遗传多态性和变异

作者：

米歇尔·弗伊(Michel Veuille)，巴黎高等研究实践学院研究主任。

遗传多态性是指存在替代状态的 DNA，决定着生物体向更高水平整合的变异。生物体中存在不同种类的基因组修饰（突变），其中研究最多的是编码区和调控区中的核苷酸替换。

1. 定义

遗传多态性是指在一个种群中，在基因组或基因座的确定位置上（基因在染色体上的位置）存在好几种替代状态的 DNA 或**等位基因** 【在群体遗传学中，一组同源基因（同源类），如果两个基因与减数分裂相匹配，则它们是同源的】。这个定义有几个方面：

（1）首先，这种特征必须由**染色体**携带并具有遗传性；

（2）然后等位基因在基因组中的位置必须是同源的【如果两条染色体或两个基因在减数分裂过程中相互匹配并相互排斥，则称其为同源的】，这排除了稍后将讨论的某些类型的变异；

（3）但由于性状是可遗传的，位置同源性也意味着等位基因在血统上是同源的； 如果它们不同，那么在将它们与最后一个共同祖先联系起来的其中一个**品系**中（至少）发生了**突变**（图 1）；

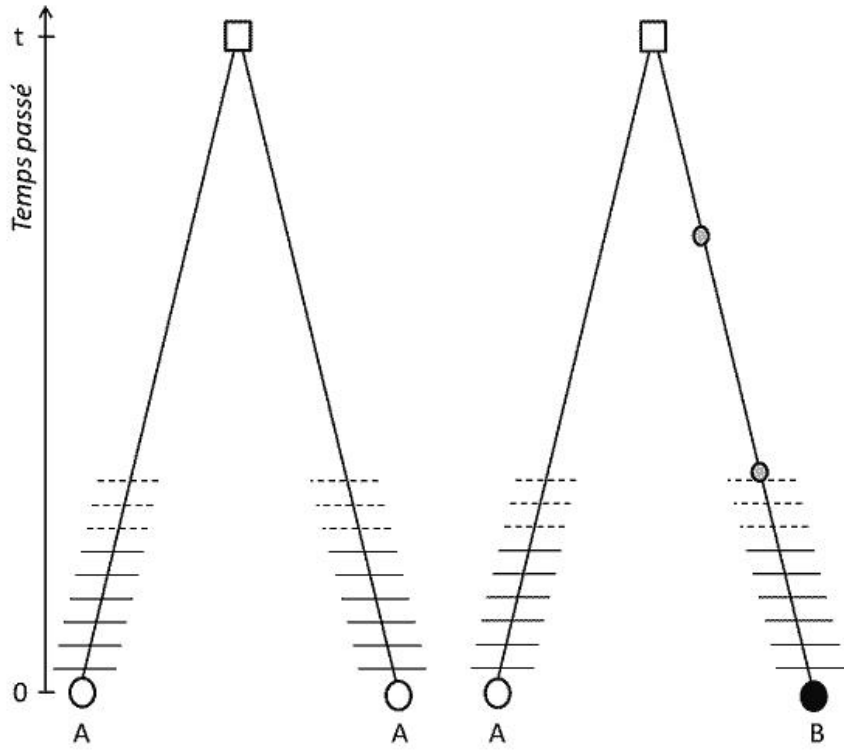

图 1. 等位性和同源性。表示两对基因通过血统线与它们的最后一个共同祖先（正方形）相连;水平的刻度代表世代，灰色圆圈代表祖先谱系的突变。如果两个基因有共同的祖先，那么它们就是同源的。它们在染色体上的位置是同源的，在后代上也是同源的。如果它们是不同的(图的右边)，它们就被认为是等位基因，这意味着自他们最后一个共同祖先的时代(t)开始，至少发生了一个突变。否则，他们被认为状态是相同的(图的左边)。两个随机基因间突变数的期望为Θ =2Tμ，其中μ为单位时间(代)的突变率，T 为种群中 t 的期望。从种群中随机选择的两个基因是等位基因，因此至少有一个突变的差异，其概率是 H≈Θ/（Θ+1）。

（图 1 Temps passé 耗费的时间）

（4）最后，遗传多态性可以在组成 DNA 的最小单位的尺度上定义：**核苷酸位点**。 因此，每个核苷酸变异都可以决定更高层次生物结构的多态性 - 个体的基因、蛋白质和表型——这些成为可以描述遗传多态性的尺度（图 2）。

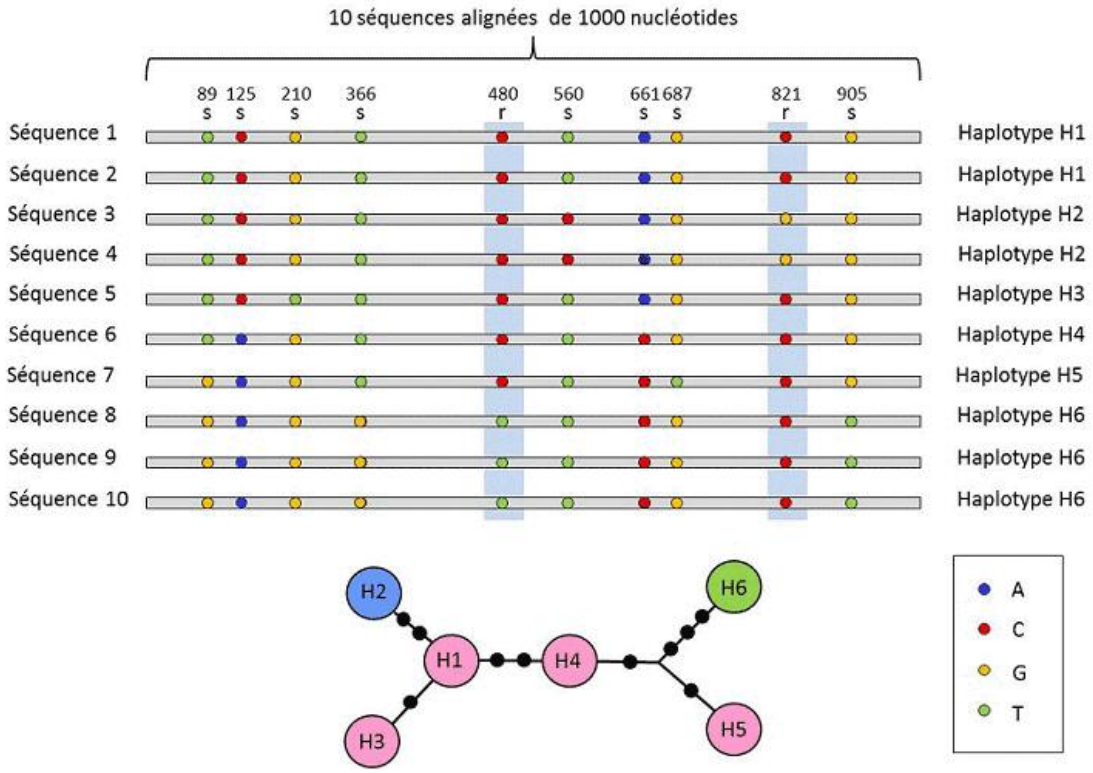

图 2. 核苷酸多态性和等位基因。示例显示编码蛋白质的 1000 个核苷酸的 10 个同源序列的比对。10 个可变位点（核苷酸 A、C、G、T）及其在序列上的位置都采用彩色标记。它们包括 8 个沉默位点（s）和 2 个替换位点（r：氨基酸替换位点，用蓝色标出），位于 480 和 821 位（因此有三个蛋白质的变体，用彩色标记标识）。有六个不同的单倍型（多态位点的线性排列，编号为 H1 至 H6）。这些单倍型按照一个无根树(图底)分组，由连接单倍型的分支组成，并携带 10 个突变(黑色圆圈)，对应于 10 个多态位点。该树是唯一的，因为没有基因内重组；否则，有多少被重组事件分开的片段就有多少树。在这个例子中，蛋白质的等位基因多样性为 H=0.42，单倍型多样性为 H=0.80；核苷酸多样性为π=0.00416（定义和公式见正文）。（图 2 10 séquences alignées de 1000 nucléotides 编码 1000 个核苷酸的 10 个同源序列；Séquence1-10 序列 1-10；Haplotype H1-H10 单倍型 H1-H10）

以 ABO 血型系统为例，它涉及到输血过程中抗原相容性【被抗体或淋巴细胞受体识别的部分抗原也称为抗原表位或抗原决定簇。同一个抗原有几个抗原表位(相同或不同)，从而引起不同的免疫反应】。这是一个存在于人类群体中的遗传多态性案例，涉及 ABO 糖基转移酶蛋白【将含有糖的残基转移到蛋白质中的酶。在 ABO 系统中，糖基转移酶 A 和 B 诱导的个体分别属于 A 型或 B 型。当两种糖基转移酶都存在时，个体属于 AB 型。】，它有三个等位基因，即 A、B 和 O。一个人可以有（AA）、（AO）；（BB）、（BO）；（OO）或（AB）基因型。它的表现型将是[A]、[A]、[A]；[B]、[B]；[O]或[AB]，在这里我们看到 A 和 B“支配”O（即它们的表达掩盖了 O 的表达），并且它们之间是共显性的（A 和 B 之间的杂合子具有可识别的表达，[AB]）

多态性可以首先在编码蛋白质的基因位点的 DNA 序列规模上加以描述。有些多态性是“同义的”，即它们不会改变蛋白质的氨基酸序列;它们通常是数量最多的多态性。其它的修饰氨基酸，称为“置换”多态性。ABO 糖基转移酶的置换多态性有两种类型:改变氨基酸而不改变抗原单位的多态性(它们不干扰个体的 ABO 表型);其他替代多态性决定了 ABO 表型【一组可观察到的个体特征】。

在核苷酸和等位基因水平之间，遗传分析考虑了中间水平的描述：基因座上可变位点的线性排列。这些称为单倍型的排列【位于同一染色体上不同位点的等位基因组通常一起遗传。单倍型（Haplotype）是英语词组单倍体基因型(haploid genotype)的缩写。位于同一染色体上的所有基因及其等位基因在减数分裂一起分离形成单倍型。 这些基因被称为“遗传连锁”的。】在进化遗传学中很有用，因为它们可以发现等位基因之间的系谱联系（图 2）。

核苷酸多态性也存在于基因间区（编码区之外）。有些影响到基因表达调控的区域，因此具有表型表达。其他的则没有已知的影响，被称为“沉默的”。

除核苷酸替换之外，一种特殊类型突变（**微卫星**突变)的多态性是指非编码 DNA 片段长度的变化(图 3)，这是由于短重复核苷酸序列的重复次数的变化，例如 CACACACA 或 TGTGTG。

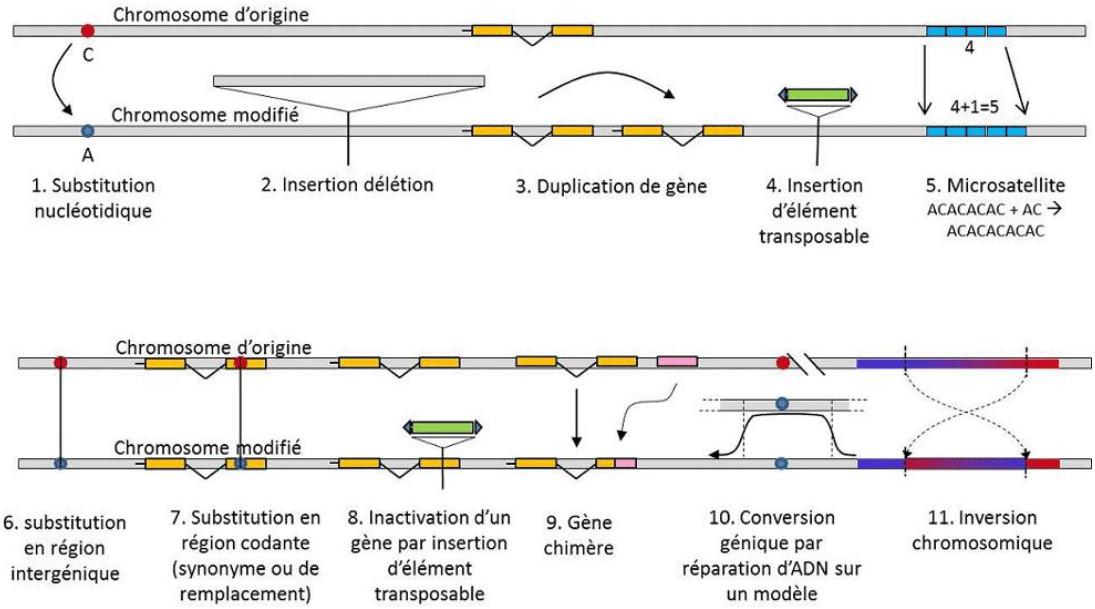

图 3. 不同类型的突变。1，核苷酸替换，影响碱基(A，C，G，或 T)；2，DNA 片段的插入或删除；3，基因的复制；4，插入转座子；5，微卫星位点的伸长或缩短；6，基因间区的替代(沉默或不沉默)；7，编码区替换(同义或氨基酸替换)；8，通过插入转座子使基因失活；9，通过结合两个编码区创造嵌合基因；10，以另一个等位基因为模型，通过修复受损 DNA

进行基因转换；11，染色体倒置。（图 3 Chromosome d'origine 原始染色体；Chromosome modifié 突变染色体）

并非所有的遗传变异都属于遗传多态性定义的范畴，因为它需要替代物的位置同源（图 3）。因此，有一些通过复制基因得到的重复序列，对它们来说，每个位置上的两个重复都不是同源的。通常，我们也不能谈论转座因子的同源性，因为它们一般在细胞世代中改变位置，可以成倍增加并侵入基因组。这样一来，两个转座因子的同源性就不能被确定。另一方面，同一基因座的两个编码序列，其中一个是完整的，另一个因转座因子插入而失活，这很符合同源情况。有一天，我们可能会考虑谈论所谓的 “表观遗传”染色体的修饰的多态性，这些修饰有时会在几代细胞中传播，包括体细胞【界定非生殖细胞或体细胞。影响体细胞基因的体细胞突变随着携带者个体消失而消失。】或生殖细胞【符合配子的条件。生殖细胞的突变可传给其子代。】，如甲基化。它们在种群进化中的重要性还有待评估。遗传多态性的概念仅限于某些类别的变异--本质上是核苷酸替换--由于后者在探索自然种群历史中的巨大效用。它们适用于进化的数学模型。

多态性这个词的意思是“多种形式”。它与**单态**相反，单态表示没有变异。在生物学词汇中，这种对立早在遗传学之前就被用来表示同一物种中几种不同类型个体的并存，例如群居昆虫的等级多态性(蚁后和工蜂)；北极一些哺乳动物的季节性多态性(毛皮变化)。这些案例不属于遗传多态性，这一较新的概念，具有更狭窄和更精确的涵义。连续变异(例如大小变异)也不属于多态性的范畴，因为它们不存在不同的替代。然而，影响大小的基因座属于这个定义。我们将在之后讨论研究最多的遗传变异——核苷酸替换，因为它们是进化中最重要的，然后我们将讨论表型变异。

2. 测量

仅仅说一个基因座变异性大或小，而不给这个判断一个定量评估是不够的。研究人员可以在不同的尺度上研究变化。如果他只对蛋白质的等位基因感兴趣，他将测量**等位基因多样性**，用“H”表示。 如果他对 DNA 多样性感兴趣，他将测量**核苷酸多样性**，用“π”来表示。

等位基因多样性 H 定义为在两次有替换的抽样中抽取两个不同等位基因的概率【在一个装有 n 个代币的盒子中进行连续的抽签，取第一个代币，读取其价值，将其放回盒子中，取第二个代币，读取其价值，将其放回盒子中，等等，直

到第 p 个代币。这意味着重复(可以多次选择相同的对象)并按顺序(选择对象的顺序很重要)在 n 个代币中选择 p 个代币。在 n 中连续抽出代币的次数为：$n \times n \times n \times ... \times n = n^p$。】。如果我们称 p_i 为等级 i 的等位基因的频率，结果表明两次抽取相同等位基因的概率为 $F = \Sigma p_i^2$。样本的等位基因多样性是其对 1 的补充，即:

$$H = 1 - \Sigma p_i^2 \quad (1)$$

该公式对于蛋白质等位基因和单倍型都适用，被称为**单倍型的多样性**。它也可以被称为**杂合度**，因为在二倍体位点的情况下，它给出了杂合子的预期频率。

核苷酸多样性π相当于每个核苷酸尺度上的 H。它的计算方法是将两两取样的样本中序列之间的核苷酸差异数的平均值（δ_{ij}），除以 DNA 片段的长度获得核苷酸的数量（L）。

$$\pi = \text{average } (\delta_{ij})/L \quad (2)$$ （参考文献[1]）

这个值因物种而异。在人类基因组的编码部分，从种群中随机选择的两条染色体平均每 1000 个核苷酸相差 1 个。在果蝇(*Drosophila melanogaster*)基因组中，这种差异大约是百分之一。因此，果蝇的可变性是人类的十倍。当这些值与编码区域的大小(果蝇大约有 15500 个基因，人类至少有 22000 个基因)，甚至基因组的大小(果蝇每个单倍体基因组有 1.4 亿个碱基对，人类大约有 32 亿个碱基对，比果蝇多 20 倍)相关联时，多态位点的数量是天文数字，导致任何物种中的有性一代出生的与过去、现在或将来的另一个在基因上没有相同之处。DNA 多态性的信息力量是巨大的。在法医学上，侦探可以通过 16 个微卫星位点来识别任何嫌疑人。

3.历史

“**变异**”一词出现在达尔文的著作《物种起源》(1859)的前两章的标题中。在生物遗传规律仍然是个谜的时候，达尔文将这一概念引入自然科学，并专门撰写了另一本重要著作《动植物变异》（1868）。他深信，进化是关于微小的变化，这些变化对生物体适应其生活环境的影响不大，因此他非常重视微小的数量变化，这促使他的继任者(尤其是卡尔·皮尔森)创立了生物统计学【生命测量科学。广义上指对生物的定量研究。】。但是 1900 年孟德尔定律的重新发现将人们的

兴趣转向了不连续变化。

从 1908 年到 1930 年，群体遗传学本质上是一门试图调和达尔文进化论与孟德尔遗传学【基于单一基因在显性、隐性或性染色体连接模式 X（或 Y）下的传递的遗传。指具有简单决定性的遗传性状，由一对或少数几对基因决定。】的理论学科，而且概率在其中发挥了主要作用。遗传学是反直觉的。它预测在后代中不存在亲本的**复制**，因为后代的基因型是由两个亲本半基因组融合前抽签式等位基因分离的结果。人们意识到，这些是代际间传递的等位基因频率【在种群中发现变异等位基因的频率。以比例或百分比表示。种群中一个基因的所有等位基因的等位基因频率之和因此定义等于 1。在群体遗传学中，等位基因频率代表种群或物种水平的遗传多样性。】，而不是基因型或表现型。这些频率从一代到下一代或多或少都是稳定的，除了个体之间的多重亲缘纠缠之外，还会**产生**相同的基因型频率【种群的遗传结构。由等位基因频率决定的】（图 4）。因此，与变异数相关联的群体基因型分布是唯一可预测的因素。1930 年左右，三位理论家罗纳德-费希尔、JBS-霍尔丹和苏厄尔-莱特帮助创建了孟德尔种群【遗传服从孟德尔定律的种群。】的概念[2]。在这种情况下，进化必然联系到三个结构层次：基因、个体和种群。这种三方面的关联可以用以下公式来概括：种群进化是等位基因频率的变化（基因尺度，也是群体尺度），这种变化取决于选择对表型的分类（个体尺度）。经验群体遗传学研究随后发展起来，但在很长一段时间内，由于我们对染色体的工作原理一无所知而受到限制（DNA 的结构在 1953 年被了解，其测序工作在 1977 年开始），不得不依赖研究少数可见的多态性，如瓢虫鞘翅或蜗牛外壳的色泽（见焦点——伟大的蜗牛辩论）。对于自然种群是普遍多态的还是单态的，以及多态性本身是否有益，遗传学家之间存在很多争论。在回答这些问题之前，对于大分子尺度变异的研究不得不等到 1966 年（对于蛋白质来说）和 1983 年（对于 DNA 来说）。

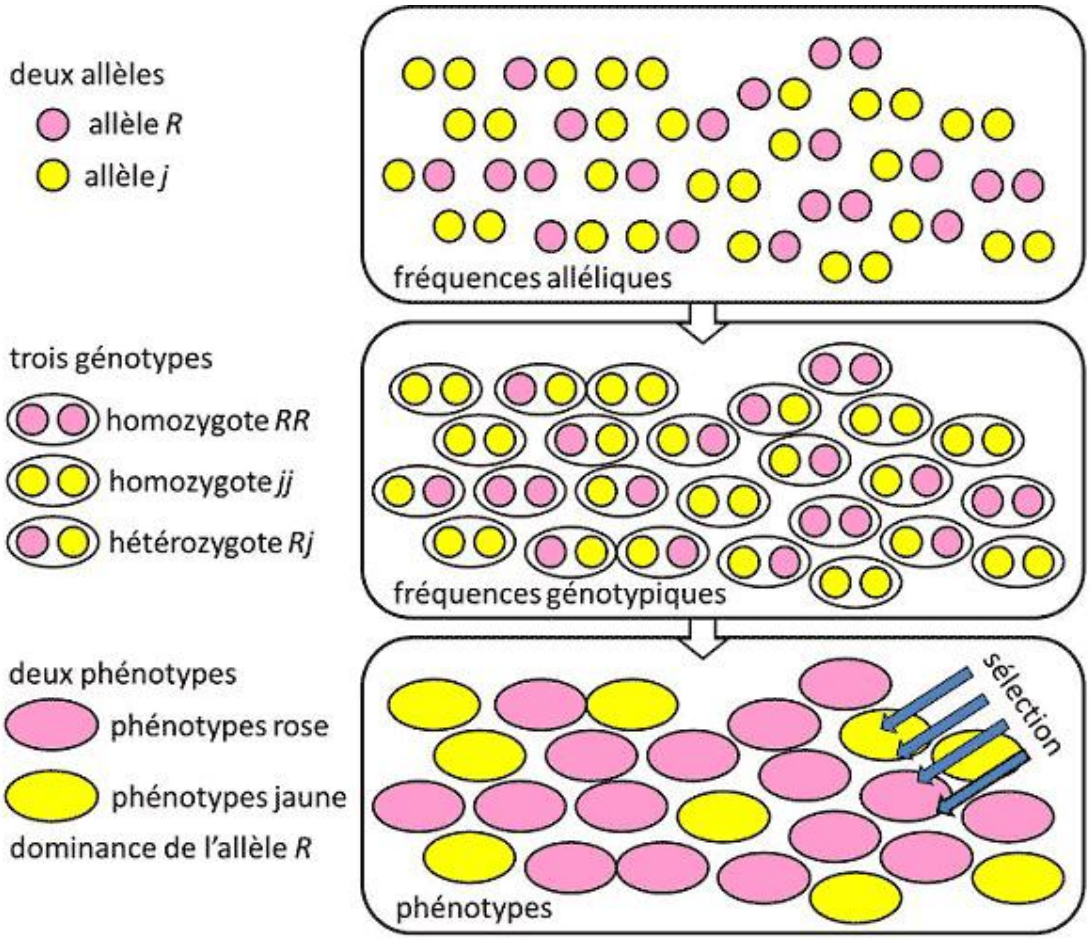

图 4. 基因、基因型和表型：由一对等位基因说明种群遗传学的三个结构层次。等位基因频率分别为 p=0.40（粉色等位基因或 R）和 q=0.60（黄色等位基因或 j），基因型频率分别为 x=0.16（RR 纯合子），y=0.36（jj 纯合子）和 z=0.48（Rj 杂合子），其中 p+q=1，x+y+z=1。在这个例子中，用于基因型频率 x，y 和 z 的值是在配子随机结合情况下预测的理论值：这些是所谓的哈迪–温伯格比例，其中纯合子频率分别是 p^2 和 q^2，以及杂合子频率 $2pq$。（图 4 deux allèles 一对等位基因；allèle R 等位基因 R；allèle j 等位基因 j；fréquences alléliques 等位基因频率；trois génotypes 三种基因型；homozygote RR 纯合子 RR；homozygote jj 纯合子 jj；hétérozygote Rj 杂合子 Rj；fréquences génotypiques 基因型频率；deux phénotypes 两种表型；phénotypes rose 粉色表型；phénotypes jaune 黄色表型；dominance de l'allèle R 等位基因 R 的主导地位；phénotypes 表型；sélection 选择）

4.数量变异和遗传力

当我们环顾四周时，我们看到不同人的身体差异，有些是复杂的，如面部特征，有些是容易测量的，如体重或身高。常识表明，它们有一部分是可以遗传的，尽管遗传的方式很难明确。这些性状通常是多基因决定的【依赖于许多基因。我们讨论多基因遗传。糖尿病是多基因遗传疾病。】，这意味着它们受到许多基因位点的影响。发育遗传学及其在物种比较中的应用（**evo-devo**，意为 “发育的进化”）揭开了复杂的相互作用网络，这些相互作用使身体细胞在发育过程中都

具有相同的遗传包裹，通过后生作用分化自己产生不同的组织。目前还不清楚复杂的性状是如何形成的。基于统计分布的特性，可测量的表型可以用**生物统计学**来研究。一个具有有用属性的值是方差【在一个个体样本中，一个特定的性状被测量，方差是值的平方的均值与均值平方之间的差异。这种测量总是正的，表明个体的分散程度。】[3]。当几个独立的原因决定了一个品系的变异时，它们的变异是可累加的，这些变异的和给出这个品系的方差。如果它们不是独立的，则协方差之和要加到方差之和上。一个表型性状的总方差 V_t 是：

$$V_t=V_{ga}+V_{gd}+V_{gd}+V_{gi}+V_e$$

V_{ga}，加性遗传方差，是由每个独立基因座产生的变异的总和；V_{gd}，显性遗传方差，是同一基因座的两个等位基因相互作用的结果；V_{gi}，基因座之间的互作方差，是由于同一个体的基因座之间相互作用的结果；V_e，环境方差，这里假设独立于遗传方差。

显性遗传变异对于同一父母的两个孩子来说是很常见的，因为在一个特定的基因座上他们从父母双方获得相同的等位基因，他们共享显性效应，这发生在他们四个位点中的一个。他们没有与父母共享这些影响，所以他们彼此之间比他们与父母之间更“相似”，尽管事实上他们彼此之间共享如此多的基因。当然，一个亲本可能偶然地在某个位点上获得了与他的孩子相同的两个等位基因。这种机会取决于种群中等位基因的频率。我们看到，一个孩子和他或她的父母在某一特定特征上的相似程度是一个公共属性。

更普遍的是，遗传变异的所有组成部分($V_{ga}+V_{gd}+V_{gi}$)在群体间都存在差异。在繁殖系中，一种常见的情况是近亲繁殖，这耗尽了遗传变异：结果，一个性状在两个种群中可能有相同的平均值，但有不同的遗传变异。因此，育种家选择一种性状的能力是选择品系的特性，而不是性状的特性。

父母与子女之间的关系由**遗传力** h^2 衡量。它被定义为加性遗传方差与总方差的比值[4]：

$$h^2 = V_{ga}/Vt$$

图 5 显示了在各种假设下的情况。遗传力是决定一个特征是否可以被选择的能力。图 5-3 显示，一个育种者如果选择了数值为 a 的个体，就能在一个杂交世

代内将一个特征的平均值从 0 移到 b。我们可以证明，$b=h^2a$。为了选择一个性状，通过使环境标准化（从而减少环境方差）和将自己置于有利于性状出现的环境条件中有利于提高遗传率。遗传学家沃丁顿已经证明，在自然界中，环境的变化可以揭示在进化过程中将被选择的新性状。

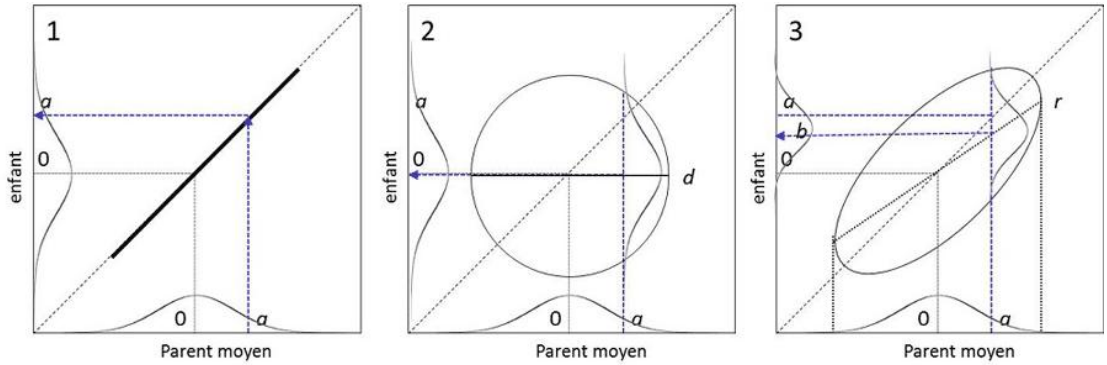

图 5. 亲子关系的定量特征。亲代平均数与子代不同理论情况的相关性大小如下：(1) 相关性为 1，其中子代的身高与亲代的身高完全相同，均值和方差没有差异。边际分布是指亲代均值和子代均值的大小。这些点（亲代、子代）的坐标位于坐标轴的等分线上。一个有价值的亲代均值产生一个有价值的子代 a。(2) 相关性为 0，即子代的大小与亲代的大小无关。在子代的分布中，一个有价值的亲代均值产生了一个无价值的子代。各点的坐标位于一个圆上；其期望值位于与亲代轴线平行的线上（d）。(3) 0 和 1 之间的中间亲子相关性：点的坐标位于一个椭圆上。期望值不再是位于坐标轴的平分线上（这也是椭圆分布的长轴），但在回归线 r 上。一个数值的均值亲代产生了一个具有一定方差的数值 b 的子代（在 0 和 a 之间）。值得注意的是，案例 (3) 介于案例 (1) 和案例 (2) 之间。（图 5 Parent moyen 亲代均值；enfant 子代）

参考文献和注释

封面照片：雷默瑞丽蜗牛和花园葱蜗牛的壳（*Cepaea nemoralis & Cepaea hortensis*）安德烈·昆泽尔曼（André Künzelmann），UFZ.

[1] 这个公式可以用更常规的方式写成： $\pi = \frac{n(n-1)}{2L}\sum_{i=1}^{n-1} \quad \sum_{j=i+1}^{n} \delta ij$

[2] 费舍尔（Fisher R.A.）(1930)《自然选择的遗传理论》，克拉伦登出版社，牛津；霍尔丹（Haldane J.B.S.） (1932) 《进化的原因》。伦敦:哈珀兄弟(Harper & Brothers);《进化与群体遗传学》1,2,3,4;新版。芝加哥大学出版社，1984(再版作者主要成果)。

[3] 作为测量研究性状的个体样本，方差可以定义为数值平方的平均值与数值平均值的平方

之间的差异。这个度量总是正值，表明个体的分散性。

[4] 这是严格意义上的遗传力。广义上的遗传力是遗传方差之和与总方差之比。 这个公式给出了父母双方（“平均父母”）的遗传力；只有单亲的估计值将给出 $h^2/2$。

译者：石鸿宇　　校正：郎明林　　责任编辑： 胡玉娇