. Determining phase is determining which of the 4 possibilities 01111 and 00100, 01110 and 00101, 01101 and 00110, 01100 and 00111 holds.

For convenience we may write an unphased genotype as (01111,00100) and the phased version as (01111/00100).

For large samples and/or small numbers of SNPs we may use EM algorithm to estimate haplotype frequencies, and this also provides probabilities of phasings, given the estimated frequencies. However, for large numbers of SNPs this does not work well: many sample haplotype freqs are 0, and likelihood surface is high-dimensional and multimodal.

1.6.3 HAPLOTYPING: TWO (OTHER) ALGORITHMS

Clarke's algorithm: note where individuals are homozygous, haplotyping is trivial. Also trivial if heterozygous at just 1 locus.

Use individuals heterozygous at at most 1 locus to identify haplotypes that must be present. Assuming these, see which other individuals can be explained by one of these haplotypes, plus a new one – add these new ones to the collection, and continue for as long as possible.

Problems: May not be able to start. May not be able to finish. Final guess may depend on order one adds haplotypes to the pool.

Stephens' algorithm (PHASE): use a model that summarizes similarities of haplotypes in a population – the idea is that haplotypes should look like each other "in chunks". Use Monte Carlo to simulate alternative phasings under the model. Produces "probable phasings" with estimates probabilities. (Now also FastPHASE.)

**1.7 MAINTAINING VARIATION:
1.7.1 MUTATION AND SELECTION**

Mutation provides new variation.

Directional selection removes variation.

In equilibrium, “loss” = “gain”.

Hence, indirect estimates of mutation rates.

For example, recessive with selection coefficient s :
we lose $2A$ alleles, with prob, s , for each AA individual.
we gain μ A alleles, in each of $2N$ meioses (approx.)
So $Ns2q^2 = 2N\mu$, or $\mu = sq^2$.

1.7.2 RANDOM GENETIC DRIFT

Real populations are finite (and have structure, and history, ...). Let $X(t)$ be number of A alleles at time t in popn size $2N$ genes. Suppose $(X(t)|X(t-1))$ is $Bin(2N, X(t-1)/2N)$ (Wright-Fisher model). Then

$$\begin{aligned} E(X(t)) &= E(E(X(t)|X(t-1))) = E(2N \frac{X(t-1)}{2N}) \\ &= E(X(t-1)) = \dots = X(0) \end{aligned}$$

Hence $E(X(\infty)) = X(0)$ so $\Pr(X(\infty) = 2N) = X(0)/2N$.

HOMOZYGOSITY

$$\begin{aligned} \text{Note } E(X^2) &= \text{var}(X) + (E(X))^2 \\ \text{So } E(X(t)^2) &= E(E(X(t)^2|X(t-1))) \\ &= E(\text{var}(X(t)|X(t-1)) + E(X(t-1))^2) \end{aligned}$$

Homozygosity increases relative to time 0, because the allele frequency has increasing chance of being closer to 0 or 1, but population is still in HWE.

POPULATIONS DIVERGE

Let $V_t = \text{var}(X(t))$ and $X_1(t)$ and $X_2(t)$ counts in two indep popns with same $X(0)$

$$\begin{aligned} E((X_1(t) - X_2(t))^2) &= E(X_1^2) - 2E(X_1X_2) + E(X_2^2) \\ &= (V_t + X(0)^2) - 2X(0)^2 + (V_t + X(0)^2) \\ &= 2V_t \approx (4Nt)(X(0)/2N)(1 - X(0)/2N) \end{aligned}$$