第5章
遗传多样性的分子理论

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

本章的主要内容

- §5.1 遗传变异的分子基础
- §5.2 基因融合和基因树
- §5.3 中性突变理论
- §5.4 近交系数计算方法小结

§5.1 遗传变异的分子基础

- §5.1.1 DNA序列的多态性
- §5.1.2 无限等位基因模型

变异的类型

- 遗传变异可以反映在性状的表型、蛋白质的氨基酸序列和DNA的碱基序列等不同层次上。
- 性状在表型上的差异，除遗传因素外还受环境的影响，有时还存在不同基因座位间、同一座位内不同等位基因间的互作，表型上的差异很多时候难以完全反映遗传差异。
- 在DNA的转录和翻译过程中，64种三联密码子控制着20种氨基酸的合成，大多数氨基酸是由两个或两个以上的三联密码子控制。DNA序列上的差异也不一定都能在蛋白质的氨基酸序列上体现出来。
- 基因组DNA序列上还有大量的内含子和非编码区域，如哺乳动物的基因组只有大约1.5%的序列编码各种功能蛋白质。大量非编码区域上DNA序列的改变，不会影响氨基酸的合成和蛋白质的功能，也就不会产生新的表型。

DNA序列的多态性

- 前面所说的那些差异都能通过DNA序列的比对检测出来。因此，DNA序列数据包含了比蛋白质的氨基酸序列和性状的表型更加丰富的遗传变异信息。
- DNA序列上的差异是所有其他层次上遗传差异的基础，最小的差异单元是单核苷酸多态性（single nucleotide polymorphism，简称SNP）。

无限等位基因模型

- 编码一个功能蛋白的DNA序列长度一般都有数千碱基对（bp），每个核苷酸位置上都有A、T、C、G 4种可能。可能DNA序列的个数是一个非常大的数字。因此，有理由认为每次单核苷酸改变产生的突变，都是群体中不存在的新等位基因，称为突变的无限等位基因模型（infinite-alleles model of mutation）。

一个长度为500bp的座位上5个等位基因的DNA序列比对

等位基因	核苷酸物理位置/bp（其他位置上无差异，因此未给出）									
	132	142	162	192	198	201	207	240	246	
a	T	T	A	C	C	C	G	T	A	
b	T	C	T	A	A	C	G	T	T	
c	C	C	C	C	C	C	G	T	A	
d	C	C	C	C	C	C	A	T	T	
e	C	C	C	C	C	T	G	C	A	
非匹配数	6	4	7	4	4	4	4	4	6	
平均非匹配	4.3									
杂合基因型	ab	ac	ad	ae	bc	bd	be	cd	ce	de
非匹配数	5	3	5	5	5	5	7	2	2	4
平均非匹配	4.3									

同义突变和非同义突变

- 从64种三联密码子与20种氨基酸的对应关系（Hartl and Jones 2005）可以看到，密码子中第3个碱基的变化，大多不影响最终的生化合成产物。可以预期，前表中的9个SNP，大多属于同义多态性（synonymous polymorphism），同义多态性不会引起蛋白质序列上氨基酸的替换。
- 与同义多态性相对应，如果DNA在序列上的差异引起了蛋白质在氨基酸序列上的替换，这样的多态性称为非同义多态性（nonsynonymous polymorphism）。

DNA序列多态性的度量

- 对前表5个等位基因序列进行成对比对，相当于对所有可能杂合基因型携带两个等位基因的DNA序列进行比对，比对的结果可以得到碱基的非匹配数（nucleotide mismatches），列于前表最后三行。
- 例如，等位基因*a*和*b*在5个位置上存在非匹配，*a*和*c*在3个位置上存在非匹配，*a*和*d*在5个位置上存在非匹配等等，由此得到平均的非匹配数Π=4.3。

DNA序列多态性的度量

- 一般情况下，如果有n条DNA序列，多态性位点有S个，多态性位点上A、T、C、G 4种碱基所在的序列数用n_A、n_T、n_C、n_G表示，下面的公式给出平均非匹配数Π的一般计算方法。

$$\Pi = \frac{2}{n(n-1)S}\sum(n_A n_T + n_A n_C + n_A n_G + n_T n_C + n_T n_G + n_C n_G)$$

- 后面将会看到，多态性位点数S和平均非匹配数Π这两个参数，在中性突变理论的研究和检验中起重要作用。

§5.2 基因融合和基因树

- §5.2.1 几何分布及其性质
- §5.2.2 基因融合模型

负二项分布

- 一次Bernoulli试验中，记事件A发生的概率为p（$0<p<1$）。二项分布B(n, p)给出了n次独立Bernoulli试验中，事件A发生次数k的概率。
- 有时关心的可能不是固定次数试验中事件A发生多少次的概率，而是事件A发生k次需要的试验次数。这时，事件A的发生次数k是一个固定的数值，试验次数则是一个随机变量。为了与中的n区分，这里的试验次数用T表示，服从的分布称为负二项分布（negative binomial distribution），T 的取值范围是大于或等于k的所有正整数。

几何分布

- 事件A发生1次的试验次数T是负二项分布的一种特例，称为几何分布（geometric distribution），用符号$G(p)$表示。
- 在$T=t$时事件A发生了1次的概率，等于前t-1次试验中事件A均未发生的概率 与第t次试验中事件A发生的概率 的乘积。因此得到几何分布$G(p)$的概率计算公式。

$P(T=t)=p(1-p)^{t-1}$，其中t=1, 2, …，$0<p<1$

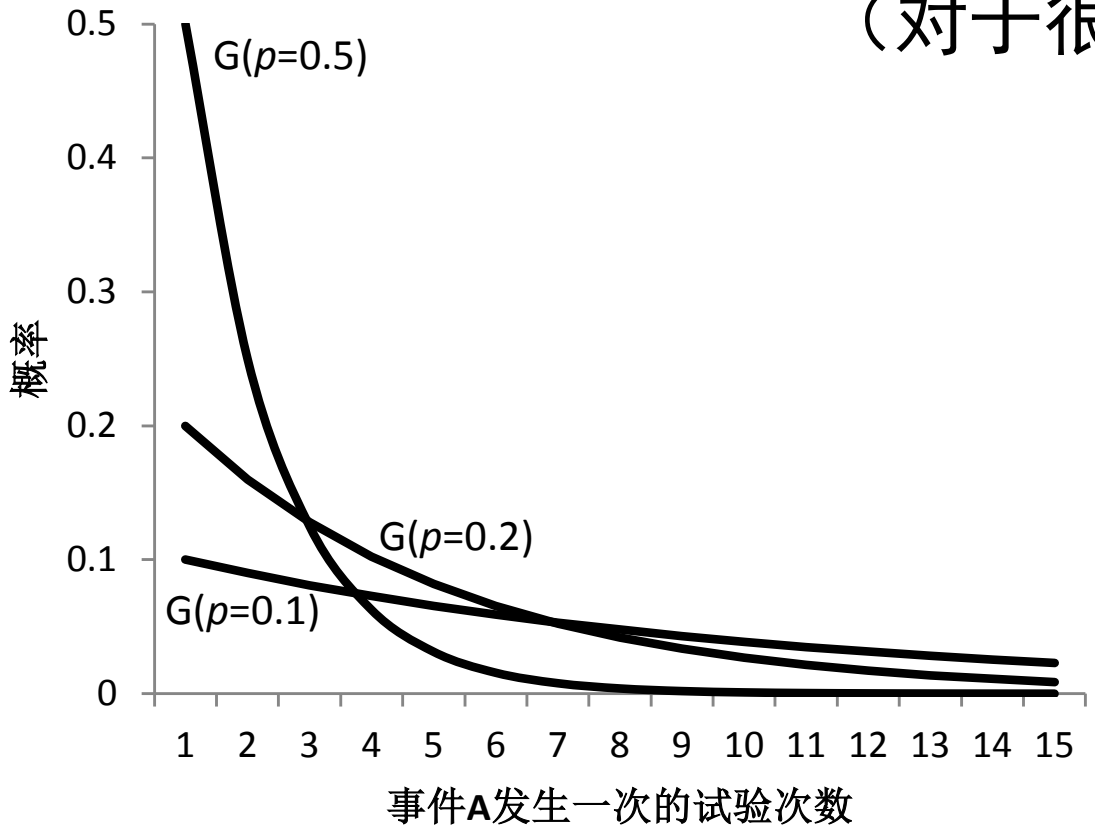
几何分布的性质

- 几何分布具有无记忆性，即$T=t$时事件A是否发生，与这个时间之前事件A发生与否、发生了多少次没有关系。
- 每次试验中，事件A是否发生都服从同一个Bernoulli分布。同时，$T=t$时事件A发生1次的试验次数，与这个时间之前事件A发生的次数也没有关系。

几何分布的期望和方差

$$E(T)=\frac{1}{p} \qquad V(T)=\frac{1}{p}\left(\frac{1}{p}-1\right)\approx\frac{1}{p^2}$$

（对于很小的p）



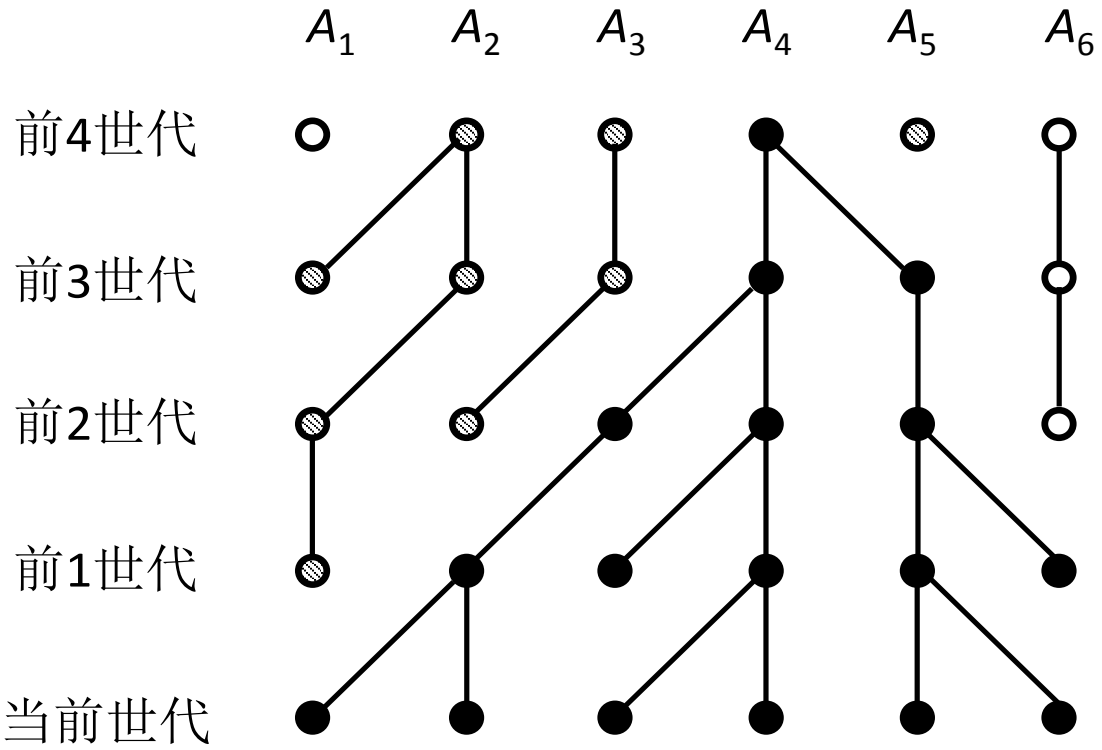
基因在祖先世代中的融合

- 第3章的Wright-Fisher模型，是一种从前向后或由过去推测现在的基因谱系研究方法。还有一种从后向前或从现在推测过去的研究方法，有时也同样重要。
- 当前群体的一组有限样本不仅可以用来估计基因和基因型频率，群体中的两个等位基因还可能是几个世代前的同一个基因传递而来，也就是说它们具有共同的祖先，称这两个基因在祖先世代中发生了融合（coalescence）。

基因融合模型

- 例如，当前群体的一组样本中包含k个基因，如果之前的某个祖先世代中发生了一次基因融合，那么当前的k个基因就只有k-1个亲本来源。发生两次基因融合后，当前的k个基因就只有k-2个亲本来源。如此下去，发生k-1次基因融合后，当前的k个基因就只有一个亲本来源。发生一次基因融合，称为一个融合事件（coalescent event）。
- 基因融合模型通过研究融合事件之间的时间间隔，提供了一种研究基因进化的方法，有时还会比从前向后的Wright-Fisher模型更加便利。

理想群体中6个等位基因的谱系图



- 前4世代的6个等位基因用A_1~A_6表示。前1世代中发生了3次融合事件，前2世代中没有发生融合，前3世代中发生了1次融合事件，前4世代中发生了1次融合事件。经过5次基因融合事件后，当前世代的6个等位基因均来自前4世代的同一个基因A_4。

基因融合模型的优点

- 从前面的基因谱系图还可以看到，如采用从过去到现在的研究方法，需要追踪祖先群体中所有6个等位基因的传递过程。
- 如采用从现在到过去的研究方法，则只需追踪当前群体中固定下来的等位基因A_4，即图5.2中实心圆表示的17个基因，而不需要考虑在当前世代中已经丢失的那些基因。
- 因此，基因融合模型有时显得更为有效、更为便利。

两个基因融合发生的时间T_2

- 在大小为N的理想群体中，如果当前世代的两个基因来自前1世代的同一个基因，说明在前1世代发生了融合事件，发生概率为$1/2N$，用下面的公式表示

$$\Pr\{T_2 = 1\} = \frac{1}{2N}$$

- 时间T_2 服从几何分布$G(p=1/2N)$，概率分布为

$$\Pr\{T_2 = t\} = \frac{1}{2N}\left(1 - \frac{1}{2N}\right)^{t-1}$$ ，其中t=1，2…

两个基因融合时间T_2的期望和方差

- 时间T_2 服从几何分布$G(p=1/2N)$，概率分布为

$$\Pr\{T_2 = t\} = \frac{1}{2N}(1-\frac{1}{2N})^{t-1}$$ ，其中t=1，2…

- T_2 的期望和方差分别为

$$E(T_2) = 2N \quad V(T_2) = 2N(2N-1)$$

k个基因的融合

- k个基因在前一个世代没有发生基因融合的概率为

$$(1-\frac{1}{2N})(1-\frac{2}{2N})\cdots(1-\frac{k-1}{2N})$$

- 当$2N$较大、k较小时，

$$\prod_{i=1}^{k-1}(1-\frac{i}{2N})\approx 1-\sum_{i=1}^{k-1}\frac{i}{2N}=1-\frac{k(k-1)}{4N}$$

- k个基因在前一个世代发生基因融合的概率为

$$\Pr\{T_k=1\}=1-\prod_{i=1}^{k-1}(1-\frac{i}{2N})\approx\frac{k(k-1)}{4N}$$

k个基因融合时间T_k的期望和方差

- 时间T_k 服从几何分布$G(p=k(k\text{-}1)/4N)$，概率分布为

$$\Pr\{T_k = t\} = \frac{k(k-1)}{4N}[1-\frac{k(k-1)}{4N}]^{t-1}$$ ，其中t=1，2…

- T_k 的期望和方差分别为

$$E(T_k) = \frac{4N}{k(k-1)} \qquad V(T_k) = \frac{(4N)^2}{[k(k-1)]^2}$$

k个基因发生k-1次融合的时间T_1

- 对于当前群体中的k个基因，经过k-1次融合之后，这k个基因就可以追踪到同一个祖先基因，需要的时间自然也是一个随机变量，用T_1表示。
- 第1次融合时间的随机变量为T_k。第一次融合发生后，等位基因个数减少到k-1。因此，第2次融合时间的随机变量为$T_{k\text{-}1}$。依次类推，最后一次融合时间的随机变量为T_2。
- 这些随机变量服从前面定义的几何分布。因此，融合为一的时间T_1正好是这k-1次融合时间随机变量之和。

k个基因完全融合的时间T_1

- k个基因融合为一时间T_1的分布

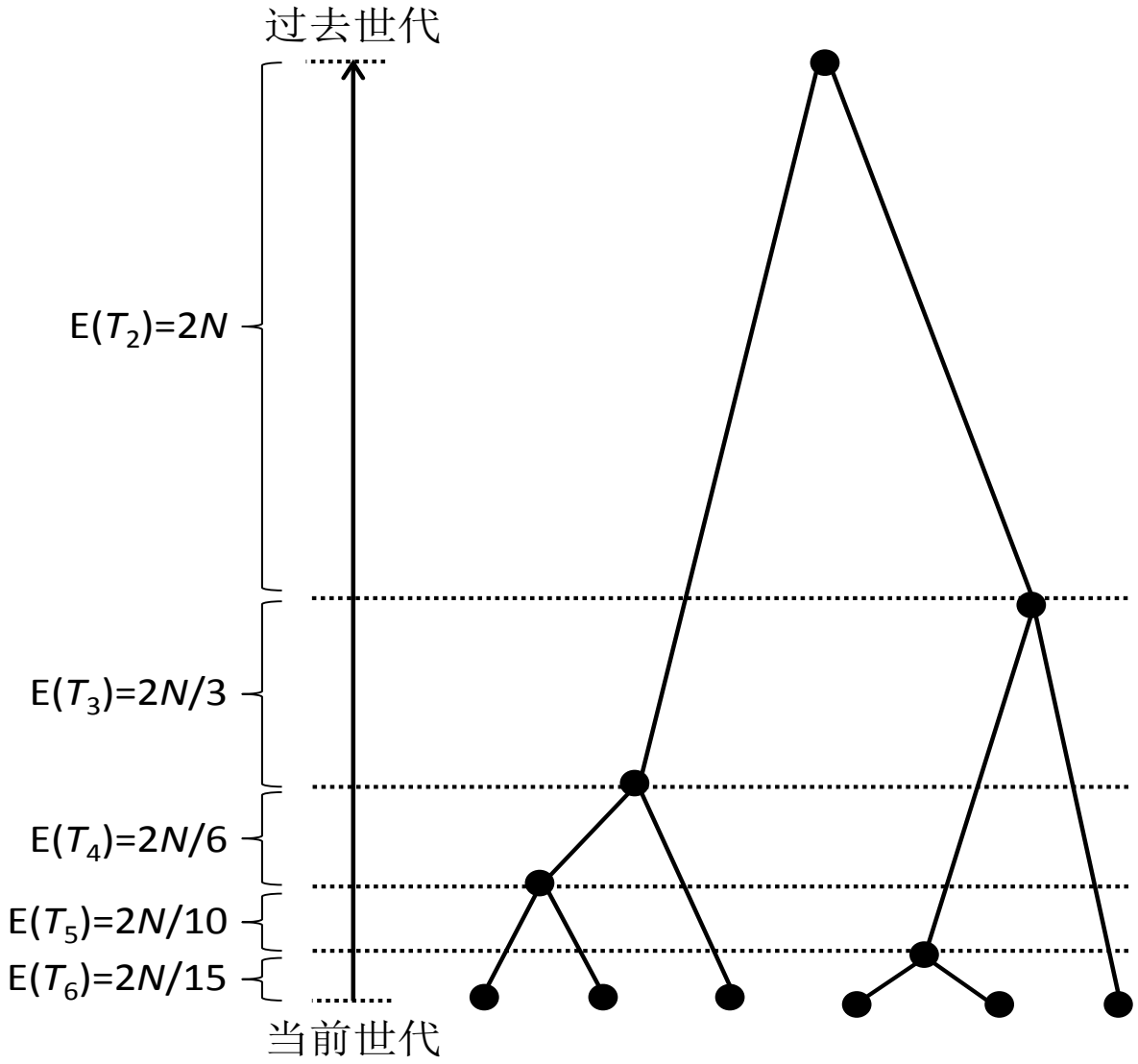
$$T_1 = T_k + T_{k-1} + \cdots + T_2 = \sum_{i=2}^{k} T_i$$

- T_1的期望和方差分别为

$$E(T_1) = \sum_{i=2}^{k} E(T_i) = 4N \sum_{i=2}^{k} \frac{1}{i(i-1)} = 4N \sum_{i=2}^{k} (\frac{1}{i-1} - \frac{1}{i}) = 4N(1 - \frac{1}{k})$$

$$V(T_1) = \sum_{i=2}^{k} V(T_i) = 16N^2 \sum_{i=2}^{k} [\frac{1}{i(i-1)}]^2$$

当前群体中6个基因的融合过程



基因树及其特性

- 前图表示的基因融合过程也称为基因树（gene tree）。在基因树中，越往上走，需要更长的时间才能再次出现基因融合事件。或者说，越往根部走（倒看），分支变得越长。从T_1的期望公式可以看出，当k较大时，融合为一的平均时间大约是$4N$。这个时间也正好是一个新突变基因、如果不丢失在群体中被固定下来的时间。

基因树及其特性

- 最后一次融合的平均时间是$2N$，几乎占整个融合时间的一半。这一现象其实也可以从融合后基因频率的变化来解释。在当前世代中，6个等位基因各占1/6，频率都比较低，随机漂移中丢失一个的概率就很大，所以第一次融合的时间最短。在群体大小稳定不变的情况下，随着丢失基因的增多，未丢失基因的频率逐渐升高，基因再次丢失的速度就会越来越慢，因此再一次融合的时间就变得越来越长。

基因树中所有分支的总长度

- 以前面的基因树为例，可认为这个基因树中包含6个长度为T_6的分支、5个长度为T_5的分支、4个长度为T_4的分支、3个长度为T_3的分支、2个长度为T_2的分支。这些分支可以从图中相邻平行线之间的实线段看出来。因此，得到所有分支总长度的计算公式。

$$E(T)=E[\sum_{i=2}^{k}iT_i]=\sum_{i=2}^{k}iE(T_i)=4N\sum_{i=2}^{k}\frac{1}{i-1}=4N\sum_{i=1}^{k-1}\frac{1}{i}$$

基因融合模型的作用

- 分支的总长度其实就是经历的总世代数，如果每个世代的突变频率为u，那么在无限等位基因模型下，分支总长度乘以突变频率，就等于多态性位点的个数S，即

$$E(S) = uE(T) = 4Nu\sum_{i=1}^{k-1}\frac{1}{i} = \theta\sum_{i=1}^{k-1}\frac{1}{i}$$

- 利用基因融合模型可以研究亲缘关系、估计一个座位上的等位基因个数、选择对该座位的影响、估计突变频率等等。

§5.3 中性突变理论

- §5.3.1 中性突变与有限随机交配群体
- §5.3.2 等位基因个数的估计
- §5.3.3 中性突变理论的*Tajima D*检验
- §5.3.4 迁移和突变的联合作用

中性突变

- 分子遗传学和进化的大量研究结果表明，新发生的突变中，大多是有害的，只有极少数是有益的。由于选择的作用，有害或有益的突变能够很快在群体中消失或固定，隐性有害基因即使不能完全消失，它存在的频率也不会太高。
- 有害或有利突变对多态性的贡献都很有限。从§5.1看到，基因组中存在一部分没有明显表型效应、可以忽略选择作用的突变，称为中性突变（Neutral Mutation）。

中性理论

- 中性突变在随机漂移的作用下，会维持一种平衡状态，这种平衡能够解释自然群体中观察到的大部分多态性。这一解释群体多态性的理论称为中性理论（Neutral Theory）。
- 中性理论认为，相当一部分突变对个体的生存和繁殖几乎没有影响，它们在群体中存在的频率不是由选择决定的，而是随机漂移的结果。

中性突变与有限随机交配群体

- 在无限等位基因模型的假定下，中性突变每次在群体中产生出一个新的等位基因，随机漂移决定这些基因的命运是固定还是丢失。经过长期的随机漂移，单个亚群体最终被固定在一个基因上。最初的$2N$个基因有着相同的固定概率，即各占$1/2N$，这也是一个特定突变基因在小群体中的固定概率。因此，一个新突变基因的丢失概率为$1-1/2N$。
- 尽管丢失的概率要远高于固定的概率，长期的突变和随机漂移也能达到一种平衡状态。处于平衡状态时，突变产生的基因正好可以弥补因漂移而丢失的基因。虽然这些基因在表型性状的遗传研究中，可能不算很合适，却包含着物种间、亚群体间在DNA序列上的大量亲缘系谱信息。因此，中性理论在现代群体遗传和进化研究中占有十分重要的位置。

中性突变与遗传漂变中的近交系数

- 假定突变的发生频率为u。下面的公式给出理想群体相邻世代间近交系数的关系。

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

- 对于后裔同样 的两个基因来说，不论是其中的一个发生了突变，还是二者同时发生突变，它们将不再是后裔同样。因此，在发生突变的情况下，维持后裔同样的概率等于$(1-u)^2$，即只有两个后裔同样基因均未发生突变时，才能继续保持它们之间的后裔同样状态。

中性突变与遗传漂变的平衡近交系数

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

- 从上面的公式可以看到，由于突变的存在，最终的近交系数不再趋近于1，而是一个小于1的数。
- 达到平衡状态的近交系数的计算如下

$$\frac{1}{2N} + (1 - \frac{1}{2N})\tilde{F} = \frac{1}{(1-u)^2}\tilde{F} \approx (1+2u)\tilde{F}$$

$$\tilde{F} = \frac{1}{1+4Nu} = \frac{1}{1+\theta}$$

$\theta = 4Nu$ 是群体遗传学中的另一个十分重要的参数

平衡群体的杂合度

- 在突变和漂变达到平衡状态时，前面的平衡近交系数表示两个基因的后裔同样概率。在无限等位基因模型下，两个状态相同但没有亲缘关系的基因结合在一起产生纯合基因型的概率为0。
- 从另外一方面讲，纯合基因型中的两个基因一定是后裔同样的。因此，平衡近交系数也就等于平衡群体中纯合基因型的频率。群体中的杂合基因型频率，即杂合度H与平衡近交系数的关系是

$$H = 1 - \tilde{F} = \frac{\theta}{1+\theta} = \frac{4Nu}{1+4Nu}$$

- 因此，可以从杂合度H估计$\theta=4Nu$这一重要参数。

参数$\theta=4Nu$的估计

- 假定一个座位上有k个等位基因，用p_i表示等位基因i的频率，$i=1, 2, \ldots, k$。

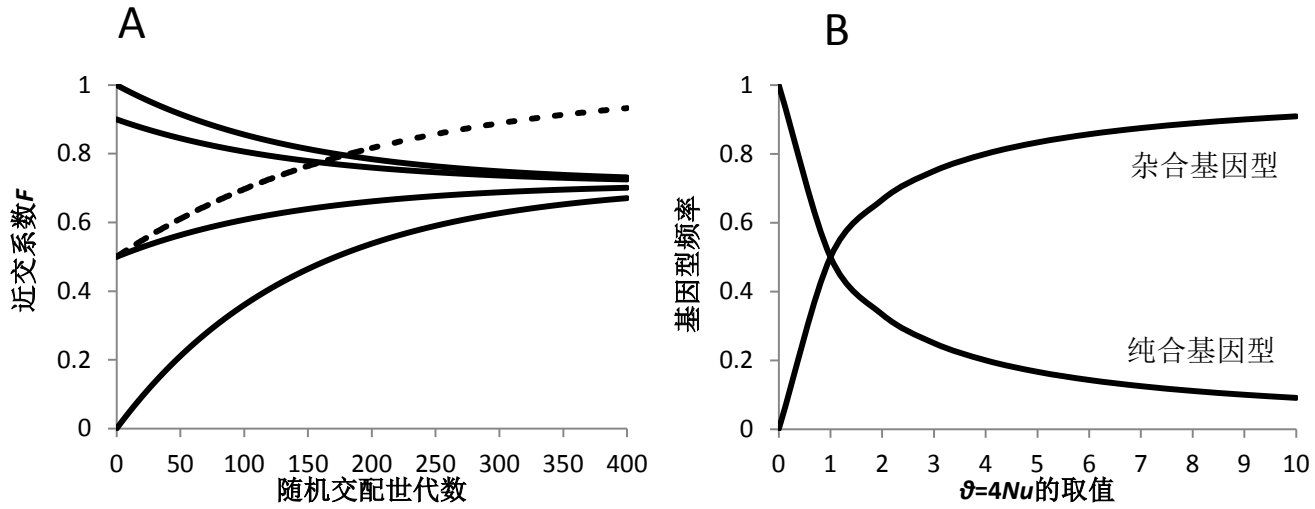
$$\tilde{F} = \sum_{i=1,\cdots,k} p_i^2 \qquad 1-\tilde{F} = H = 1-\sum_{i=1,\cdots,k} p_i^2$$

$$\theta = \frac{1-\tilde{F}}{\tilde{F}} = \frac{1-\sum_{i=1,\cdots,k} p_i^2}{\sum_{i=1,\cdots,k} p_i^2}$$

例子

- 例如，一个座位上存在4个等位基因，它们的频率分别为0.56、0.31、0.11和0.02。HW平衡时的纯合基因型频率为$0.56^2+ 0.31^2+ 0.11^2+ 0.02^2= 0.4222$，杂合度为$H$=1-0.4222 = 0.5778。因此，$\theta$= 0.5778/0.4222 = 1.3685。
- 由于$\theta=4Nu$，得到θ和N估计值后，就能估计突变频率。当然，得到θ和u估计值后，也就能估计第4章中介绍的有效群体大小N_e。

突变和漂移对群体结构的共同影响



- 图A中，N=100，u=0.001，θ=0.4，平衡时的近交系数=0.7143。四条实线自上而下代表1、0.9、0.5和0四种起始的近交系数，它们随世代的增加趋近于平衡近交系数，但趋近平衡值的速度是很缓慢的。图A还用虚线给出从0.5开始的无突变条件下近交系数的变化，这条虚线最终趋近于1。
- 图B为突变和漂移达到平衡状态时，不同θ取值对应的杂合基因型和纯合基因型的频率。一般来说，在0.2和0.8之间的杂合度已经很高了，它们对应的θ大约在0.25和4之间。

突变和漂移达到平衡状态的特点

- 在突变和漂移相互作用下达到平衡状态的群体中，突变的作用并不会因此而停止，仍然会不停地产生出新的等位基因；随机漂移的作用也不会因此而停止，已有的等位基因仍然会因为漂移而丢失，甚至原来已经被固定下来的基因也有可能丢失掉，原来已经丢失的基因也有可能被固定下来。
- 这一点从前图A可以明显看出，当突变和随机漂变两种因素同时存在时，近交系数为1的亚群体并不能永远保持在$F=1$的状态。由于突变和漂移的联合作用，近交系数高于平衡频率时就会不断下降。在近交系数下降的过程中，原来被固定下来的基因就又可能丢失，原来丢失的基因又可能被固定。

突变和漂移达到平衡状态的特点

- 突变和漂移的相互作用最终达到一种动态平衡状态，群体中的各种等位基因仍处在变动之中，单个等位基因的频率也仍在变化之中。
- 由于突变产生的基因正好弥补了因为漂移而丢失的基因，群体中等位基因的个数、等位基因的频率分布、纯合基因型的频率、杂合度等保持相对稳定。

平衡状态下的等位基因个数

- 假定一个突变和漂移平衡群体中，某座位上有k个等位基因，按照频率从高到低的顺序排列，最常见等位基因的频率用p_1表示、接下来的用p_2表示等等。一个等位基因是否最常见，也不是一成不变的。
- 等位基因频率相等时的近交系数最高。换句话说，只有在等位基因频率存在差别时，才会有较低的近交系数，群体才会有较高的杂合度。因此，对于特定的平衡群体，将k个等位基因频率按照从高到低的顺序排列时，就会表现出一定的规律性。正是这种规律性，提供了一种检验中性突变理论的方法。

平衡状态下的等位基因构成

- 假定从一个突变和漂移的平衡群体中，随机抽取大小为$n=2N=20$的一组基因样本，调查共发现$k=10$个等位基因，一个出现了7次、一个出现了3次、两个出现了2次、其它六个各出现了1次。在这个样本群体中，按照从高到低顺序排列的等位基因观测个数，称为等位基因构成（allelic configuration）。
- 在中性突变和无限等位基因模型下，等位基因构成呈现出一定的规律性。这种规律性可以用来检验中性理论在一个自然群体中的适用性。

等位基因个数的估计

- Kimura and Crow (1964) 给出了等位基因个数相对于等位基因频率的概率密度函数公式：

$$\Phi(x) = \theta(1-x)^{\theta-1}x^{-1}$$

$\Phi(x)\mathrm{d}x$ 表示频率在x和$x+\mathrm{d}x$之间的等位基因个数

- 基因样本量为n（$=2N$，N为二倍体个体数）的群体中，等位基因个数k的期望为：

$$E(k) = \int_{\frac{1}{n}}^{1} \Phi(x)dx = \theta\int_{\frac{1}{n}}^{1} (1-x)^{\theta-1}x^{-1}dx$$

特殊条件下的等位基因个数

- θ=1时： $E(k \mid \theta = 1) = \ln n$
- θ=2时： $E(k \mid \theta = 2) = 2(\ln n + \frac{1}{n} - 1)$
- Ewens（1972）近似公式：

$$E(k) = \frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \cdots + \frac{\theta}{\theta+n-1}$$

不同大小样本群体中的平均等位基因个数

$2N$	$\theta=4Nu$							
	0.1	0.5	1	2	5	10	1	2
10	1.27	2.13	2.93	4.04	5.84	7.19	2.30	2.81
20	1.34	2.48	3.60	5.29	8.46	11.33	3.00	4.09
50	1.43	2.94	4.50	7.04	12.46	18.34	3.91	5.86
100	1.50	3.28	5.19	8.39	15.72	24.44	4.61	7.23
250	1.59	3.74	6.10	10.21	20.17	33.07	5.52	9.05

样本群体中观测等位基因构成的概率

- Karlin和McGregor（1972）给出了样本群体中观测等位基因构成概率的计算方法：

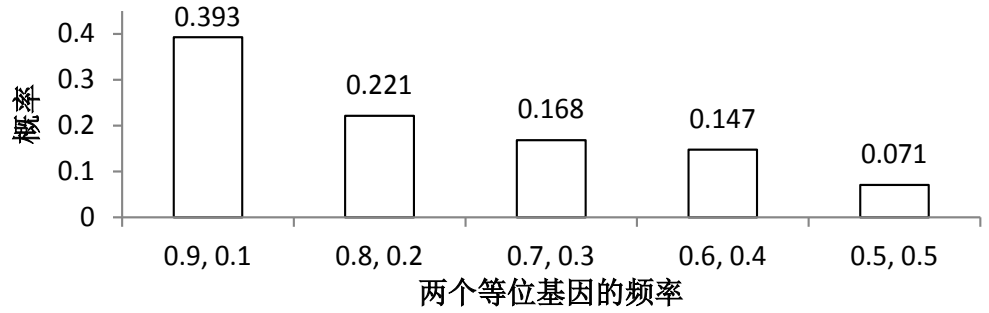
$$\Pr\{n_1, n_2, \cdots, n_k\} = \frac{n!\theta^k}{k! n_1 n_2 \cdots n_k S_n(\theta)}$$

- 其中， $n=2N$，k为观察到的等位基因个数，$n_1, n_2, \ldots, n_k$为各种等位基因的观测个数，

$$S_n(\theta) = \theta(\theta+1)(\theta+2)\cdots(\theta+n-1)$$

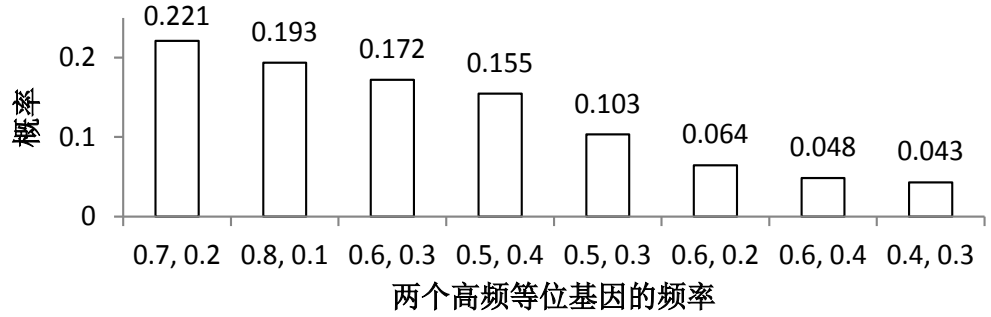
例子：N=5，n=10

A



两个等位基因频率均等的概率最低

B



三个等位基因频率均等的概率最低

- 不考虑等位基因的顺序，两个等位基因存在时只有5种可能的频率，三个等位基因存在时只有8种可能的频率。
- 图B中，第三个等位基因的频率等于1减去前两个的频率之和，在X轴上未给出第三个的频率。

利用等位基因构成检验中性理论

- 如果样本群体的观测频率与理论频率有较大的差异，如高频等位基因的频率过高或过低，中性理论可能不适合被考察的群体。高频等位基因的频率过高，说明这个等位基因对选择可能是有利的；高频等位基因的频率过低，说明这个等位基因可能对选择是不利的。
- 实际数据的样本量n都很大，可能的等位基因个数$n_1, n_2, \ldots, n_k$非常多，难以通过穷举的方法计算理论频率，一般都是采用计算机模拟的方法获得平衡群体的理论频率。

参数θ的两种估计方法

$$E(S) = uE(T) = 4Nu\sum_{i=1}^{k-1}\frac{1}{i} = \theta\sum_{i=1}^{k-1}\frac{1}{i}$$

$$V(S) = \sum_{i=1}^{n-1}[\frac{\theta}{i} + (\frac{\theta}{i})^2]$$

- 因此得到θ的一个无偏估计

$$\hat{\theta} = \frac{S}{a}\text{，其中 } a \hat{=} \sum_{i=1}^{n-1}\frac{1}{i}$$

参数θ的两种估计方法

- Tajima（1983）给出：

$$E(\Pi) = \theta$$

$$V(\Pi) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2$$

- 因此得到θ的另一个无偏估计

$$\hat{\theta} = \Pi$$

中性突变理论的Tajima *D*检验

- 在选择的作用下，有害突变在群体中的存在频率会很低。Tajima（1989）指出，在计算多态性位点的个数时，没有考虑等位基因频率高低这一因素，会放大有害突变对多态性位点个数的影响。另外，在计算非匹配平均数时，考虑到等位基因频率这一因素，低频有害突变的存在对非匹配平均数的影响不会很大。

Tajima *D*检验统计量

- 参数θ两个估计值之间的差异，在某种程度上反映了选择的效应，显著差异预示着选择的存在。因此，二者之间的差异可以用于中性理论的检验。这一检验方法称Tajima *D*检验，自1989年提出之后得到广泛应用。

$$D = \frac{\Pi - S/a}{\sqrt{V(\Pi - S/a)}}$$

例子（表5.1中的5条DNA序列）

- 多态性位点数S=10，非匹配数Π=4.3。
- a=1+1/2+1/3+1/4=2.0833，S/a=4.32。
- 两种估计值之间的差别不是很大。因此，可以认为这个基因座位上表现出的多态性符合中性理论。
- 需要说明的是，这里只是用了一个很小的样本，实际数据的样本要大得多。

迁移和漂变的联合作用

- 在图2.3表示的大陆群体向岛屿群体单向迁移模型中，假定岛屿群体有固定大小N，每个世代的迁入个体比例为m，即迁入个体数等于Nm。分别用F_t和$F_{t\text{-}1}$表示当前世代t和前一世代t-1的近交系数。无迁移发生时，公式3.31给出了相邻两个世代近交系数之间的关系。即：

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

迁移和漂变的联合作用

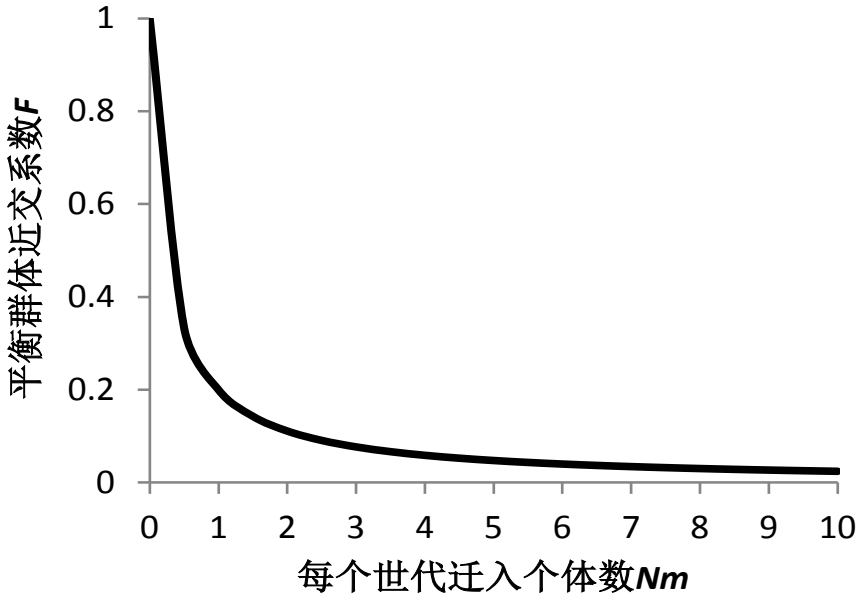
- 当迁移存在时，只要两个基因中有一个是迁移而来，它们就不再是后裔同样。换句话说，要保持两个基因的后裔同样状态，必须要求它们都不是迁入过来的，该事件发生的概率为$(1-m)^2$。因此，公式3.31的右端项乘以$(1-m)^2$之后就得到当前世代t的近交系数，即

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

迁移与突变之间的相似性

- 突变和漂变的共同作用 $F_t = [\frac{1}{2N} + (1 - \frac{1}{2N})F_{t-1}](1-u)^2$
- 迁移和漂变的共同作用 $F_t = [\frac{1}{2N} + (1 - \frac{1}{2N})F_{t-1}](1-m)^2$
- 对比发现，除了一个地方是突变频率u、一个地方是迁移比例m外，两个公式完全相同。迁移与漂变的平衡近交系数为 $\tilde{F} = \frac{1}{1+4Nm}$
- 与突变类似，纯合基因型的两个等位基因一定是后裔同样，上面公式给出的近交系数，同时也等于群体中纯合基因型频率之和，因此有时也称为固定系数（fixation index）。

平衡群体的近交系数随迁入个体数的变化曲线



- 每个世代迁入0.25个个体，平衡群体的固定系数就从1下降到0.5；每个世代迁入1个个体，固定系数下降到0.2；每个世代迁入2个个体，固定系数下降到0.11。每个世代迁入5个或更多个个体时，固定系数就低于5%。也就是说，平衡群体的近交程度变得很低。
- 迁移对群体的影响与突变是类似的，只存在程度上的差异。迁移比例m往往远大于突变频率u，因此迁移对群体的影响程度要远大于突变。但最终的结果，都是降低了群体的固定系数，提高了群体的杂合度和遗传多样性。

可逆突变和漂变的联合作用

- 对于向前和向后同时发生的突变，分别用u和v表示两个方向的突变频率。这时，不论哪个基因发生了突变，也不论突变的方向是什么，都会打破两个基因的后裔同样状态。因此，维持两个后裔同样基因的概率为$(1-u-v)^2$。

$$F_t = [\frac{1}{2N} + (1 - \frac{1}{2N})F_{t-1}](1-u-v)^2$$

$$\tilde{F} = \frac{1}{1+4N(u+v)}$$

可逆突变、迁移和漂变的联合作用

- 如果同时还有迁移存在，每个世代迁入个体占的比例为m。这时，不论哪个基因发生了突变，也不论突变的方向是什么，不论哪个基因是迁移而来的，都会打破两个基因的后裔同样状态因此，维持两个后裔同样基因的概率为$(1-u-v-m)^2$ 。

$$F_t = [\frac{1}{2N} + (1 - \frac{1}{2N})F_{t-1}](1-u-v-m)^2$$

$$\tilde{F} = \frac{1}{1+4N(u+v+m)}$$

§5.4 近交系数计算方法小结

- §5.4.1 近交系数计算方法
- §5.4.2 近交系数应用中需要注意的一些问题

不同交配系统的近交系数

交配系统	群体大小	世代t与邻近世代近交系数的关系	平衡近交系数F
随机交配	无限大	$F_t = 0$	F=0
自交系统	任意	$F_t = \frac{1}{2} + \frac{1}{2}F_{t\text{-}1}$	F=1
随机交配	大小为N的理想群体	$F_t = \frac{1}{2N} + (1 - \frac{1}{2N})F_{t-1}$	F=1
随机交配	有效大小为N_e的非理想群体	$F_t = \frac{1}{2N_e} + (1 - \frac{1}{2N_e})F_{t-1}$	F=1
全同胞系统	任意	$F_t = \frac{1}{4}(1 + 2F_{t\text{-}1} + F_{t\text{-}2})$	F=1
亲子系统	任意	$F_t = \frac{1}{4}(1 + 2F_{t\text{-}1} + F_{t\text{-}2})$	F=1
半同胞系统	任意	$F_t = \frac{1}{8}(1 + 6F_{t\text{-}1} + F_{t\text{-}2})$	F=1
回交系统，A表示轮回亲本	任意	$F_t = \frac{1}{4}(1 + F_{\mathrm{A}} + 2F_{t\text{-}1})$	$F = \frac{1}{2}(1 + F_{\mathrm{A}})$
混合自交和异交，C为异交率	无限大	$F_t = \frac{1}{2}(1 + F_{t\text{-}1})(1 - C)$	$F = \frac{1-C}{1+C}$
阶梯结构群体（阶梯数为n）	$F_{ST} = 1 - (1 - F_n)(1 - F_{n-1})\cdots(1 - F_1)$，$F_i$为阶梯$i$的平均近交系数		

一种或多种遗传因素作用下，理想群体的近交系数，不考虑选择的作用

遗传因素	世代t与邻近世代近交系数的关系	平衡近交系数F
向前突变（频率为u）	$F_t=[\frac{1}{2N}+(1-\frac{1}{2N})F_{t-1}](1-u)^2$	$F=\frac{1}{1+4Nu}$
双向突变（频率分别为u和v）	$F_t=[\frac{1}{2N}+(1-\frac{1}{2N})F_{t-1}](1-u-v)^2$	$F=\frac{1}{1+4N(u+v)}$
迁移（频率为m）	$F_t=[\frac{1}{2N}+(1-\frac{1}{2N})F_{t-1}](1-m)^2$	$F=\frac{1}{1+4Nm}$
迁移（频率为m）与向前突变（频率为u）	$F_t=[\frac{1}{2N}+(1-\frac{1}{2N})F_{t-1}](1-m-u)^2$	$F=\frac{1}{1+4N(u+m)}$
迁移（频率为m）与双向突变（频率分别为u和v）	$F_t=[\frac{1}{2N}+(1-\frac{1}{2N})F_{t-1}](1-u-v-m)^2$	$F=\frac{1}{1+4N(u+v+m)}$

近交系数在群体遗传中的作用

- 从第3章开始，我们已经看到近交系数在群体遗传学研究中发挥的重要作用。通过近交系数这一概念，可以把自然条件下的非理想群体、人工控制条件下的规则交配系统转化为理想群体进行研究；在给定基因频率的条件下，通过群体的近交系数可以估计群体的基因型频率；在中性突变和无限等位基因模型下，通过群体的近交系数可以估计基因的突变频率和杂合度；在迁移和漂变作用下的平衡群体中，可以通过近交系数估计迁移比例和杂合度等等。
- 自然条件和人工控制条件下，不同交配系统产生的各种各样遗传群体，通过计算它们的近交系数，就能很多程度上认识这些群体的遗传构成。因此，近交系数有机地统一了各式各样的交配方式和遗传群体。

近交系数的相对性

- 与基因频率和基因型频率类似，近交系数也是一个基于群体的遗传参数。离开遗传群体谈论近交系数是没有意义的。
- 共祖先系数衡量一个遗传群体中个体之间亲缘关系的远近，近交系数衡量子代群体中两个等位基因亲缘关系的远近，二者在遗传上其实是一回事，即亲代群体的共祖先系数等于子代群体的平均近交系数。
- 由于近交系数是一个群体参数，严格地讲，谈论单个个体的近交系数也是没有意义的，除非后代群体中只有一种基因型。一般来说，后代都是多种基因型按照一定频率构成的群体，近交系数衡量的是后代携带两个基因的平均亲缘关系。

度量近交系数的基础群体

- 近交系数在群体和数量遗传学研究中均发挥重要的作用。应用中要注意，近交系数是一个相对的概念，离开一个基础群体或参考群体，也就难以衡量个体之间的亲缘关系，近交系数也就失去了意义。
- 基础群体大致可以分为自然交配和人工控制交配两大类型。如果基础群体来自自然条件下一个物种的种群，根据物种的繁殖方式还可分为以下4种类型。

自然无性系群体

- 单个无性系群体中，不同个体、以及它们的亲代和子代有着完全相同但高度杂合的基因型，用Z表示单个无性系群体中的个体，则$F_Z=0$。
- 在单个无性系群体内，个体之间的共祖先系数等于Z与其自身的共祖先系数。因此，个体间的共祖先系数$f_{ZZ}=(1+F_Z)/2=1/2$。如果无性系也能发生自交，则这个共祖先系数当然就是自交一代群体的近交系数。
- 多个无性系构成的自然群体等同于一个随机交配群体，可以将单个无性系等价地视为随机交配群体中的单个个体。因此，群体的近交系数为0，不同无性系之间的共祖先系数也为0。

自然自交群体

- 单个自交系群体中，不同个体、以及它们的亲代和子代，有着完全相同的纯合基因型。同时，纯合基因型中的两个等位基因还是后裔同样。用Z表示单个自交系群体中的个体，则$F_Z=1$，个体之间的共祖先系数$f_{ZZ}=(1+F_Z)/2=1$ 。
- 多个未知来源自交系构成的自然群体中，近交系数也为1，但是不同自交系之间的共祖先系数为0。

自然异交群体

- 一般认为群体无限大，用Z表示单个个体，则F_Z=0。任意两个个体X与Y之间的共祖先系数f_{XY}=0。个体Z与它自身的共祖先系数f_{ZZ}=(1+F_Z)/2=1/2。
- 如果发生自交，则这个共祖先系数也就是自交一代群体的近交系数。

自然混合自交和异交群体

- 一般认为群体无限大，用Z表示单个个体，C表示异交率，则$F_Z=(1-C)/(1+C)$，个体Z与它自身的共祖先系数$f_{ZZ}==(1+F_Z)/2=1/(1+C)$。
- 由于$F=(1-C)/(1+C)$是平衡点，亲代和子代的近交系数都应该等于 $(1-C)/(1+C)$ 。
- 把Z看作子代，则任何两个个体X与Y之间的共祖先系数为$f_{XY}=(1-C)/(1+C)$ 。

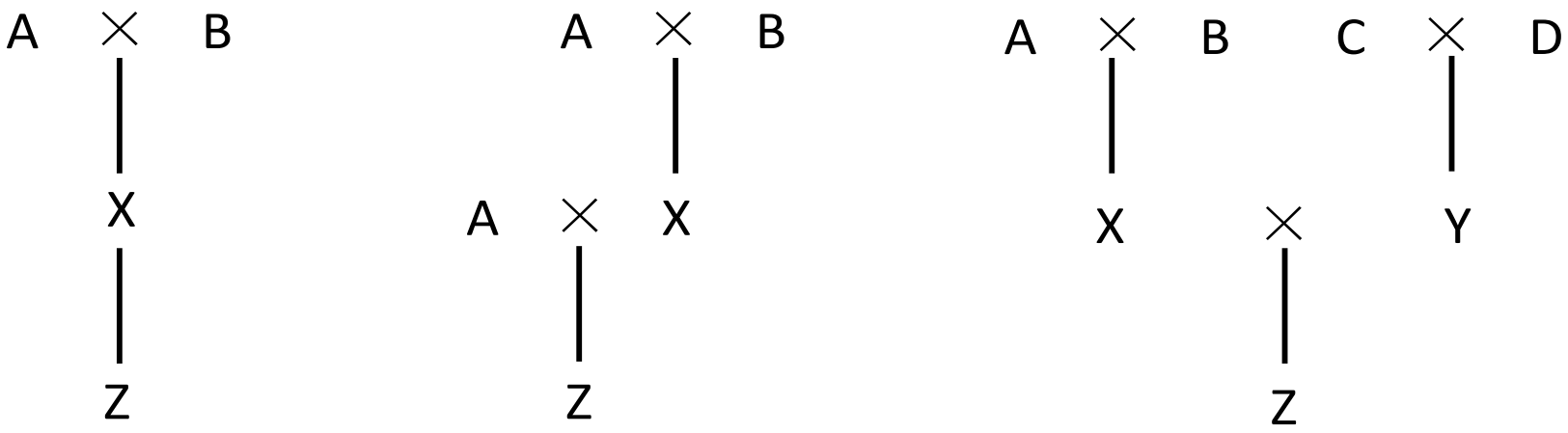
人工控制杂交产生基础群体

- 前面列出的4种自然群体类型，其实代表着自然界的4种繁殖方式。根据物种的生物学特性，人工控制条件下的繁殖方式更多（详见§1.3.1）。第4章介绍的一些规则近交系统，是常见的人工控制繁殖方式，这些交配系统中，近交系数的变化都有一定的规律性。
- 利用这些群体时，要对它们的近交系数有所了解，以便采取适当的分析方法，开展目的明确的遗传研究，获得更有价值的遗传研究结果。在规则近交系统中，组成世代0的个体是后续近交世代的原始亲本，这些亲本个体的来源就是基础群体。

自交系及其互交产生的群体

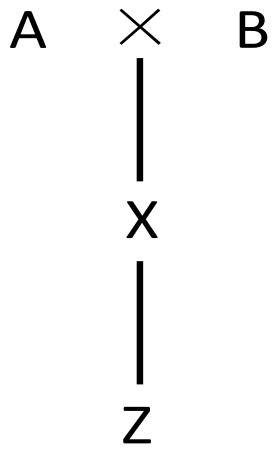
- 植物育种中经常利用自交系统产生大量的自交系，在此基础上开展品种选育或遗传研究。
- 有时又利用两个或多个来源广泛的自交系，开展相互杂交，利用互交群体开展遗传研究或轮回选择育种。
- 共祖先系数和近交系数之间的各种关系，在这些群体中也是适用的。

自交系之间互交产生的基础群体



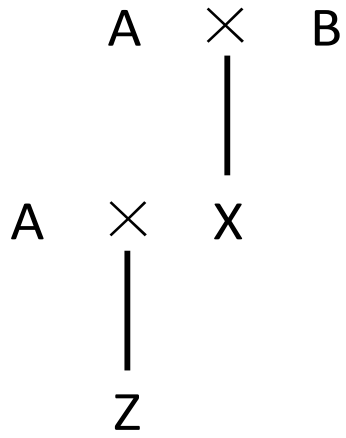
- 左边为两个自交系A和B杂交后代的系谱，X表示杂种一代F_1，Z表示杂种二代F_2。
- 中间为两个自交系A和B杂交和回交后代的系谱，X表示杂种一代F_1，Z表示回交一代BCF_1。
- 右边为4个自交系A、B、C、D杂交后代的系谱，X表示自交系A和B的杂种一代，Y表示自交系C和D的杂种一代，Z表示双交一代DCF_1。

杂种F_2群体的近交系数



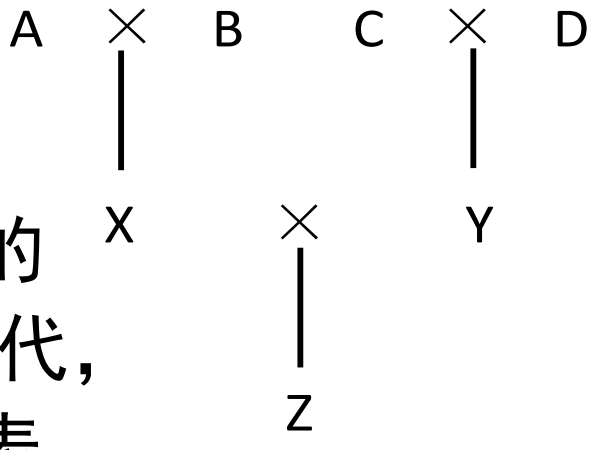
- 两个自交系A和B的杂交后代系谱，X表示杂种一代F_1，Z表示杂种二代F_2。显然，$F_A=F_B=1$。如果自交系A和B之间没有亲缘关系，则$f_{AB}=0$，$F_X=f_{AB}=0$。
- 群体X只有一种基因型，一般不适合开展遗传和育种研究。根据自交系A和B差异的大小，X的自交后代Z存在不同程度的基因型分离，适宜于遗传和育种研究。若作为基础群体，群体Z的近交系数$F_Z=f_{XX}=(1+F_X)/2=1/2$。

回交群体的近交系数



- 两个自交系A和B的杂交和回交的系谱，X表示杂种一代F_1，Z表示回交一代BCF_1。如果自交系A和B之间没有亲缘关系，则$F_A=F_B=1$，$f_{AB}=0$。
- 群体X与自交系A的回交后代Z存在不同程度的基因型分离，若作为基础群体，群体Z的近交系数$F_Z=f_{AX}=(f_{AA}+f_{AB})/2=1/2$。
- 可见，回交一代的近交系数等于自交一代的近交系数，它们其实都等于群体中纯合基因型的频率。

四交群体的近交系数



- 四个自交系A、B、C、D杂交后代的系谱，X表示自交系A和B的杂种一代，Y表示自交系C和D的杂种一代，Z表示双交一代。显然，$F_A=F_B=F_C=F_D=1$。
- 如果4个自交系没有亲缘关系，则$f_{AB}= f_{AC}= f_{AD}= f_{BC}= f_{BD}= f_{CD}=0$，$F_X=F_Y=0$。X与Y之间的共祖先系数$f_{XY}= (f_{AC}+f_{AD}+f_{BC}+f_{BD})/4=0$。因此，双交一代的近交系数$F_Z=f_{XY}=0$。
- 与两个自交系的杂种F_1代不同的是，双交一代中存在基因型的分离，可以用于遗传和育种研究。若作为基础群体，双交F_1群体Z的近交系数为0。