第3章
有限大小的随机交配群体

王建康
中国农业科学院作物科学研究所
wangjiankang@caas.cn
http://www.isbreeding.net

随机漂移引起的群体分散过程

- 如果从随机交配大群体（称其为基础群体）中抽取一个较小的样本，即使不存在突变、迁移和选择等因素，繁衍得到的新群体与基础群体的基因频率也可能存在差异。也就是说，随机抽样也会引起群体结构的改变，但改变的方向往往是不可预测的，改变的幅度是可以预测的，这一过程称为分散过程（disperse process）。
- 分散过程造成基因频率随机波动的现象，称为随机漂移（random drift）。随机漂移是不可逆的，它同样决定着后代群体的遗传结构。因此，随机漂移有时也称为遗传飘移（genetic drift）。

分散过程的四大特点

- 随机飘变 (random drift)
- 亚群体间分化 (differentiation of sub-populations)
- 亚群体内遗传同质 (genetic uniformity within sub-populations)
- 纯合基因型频率的增加 (overall increase of homozygotes as a consequence of the dispersion of gene frequencies)

本章的主要内容

§ 3.1 离散型随机变量及其遗传应用

§ 3.2 近交和近交系数

§ 3.3 理想有限大小群体的遗传构成

§ 3.4 自然群体的分化

§3.1 离散型随机变量及其遗传应用

- §3.1.1 离散型随机变量
- §3.1.2 二项分布和多项分布
- §3.1.3 精确检验
- §3.1.4 泊松分布

常量和变量

- 自然界有常量和变量之分。常量又称常数（constant number），是一个固定不变的数值，如圆周率π=3.14159…，自然对数的底e=2.71828…。
- 根据是否可控，变量又分为固定变量（在不易混淆的情况下简称为变量）和随机变量。
- 例如，作物生产中的播期、播量和施肥量等等，它们虽然是可变的，但人们能够根据需要控制它们的取值，因此是固定变量（fixed variable）。
- 一个作物品种在一定播期、播量和施肥量等栽培方式下的产量也是可变的，它的取值事先是不能控制的，因此是随机变量（random variable）。

随机变量的分类

- 遗传研究中有很多随机变量，如一个群体的基因和基因型频率、两个座位间交换次数和重组率、各种数量性状的表型观测值等等。用X表示随机变量，X取不同数值表示不同的概率事件。虽然对于X取何值是事先不知道的，但是随机变量一般都服从一定的分布，不同取值都有一定的概率。
- 根据取值情况，又可以把随机变量分为离散和连续两类。如果取值是有限个或可数个（可以按一定顺序一一列举出来），则称为离散型随机变量（discrete random variable）。
- 如果取值为无穷不可数个，则称为连续型随机变量（continuous random variable）。

离散型随机变量

- 设离散型随机变量X的可能取值是1, 2, …, k…。为了完整地描述随机变量X，只知道它的可能取值是远远不够的，更重要的是要知道各种可能取值的概率。下面给出的概率也称为随机变量X的概率分布，它清楚而完整地表示了X各种可能取值概率的大小。

$$P(X = x_k) = p_k \quad (k=1, 2\ldots)$$

离散型随机变量概率分布的性质

- 离散型随机变量X有两个基本性质，一个说明任何概率都具有非负性，另一个说明随机变量X遍历所有可能的取值时，得到的是一个概率为1的必然事件。

$$p_k \geq 0 \quad (k=1, 2\ldots)$$

$$\sum_{k=1}^{\infty} p_k = 1$$

随机变量的数字特征

- 随机变量的概率分布完整地描述了随机变量的取值规律，而且往往只依赖于少数的几个参数，这些参数称为随机变量的数字特征。
- 确定了随机变量的数字特征，也就确定了随机变量的分布。
- 均值和方差是两个最重要的数字特征。

离散随机变量的均值和方差

- 均值是随机变量各种取值按概率的加权平均，用来衡量随机变量的平均取值。
- 方差是随机变量各种取值与均值的离差平方和按概率的加权平均，用来衡量随机变量取值的集中程度。
- 离散随机变量的均值和方差如下：

$$E(X)=\sum_{k=1}^{\infty}p_k k$$

$$V(X)=\sum_{k=1}^{\infty}p_k[k-E(X)]^2=\sum_{k=1}^{\infty}p_k k^2-[E(X)]^2$$

Bernoulli分布（两点分布）

- 概率论把一次试验中，事件A发生概率为p（$0<p<1$）、不发生概率为q（$q=1-p$）的分布，称为两点分布或Bernoulli分布（Bernoulli distribution）。
- 如用X表示Bernoulli分布中事件A发生的次数，则X只有0和1两种取值，取0的概率为q，取1的概率为p。
- 那么，n次独立的Bernoulli分布试验中，事件A发生的次数k就服从的分布称为二项分布。

二项分布

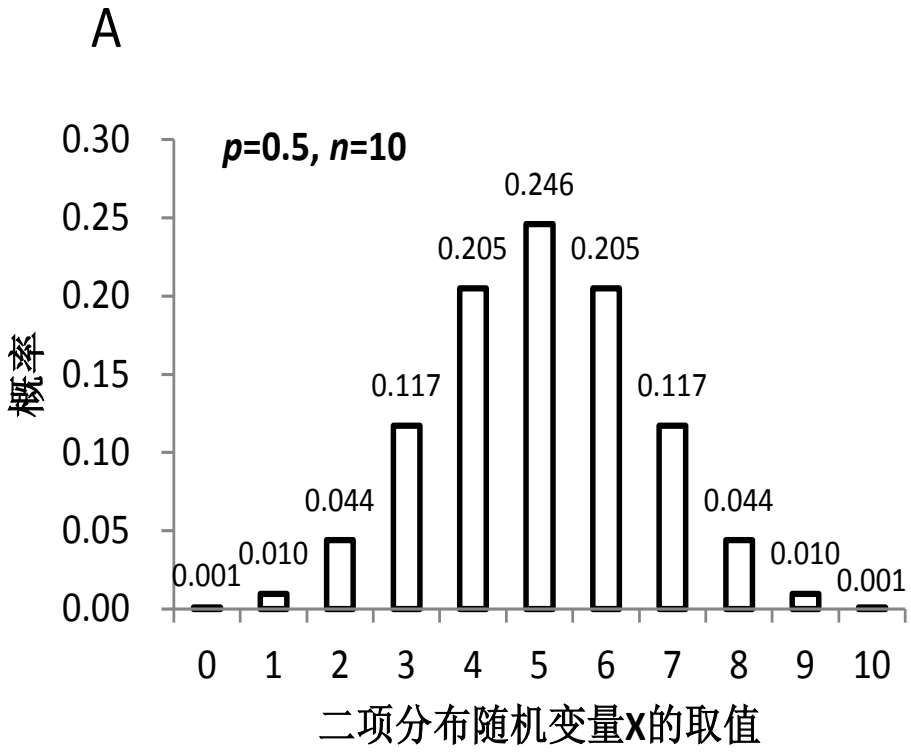
- 如果随机变量X的概率分布由下面的公式给出，则称X服从二项分布（binomial distribution），用符号 $B(n, p)$ 表示。

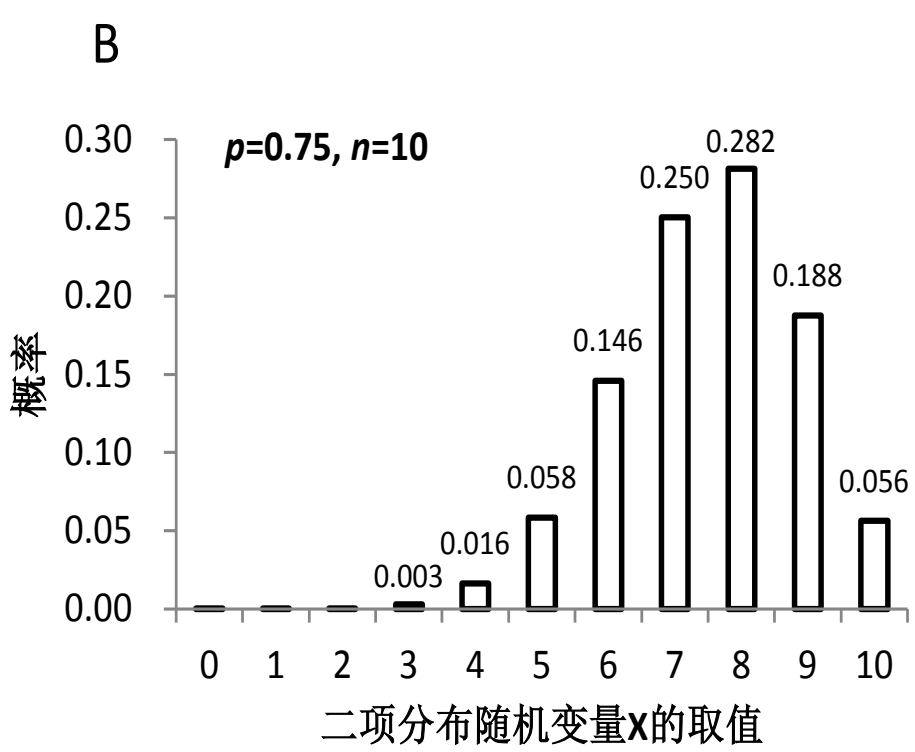
$$P(X=k)=C_n^k p^k q^{n-k}=\frac{n!}{k!(n-k)!}p^k q^{n-k}$$

其中 k=0、1、…、n；0<p<1；q=1-p

回交（A）和F_2（B）群体中，红花表型个体数的概率分布

(两个群体的群体大小均为10)





二项分布的均值和方差

$$E(X) = np$$

$$V(X) = npq$$，其中q=1-p

二项分布的参数估计

- 一个二项分布的n次试验中，观察到事件A发生了k次。参数p无偏估计及其方差分别为

$$\hat{p} = \frac{k}{n} \qquad V(\hat{p}) = V(\frac{k}{n}) = \frac{1}{n^2} V(k) = \frac{pq}{n}$$

- 传统频率派统计学（frequentist statistics）认为，p是一个未知的但可以通过样本观测值进行估计的参数，其真实值可能永远不知道。上式给出的是未知参数p的一个估计值，不同样本观测值会得到不同的估计值，因此分布参数的估计是随机变量。

多项分布

- 多项分布（multinomial distribution）是二项分布的推广。在一次试验中，有多于两种的取值，如m个。各种可能取值的概率用$p_1, p_2, \ldots, p_m$表示，用$X_1, X_2, \ldots, X_m$表示不同取值的随机变量。
- 一次试验中，$X_i=1$（$i=1, 2, \ldots, m$）的概率为p_i，$X_1, X_2, \ldots, X_m$满足和为1的限制条件。在n次独立试验中，$X_1, X_2, \ldots, X_m$（满足和为n的限制条件）这些随机变量就服从一个m项分布，用符号 $B(n, p_1, p_2, \ldots, p_m)$ 表示。

多项分布的取值概率、均值和方差

$$P(X_1 = k_1, \cdots, X_m = k_m \mid k_1 + \cdots + k_m = n) = \frac{n!}{k_1! \cdots k_m!} p_1^{k_1} \cdots p_m^{k_m}$$

$$E(X_i) = np_i \text{ , } i\text{=1, 2, …, } m$$

$$V(X_i) = np_i(1 - p_i) \text{ , } i\text{=1, 2, …, } m$$

多项分布的性质

- 从前面的分布概率公式，可以看出多项分布的三个重要性质。
- 首先，二项分布$B(n, p)$ 可以看作$m=2$的多项分布$B(n, p_1=p, p_2=1-p)$ 。因此，二项分布是多项分布的一种特殊情况。
- 其次，多项分布的任何一个随机变量X_i可以看作二项分布$B(n, p_i)$ 。
- 最后，多项分布的任何两个随机变量X_i和X_j之和，可以看作二项分布$B(n, p_i+p_j)$ 。

多项分布的协方差和相关

- 利用前面的性质，可以利用二项分布的均值和方差公式，计算多项分布的协方差和相关系数。

$$\mathrm{Cov}(X_i, X_j) = -np_i p_j$$

$$\mathrm{Corr}(X_i, X_j) = -\frac{p_i p_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}}$$

Fisher精确检验

- 双亲群体的孟德尔分离比检验、以及随机交配群体的HW平衡检验过程中，当样本量较小时，χ^2检验的效果可能不是很好。一种办法是对χ^2统计量进行矫正。
- 还有一种办法是在固定样本量的情况下，直接计算各种可能取值的概率。然后，把各种可能取值的概率按照从小到大排序，从中计算累计概率并进行统计推断。这种方法称为精确检验（exact test），由Fisher（1935）首先提出。因此，也称为Fisher精确检验。

例子

- 以10个F_2单株观察到6个红花、4个白花的数据为例，现在需要检验两种表型是否服从3:1的期望孟德尔分离比。

在3:1期望分离比下，所有红花和白花可能取值的分布概率（从小到大排列）

红花个体	白花个体	概率	累积概率
0	10	0.0000	0.0000
1	9	0.0000	0.0000
2	8	0.0004	0.0004
3	7	0.0031	0.0035
4	6	0.0162	0.0197
10	0	0.0563	0.0760
5	5	0.0584	0.1344
6	4	0.1460	0.2804
9	1	0.1877	0.4682
7	3	0.2503	0.7184
8	2	0.2816	1.0000

精确检验的检验方法

- 在零假设成立的前提下，计算变量的各种取值概率，按照从小到大排序，并计算累积概率。根据累积概率就可以进行统计推断。
- 如果出现4个红花和6个白花，把前表给出的累计概率0.0197视为显著性概率，由于它低于0.05的显著性标准，这时就说观测群体的两种表型不符合3:1分离比。这一推断犯第一类错误的概率将低于0.0197。
- 如果出现6个红花和4个白花，其显著性概率为0.2804，高于0.05的显著性标准。这时，就说观测群体的两种表型符合3:1分离比，观察频率与期望频率之间的差异可能完全是由抽样误差引起的。

HW平衡的精确检验方法

- 当样本量较小时，随机交配群体的HW平衡也可利用精确检验，只不过要比双亲群体已知基因频率的孟德尔分离比检验复杂些，需要首先从基因型观测值估计基因频率，精确检验时需要同时考虑基因型样本量和等位基因样本量两个限制条件。
- 例如，基因型的总样本量为N，则配子的总样本量为$2N$。考虑一个座位上的两个等位基因A和a，三种基因型AA、Aa、aa的观测值分别用N_{11}、N_{12}、N_{22}表示，$N_{11}+N_{12}+N_{22}=N$。因此，等位基因A和a的个数分别为$n_1=2N_{11}+N_{12}$和$n_2=2N_{22}+N_{12}$。显然，$n_1+n_2=2N$。

HW平衡的精确检验方法

- 等位基因个数n_1、n_2的分布概率

$$P(n_1,n_2)=\frac{(2N)!}{n_1!n_2!}(p)^{n_1}(q)^{n_2}$$

- 基因型个数X_{11}、X_{12}、X_{22}的分布概率

$$P(X_{11},X_{12},X_{22})=\frac{N!}{X_{11}!X_{12}!X_{22}!}(p^2)^{X_{11}}(2pq)^{X_{12}}(q^2)^{X_{22}}$$

- 样本总量等于N和等位基因样本量等于$2N$两个限制条件下，各种可能基因型个数的分布概率

$$P(X_{11},X_{12},X_{22}\,|\,n_1,n_2)=\frac{n_1!n_2!N!}{X_{11}!X_{12}!X_{22}!(2N)!}\times 2^{X_{12}}$$

基因型个数的取值范围

- 在实际数据中，可以利用下面的方法确定三种基因型AA、Aa、aa个数的可能取值。
- 首先，利用观测值N_{11}、N_{12}、N_{22}计算等位基因的个数$n_1=N_{11}+2N_{12}$和$n_2=N_{22}+2N_{12}$，并固定n_1和n_2（相当于得到基因频率的估计值）。
- 三种基因型的取值既要满足和为N这一限制条件，又要满足等位基因A的个数为n_1这一限制条件。因此，如果知道了杂合型Aa的个数X_{12}，就能从等位基因个数计算公式$n_1=2X_{11}+X_{12}$和$n_2=2X_{22}+X_{12}$分别求得基因型AA和aa的个数。

基因型个数的取值范围

- 由于X_{11}和X_{22}不能为负数。从等式$n_1=2X_{11}+X_{12}$和$n_2=2X_{22}+X_{12}$可以看出，杂合型的个数要满足条件X_{12}小于或等于min（n_1, n_2）的条件。又由于$n_1+n_2=2N$，因此n_1和n_2要么同是奇数，要么同是偶数。
- 因此，如果n_1和n_2是奇数，X_{12}的可能取值就是0与min（n_1, n_2）之间的所有奇数；如果n_1和n_2是偶数，X_{12}的可能取值就是0与min（n_1, n_2）之间的所有偶数。

例子

- 例如，一个群体的三种基因型AA、Aa、aa观测值分别为5、4、13，现要检验这个群体是否符合HW平衡。
- 从观测值得到等位基因A和a的个数分别为n_1=14和n_2=30，均为偶数。因此，X_{12}的可能取值是0与14之间的所有偶数。

基因型*AA*、*Aa*、*aa*观测值分别为5、4、13的样本群体不满足HW平衡定律

AA	Aa	aa	概率		AA	Aa	aa	概率	累积概率
7	0	15	0.0000		7	0	15	0.0000	0.0000
6	2	14	0.0003		6	2	14	0.0003	0.0003
5	4	13	0.0087		**5**	**4**	**13**	**0.0087**	**0.0090**
4	6	12	0.0756		0	14	8	0.0456	0.0546
3	8	11	0.2592		4	6	12	0.0756	0.1302
2	10	10	0.3802		1	12	9	0.2304	0.3606
1	12	9	0.2304		3	8	11	0.2592	0.6198
0	14	8	0.0456		2	10	10	0.3802	1.0000

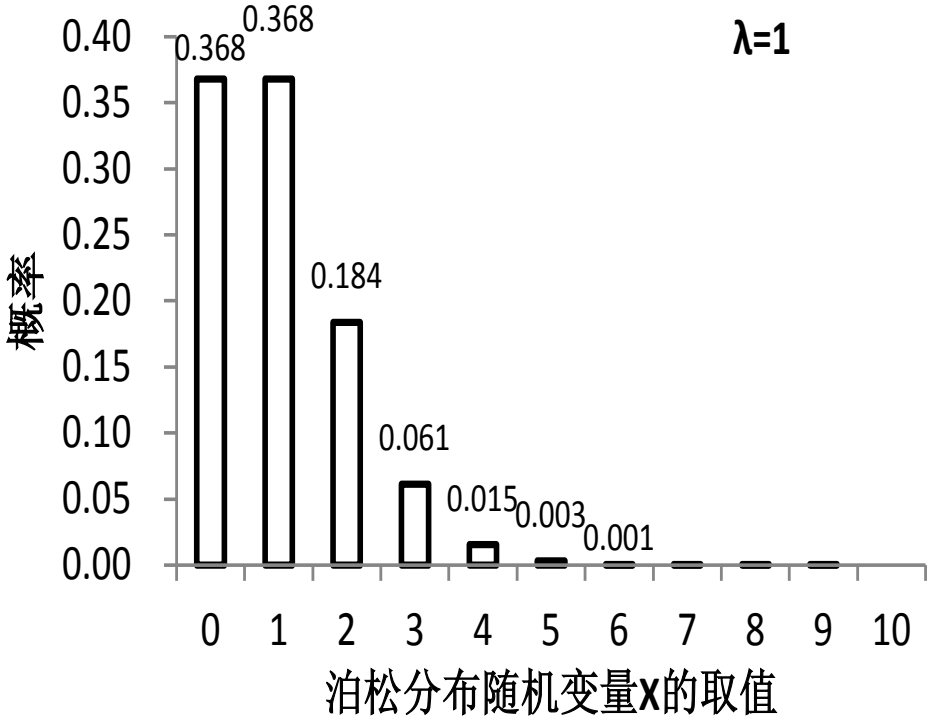
泊松分布

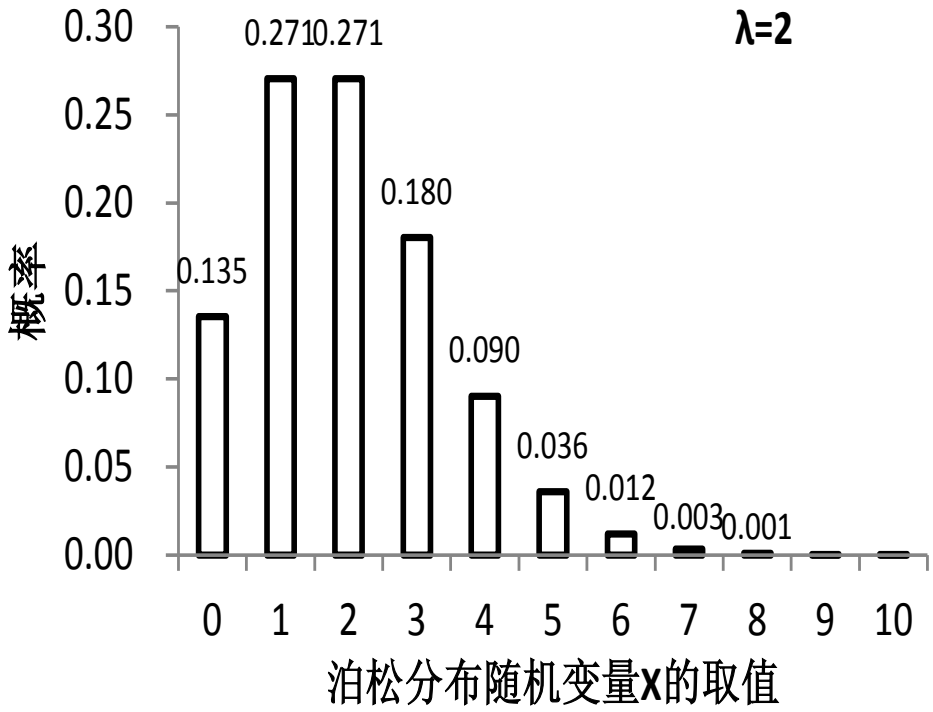
- 一个商店每小时到访的顾客数，一台电话每小时打进来的次数，一定面积田间地块中的害虫个数，一种稀有疾病在一类人群中每年的发病人数等等。
- 这些随机变量都服从或近似服从泊松分布。泊松分布的概率函数、均值和方差分别为

$$p(X=k)=\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} \quad (k=0，1，2，\ldots，\lambda>0)$$

$$E(X)=V(X)=\lambda$$

参数λ=1和λ=2的泊松分布





二项分布的泊松近似

- 当样本量n比较大时，二项分布概率的计算还是挺麻烦的。而泊松分布的概率公式中只有一个参数，计算起来要容易得多。
- 例如，某疾病的发病率只有p=0.001，在一个2000人组成的人群中，病人个数服从n=2000，p=0.001的二项分布。从二项分布概率公式，计算人群中1名患者或2名患者的概率不太容易。在这种情况下，$\lambda=0.001\times2000=2$的泊松分布就可以很好地被用来近似这个二项分布。
- 一般地讲，如果一个二项分布中，np是一个固定的数，当n充分大时，二项分布$B(n, p)$ 的$X=k$取值概率近似等于参数$\lambda=np$的泊松分布概率。

单个基因的在下一代拷贝数的分布

- 假定一个随机交配群体，在上下代之间保持恒定的群体大小N，等位基因的个数为$2N$。如果某个基因突变成为一个新的等位基因，在刚发生突变的群体中，它的频率为$1/2N$，其它基因的频率为$1-1/2N$。
- 该基因在下一代群体中存在的次数k服从二项分布$B(n=2N, p=1/2N)$，分布概率对应于二项式的展开 $\left(\frac{1}{2N}+(1-\frac{1}{2N})\right)^{2N}$

单个基因的丢失是大概率事件

- 近似服从参数λ=1的泊松分布。
- 对应于二项分布$B(n=2N, p=1/2N)$ 的$k=0$概率，即为丢失概率θ。当$2N$较大时，这一概率趋近于λ=1泊松分布的$k=0$概率。

$$\theta = p^0 q^{2N} = (1-\frac{1}{2N})^{2N} \to e^{-1} = 0.3679 = \frac{\lambda^k}{k!} e^{-\lambda} \mid_{\lambda=1,k=0}$$

- 如世代t的丢失概率为θ_t，世代t未发生丢失，则世代$t+1$的丢失概率为θ_{t+1}为 $\theta_{t+1} = e^{\theta_t - 1}$

§3.2 近交和近交系数

- §3.2.1 小群体中基因的随机漂移
- §3.2.2 祖先关联和近交
- §3.2.3 状态相同基因和后裔同样基因
- §3.2.4 共祖先系数和近交系数

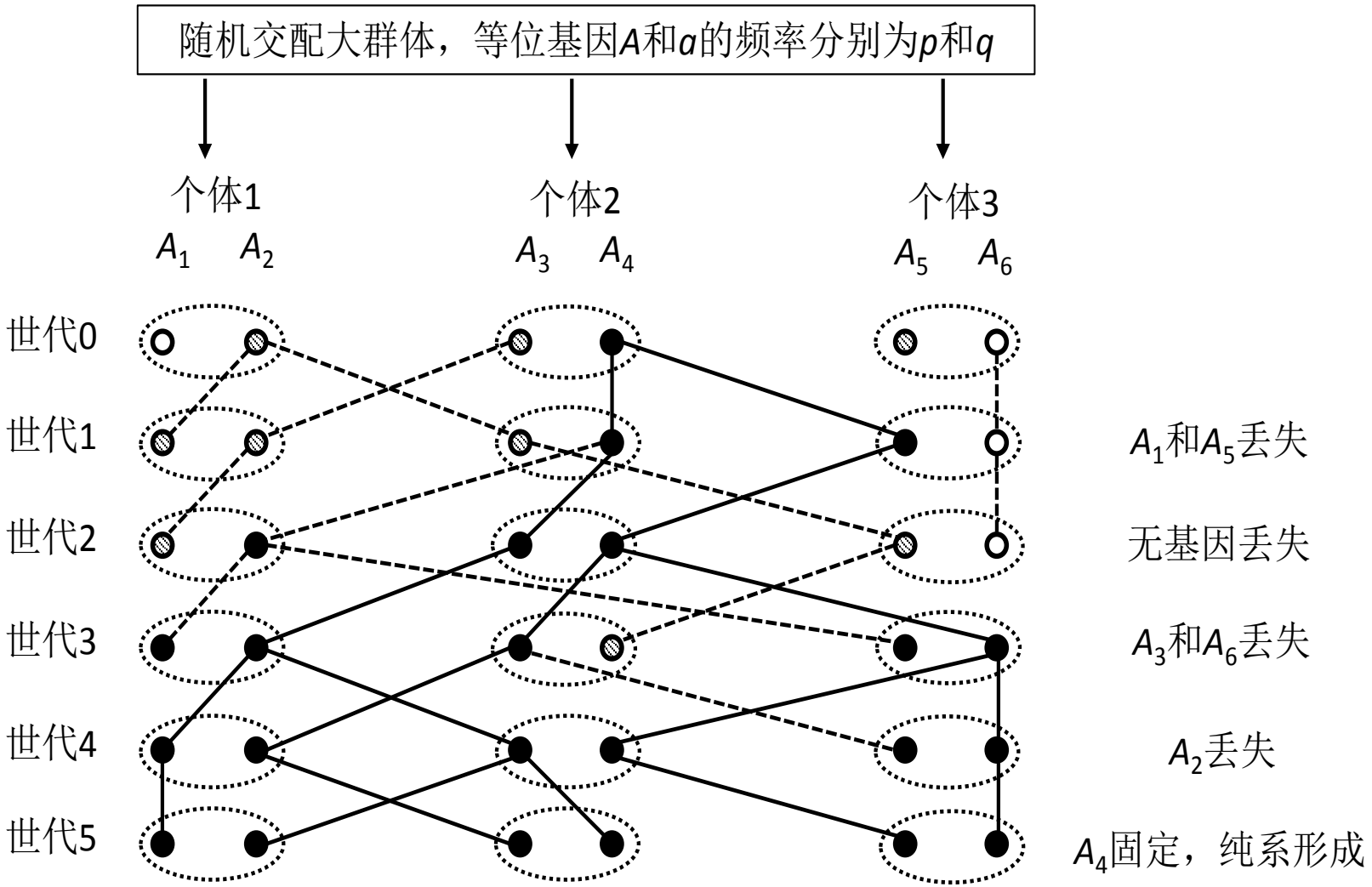
有限的群体与无穷的配子

- 几乎所有的生物物种，都能产生远远超过生存所需要的雌雄配子。但是，只有有限个雌雄配子能够结合在一起形成下一代合子。
- 如下一代的群体大小为N，则共需要N个雌配子和N个雄配子。它们可以看作无限大雌雄配子群体的$2N$个样本。
- 一个等位基因在这$2N$个样本中存在的次数k，是一个服从二项分布$B(n, p)$的随机变量，其中的n就是配子的个数$2N$，p就是这个等位基因在亲本群体中的存在频率。

小群体中基因的随机漂移

- 一次抽样构成的群体中，等位基因的频率会偏离无限大配子群体的频率。抽样频率既可能高于原来的频率、也可能低于原来的频率，偏离的方向事先是无法预测的。
- 这种由于随机抽样引起的基因频率波动，称为随机漂移（random drift），随机漂移最终导致群体被固定在某一个等位基因上。但是，不同的抽样群体可能被固定在不同的等位基因上。

大小为三的随机交配群体中基因丢失和固定的示意图



基因的固定或丢失导致纯系

- 长期的随机漂移，最终的群体中只包含一种纯合基因型，并且所有的基因都来自世代0的同一个基因A_4，这样的群体又称为纯系（pure line）。
- 纯系群体在没有迁移、突变等因素存在的情况下，群体结构也不会再发生任何改变。在所考察的座位上，纯系群体中只存在一种等位基因。因此，纯系群体中不存在基因的多态性。

纯系的产生也具有随机性

- 亲本基因在后代中的传递具有随机性，如果能够重复前面的随机漂移过程，则最终的群体也可能被固定在A_4之外的其它基因上。因此，在多个这样的纯系群体中，有的具有基因型AA，有的具有基因型aa。
- 在§3.3中我们还会看到，纯系AA和纯系aa的比例正好等于世代0中等位基因A和a的频率。如考虑多个座位，如l个，每个座位上有2个等位基因，所有可能的纯合基因型有2^l种，如$2^{10}=1024$，$2^{20}=1,048,576$。
- 因此，不同小群体随机交配多代，由随机漂移形成的纯系就会具有互不相同的纯合基因型。

祖先关联

- 在一个遗传群体中，如果知道亲代和子代的系谱关系，这时就可以研究个体之间的亲缘关系。如果两个个体具有共同的祖先，或者说至少具有一个共同的祖先，则称这两个个体是祖先关联（related by ancestry）。
- 从进化意义上讲，遗传群体中的任何两个个体最终都能追踪到一个或多个共同祖先。因此，群体遗传学的祖先关联是一个相对概念。如果两个个体向前追踪若干个世代，仍然没有找到共同的祖先，一般就认为这两个个体之间不存在祖先关联。
- 无限大的随机交配群体中，可以认为任何两个个体都不存在祖先关联。

近交

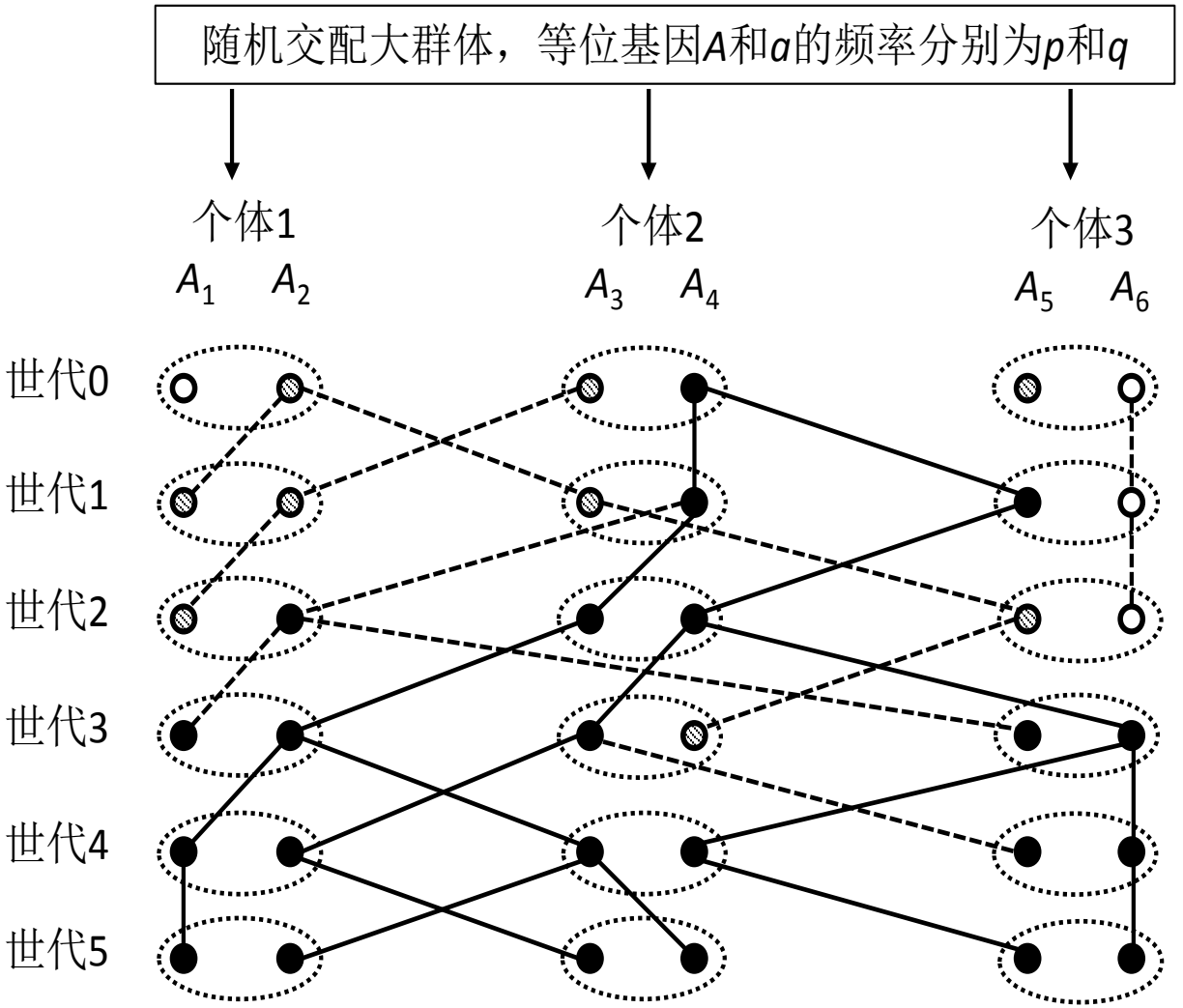
- 有祖先关联的两个个体之间的交配称为近交（inbreeding）。后面可以看到，近交会导致近交系数（coefficient of inbreeding）的增加，近交系数的增加意味着两个等位基因具有相同来源的概率的增加。
- 如果共同的亲本十分遥远，由共同亲本造成的近交效应可以忽略。因此，近交的研究往往是相对于一个特定时期的亲本群体而言，这个群体有时也称为基础群体（base population）或参照群体（reference population）。基础群体中，个体间假定是没有祖先关联的，因此也不存在近交。

状态相同基因和后裔同样基因

- 为了度量祖先关联和近交的程度，需要区分一个座位上两个等位基因的来源。例如，有两个基因A_1和A_2，如果他们具有相同的物理结构（即有相同的DNA序列、相同的功能等）和表型效应，则这两个基因称为状态相同（alike in state）。例如A_1的一个拷贝与A_1的其他拷贝一定是状态相同的。
- 两个状态相同的基因，可能来自基础群体中同一个个体中的同一个基因，也可能来自不同个体中的不同等位基因。如果两个基因是共同亲本中同一个基因的拷贝，则称它们是后裔同样的（identical by descent）。后裔同样的两个基因X与Y用符号$X \equiv Y$表示。

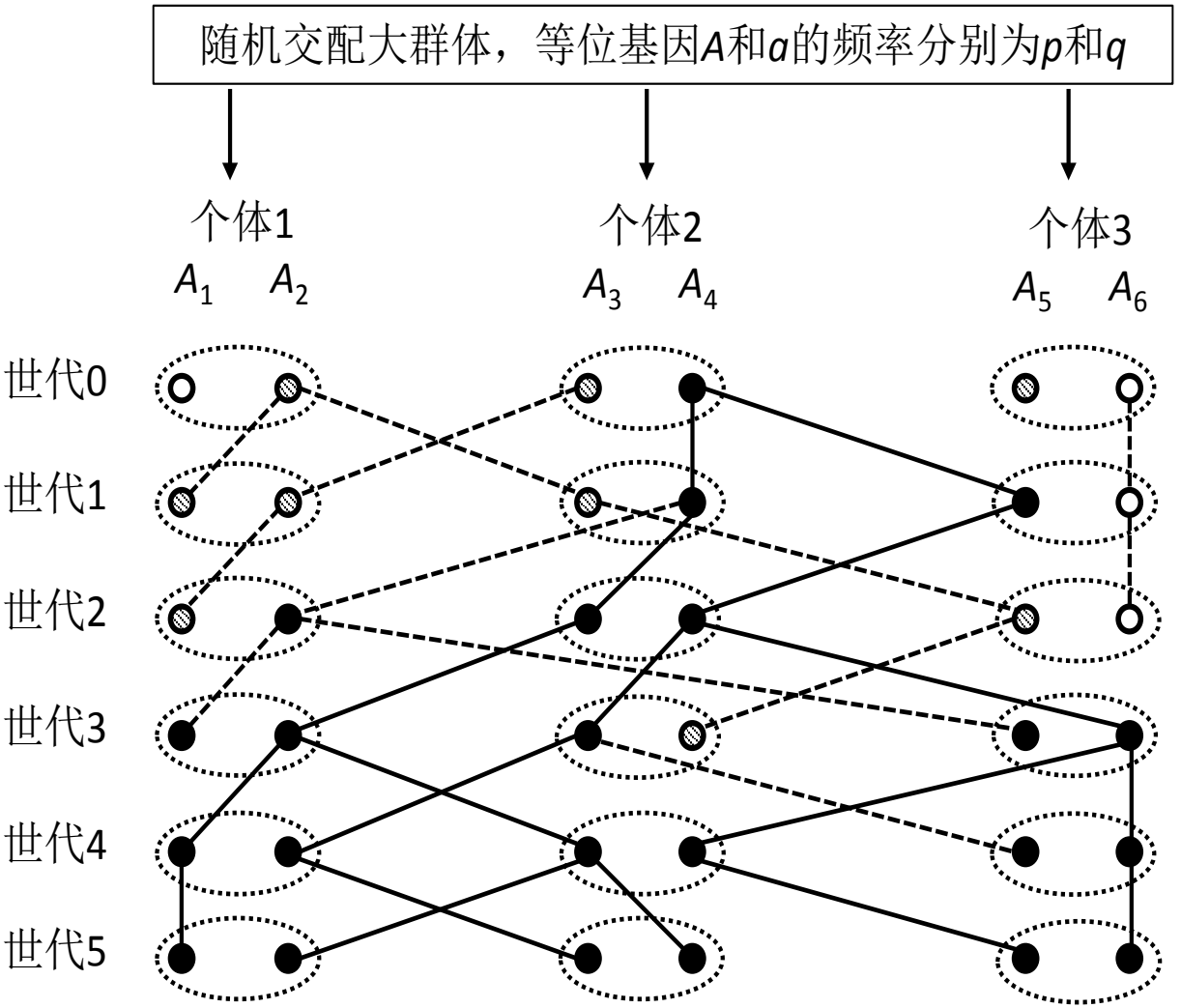
状态相同基因和后裔同样基因

- 两右图中，个体01携带两种不同的等位基因，它们不是状态相同基因。个体02携带的完全相同的两个等位基因，它们是状态相同基因，但不是后裔同样基因。



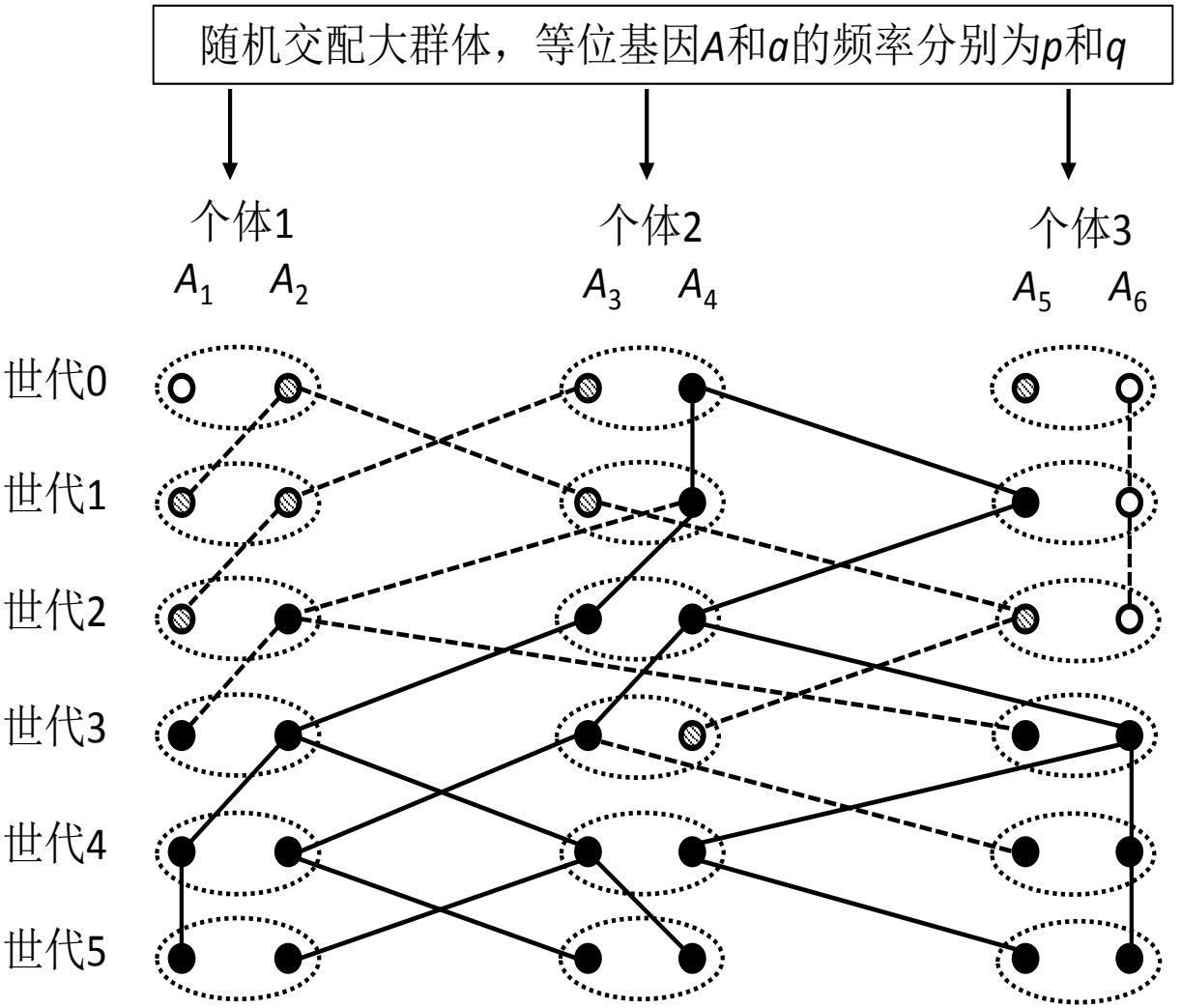
状态相同基因和后裔同样基因

- 个体32的两个A一个是A_2传递而来，另一个是A_4传递而来。因此，个体32携带的两个基因不是由祖先世代0同一个基因传递下来的，它们仅仅是状态相同，而不是后裔同样。



状态相同基因和后裔同样基因

- 个体31、33以及之后的所有个体，它们携带的两个基因是同一个祖先（即个体02）中的同一个基因（即A_4）传递下来的，这些基因全部都是后裔同样的。



共祖先系数

- 群体遗传学中常用共祖先系数（coefficient of coancestry）来度量两个个体的祖先关联程度，共祖先系数有时也称为亲缘系数或亲本系数（coefficient of parentage）。
- 在一个基因座位上，从个体X中随机抽取一个等位基因与个体Y中随机抽取的一个等位基因是后裔同样的概率，定义为个体X和Y的共祖先系数，用符号f_{XY}表示。

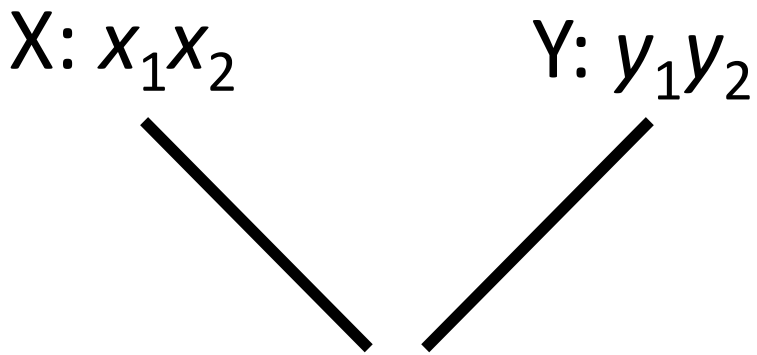
近交系数

- 祖先关联个体间的交配为近交。因此，祖先关联的程度越高，近交的程度也越大。群体遗传学中常用近交系数（coefficient of inbreeding）来度量近交的程度。
- 一个基因座位上，个体Z携带的两个等位基因是后裔同样的概率，定义为个体Z的近交系数，用符号F_Z表示。

共祖先系数和近交系数的取值范围

- 从上面的定义来看，共祖先系数针对两个个体而言，而近交系数则针对某一个个体。Z的近交系数F_Z=0意味着无近交。
- 如果两个亲本X和Y很多后代的近交系数都是0，这时可能意味着X和Y之间不存在祖先关联，即亲本的共祖先系数f_{XY}也是0。
- 近交系数F_Z=1意味着完全近交。如果X和Y很多后代的近交系数都是1，这时可能意味着他们的共祖先关联系数也是1。

共祖先系数和近交系数的关系



¼ x_1y_1 + ¼ x_1y_2 + ¼ x_2y_1 + ¼ x_2y_2

$f_{XY} = P\{x \equiv y\}$, 其中x表示x_1或x_2，y表示y_1或y_2

$F_Z = P\{z_1 \equiv z_2\}$

共祖先系数和近交系数的关系

$$F_{\mathrm{Z}} = \frac{1}{4}[P\{x_1 \equiv y_1\} + P\{x_1 \equiv y_2\}$$

$$+ P\{x_2 \equiv y_1\} + P\{x_2 \equiv y_2\}] = f_{\mathrm{XY}}$$

- 从中可以看到，后代的近交系数其实就是后代群体中四种可能基因型的近交系数的均值。也就是说，后代的平均近交系数等于亲本的共祖先系数。
- 如果四种后代基因型的近交系数不完全相等，这时就不能用一种后代基因型的近交系数作为双亲的共祖先系数。
- 同理，如果用共祖先系数作为后代的近交系数，这个近交系数其实是后代群体的平均近交系数，单个后代个体的近交系数之间可以存在差异。

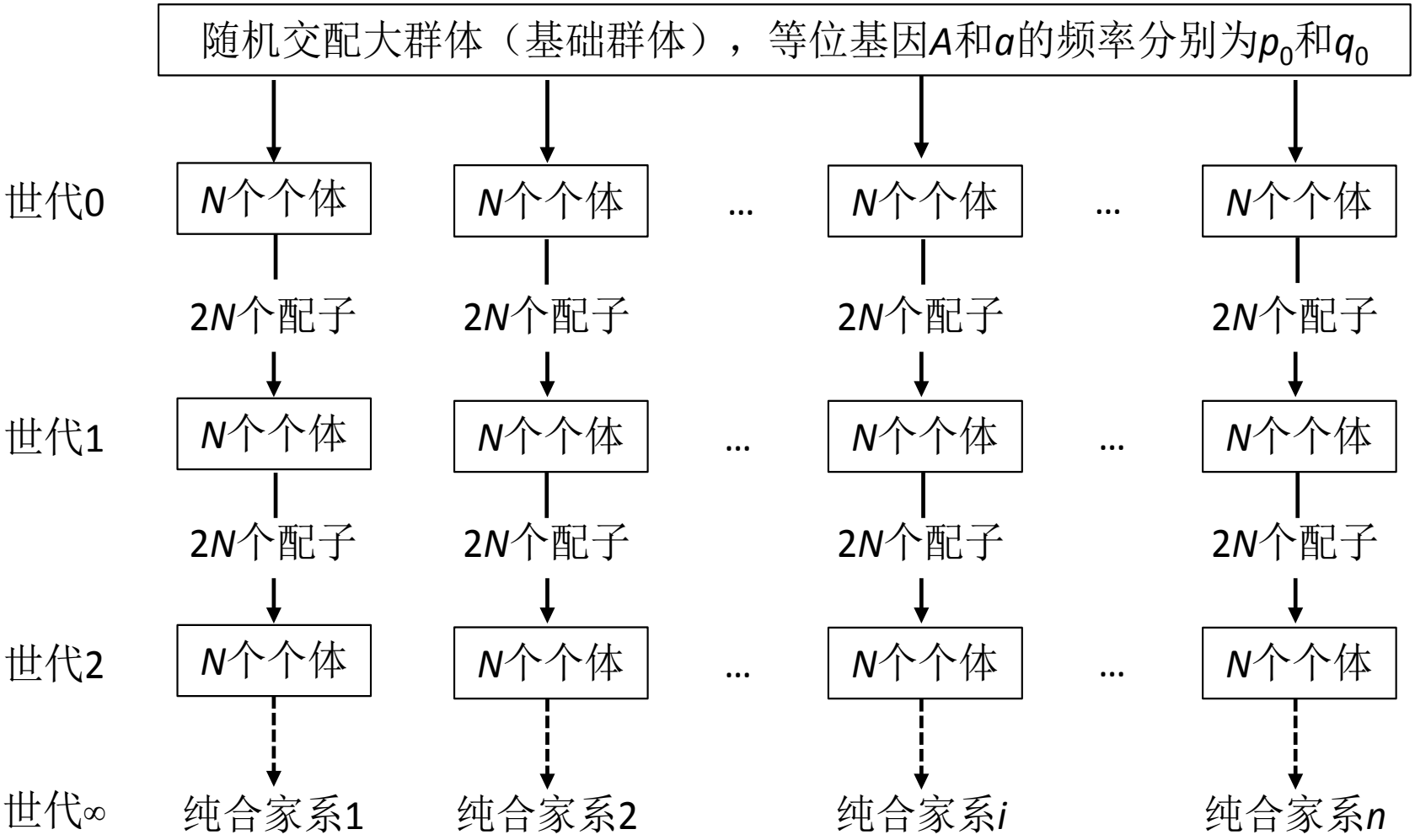
§3.3 理想有限大小群体的遗传构成

- §3.3.1 理想的有限大小随机交配群体
- §3.3.2 随机漂移的Wright-Fisher模型
- §3.3.3 理想群体中的近交
- §3.3.4 理想群体中基因频率的波动
- §3.3.5理想群体中基因型频率的波动

理想的有限大小群体

- 理想群体可以是随机交配大群体由于自然条件、地理或生活环境的变化，而变为许多个亚群体，也可以是人为创造的育种或实验群体。原来的随机交配大群体叫基础群体或起始群体，而亚群体叫家系或品系。

一个随机交配大群体分化为n个大小为N的亚群体或家系过程示意图



理想群体满足的条件

（1）交配只在亚群体内部发生，亚群体之间没有相互迁移；

（2）上下代区分明显，没有世代重叠；

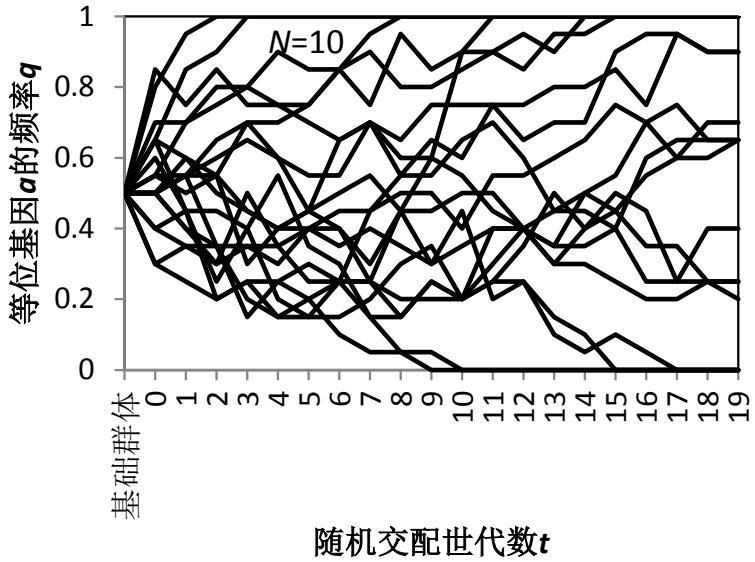
（3）各个亚群体中，每个世代具有相同数量的个体数；

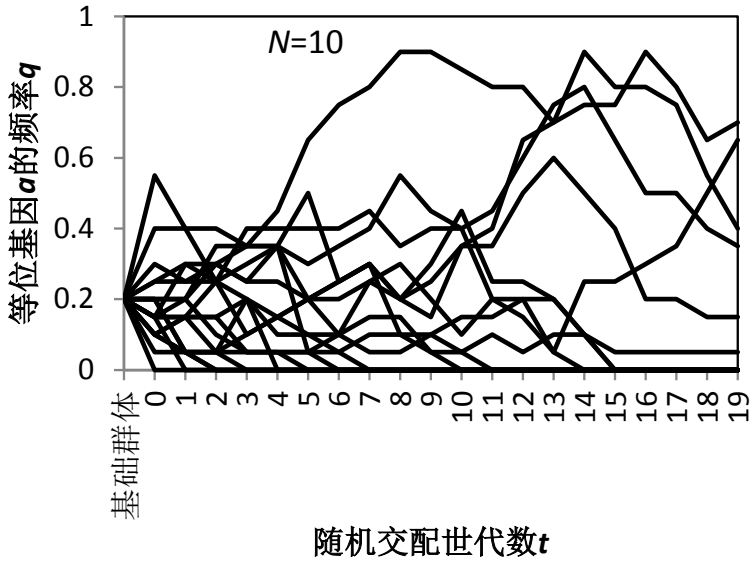
（4）每个亚群体内的交配是随机的，不存在性别差异（即允许自交）；

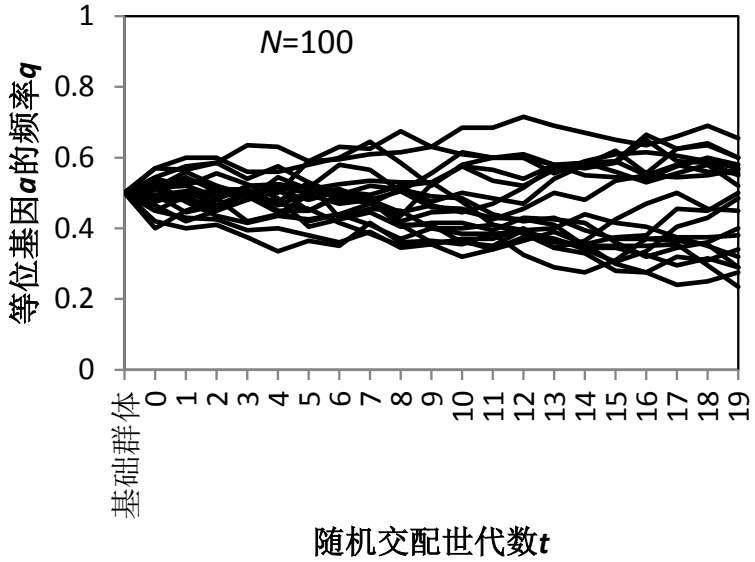
（5）个体有相同的适合度，即不受选择的影响；

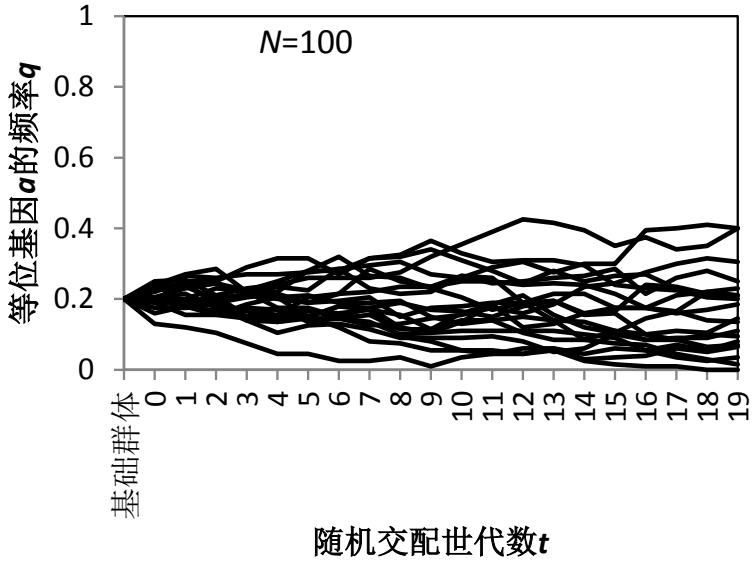
（6）不考虑基因突变。

有限大小群体的随机飘变 $\sigma_q^2 = \sigma_{\Delta q}^2 = \frac{q_0(1-q_0)}{2N}$









随机漂移的Wright-Fisher模型

- 利用随机漂移的马尔可夫模型，可以从理论上说明，前图观察到的基因频率波动以及基因的固定和丢失并不是偶然的，而是一种必然现象。
- 利用马尔可夫链描述随机漂移过程，是由Fisher和Wright分别在1930和1931年首先提出来，因此也称为Wright-Fisher模型。

世代0群体中等位基因A的个数X_0

- 第0世代中，单个亚群体的N个个体是无限大随机交配群体的一个随机样本。他们携带的等位基因有$2N$个，相当于等位基因A、a频率分别为p_0、q_0的无限大配子群体中随机抽取的$2N$个样本。这$2N$个样本中，等位基因A的个数是一个随机变量，用$\boldsymbol{X_0}$表示（k表示它的不同取值），服从二项分布 $B(n=2N, p=p_0)$。取值概率为

$$P(X_0 = k) = \frac{(2N)!}{k!(2N-k)!} p_0^k q_0^{2N-k}$$ ，其中k=0, 1, …, $2N$

世代t-1到世代t的转移概率矩阵

- 用X_{t-1}、X_t表示世代t-1、t中等位基因A的个数，k和j表示它们的取值。
- 产生第t个世代N个个体的过程，相当于从第t-1个世代N个个体产生的无穷大配子群体中随机抽取2N个配子。

$$T_{jk} = P(X_t = j \mid X_{t-1} = k) = \frac{(2N)!}{j!(2N-j)!}\left(\frac{k}{2N}\right)^j\left(1-\frac{k}{2N}\right)^{2N-j}$$

其中，j=0, 1, …, 2N；k=0, 1, …, 2N

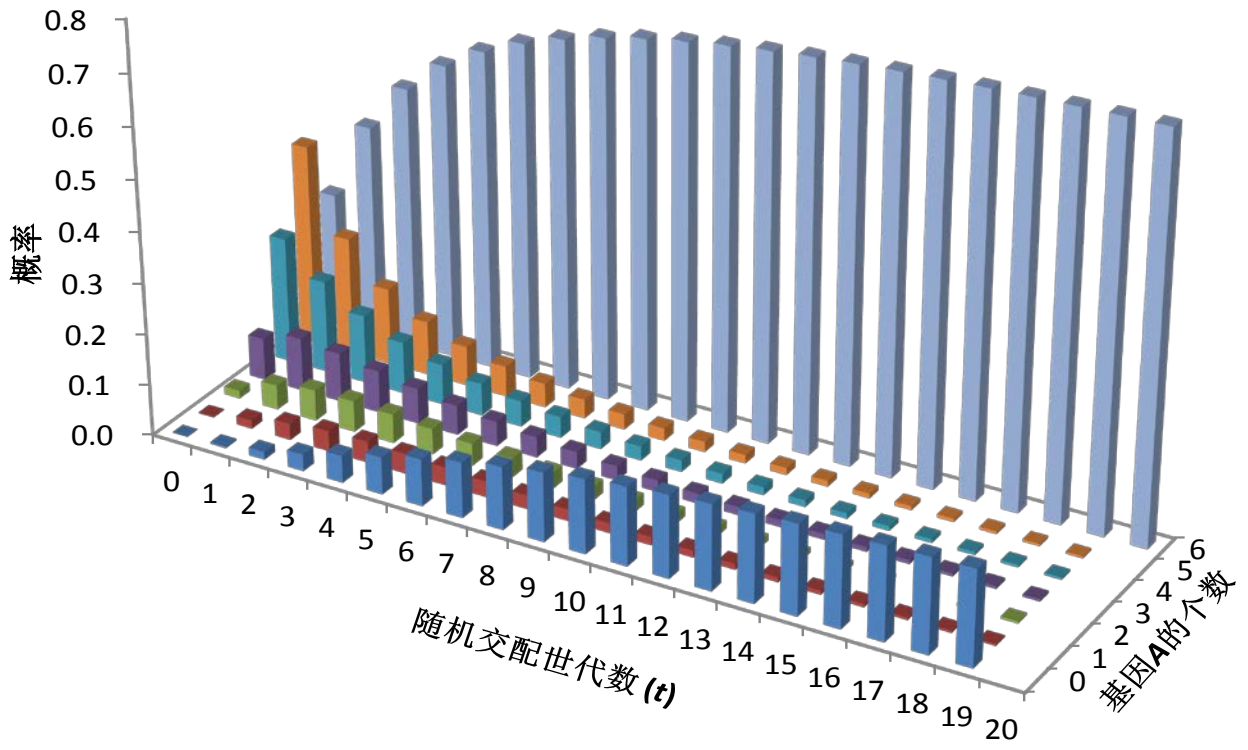
大小为3理想群体的转移矩阵

$$\mathbf{T}=\begin{bmatrix} 1 & 0.3349 & 0.0878 & 0.0156 & 0.0014 & 0.0000 & 0 \\ 0 & 0.4019 & 0.2634 & 0.0938 & 0.0165 & 0.0006 & 0 \\ 0 & 0.2009 & 0.3292 & 0.2344 & 0.0823 & 0.0080 & 0 \\ 0 & 0.0536 & 0.2195 & 0.3125 & 0.2195 & 0.0536 & 0 \\ 0 & 0.0080 & 0.0823 & 0.2344 & 0.3292 & 0.2009 & 0 \\ 0 & 0.0006 & 0.0165 & 0.0938 & 0.2634 & 0.4019 & 0 \\ 0 & 0.0000 & 0.0014 & 0.0156 & 0.0878 & 0.3349 & 1 \end{bmatrix} \quad \mathbf{P}_1=\begin{bmatrix} 0.0001 \\ 0.0015 \\ 0.0154 \\ 0.0819 \\ 0.2458 \\ 0.3932 \\ 0.2621 \end{bmatrix}$$

$$\mathbf{P}_t = \mathbf{T}\mathbf{P}_{t-1}$$

大小为3理想群体的漂移过程

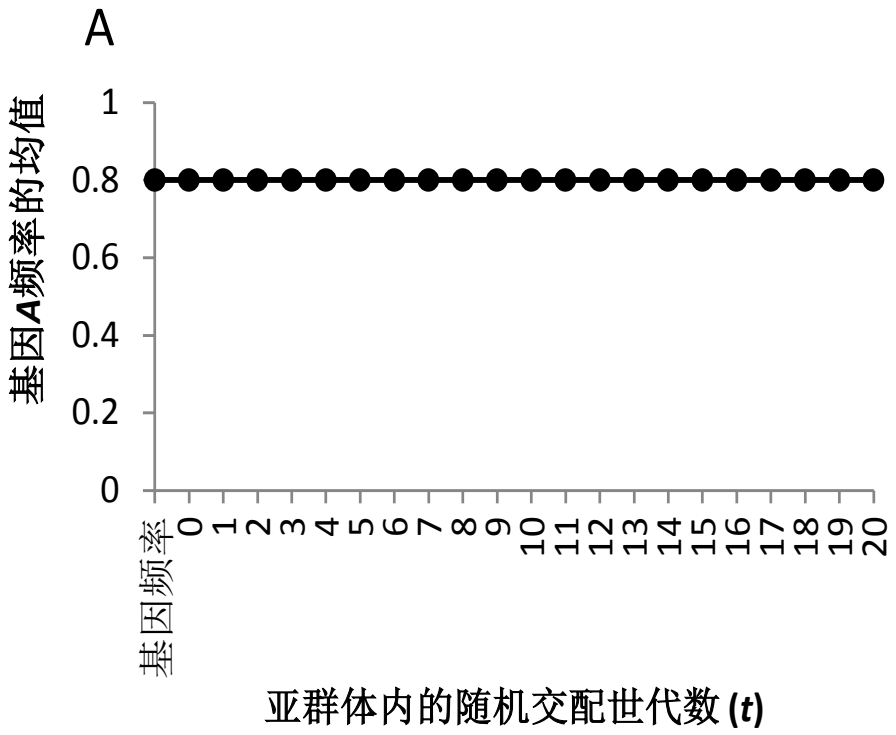
- 基础群体中，等位基因A的频率为0.8，等位基因a的频率为0.2。X轴为随机交配世代数表示的时间，Y轴为等位基因A的个数，Z轴为概率。

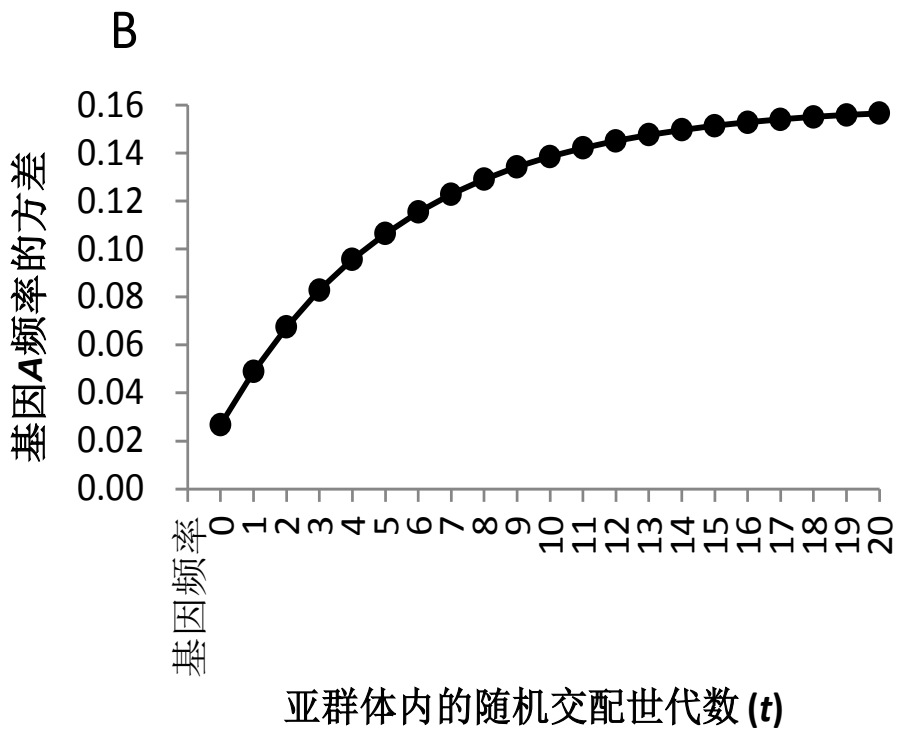


马尔可夫链的平稳性

- 马尔可夫链有一条重要性质，就是经过长时间的状态转移后，不论起始状态如何，都会进入一个相对平稳的状态，相当于$\mathbf{P}_t=\mathbf{P}_{t-1}$。
- 在随机漂移过程中，这个平稳状态就是$k=0$的概率接近于q_0，$k=2N$的概率接近于p_0，k在各种非吸收态上的概率近似相等。
- 这一点其实在前图中也可以看到。当$N=3$时，经过20个世代的随机漂移，$k=0$的概率为0.1912，$k=6$的概率为0.7912，k取1、2、3、4、5的概率在0.0032~0.0038，近似相等。

大小为3的理想群体随机漂移过程中基因频率的均值（A）和方差（B）的变化





随机漂移改变亚群体的结构

（1）单个亚群体中，基因频率在世代间的变化是完全随机的，没有任何恢复到基础群体基因频率的倾向；

（2）长期的随机漂移，导致不同亚群体之间的分化，不同的亚群体有着不同的基因频率和基因型频率；

（3）随机漂移导致亚群体内的一致性，单个亚群体内的个体之间，具有越来越强的祖先关联和近交，最终趋于一致；

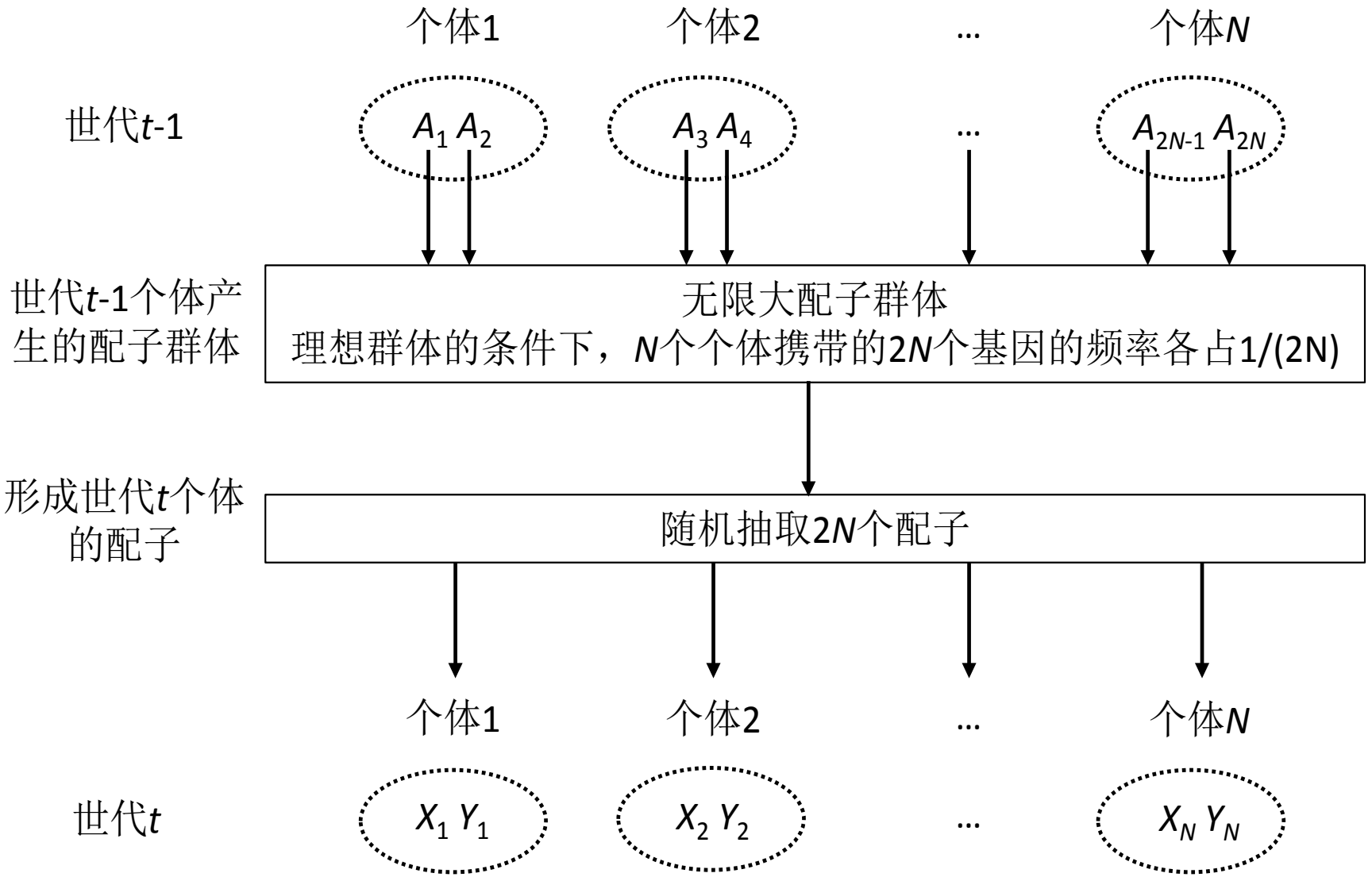
（4）如果所有亚群体同时考虑，则纯合基因型的频率不断增加，杂合基因型的频率不断下降。

理想群体中的近交

- 考虑某一位点的两个等位基因。基础群体的近交系数为0，即F_0=0。
- 小群体的容量为N，一共有$2N$个等位基因。在雌雄交配中，每个配子与其本身完全一样的配子结合的概率为$1/2N$。
- $1/2N$定义为理想群体在世代1的近交系数

$$F_1 = \Delta F = \frac{1}{2N}$$

理想群体随机交配产生下一代示意图



相邻世代间近交系数的关系

$$F_t = \frac{1}{2N} + (1 - \frac{1}{2N})F_{t-1}$$

- 等式右边的第一项可以看作新增加的近交，第二项可以看作前一个世代保留下来的近交。

近交系数的递推公式

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

$$F_t=1-(1-\Delta F)^t(1-F_0)$$

$$F_t=1-(1-\Delta F)^t\rightarrow 1$$

长期漂移的结果

- 对于有限的群体大小N，只要t足够大，群体的近交系数就会接近于1。这时，每个家系与一个纯系有着相同的遗传结构，即群体中所有个体的基因型都是纯合一致的，所有基因都是后裔同样的。但是，不同家系有着不同的基因型。
- 当座位数和等位基因数较大时，几乎不可能存在基因型完全相同的两个家系。这些亚群体再经过长期的进化过程和生殖隔离，不同的家系就可能会形成截然不同的新物种。因此，随机漂移是群体遗传学和物种形成的重要研究对象。

纯系的基因型与随机交配大群体中纯合基因型的区别

- 随机漂移以及重复自交产生纯系的基因型，与HW平衡大群体中的纯合基因型有着本质区别。
- 纯系的两个等位基因不仅状态相同，而且后裔同样。随机交配大群体中，纯合基因型的两个等位基因仅仅状态相同，但不是后裔同样。
- 亲缘关系是由后裔同样基因决定的，而不是由状态相同基因决定的。基因型是否纯合与近交系数之间没有必然联系，后裔相同的纯合型才是近交系数的决定因素。

小样本理想群体中近交系数的变化

群体大小	随机交配次数（世代数等于随机交配次数加上1）								
	0	1	2	3	5	10	20	50	100
1	0	0.5	0.75	0.875	0.9688	0.9990	1.0000	1.0000	1.0000
10	0	0.05	0.0975	0.1426	0.2262	0.4013	0.6415	0.9231	0.9941
20	0	0.025	0.0494	0.0731	0.1189	0.2237	0.3973	0.7180	0.9205
30	0	0.0167	0.0331	0.0492	0.0806	0.1547	0.2855	0.5684	0.8138

近交系数的增长速率

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

$$\Delta F = \frac{F_t - F_{t-1}}{1 - F_{t-1}}$$

- 公式中的分子代表相邻两个世代近交系数的增加量，分母是上个世代近交系数与纯系近交系数1之间的差距。$\Delta F = 1/2N$正好度量了相对于纯系近交系数的增长速率（rate of inbreeding）。
- 下一章将看到，这一公式可用于估计非理想情况下的有效群体大小。

理想群体中基因频率的波动

- 世代1群体中，等位基因A的个数X_1服从二项分布$B(n=2N, p=p_0)$。因此，

$$V(X_1) = 2Np_0q_0$$

$$p_1 = \frac{X_1}{2N} \qquad \Delta p_1 = p_1 - p_0$$

$$\sigma^2_{p_1} = \sigma^2_{\Delta p_1} = \frac{p_0q_0}{2N} = p_0q_0F_1$$

基因频率方差与近交系数的关系

- 利用Wright-Fisher模型的性质，还可以证明世代1基因频率方差与近交系数的关系对于所有世代都是成立的，即：

$$\sigma_{p_t}^2 = \sigma_{\Delta p_t}^2 = p_0 q_0 F_t$$

- 因此，随机交配世代数足够多时，近交系数趋于1，基因频率的方差趋于$p_0 q_0$，也就是说只与起始频率有关。

容量N=3的理想群体中，等位基因个数的概率分布、基因频率均值和方差

基因A的个数	基因A的频率	世代					
		0	1	2	3	4	5
0	0.0000	0.0001	0.0036	0.0152	0.0330	0.0532	
1	0.1667	0.0015	0.0166	0.0322	0.0401	0.0417	
2	0.3333	0.0154	0.0480	0.0602	0.0610	0.0567	
3	0.5000	0.0819	0.1041	0.0959	0.0824	0.0694	
4	0.6667	0.2458	0.1804	0.1342	0.1022	0.0796	
5	0.8333	0.3932	0.2307	0.1508	0.1059	0.0782	
6	1.0000	0.2621	0.4167	0.5115	0.5754	0.6212	
基因频率的均值		0.8	0.8	0.8	0.8	0.8	
基因频率的方差		0.0267	0.0489	0.0674	0.0828	0.0957	
近交系数F			0.1667	0.3056	0.4213	0.5177	0.5981
利用F计算的基因频率方差		0.0267	0.0489	0.0674	0.0828	0.0957	

理想群体中基因型频率的波动

- 用p表示单个亚群体中基因A的频率、p^2表示基因型AA的频率。不同的亚群体有不同的p和p^2值。利用$E(p)=p_0$以及方差计算公式

$$V(p) = E(p^2) - [E(p)]^2 = p_0 q_0 F$$

- 得到基因型AA频率p^2的期望为

$$E(p^2) = p_0^2 + p_0 q_0 F$$

- 类似得到基因型Aa*和*aa频率的期望分别为

$$E(q^2) = q_0^2 + p_0 q_0 F \qquad E(2pq) = 2 p_0 q_0 (1 - F)$$

近交降低杂合型的频率

- 从前面的公式可以看到，近交引起纯合型频率的增加、杂合型频率的减少。第1章介绍的自交是近交的极端情况，纯合型频率增加、杂合型频率减少的速度更快。因此，杂合型频率相对于HWE群体的变化，可以看作是度量近交程度的另外一种方法。即：

$$F = \frac{2p_0q_0 - E(2pq)}{2p_0q_0}$$

- 公式中的$2p_0q_0$是HWE群体的杂合基因型频率，$E(2pq)$是漂变产生众多亚群体中杂合基因型的平均频率。

利用杂合型频率的降低量估计近交系数

- 一般地，用H_T表示给定基因频率在HW平衡时的杂合型频率或杂合度，用H_S表示近交发生后的杂合型频率或杂合度，下面的公式给出利用杂合型频率变化的近交系数计算方法：

$$F_{ST} = \frac{H_T - H_S}{H_T} = 1 - \frac{H_S}{H_T}$$

- 为与之前的近交系数相区分，上面公式给出的近交系数用F_{ST}表示。

从近交系数计算杂合度的下降程度

- 例如，基础群体中等位基因*A*和*a*的频率均为0.5，亚群体的容量*N*=8。在理想群体的条件下，得到5个世代随机交配后的近交系数*F*=0.276。
- 利用前面的公式，得到亚群体之间等位基因频率的方差为0.069，三种基因型*AA*、*Aa*、*aa*的频率均值分别为0.319、0.362、0.319。基础群体的三种基因型频率分别为0.25、0.5、0.25。因此，经过5个世代的漂移后，亚群体的杂合度平均下降了27.6%。

从近交系数计算各种基因型的期望频率

- 在一个无限大随机交配群体中，等位基因*A*和*a*的频率分别为0.6和0.4，从中随机抽取30个个体。计算在这样的有限群体中，随机交配10代后各种基因型的频率。
- 随机交配10代的近交系数F=0.155，由此得到杂合基因型*Aa*的频率为0.406，纯合基因型*AA*和*aa*的频率分别为0.397和0.197。
- 杂合基因型相对于基础群体降低了15.5%，纯合基因型分别增加了10.3%和23.2%。

§3.4 自然群体的分化

- §3.4.1 群体的阶梯结构
- §3.4.2 分化程度的度量
- §3.4.3 隔离拆除效应

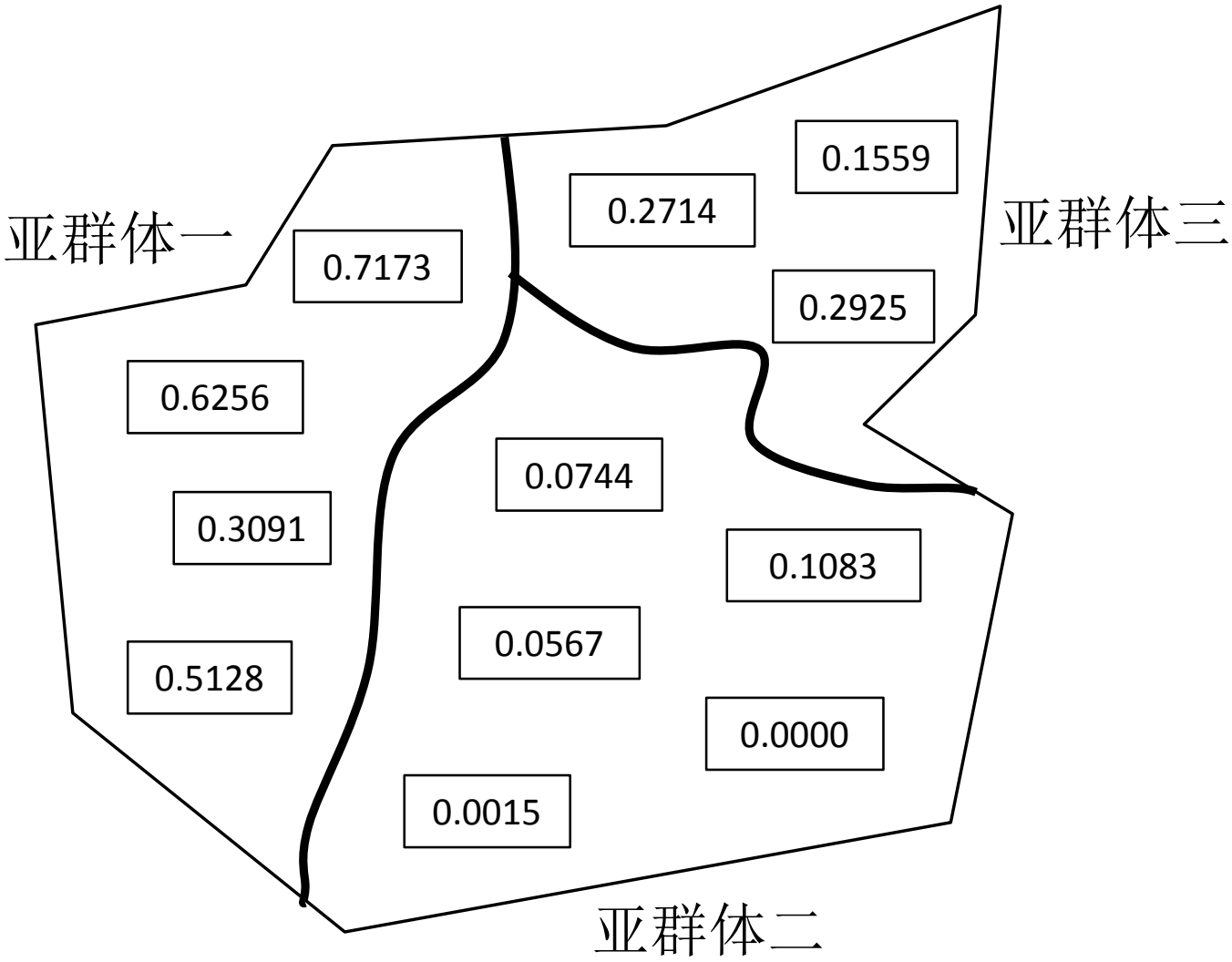
随机漂移与群体分化

- 从近交系数的计算公式可以看出，理想群体经历的时间越长，近交系数就越高，亚群体内个体间的亲缘关系愈加密切。从基因频率的方差公式可以看出，近交系数越高，亚群体之间基因频率的差异也越大。这样，经过的时间越长，亚群体之间的分化就越明显。
- 从基因型频率的计算公式可以看出，近交系数越高，杂合基因型的频率就越小。因此，随机漂移的时间越长，亚群体内部的杂合度、以及亚群体之间的平均杂合度下降也越多。
- 由于生存环境和地理距离等方面的限制，自然界中大多数物种的种群都能被划分成很多较小的亚群体，个体之间的交配繁殖往往只局限在亚群体内部。因此，尽管是一种理想状态，理想群体的分化过程在生物界具仍有广泛的代表性。

群体的阶梯结构

- 群体划分势必造成亚群体之间的遗传分化（genetic differentiation），也就是基因频率在亚群体之间的差异。不同环境下，各种基因型的适合度不尽相同，因此选择又会进一步引起亚群体在基因频率上的差异。
- 即使不存在选择或其他改变基因频率因素的情况下，仅随机漂移就能不断增加基因频率在亚群体之间的方差，造成基因频率在亚群体之间的广泛差异。
- 这种划分还可以继续下去，将亚群体进一步划分为更小的、更封闭的次级亚群体。因此，自然界中的物种种群大多表现出一种阶梯式结构（hierarchical structure）。

一个随机交配植物群体在一个小岛上的地理分布图（数字为某基因的调查频率）



群体分化程度的度量：二级亚群体的平均杂合度（H_2）

亚群体一		亚群体二		亚群体三	
基因频率	杂合度	基因频率	杂合度	基因频率	杂合度
0.7173	0.4056	0.0744	0.1377	0.2714	0.3955
0.6256	0.4684	0.0567	0.1070	0.1559	0.2632
0.3091	0.4271	0.1083	0.1931	0.2925	0.4139
0.5128	0.4997	0.0015	0.0030		
		0.0000	0.0000		
二级亚群体的平均杂合度（H_2）					
0.2762					

- 由于采用的是分层抽样方法，每个次级亚群体在整个种群中有相同的权重。根据表中基因频率后面的12个杂合度，计算它们的简单平均数得到的平均杂合度为0.2762。

群体分化程度的度量：
一级亚群体的平均杂合度（H_2）

亚群体一		亚群体二		亚群体三	
平均基因频率	平均杂合度	平均基因频率	平均杂合度	平均基因频率	平均杂合度
0.5412	0.4966	0.0482	0.0917	0.2399	0.3647
一级亚群体的平均杂合度（H_1）					
0.2949					

- 如果只是存在一级而不存在二级结构，三个一级亚群体的平均杂合度H_1应该等于0.2949（第3行三个杂合度以4、5、3为权重的加权平均）。
- 这里请注意，表中三个一级亚群体的平均杂合度等于平均基因频率在HW平衡时的杂合度，这其实就是不存在二级结构假定下的杂合度。

群体分化程度的度量：
无任何结构群体的平均杂合度（H_0）

整个种群的平均基因频率	整个种群的平均杂合度（H_0）
0.2605	0.3852

- 如果没有任何阶梯结构，等位基因A频率为0.2605时，整个种群的杂合度H_0应该等于0.3852。

不同阶梯结构群体的杂合度

- 没有任何阶梯结构的杂合度H_0=0.3852。
- 只是存在一级而不存在二级结构，三个一级亚群体的平均杂合度H_1=0.2949。
- 既存在一级又存在二级结构，平均杂合度H_2=0.2762。
- 随着阶梯结构的提高，杂合度呈逐渐下降的趋势。

阶梯结构群体的分化系数

- 次级亚群体相对于一级亚群体的平均杂合度下降引起的近交：

$$F_{2,1} = \frac{H_1 - H_2}{H_1} = 1 - \frac{H_2}{H_1} = 0.0636$$

- 一级亚群体相对于整个种群的平均杂合度下降引起的近交：

$$F_{1,0} = \frac{H_0 - H_1}{H_0} = 1 - \frac{H_1}{H_0} = 0.2344$$

- 次级亚群体相对于整个种群的平均杂合度下降引起的近交：

$$F_{2,0} = \frac{H_0 - H_2}{H_0} = 1 - \frac{H_2}{H_0} = 0.3852$$

- 三者之间的关系：$1 - F_{2,0} = (1 - F_{2,1})(1 - F_{1,0})$

*n*阶阶梯结构群体的分化系数

$$1-F_{ST}=1-F_{n,0}=(1-F_{n,n-1})(1-F_{n-1,n-2})\cdots(1-F_{2,1})(1-F_{1,0})$$

- 习惯上，把整个种群的平均杂合度用H_T表示（即前面公式中的H_0），最后一个阶梯结构的平均杂合度用H_S表示（即前面中的H_2），最后一个阶梯结构相对于整个种群的近交系数用F_{ST}表示（即前面公式中的$F_{2,0}$）。当存在n层阶梯结构时，F_{ST}相当于$F_{n,0}$，有时也称分化系数。
- 这些公式定义近交系数 的取值为0~1，这种定义方法也能被推广到复等位基因的座位。

近交系数F_{ST}的作用

- 近交系数 是一个十分重要的群体遗传学指标，广泛用于亚群体的遗传分化和进化研究。
- Wright（1978）曾建议根据近交系数F_{ST}表示的大小来划分遗传分化的程度，近交系数0~0.05对应于很低程度的遗传分化，0.05到0.15对应于中等程度的遗传分化，0.15到0.25对应于高度的遗传分化，超过0.25则对应于极高程度的遗传分化。
- 例如，人类遗传学的研究表明，不同族群之间的分化系数 在0.07左右，属于中等程度的遗传分化。果蝇不同群体之间的分化系数 在0.10左右，也属于中等程度的遗传分化。

群体杂合度的恢复

- HW平衡定律告诉我们，不论什么样的群体，只需要一代随机交配就能达到平衡状态。因此，如果把存在遗传分化的很多亚群体混合在一起形成混合群体，并进行随机交配，则混合亚群体中降低的杂合度，能够通过一代随机交配得以完全恢复。

隔离拆除效应

- 由于亚群体间生殖隔离的消除，而引起纯合基因型频率下降的现象，有时也称为隔离拆除效应（isolate breaking）。
- 举一个极端的例子说明这一现象。如有两个亚群体，一个被固定在等位基因*A*上，另一个被固定在等位基因*a*上。两个亚群体的等量混合群体中，等位基因*A*和*a*的频率均为0.5，混合群体的杂合度为0。但是，只需经过一代随机交配，后代群体就能恢复到0.5的杂合度。同时，两种纯合基因型的频率从1下降到0.5。