第9章
基因型与环境间的互作

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

基因和环境对表型的共同作用

- 环境对数量性状的影响要比对质量性状的影响大得多，因此才有育种中的多年份和多地点试验。一个玉米商业杂交种在走向生产前，往往要经过几百甚至上千个环境下的测试。同时，大多数数量性状的遗传研究，也要建立在多环境表型鉴定的基础之上。
- 基因型在不同环境下有不同的表现，植物中表现得尤为明显，这时就认为存在基因型和环境的互作。

本章的主要内容

- §9.1 宏环境、微环境和目标环境群体
- §9.2 多环境表型鉴定试验的方差分析
- §9.3 基因型的环境稳定性分析
- 附：关联分析的丢失遗传力现象

§9.1 宏环境、微环境和目标环境群体

- §9.1.1 环境的定义和类型
- §9.1.2 基因型与环境的互作模式和利用途径

宏环境和微环境

- 遗传学上，通常把环境定义为影响生物个体表现的一组非遗传因素。这些非遗传因素又可分为非生物因素和生物因素两大类。非生物因素包括土壤的物理和化学特性、气候因子（如光照，降雨量和温度）、耕作制度、栽培方式等。生物因素包含害虫、病原体、线虫和杂草等。这样定义的环境，有时又称为宏环境（macro-environment）。
- 与宏环境相对应的还有微环境（micro-environment），定义为单个植株或小区所处的生长环境。微环境的差异无处不在，两个不同的植株或小区具有同样微环境的可能性几乎是0。

宏环境和微环境在效应上的差异

- 一般来说，宏环境间的差异要比微环境间的差异大得多。
- 宏环境可以是单个栽培方式、地点或年份，也可以是不同栽培方式、不同地点和不同年份的组合。宏环境的效应一般都具有一定程度的重复性。
- 尽管可以通过适当的田间试验设计进行控制，但微环境产生的效应一般不具有重复性，通常只能视为随机误差。
- 基因型和环境互作研究中的环境一般指宏环境，一部分宏环境效应具有重复性，一部分宏环境效应不具有重复性。

环境效应的部分可预测性

- 例如，同一个地点在不同年份间，存在相对稳定不变的一些生物和非生物因素，但这些生物和非生物因素在年份间又会有差异。相对稳定的环境因素所产生的效应，在时间上是可以重复的，或者说，可以用过去的效应来很好地预测未来的效应。因此，有时也把环境变异分为可预测的环境变异和不可预测的环境变异。
- 可预测的环境变异包括一些永久性质的环境因素，如气候的周期性变化、土壤类型、日照时间等。一些耕作制度和栽培措施也可被看作可预测的环境变异，如轮作、播种时间、播种密度、施肥水平、收获方式等。
- 气候因素的随机变动是不可预测和不能重复的，如一个地点降雨量和温度的随机波动、以及病虫害的侵袭等，这类变异往往也被归结为随机效应。

目标环境群体

- 开展多个基因型、在多种环境条件下的表型鉴定试验，是植物育种中选择优良基因型的重要环节。生物个体在生长过程中，面临着各种各样的环境条件。即使在同一个地点种植，不同年份间的各种生物和非生物因素也会有很大差异。生物个体生长过程中面临的所有可能环境，构成了一个目标环境群体（TPE，target population of environments）。
- 与其他统计总体一样，TPE也是不能穷尽的。基因型的多环境试验，只能在有限的环境下进行，开展表型鉴定试验的环境只是TPE的一组有限样本。与任何统计样本一样，试验环境也要求具有代表性，即试验环境要能代表生物个体的TPE。只有这样，从试验环境中得到的观察值，才能代表或者用来预测个体在TPE中的表现；通过多环境试验选择到的优良基因型，才能在农业生产中发挥作用。

基因型和环境的互作研究

- 基因型和环境的互作研究中，基因型一般都包含一组遗传材料（或基因型），如一组全同胞或半同胞家系、一组重组近交系、一组测交组合或一组杂交种等。它们并非只在一个基因座位上存在差异，而是在很多座位上都有差异，它们在一起构成了一个育种或遗传群体。
- 假定这些基因型在多个环境下种植，每个环境下有若干次重复，用y_{ijk}表示第i个基因型在第j个环境下的第k次重复观测值。

基因型和环境的互作模型

- 表型值可被分解为：

$$y_{ijk} = \bar{\mu} + G_i + E_j + GE_{ij} + \varepsilon_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

- 其中 $\bar{\mu}$ 称为总平均数，G_i称为第i个基因型的效应，E_j称为第j个环境的效应，GE_{ij}称为第i个基因型和第j个环境的互作效应，ε_{ijk}是随机误差。
- 公式中第二个等号后面的，μ_{ij}称为第i个基因型在第j个环境下的平均表现，它是一个可估计的未知参数。

基因和环境对表型的贡献

- 前面的公式说明了表型是基因型和环境共同作用的结果，是最一般也是最常用描述表型、基因型、环境三者关系的线性模型。
- 如何提高公式中的基因型效应，如何从已有育种群体中把基因型效应最好的个体鉴定出来，是育种家的主要任务。
- 同时，好的环境也可以改变生物个体的表型，这里的环境指的是可重复的宏环境。如何通过环境的改变以提高环境效应、最终提高一个基因型的表现，是栽培生理学家的主要任务。
- 严格区分遗传改良和环境改良对农业生产的贡献是很困难的，二者贡献各占50%可能是大多数人都愿意接受的数字。

基因和环境互作的利用

- 育种和农业生产过程中经常观测到，一些基因型对环境的变化表现得很敏感，而另外一些基因型对环境变化的反应却很迟钝。甚至还会出现，一些环境下表现很好的基因型，但在其他环境下的表现却很差；一些环境下表现较差的基因型，在其他环境下的表现却很好。这时，基因型和环境之间就存在互作。
- 与§7.4的基因间上位性互作一样，当互作达到一定程度时，最高的基因型效应与最高的环境效应结合在一起，不一定会得到最高的表型。
- 基因型和环境互作是一个普遍的遗传学现象，农业生产中强调的“良种配良法”，其实就是期望通过提高遗传效应、改进环境效应、利用基因型和环境互作，来共同达到提高产量这一重要的目标。

两个基因型在两个环境下的平均表现及基因型和环境效应的计算

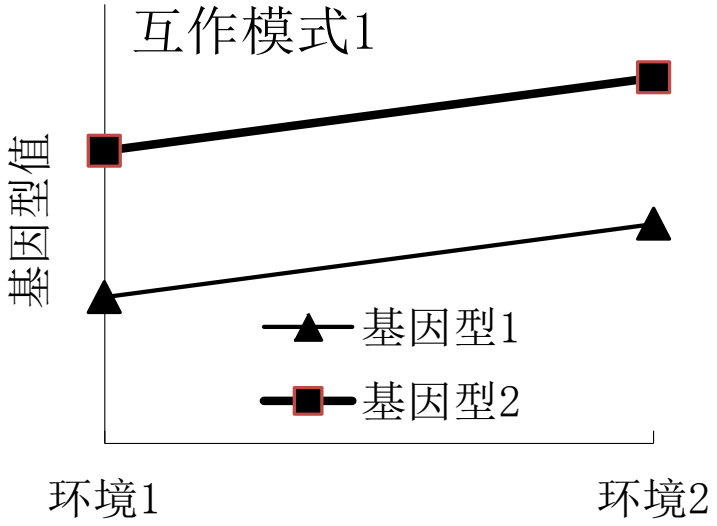
	环境 1	环境 2	行平均	基因型效应
基因型 1	μ_{11}	μ_{12}	$\bar{\mu}_{1\cdot}=\frac{1}{2}(\mu_{11}+\mu_{12})$	$G_1=\bar{\mu}_{1\cdot}-\bar{\mu}$
基因型 2	μ_{21}	μ_{22}	$\bar{\mu}_{2\cdot}=\frac{1}{2}(\mu_{21}+\mu_{22})$	$G_2=\bar{\mu}_{2\cdot}-\bar{\mu}$
列平均	$\bar{\mu}_{\cdot 1}=\frac{1}{2}(\mu_{11}+\mu_{21})$	$\bar{\mu}_{\cdot 2}=\frac{1}{2}(\mu_{12}+\mu_{22})$	$\bar{\mu}=\frac{1}{4}(\mu_{11}+\mu_{12}+\mu_{21}+\mu_{22})$	
环境效应	$E_1=\bar{\mu}_{\cdot 1}-\bar{\mu}$	$E_2=\bar{\mu}_{\cdot 2}-\bar{\mu}$		

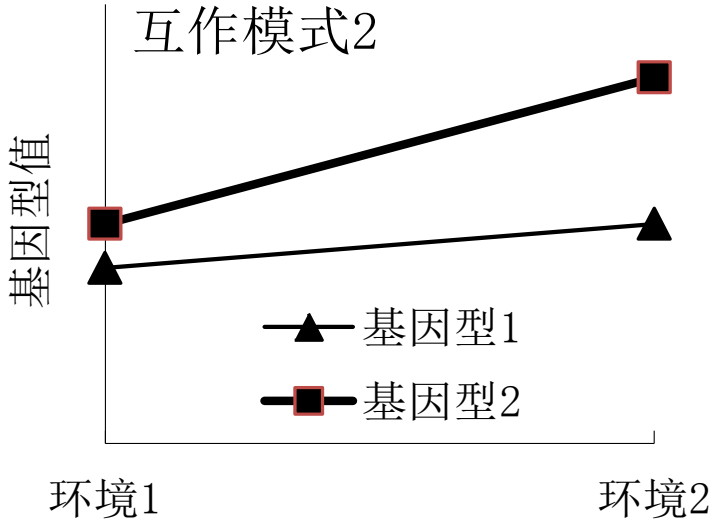
两个基因型在两个环境下基因型和环境互作效应的计算

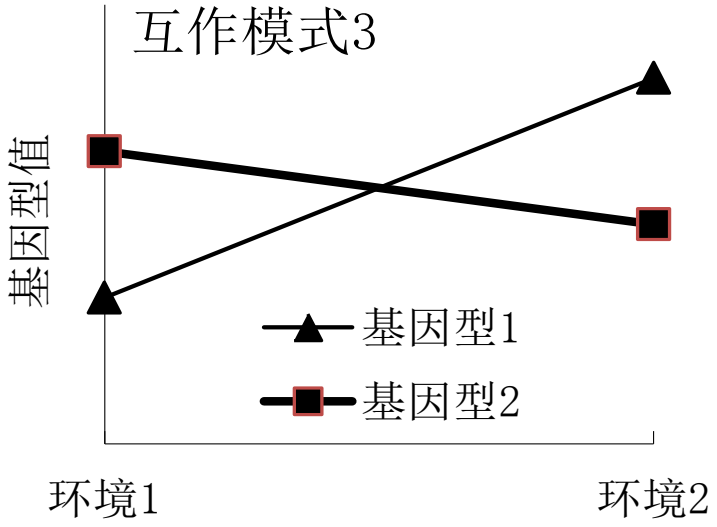
	环境 1	环境 2
基因型 1	$GE_{11} = \mu_{11} - \bar{\mu}_{1\cdot} - \bar{\mu}_{\cdot 1} + \bar{\mu}$	$GE_{12} = \mu_{12} - \bar{\mu}_{1\cdot} - \bar{\mu}_{\cdot 2} + \bar{\mu}$
基因型 2	$GE_{21} = \mu_{21} - \bar{\mu}_{2\cdot} - \bar{\mu}_{\cdot 1} + \bar{\mu}$	$GE_{22} = \mu_{22} - \bar{\mu}_{2\cdot} - \bar{\mu}_{\cdot 2} + \bar{\mu}$

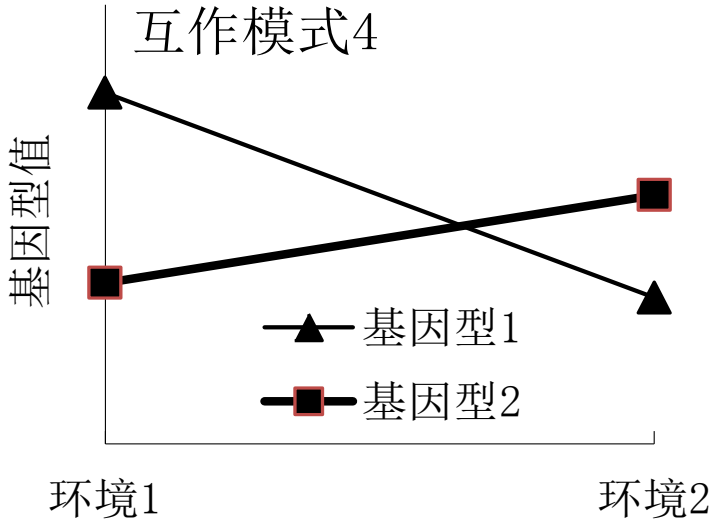
- 把互作效应排成一个双向表，这些效应满足行和等于0、列和等于0、总和等于0的约束条件，独立参数的个数等于基因型个数减1与环境个数减1之积。

基因型与环境交互作用的4种模式









非交叉基因型与环境交互作用

- 在模式1下，一个基因型在两个环境下都优于另一个基因型，同时，基因型间的差异在2个环境下是相等的。如果把基因型在两个环境下的表现用直线连接起来，则代表基因型的两条直线是平行的。在这种模式下，基因型在两个环境下的差异完全由环境效应决定，所有互作效应均为0，即不存在基因型和环境间的互作。

非交叉基因型与环境交互作用

- 在模式2下，一个基因型在两个环境下都优于另一个基因型，基因型间的差异在两个环境下不相等，基因型2随着环境效应的增加表现出更大的优势。在这种模式下，基因型之间的差异因环境而变，也就是说存在基因型和环境间的互作。但是，这样的互作并没有改变基因型优劣的顺序，有时也称为非交叉互作（non-crossover interaction）。
- 对于无互作的模式1以及无交叉互作的模式2，在不关心环境效应的情况下，通过单个环境的表型鉴定，育种家便可知道不同基因型的优劣并进行选择，因此可以大大减少育种家的工作量。

交叉基因型与环境交互作用

- 模式3和4代表的均是交叉互作（crossover interaction），基因型的表现因环境而异。在环境1中，基因型2优于基因型1，但随着环境的变化其优势愈来愈小，最终在环境2下，基因型1优于基因型2。
- 在交叉互作的模式3中，基因型间差异的绝对值在两个环境下是相等的，这时的基因型效应为0，只存在环境效应和互作效应。
- 在交叉互作的模式4中，基因型间差异的绝对值在两个环境下不相等，这时，表9.1的基因型效应和环境效应、以及表9.2的互作效应均不为0。

交叉基因型与环境交互作用

- 对于模式3和模式4来说，一个环境下基因型的优劣不能代表另一个环境下基因型的优劣，必须通过多环境的表型鉴定，才能全面评价基因型的好坏。
- 对于多个基因型的多环境表型数据，图9.1中的4种模式可能会同时出现。各种互作模式的并存，显示了多环境试验在遗传研究和育种中的必要性。

两个小麦品种在2个环境下的赤霉病感染率（%）

基因型	E_1：环境1	E_2：环境2	行平均	基因型效应
G_1：A_1A_1	10	20	**15**	**-22.5**
G_2：A_2A_2	50	70	**60**	**22.5**
列平均	**30**	**45**	**37.5**	
环境效应	**-7.5**	**7.5**		

- 互作效应的计算

$$GE_{11} = (10 - 37.5) - G_1 - E_1 = 2.5 \quad GE_{12} = (20 - 37.5) - G_1 - E_2 = -2.5$$

$$GE_{21} = (50 - 37.5) - G_2 - E_1 = -2.5 \quad GE_{22} = (70 - 37.5) - G_2 - E_2 = 2.5$$

基因型和环境互作的利用途径

- TPE的大小和同质程度，会影响公式9.1中各种方差成分。
- 对于一个较小的TPE来说，互作方差一般也较小， 会在表型方差中占较大的比例；对于一个很大的TPE来说，互作效应就会很高， 可能是表型方差的主要成分。
- 基因型与环境交互作用是广泛存在的，实际中有以下三种利用基因型与环境之间交互作用的方式。

方式1：
忽略基因型与环境互作（ignore it）

- 这种方法其实并不否认基因型与环境交互作用的存在；相反，它也承认基因型与环境交互作用是存在的，并在广泛的环境下测试基因型的表现，优异基因型的推荐或选择所依据的是基因型在所有环境下的平均表现。
- 这种方式强调的是基因型对环境的一般适应性。如果互作以图9.1的模式3和4为主，推荐的基因型从平均数的角度来说是最好的，但对某一特定的环境来说就不一定了。标准方差分析得到的互作方差 和随机误差方差 可以用来优化多环境试验中的资源配置。

方式2：
降低基因型与环境互作（reduce it）

- 一般来说，较大TPE的环境异质程度也越高，基因型和环境互作就可能越大，出现交叉互作（图9.1的互作模式3和4）的可能性也越高；较小TPE的环境异质程度也较低，基因型和环境互作就越小。
- 一个较大的TPE可以被划分为几个较小的、相对同质的亚环境群体。每个亚群体内有着相似的非生物和生物特性，因此出现交叉互作的可能性较小。亚群体内，少数环境的平均表现可以较好地反映基因型在整个亚群体内的表现。对不同的亚环境群体，根据平均表现推荐不同的基因型。
- 常用的划分TPE的方法有聚类分析和主成分分析。

方式3：
利用基因型与环境互作（exploit it）

- 这种方式的主要目的在于鉴定出特定环境下的最好基因型，强调的是特定基因型对特殊环境的适应性。
- 互作利用方式3和2并不是孤立的，方式2中通过TPE的划分来减小互作方差，其实也是在利用交互作用。
- 方式3的分析方法包括稳定性分析和乘积模型。

§9.2 多环境表型鉴定试验的方差分析

- §9.2.1 表型值的线性分解
- §9.2.2 多环境表型数据的方差分析
- §9.2.3 多环境基因型值和广义遗传力的估计
- §9.2.4 异质误差方差条件下的最优无偏线性估计
- §9.2.5 评价基因型的适宜环境数和重复数

表型的分布

- 假定对g个基因型在e个环境条件下开展表型鉴定试验，每个环境设置r次重复，μ_{ij}表示第i个基因型在第j个环境下的平均表现，是一个待估计的未知参数。
- 在观测误差服从均值是0、方差是σ_ε^2的正态分布、且相互独立的假定下，第i个基因型在第j个环境下的第k个表型值y_{ijk}服从下面的正态分布。

$$y_{ijk} \sim N(\mu_{ij},\ \sigma_\varepsilon^2)$$

其中i=1, …, g，j=1, …, e，k=1, …, r

总平均、基因型平均和环境平均

- g个基因型在e个环境下的总平均表现

$$\bar{\mu}.. = \frac{1}{ge}\sum_{i,j}\mu_{ij}$$

- 单个基因型在环境间的平均表现

$$\bar{\mu}_i. = \frac{1}{e}\sum_{j}\mu_{ij}$$

- 单个环境的平均表现

$$\bar{\mu}._j = \frac{1}{g}\sum_{i}\mu_{ij}$$

基因型效应、环境效应和互作效应

- 基因型效应 $G_i \hat{=} (\overline{\mu}_i. - \overline{\mu}..)$
- 环境效应 $E_j \hat{=} (\overline{\mu}._j - \overline{\mu}..)$
- 互作效应 $GE_{ij} \hat{=} \mu_{ij} - \overline{\mu}_i. - \overline{\mu}._j + \overline{\mu}..$

基因型值和表型值的线性分解模型

- 基因型值的线性分解模型

$$\mu_{ij} = \bar{\mu}.. + G_i + E_j + GE_{ij}$$

- 表型值的线性分解模型

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \bar{\mu}.. + G_i + E_j + GE_{ij} + \varepsilon_{ijk}$$

遗传方差、环境方差和互作方差

- 遗传方差 $\sigma_G^2 \hat{=} \frac{1}{g-1}\sum_i G_i^2$
- 环境方差 $\sigma_E^2 \hat{=} \frac{1}{e-1}\sum_j E_j^2$
- 互作方差 $\sigma_{GE}^2 \hat{=} \frac{1}{(g-1)(e-1)}\sum_{i,j} GE_{ij}^2$

完全随机区组的多环境表型数据

- 如果每个环境下，田间试验均采用完全随机区组设计，并且区组之间有显著的差异，线性模型公式9.7中还应该包含区组的效应，即：

$$y_{ijk} = \mu_{ij} + B_{k/j} + \varepsilon_{ijk} = \bar{\mu}.. + B_{k/j} + G_i + E_j + GE_{ij} + \varepsilon_{ijk}$$

- 利用这一模型时要特别注意，每个环境都包含r个区组效应，整个试验共包含re个区组。因此，模型中共包含re个区组效应，区组可以看作是嵌套在每个环境中。因此，区组的效应也有re个，而不是r个，区组效应一般用$B_{k/j}$，而不是B_k表示。

利用表型数据的平均数估计各种效应

- 利用重复平均数估计基因型*i*在环境*j*的平均表现

$$\bar{y}_{ij}. = \frac{1}{r}\sum_{k} y_{ijk} \qquad \mu_{ij} = \bar{y}_{ij}.$$

- 利用总平均数估计总平均表现

$$\bar{y}... = \frac{1}{ger}\sum_{i,j,k} y_{ijk} \qquad \bar{\mu}.. = \bar{y}...$$

- 利用基因型*i*的环境和重复平均数估计遗传效应

$$\bar{y}_{i}.. = \frac{1}{er}\sum_{j,k} y_{ijk} \qquad G_i = \bar{y}_{i}.. - \bar{y}...$$

利用表型数据的平均数估计各种效应

- 利用环境*j*的基因型和重复平均数估计环境效应

$$\bar{y}_{\cdot j\cdot} = \frac{1}{gr}\sum_{i,k} y_{ijk} \qquad E_j = \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}$$

- 基因型和环境互作的估计

$$GE_{ij} = \mu_{ij} - \bar{\mu}_{\cdot\cdot} - G_i - E_j = \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}$$

- 说明：前面公式给出的其实是公式9.11中各种效应的最小二乘估计。严格地讲，一个参数和它的估计是有区别的，统计学中一般在一个参数的上方加以符号‘^’表示这个参数的估计。如：

$$\hat{\mu}_{ij} = \bar{y}_{ij\cdot} \qquad \hat{G}_i = \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}$$

多环境重复表型观测值的方差分析表（不考虑区组效应）

变异来源	自由度	平方和	均方	期望均方（固定效应模型）
基因型	g-1	SS_G	MS_G	$\sigma_\varepsilon^2 + er\,\sigma_G^2$
环境	e-1	SS_E	MS_E	$\sigma_\varepsilon^2 + gr\,\sigma_E^2$
基因型与环境互作	$(g-1)(r-1)$	SS_{GE}	MS_{GE}	$\sigma_\varepsilon^2 + r\,\sigma_{GE}^2$
随机误差	$ge(r-1)$	SS_ε	MS_ε	σ_ε^2
总和	ger-1	SS_T		

多环境试验的区组效应估计

- 如方差分析模型中包含区组效应，下面的公式给出每个环境下r个区组效应的最小二乘估计。其它效应的最小二乘估计与不含区组效应的模型完全相同。

$$\bar{y}_{\cdot jk} = \frac{1}{g}\sum_{i} y_{ijk} \qquad B_{i/j} = \bar{y}_{\cdot jk} - \bar{y}_{\cdot j\cdot}$$

多环境重复表型观测值的方差分析表（考虑区组效应）

变异来源	自由度	平方和	均方	期望均方（固定效应模型）
环境内区组	$e(r-1)$	SS_R	MS_R	$\sigma_\varepsilon^2 + g\,\sigma_R^2$
基因型	$g-1$	SS_G	MS_G	$\sigma_\varepsilon^2 + er\,\sigma_G^2$
环境	$e-1$	SS_E	MS_E	$\sigma_\varepsilon^2 + gr\,\sigma_E^2$
基因型与环境互作	$(g-1)(r-1)$	SS_{GE}	MS_{GE}	$\sigma_\varepsilon^2 + r\,\sigma_{GE}^2$
随机误差	$(g-1)e(r-1)$	SS_ε	MS_ε	σ_ε^2
总和	$ger-1$	SS_T		

平均表现的估计值及其方差

- 从表型分布看出，观测值y_{ijk}包含了第i个基因型在第j个环境下平均表现μ_{ij}的信息。重复平均数是基因型平均表现 的最优线性无偏估计（BLUE）。

$$\hat{\mu}_{ij} = \frac{1}{r}\sum_{k} y_{ijk} = \bar{y}_{ij}. \quad V(\hat{\mu}_{ij}) = \frac{1}{r}\sigma_{\varepsilon}^{2}$$

- 基因型i在环境间平均表现 的BLUE及估计值的方差：

$$\bar{\mu}_{i}. = \frac{1}{e}\sum_{j} \bar{y}_{ij}. = \frac{1}{er}\sum_{j,k} y_{ijk} \quad V(\bar{\mu}_{i}.) = \frac{1}{er}\sigma_{\varepsilon}^{2}$$

- 在误差方差未知的情况下，可以用方差分析中的误差均方来代替误差方差，以估计上面BLUE的方差。

表型方差的构成

- 多环境下，一个基因型或家系的表现等于基因型效应、环境效应、互作效应与随机误差之和，即：

$$P = \mu + G + E + GE + \varepsilon$$

- 在误差项独立且服从同一正态分布的假定下，表型方差等于基因型效应产生的方差、环境效应产生的方差、互作效应产生的方差及误差方差之和。下面公式右端各种方差成分定义在公式9.7~9.10中，通过方差分析的期望均方，可以得到他它们的无偏估计。

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 + \sigma_{GE}^2 + \sigma_\varepsilon^2$$

广义遗传力估计

- 环境方差 来源于一些非遗传的因素，在遗传力估计时不考虑这部分方差。广义遗传力H^2如下面的公式计算。

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{GE}^2 + \sigma_\varepsilon^2}$$

- 这一估计可被视为单个观测表型的遗传力。

重复平均数的方差构成

- 遗传研究一般基于基因型估计值。下面的公式给出基因型在环境和区组间平均数的方差，其中，遗传方差与公式9.21的遗传方差相同，互作方差只有公式9.21中互作方差的1/*e*，误差方差只有公式9.21中误差方差的1/*er*。

$$\sigma^2_{\bar{P}} = \sigma^2_G + \frac{1}{e}\sigma^2_{GE} + \frac{1}{er}\sigma^2_{\varepsilon}$$

重复平均数的遗传力

- 如果利用环境间重复平均数作为性状的选择标准或开展遗传研究，可以根据下面的公式计算它的遗传力：

$$H^2 = \frac{\sigma_G^2}{\sigma_{\bar{P}}^2} = \frac{\sigma_G^2}{\sigma_G^2 + \frac{1}{e}\sigma_{GE}^2 + \frac{1}{er}\sigma_\varepsilon^2}$$

- 这一估计又称为环境和重复平均数的遗传力。

重复平均数的遗传力

- 如果利用环境间重复平均数作为性状的选择标准或开展遗传研究，可以根据下面的公式计算它的遗传力：

$$H^2 = \frac{\sigma_G^2}{\sigma_{\bar{P}}^2} = \frac{\sigma_G^2}{\sigma_G^2 + \frac{1}{e}\sigma_{GE}^2 + \frac{1}{er}\sigma_\varepsilon^2}$$

- 这一估计又称为环境和重复平均数的遗传力。

表7.3 水稻双亲衍生的10个RIL家系在三个环境下的直链淀粉含量（%）

基因型	环境I		环境II		环境III	
	重复1	重复2	重复1	重复2	重复1	重复2
RIL1	15.3	15.1	14.4	14.6	14.5	14.8
RIL2	14.5	15.0	15.8	15.7	16.3	15.7
RIL3	14.0	14.9	15.9	15.8	15.2	16.1
RIL4	13.2	14.0	16.0	16.8	15.0	15.5
RIL5	15.4	15.9	16.7	16.6	15.4	15.6
RIL6	15.5	15.6	16.1	16.7	16.0	17.0
RIL7	13.2	14.1	14.3	14.9	14.1	14.5
RIL8	11.9	12.6	14.2	14.4	15.2	16.1
RIL9	12.8	13.5	14.5	14.6	15.3	15.5
RIL10	12.8	13.6	14.6	15.5	14.2	14.2

三个环境的联合方差分析
（不含区组效应）

变异来源	自由度	平方和	均方	F值	显著概率	方差估计值
基因型	9	31.4260	3.4918	20.38	<0.0001	0.5534
环境	2	19.6924	9.8462	57.47	<0.0001	0.4837
基因型与环境互作	18	16.7310	0.9295	5.43	<0.001	0.3791
随机误差	30	5.1400	0.1710			0.1713
总和	59	72.9893				

三个环境的联合方差分析
（不含区组效应）

- 方差分析的结果表明，直链淀粉含量在基因型间、环境间均存在极显著的差异，基因型和环境间的互作也达到极显著水平。
- 从方差的估计值看，遗传方差的估计值远高于随机误差方差的估计值，也远高于互作方差的估计值。
- 将最后一列的方差估计值代入公式9.22，得到小区水平的广义遗传力为50.14%；代入公式9.24，得到环境和重复平均数的广义遗传力为78.13%。

三个环境的联合方差分析
（含区组效应）

变异来源	自由度	平方和	均方	F值	显著概率	方差估计值
环境内区组	3	2.8270	0.9423	11.00	<0.001	0.0857
基因型	9	31.4260	3.4918	40.76	<0.0001	0.5677
环境	2	19.6924	9.8462	114.94	<0.0001	0.4880
基因型与环境互作	18	16.7310	0.9395	10.85	<0.0001	0.4219
随机误差	27	2.3130	0.0857			0.0857
总和	59	72.9893				

三个环境的联合方差分析
（含区组效应）

- 区组效应的自由度为3，基因型自由度仍为9，环境自由度为仍2，基因型和环境互作自由度仍为18，总自由度仍为59；误差效应的自由度降低到27。从*F*检验可以看出，区组之间也存在极显著的差异。
- 将最后一列的方差估计值代入公式9.22，得到小区水平的广义遗传力为52.79%；代入公式9.24，得到环境和重复平均数的广义遗传力为78.56%。由于此例中存在显著的区组效应，方差分析的线性模型在包含区组效应之后，降低了随机误差方差，因此得到较高的遗传力估计值。

误差方差的异质性

- 同一环境条件下，一般都会默认误差效应具有相同的方差。有时由于种植条件的差异，环境间的误差方差可能会有较大差异。
- 例如，栽培条件优良的环境下，由于有良好的灌溉设施、适当的土壤肥力、有效的病虫害防治措施，误差方差一般比较小，表型观测值更接近基因型的平均表现。
- 干旱或雨养环境条件下，误差方差一般会比较大，表型观测值与基因型的平均表现之间有较大的偏离。

异质误差方差的表型分布

- 如果一个基因型在异质环境条件下进行种植，基因型的平均表现为μ，第j个环境的误差方差为$\sigma_{\varepsilon j}^2$，y_j表示第j个环境下的表型观测值。表型观测值的线性模型如下：

$$y_j = \mu + \varepsilon_j$$

$\varepsilon_j \sim N(0,\ \sigma_{\varepsilon_j}^2)$，$j$=1, 2, …, e，且相互独立

异质误差方差的样本均值及其方差

- 样本均值及其方差为：

$$\bar{y}. = \frac{1}{e}\sum_{j} y_j \qquad V(\bar{y}.) = \frac{1}{e^2}(\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2} + \cdots + \sigma^2_{\varepsilon_e})$$

- 统计上称具有最低方差的无偏估计为最优无偏估计。当误差方差在环境间不相等，或者误差方差异质时，上面的样本均值不再是最优无偏估计。也就是说，存在比简单平均数的方差更小的线性无偏估计。

异质误差方差的最优线性无偏估计

- 构造y_j（j=1、2、…、e）的线性组合，并计算其方差。可以发现，在环境方差异质时，下面公式给出的估计比简单平均数具有更小的方差：

$$\hat{\mu} = \sum_j w_j y_j \qquad w_j = \frac{\frac{1}{\sigma^2_{\varepsilon_j}}}{\frac{1}{\sigma^2_{\varepsilon_1}} + \frac{1}{\sigma^2_{\varepsilon_2}} + \cdots + \frac{1}{\sigma^2_{\varepsilon_e}}}$$

$$V(\hat{\mu}) = \frac{1}{\frac{1}{\sigma^2_{\varepsilon_1}} + \frac{1}{\sigma^2_{\varepsilon_2}} + \cdots + \frac{1}{\sigma^2_{\varepsilon_e}}}$$

异质误差方差的最优线性无偏估计

- **异质误差方差的最优线性无偏估计**是一个加权平均数，误差方差越大的环境，其权重越低；误差方差越小的环境，其权重越高。
- 这一点从直观上也是可以理解的，误差方差较大的观测值，其偏离真实值的程度也就越大，因此包含了较少真实值的信息；误差方差较小的观测值，其偏离真实值的程度也较小，因此包含了较多真实值的信息。
- 如果观测误差非常大，观测值对真实值也就失去了代表性，参数估计中就应该剔除掉这样的观测值。

方差同质性的Bartlett检验

- 方差同质性检验的零假设和备则假设如下。

$$H_0: \sigma_{\varepsilon_1}^2 = \sigma_{\varepsilon_2}^2 = \cdots = \sigma_{\varepsilon_e}^2$$

$$H_A: \sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_2}^2, \cdots, \sigma_{\varepsilon_e}^2 \text{ 中至少有两项互不相等}$$

- 在H_0下，首先利用所有环境计算合并误差方差：

$$\sigma_\varepsilon^2 = \frac{1}{\sum_j df_{\varepsilon_j}} \sum_j df_{\varepsilon_j} \times \sigma_{\varepsilon_j}^2$$

- Bartlett检验统计量近似服从自由度为e-1的 分布。

$$\chi^2 = (\sum_j df_{\varepsilon_j}) \ln(\sigma_\varepsilon^2) - \sum_j df_{\varepsilon_j} \times \ln(\sigma_{\varepsilon_j}^2) \sim \chi^2(e-1)$$

两个水稻亲本和它们的10个RIL家系直链淀粉含量（%）平均表现的估计值

基因型	环境1	环境2	环境3	简单平均数	加权平均数
RIL1	15.20	14.50	14.65	14.78	14.82
RIL2	14.75	15.75	16.00	15.50	15.40
RIL3	14.45	15.85	15.65	15.32	15.23
RIL4	13.60	16.40	15.25	15.08	14.99
RIL5	15.65	16.65	15.50	15.93	15.98
RIL6	15.55	16.40	16.50	16.15	16.08
RIL7	13.65	14.60	14.30	14.18	14.14
RIL8	12.25	14.30	15.65	14.07	13.77
RIL9	13.15	14.55	15.40	14.37	14.17
RIL10	13.20	15.05	14.20	14.15	14.10

多环境联合方差分析

- 从严格意义上讲，多环境的表型鉴定数据应该先进行单环境方差分析，只有在所有环境误差方差同质时，才能进行多环境联合方差分析；如果环境误差方差不同质，严格地说是不宜进行多环境联合方差分析的。
- 多环境基因型BLUE的计算并不依赖于多环境联合方差分析，只要通过单环境方差分析估计出每个环境的误差方差即可。
- 开展进一步的遗传研究，如基因定位，最好利用公式9.27计算出的BLUE，它比公式9.25的简单平均更接近真实的基因型值。

基因型的环境和重复平均数的方差

- 基因型和环境互作方差以及误差方差

$$V_{GE} = \frac{MS_{GE} - MS_{\varepsilon}}{r} \qquad V_{\varepsilon} = MS_{\varepsilon}$$

- 基因型i的所有环境和所有重复平均数的方差中，既包含随机误差方差，又包含基因型和环境互作方差：

$$V_{\bar{y}_i} = \frac{V_{\varepsilon}}{er} + \frac{V_{GE}}{e}$$

可检测到的最小显著性差异

- 利用基因型均值的方差公式，还可以计算两个基因型*i*和*j*间的最小显著性差异（least significant difference, LSD）。当基因型和环境个数较大时，0.05概率水平的LSD的近似计算公式为：

$$\mathrm{LSD}_{0.05} \approx 2.0\sqrt{2V_{\bar{y}_i}}$$

环境数和重复数对LSD的影响

- 在互作方差V_{GE}和误差方差V_{ε}恒定的情况下，增加环境、增加重复都可以降低基因型均值的方差，从而使得更小的差异也能在0.05的显著性水平下被检测出来。
- 当V_{GE}低到可以忽略不计的时候，公式9.32就以随机误差方差为主，这时，增加环境与增加重复没有明显差异。

评价基因型的适宜环境数和重复数

- 对于大多数数量性状来说，基因型和环境互作都会达到显著或极显著水平，V_{GE}可能接近甚至远超过误差方差。这时，增加环境的影响更大，能够使平均数的方差变得更低。
- 因此，在单个基因型占用小区数（等于$e \times r$，即环境数与重复数之积）相对固定的情况下，如果以检测到的最小显著性差异为标准，最优的资源配置是令重复数为1，而使环境数尽可能大。

重复的必要性

- 但从另外一个角度来说，一次重复无法对误差方差进行估计，难以开展显著性检验。
- 对于基因型个数较多、确实难以设置重复的情形，如育种中高世代材料的首次产量鉴定试验，参加试验的基因型可能有数百甚至数千个，这时可以通过一个或多个对照基因型在田间的重复观测数据估计误差方差，即对少数基因型设置重复，以达到估计误差方差的目的。

评价基因型的适宜重复数

- 对于基因型个数较少、精确度要求比较高的试验，则要考虑对所有基因型设置相同个数的重复，以更好地估计误差方差、开展显著性检验。
- 对于大多数需要设置重复的育种试验来说，两次重复可能就足够了。两次重复试验中，单个基因型重复平均数的方差等于误差方差的1/2。
- 对于精确度要求更高的试验，如品种审定前的区域试验，可以考虑设置三次重复。三次重复试验中，单个基因型重复平均数的方差等于误差方差的1/3。更多次数的重复，在遗传和育种研究中没有必要，也不切合实践。

增加环境的重要性

- 上面的结论是在互作方差V_{GE}相对固定的情况下得出的。一般来说，增加环境要比增加重复产生更多的花费，并且环境数的增加势必引起互作方差的增加。这时，在较多的环境下测试基因型会造成试验精度上的损失。
- 但总的来说，增加环境的重要性要超过增加重复的重要性，广泛的多环境表型鉴定是选育广适应性品种的必要途径。

§9.3 基因型的环境稳定性分析

- §9.3.1 目标环境群体（TPE）的分类
- §9.3.2 基因型和环境双向表
- §9.3.3 基因型环境稳定性的Finlay-Wilkinson分析方法
- §9.3.4 基因型环境稳定性的Eberhart-Russell分析方法
- §9.3.5 基因型和环境互作的乘积模型

目标环境群体（TPE）的分类

- 常用的划分环境的统计方法有分类分析、主坐标分析、以及联合分类和坐标过程。对环境的分类，是育种实践中降低并利用基因型和环境互作的一种常用方法，以下简单介绍一种环境分类的聚类分析方法。
- 根据一组环境样本之间的相似性，聚类分析（Cluster Analysis）最终产生出一个聚类图，把环境划分在不同个数的亚群中，同一亚群内的两个环境要比不同亚群的两个环境间有较大的相似性。随着不同亚群的合并，群内环境间的异质性逐渐增加。

环境间距离的度量

- 利用聚类分析，需要构建一个相似性或距离的测度。聚类分析过程中使用的数据往往来自很多地点和年份，是不平衡的。两个环境间的距离 （统计上的距离，并非物理距离），可以根据两个环境中种植的共同基因型的表现来估计：

$$D_{jj'} = \frac{1}{g}\sum_{i=1}^{g}\left(\frac{y_{ij} - \mu_j}{SE_j} - \frac{y_{ij'} - \mu_{j'}}{SE_{j'}}\right)^2$$

- 对于平衡数据有：$D_{jj'} = 2\left(1 - \frac{1}{g}\right)\left(1 - r_{jj'}\right)$

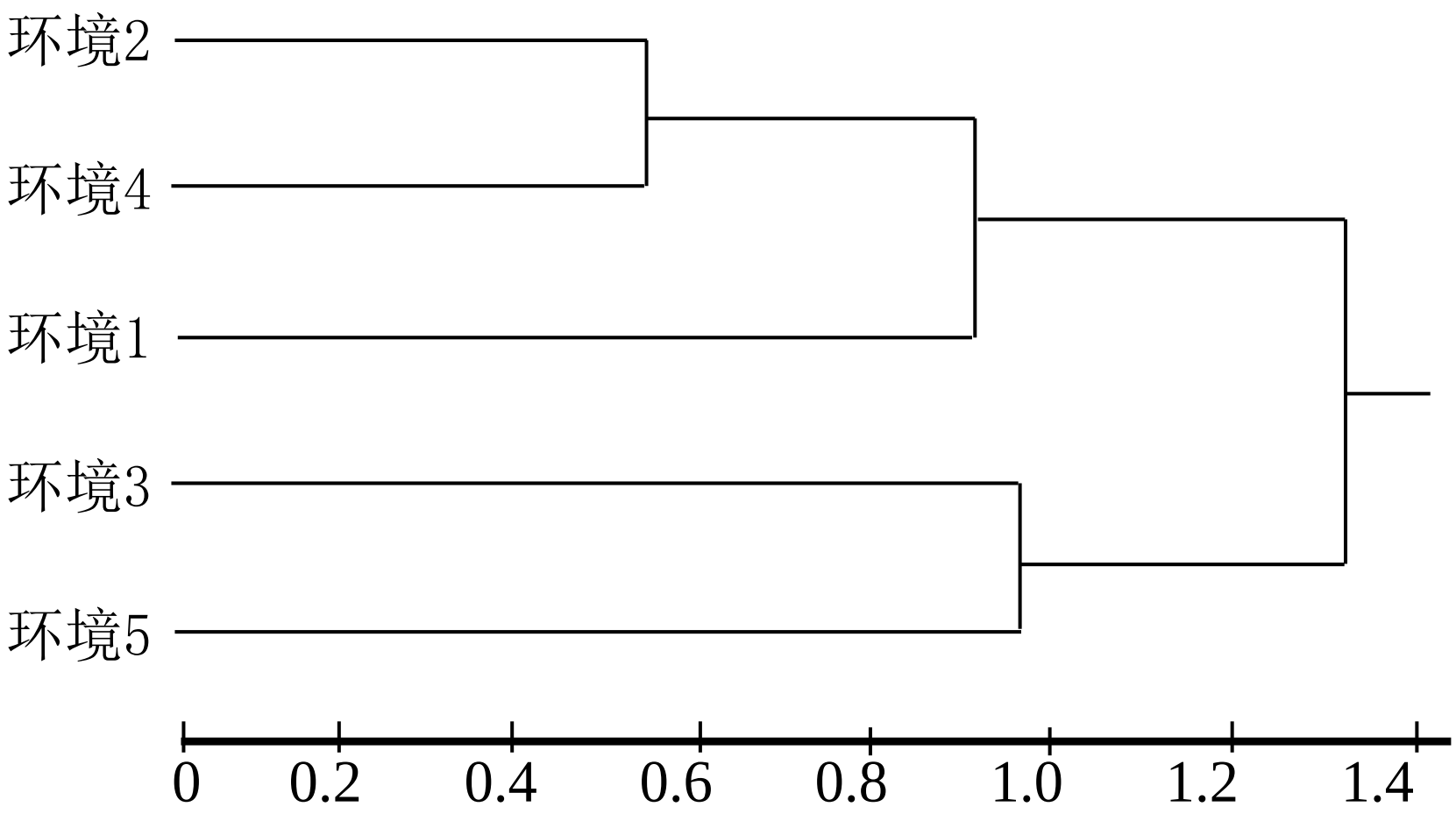
5个环境样本之间的距离矩阵

	环境2	环境3	环境4	环境5
环境1	0.88	1.56	0.95	1.20
环境2		1.18	0.55	1.28
环境3			1.63	0.95
环境4				1.33

类平均聚类方法

- 对于5个环境样本，开始时，每个环境为一个单独的亚群，聚类分析中将两个亚群合并形成一个较大的亚群有不同的方法，常用的方法是类平均（unweighted pair-group method using arithmetic average, UPGMA）。
- 这种方法把两个类间的距离定义为第一个类中的环境和第二个类中的环境间所有距离的平均。聚类过程中，优先合并平均距离最短的两个类，直到把所有的环境合并为一个类为止。

5个环境的聚类图



聚类的个数

- 当环境数较多时，将这些环境究竟分成多少个组合适，或者说聚类过程到哪一步就应该停止，统计上并没有一个严格的标准或规定。更多时候，需要考虑聚类对象的生物学意义。
- 聚类分析只是利用一定的标准，把聚类对象的远近关系用聚类图形象地表示出来。如有其他数据支持，也可将图9.3的5个环境分成3组，这时环境2、4、1在一个组中，环境3和5在另外的两个组中。

基因型和环境双向表

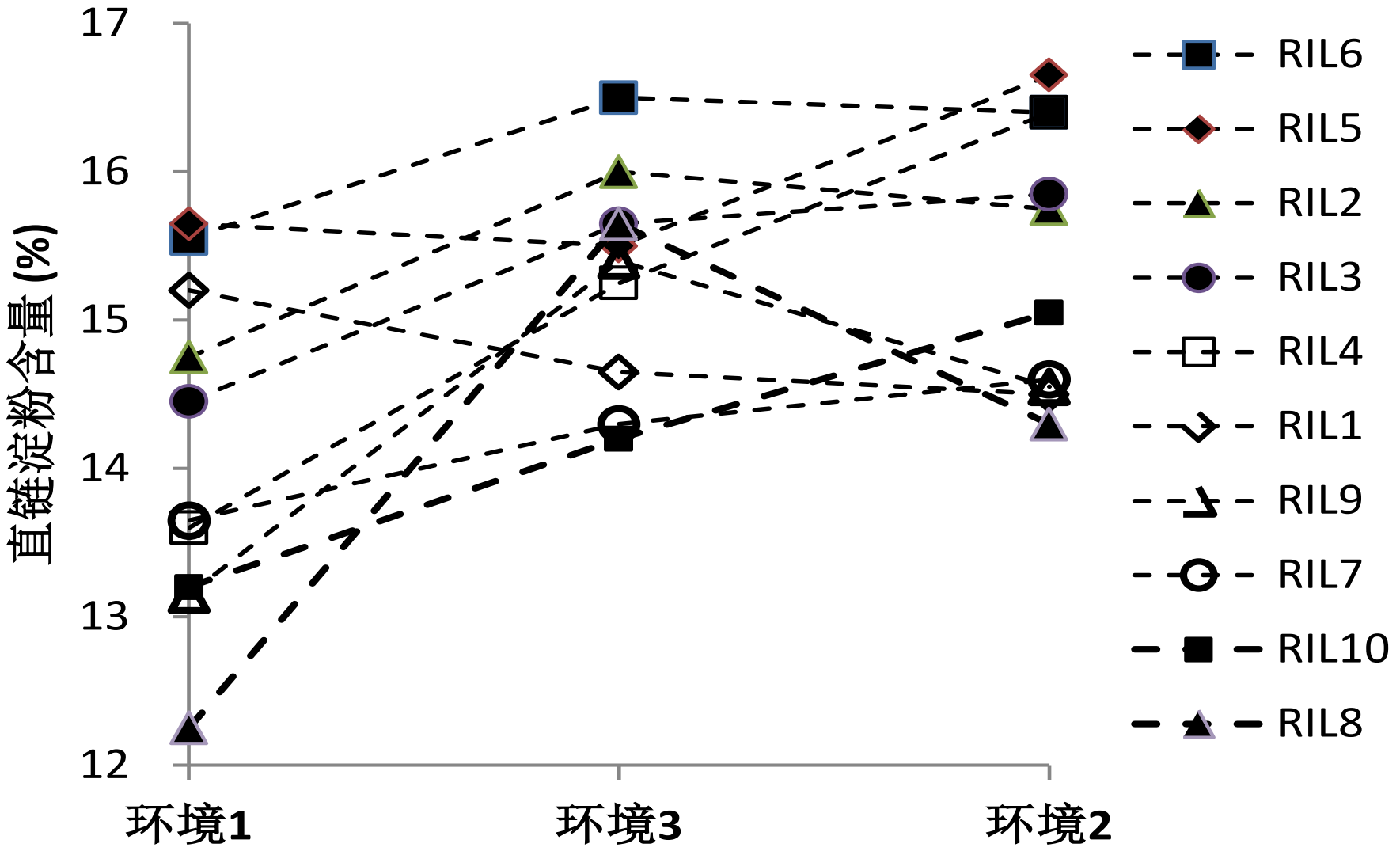
- 在固定效应模型的方差分析中，第i个基因型在第j个环境下的平均表现μ_{ij}是一个待估计的未知参数，第k个表型值y_{ijk}服从正态分布 。第i个基因型在第j个环境下k个观测值的样本均值，就是μ_{ij}的最优线性无偏估计（BLUE）。
- 所有μ_{ij}（i=1, …, g; j=1, …, e）的无偏估计在一起，可以排列成一个基因型和环境双向表，这一双向表是基因型环境稳定性分析的基础数据。

水稻10个RIL家系在3个地点的平均直链淀粉含量（%）及各种效应的估计值

列平均又称环境平均或环境指数（environmental index）

基因型	环境1	环境3	环境2	行平均	基因型效应	互作效应		
						环境1	环境3	环境2
RIL6	15.55	16.50	16.40	16.15	1.20	0.21	-0.20	-0.01
RIL5	15.65	15.50	16.65	15.93	0.98	0.53	0.26	-0.79
RIL2	14.75	16.00	15.75	15.50	0.55	0.06	-0.20	0.14
RIL3	14.45	15.65	15.85	15.32	0.36	-0.06	0.08	-0.02
RIL4	13.60	15.25	16.40	15.08	0.13	-0.68	0.86	-0.19
RIL1	15.20	14.65	14.50	14.78	-0.17	1.23	-0.74	-0.49
RIL9	13.15	15.40	14.55	14.37	-0.59	-0.41	-0.27	0.68
RIL7	13.65	14.30	14.60	14.18	-0.77	0.27	-0.04	-0.24
RIL10	13.20	14.20	15.05	14.15	-0.80	-0.14	0.45	-0.31
RIL8	12.25	15.65	14.30	14.07	-0.89	-1.01	-0.22	1.23
列平均	14.15	15.31	15.41	14.95				
环境效应	-0.81	0.36	0.45					

平均直链淀粉含量的散点图



基因型的环境稳定性分析

- 从表9.10的基因型和环境双向表中可以看到，有些基因型在环境之间的波动较小，而有些却很大。对于波动较小的基因型来说，环境的改变不会引起表型上大的变化，或者说这些基因型对环境的改变不太敏感，或者它们的环境稳定性较高。而对波动较大的基因型来说，环境的改变会引起表型上较大的变化，或者说它们对环境的改变比较敏感，或者环境稳定性较低。
- 因此，根据多个基因型在多种环境下的表型数据，就能对基因型的环境敏感性进行评价，有时又称稳定性分析（stability analysis）。利用基因型和环境双向表，可以构建出多种稳定性参数，从不同的侧面来定量评估基因型对环境的敏感程度或稳定程度。

环境稳定性的Finlay-Wilkinson分析方法

- 互作的联合回归分析（Joint regression analysis of interaction）方法首先是由Yates和Cochran（1938年）提出，当时并没有得到多少应用。Finlay和Wilkinsom（1963）重新提出并完善了这种分析方法，随后随着计算机的普及而得以较广泛的应用。
- 将第i个基因型在所有环境的平均表现对环境平均数进行一元线性回归分析，回归系数为：

$$b_i = \frac{\sum_j (\bar{y}_j - \bar{y})(y_{ij} - \bar{y}_i)}{\sum_j (\bar{y}_j - \bar{y})^2} = \frac{\sum_j y_{ij}\bar{y}_j - \frac{1}{e}(\sum_j y_{ij})(\sum_j \bar{y}_j)}{\sum_j \bar{y}_j^2 - \frac{1}{e}(\sum_j \bar{y}_j)^2}$$

稳定性的判定方法

- 回归系数越大，表示该品种对环境的反应越敏感，因而稳定性较低；回归系数越小，表示该基因型对环境的反应迟钝，因而稳定性较高。鉴于此，Finlay和Wilkinson（1963）提出以下利用回归系数评价基因型稳定性的方法。

  （1）当b_i=1，表示基因型i对环境的反应等于所有基因型的平均反应；

  （2）当b_i>1，表示基因型i对环境的反应高于所有基因型的平均反应，稳定性低；

  （3）当b_i<1，表示基因型i对环境的反应低于所有基因型的平均反应，稳定性高。

10个RIL家系对三个地点的回归系数

基因型	RIL6	RIL5	RIL2	RIL3	RIL4
基因型值	16.15	15.93	15.50	15.32	15.08
回归系数	0.73	0.40	0.91	1.08	1.88

基因型	RIL1	RIL9	RIL7	RIL10	RIL8
基因型值	14.78	14.37	14.18	14.15	14.07
回归系数	-0.52	1.46	0.67	1.21	2.17

- RIL2、RIL3、RIL10的回归系数接近于1，它们具有中等程度的环境稳定性。RIL4、RIL9、RIL8的回归系数远高于1，它们的环境稳定性较低。RIL6、RIL5、RIL7的回归系数远低于1，它们的环境稳定性较高。RIL1的回归系数为负值，但其绝对值远低于1，也具有较好的环境稳定性。

稳产与高产的关系

- 在前面评价基因型的稳定性时，我们只是用高或低来称呼，而不是好或坏。稳定性反映的是基因型对环境的敏感程度，脱离了优异平均表现的高稳定性，显得毫无价值。
- 选择平均表现高同时又具有较高环境稳定性的品种，才是育种的重要目标。这样的品种往往具有广泛的适应性，能够在更广阔的地区进行种植，在农业生产上发挥更大的作用。
- 从表9.11来看，基因型值的高低似乎和回归系数的高低没有必然联系，基因型值在育种中的重要性要远高于回归系数这一稳定性参数。

稳定性的Eberhart-Russell分析方法

- Eberhart和Russell（1966）提出利用二个参数测定品种的稳定性，一是品种对环境的反应参数（parameter of response），即前面的回归系数；二是离回归的方差，称为稳定性参数（parameter of stability）。

$$y_{ij} = \bar{\mu} + G_i + E_j + GE_{ij}$$

$$y_{ij} = \bar{\mu} + G_i + b_i E_j + \delta_{ij}$$

$$y_{ij} = \bar{\mu} + G_i + E_j + (b_i - 1)E_j + \delta_{ij}$$

离回归方差的计算

- 在得到回归系数的估计值后，对基因型i在环境j下的表现进行预测，并计算回归离差：

$$\hat{y}_{ij} = \overline{\mu} + G_i + \hat{b}_i E_j \qquad \hat{\delta}_{ij} = y_{ij} - \hat{y}_{ij}$$

- 对于重复观测数据计算离差方差：

$$s^2_{\delta(i)} = \frac{1}{e-2}\sum_j \hat{\delta}^2_{ij} - \frac{1}{r} MS_\varepsilon$$

- 较低的方差意味着回归系数解释了较多的基因型在环境间的变异。

稳定性的判定方法

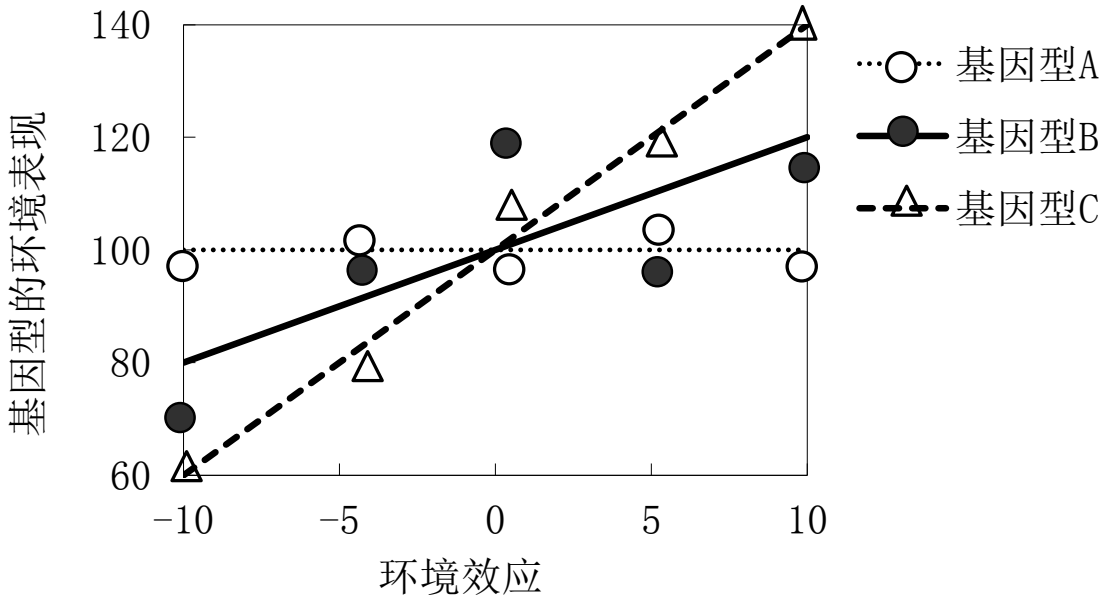
- 当b=1、s_δ^2不显著时，说明该基因型的表现比较稳定，基因型与环境的线性关系成立。
  - 对回归系数b大于1的基因型来说，环境效应每增加一个单位，表型相应增加b个单位，因而可在环境较好的条件中种植，以充分发挥其生产潜力；对于b=1的基因型，则认为具有广泛的适应性。
  - 如果回归系数b<1，则称该基因型很稳定，可种植在条件比较差的环境中。

稳定性的判定方法

- 当s_δ^2不显著时，可用回归系数b预测基因型在不同环境下的表现。凡是可以用回归系数来预测其表现的基因型，均称为稳定性基因型，因而稳定性包含了可预测（predictability）的含义。
- 如果s_δ^2与随机误差之间存在显著的差异，则表明互作效应与环境的关系并非线性，利用回归系数难以预测基因型在不同环境下的表现。
- 利用回归系数和离差平方和可以将稳定性划分出更多的类型。

基因型对环境的反应类型

- 基因型A在环境间有类似的表现，回归系数为0、离差平方和也很低，稳定性程度最高。



- 基因型B的环境回归系数为1、离差平方和较高，但从回归系数来看具有一定的稳定性，但从离差平方和来看，稳定性程度较低。
- 基因型C的环境回归系数高于1、离差平方和较低，从回归系数来看稳定性程度较低，但从离差平方和来看，稳定性程度较高。

稳定性分析的育种价值

- 实际中，究竟应该选择哪一种稳定性，不同的育种家有不完全一致的答案。一般来说，育种家可能希望基因型A的稳定性，同时又希望这种基因型在所有环境下都表现优良。
- 遗憾的是，具有基因型A那样的稳定性，它们的平均表现一般都不会太好，育种家必须考虑其它类型的稳定性。同时需要说明的是，稳定性只是衡量基因型对环境的敏感程度，如果不同时考虑基因型的平均表现，则稳定性参数对育种的指导作用是有限的。

基因型和环境互作的乘积模型

- 在大多数产量试验中，环境只是来自TPE的一个有限样本，双向表给出的基因型*i*在环境*j*下平均表现的估计值中存在随机误差；双向表给出的基因型和环境互作效应并不都是可以重复的。
- 主效相加互作相乘模型（AMMI，additive main effects and multiplicative interaction）是常用的乘积模型，其目的是从互作效应中分解出可重复的部分，以更好地预测不同基因型在不同环境下的表现。AMMI模型对主效应采用方差分析方法，对互作效应采用主成分分析。

AMMI乘积模型

$$y_{ijk} = \mu + G_i + E_j + \sum_{m=1}^{N} (\mathrm{IPCA}_m^{Gi})(\mathrm{IPCA}_m^{Ej}) + \delta_{ij} + \varepsilon_{ijk}$$

- 求和符号中的两个乘积项分别是第i个基因型在第m个坐标轴上的IPCA（主成分）得分和第j个环境在第m个坐标轴上的IPCA得分，N是主坐标的个数，δ_{ij}是基因型和环境互作效应中主成分分析后的剩余部分。

AMMI乘积模型对互作效应的分解

$$GE_{ij} = \sum_{m=1}^{N} (\mathrm{IPCA}_m^{Gi})(\mathrm{IPCA}_m^{Ej}) + \delta_{ij}$$

- AMMI认为基因型在特定环境下的表型观测值不是真实表现的最好估计，AMMI分析方法的主要目的是改进基因型在特定环境下表型的估计。
- 基因型和环境互作可以被分解为两部分，一部分是可重复的互作效应，另一部分是非重复互作。因此，在对特定环境下的表型进行估计时，如果能够扣除非重复互作这种噪音的影响，就可达到提高估计值的精确度的目的。

AMMI乘积模型对表型的预测

- AMMI模型的一个基本假定是，少数的几个主成分可以解释互作方差中大部分可重复的互作变异。忽略噪音的影响，下面的公式给出基因型i在环境j下表型的估计值，N一般取1、2或3。

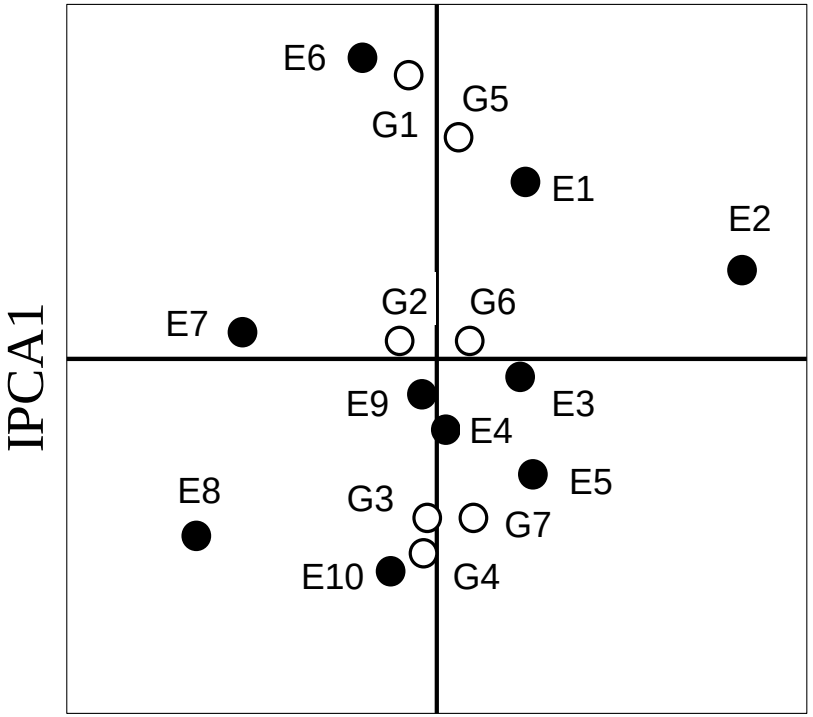
$$\hat{P}_{ij} = \mu + g_i + e_j + \sum_{m=1}^{N} (IPCA_m^{Gi})(IPCA_m^{Ej})$$

AMMI乘积模型的双标图

- AMMI分析结果一般用一个双标图（bi-plot）表示。双标图中，以基因型或环境与总平均数的离差作为*X*轴，IPCA得分作为*Y*轴。

包含7个基因型（用G1~G7表示）和10个环境（用E1~E10表示）的AMMI分析双标图

- 具有较高平均表现的基因型和环境出现在双标图的右上方，如基因型5和6、环境1和2。
- 如果某个基因型和某个环境的IPCA1得分或者同正，或者同负，这时的基因型和环境互作是有利的，如基因型3与环境8与10之间的互作。
- 如果某个基因型和某个环境的IPCA1得分一正一负，这时的基因型和环境互作是不利的，如基因型3与环境6或7之间的互作。



AMMI分析的主成分数和育种价值

- 一些经验表明，*N*=1往往就是很好的模型。AMMI的有用性体现在IPCA得分所能解释的可重复互作变异的大小。在有些试验中，第一主成分可以解释60~80%的互作变异，利用AMMI分析得到很好的结果。在有些试验中，第一主成分只可以解释10~30%的互作变异，利用AMMI分析不一定会得到很好的结果。
- AMMI预测结果也依赖于互作变异中可重复部分的大小，以及可重复变异的可重复性。利用这一信息去预测基因型在环境8中的表现的前提是，这种有利互作在年份间是可重复的。有些环境因素在年份间保持相对稳定，如土壤特性和耕作方式等，而有些环境因素，如降水量和温度等，不容易预测。因此AMMI分析利用的是年份间能够重复的互作。

附：关联分析的丢失遗传力现象

Heritability is missing in GWAS!

(B. Maher, 2008. **Nature**, 456: 18-21)

"When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. **But they were nowhere to be seen**. "

The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

If you want to predict how tall your children might one day be, a good bet would be to look in the mirror, and at your mate. Studies going back almost a century have estimated that height is 80–90% heritable. So if 29 centimetres separate the tallest 5% of a population from the shortest, then genetics would account for as many as 27 of them[1].

This year, three groups of researchers[2–4] scoured the genomes of huge populations (the largest study[4] looked at more than 30,000 people) for genetic variants associated with the height differences. More than 40 turned up.

But there was a problem: the variants had tiny effects. Altogether, they accounted for little more than 5% of height's heritability — just 6 centimetres by the calculations above. Even though these genome-wide association studies (GWAS) turned up dozens of variants, they did "very little of the prediction that you would do just by asking people how tall their parents are", says Joel Hirschhorn at the Broad Institute in Cambridge, Massachusetts, who led one of the studies[2].

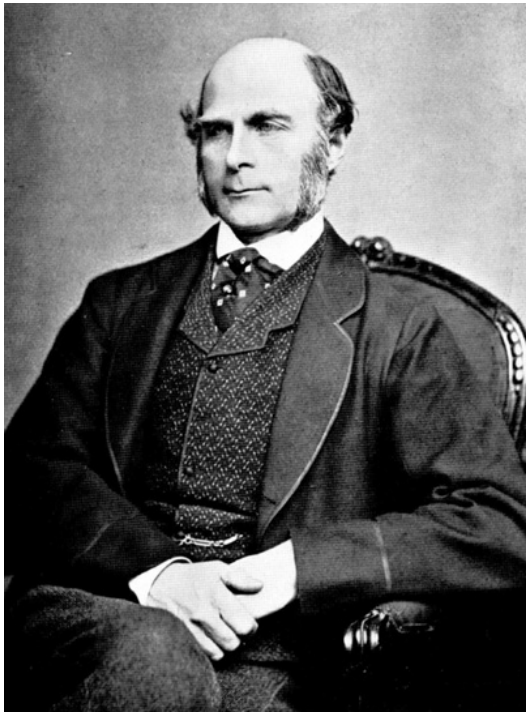
Height isn't the only trait in which genes have gone missing, nor is it the most important. Studies looking at similarities between identical and fraternal twins estimate heritability at more than 90% for autism[5] and more than 80% for schizophrenia[6]. And genetics makes a major contribution to disorders such as obesity, diabetes and heart disease. GWAS, one of the most celebrated techniques of the past five years, promised to deliver many of the genes involved (see 'Where's the reward?', page 20). And to some extent they have, identifying more than 400 genetic variants that contribute to a variety of traits and common diseases. But even when dozens of genes have been linked to a trait, both the individual and cumulative effects are disappointingly small and nowhere near enough to explain earlier estimates of heritability. "It is the big topic in the genetics of common disease right now," says Francis Collins, former head of the National Human Genome Research Institute (NHGRI) in Bethesda, Maryland. The unexpected results left researchers at a point "where we all had to scratch our heads and say, 'Huh?'", he says.

Although flummoxed by this missing heritability, geneticists remain optimistic that they can find more of it. "These are very early days, and there are things that in the next year or two that may well explain another sizeable chunk of heritability," says Hirschhorn. So where might it be hiding?

ILLUSTRATIONS BY D. PARKINS

1. How was Heritability found, defined and estimated 100 years ago?

Sir Francis Galton (1822-1911)



Hereditary Stature by F. Galton (1886)

HEREDITARY STATURE [1]

IT will perhaps be recollected that, at the meeting last autumn of the British Association in Aberdeen, I chose for my Presidential Address to the Anthropological

[1] Extracts from Mr. F. Galton's Presidential Address to the Anthropological Institute, January 26.

almost absurdly simple, and not only so, but it is explained most easily by a working model that altogether supersedes the trouble of calculation. I exhibit one of these: it is a large card ruled with horizontal lines 1 inch apart, and numbered consecutively in feet and inches, the value of 5 feet 8 inches lying about half way up. A pin-hole is bored near the left-hand margin at a height corresponding to 5 feet $8\frac{1}{4}$ inches. A thread secured at

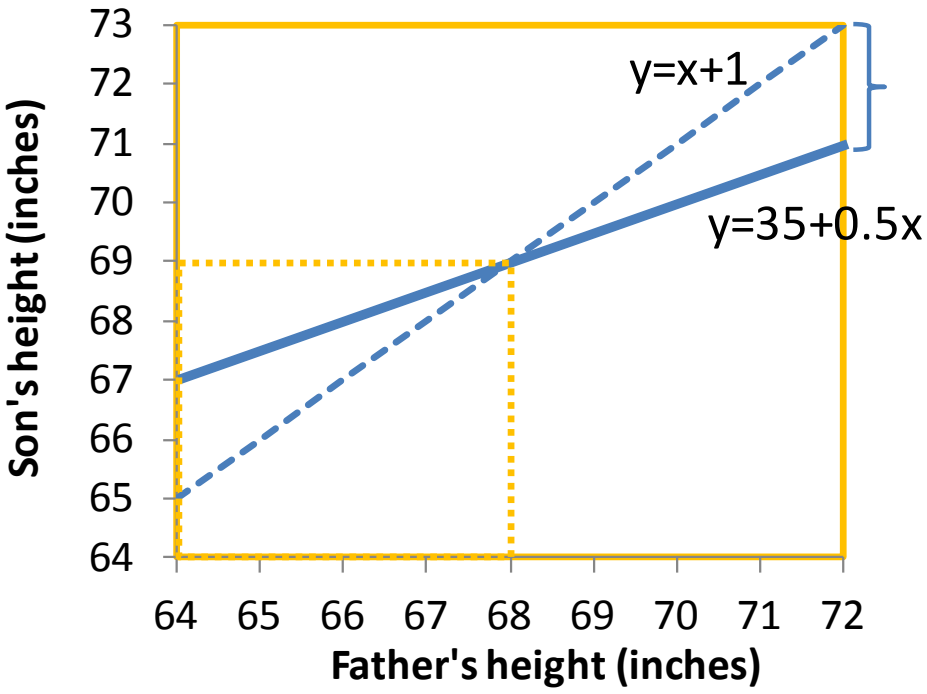
the back of the card is passed through the hole; when it

already explained, we shall see from the divisions on the

- 1078 pairs of son (y) and father (x)
- Average of sons: m(y) = 69 inches
- Average of fathers m(x) = 68 inches
- On average, taller father has taller son
- Can we use y=x+1 to predict son's stature?

Regression of son on father's height

- When grouping on fathers
  - For fathers **x=72** [4 in. taller than m(x)], **y=71** (2 in. shorter than x+1 and 1 in. shorter than x);
  - For fathers **x=64** [4 in. shorter than m(x)], **y=67** (2 in. taller than x+1 and 3 in. taller than x;



Regression of offspring on mid-parent height

- Slope from offspring and mid-parent is higher than slope from son and father!

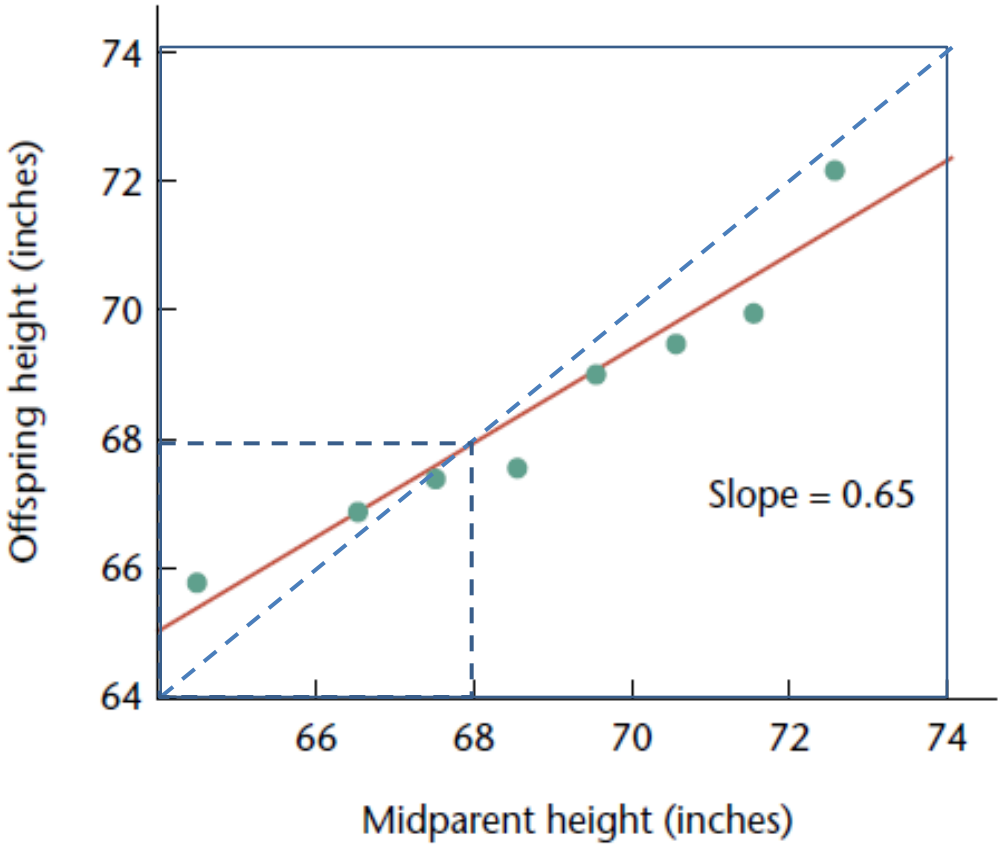


Figure 1 Galton's 1889 plot of average parental height versus average height of offspring.

Components of genetic value and components of genetic variance

Effect	Variance
Phenotype value (P)	Phenotype variance (V_P)
Genotype value (G)	Genotype variance (V_G)
Breeding value (A)	Additive variance (V_A)
Dominant deviation (D)	Dominant variance (V_D)
Epistatic deviation (I)	Epistatic variance (V_I)
Genotype by environment interaction (GE)	Genotype by environment variance (V_{GE})
Random error (ε)	Random error variance (V_ε)

$$V_P = V_G + V_E + V_{GE} + V_\varepsilon$$

$$V_G = V_A + V_D + V_{AA} + V_{AD} + V_{DA} + V_{DD} + \cdots$$

Estimation of Heritability in animals

(Falconer and Mackay "Introduction to Quantitative Genetics", p162)

Species	Trait	h^2 (%)	Species	Trait	h^2 (%)
Man	Stature	65	**Poultry**	Body weight at 32 wks	55
	Serum immuno-globulin (IgG) level	45		Egg weight at 32 wks	50
Cattle	Body weight (adult)	65		Egg production to 72 wks	10
	Butterfat (%)	40	**Mice**	Tail length at 6 wks	40
	Milk yield	35		Body weight at 6 wks	35
Pigs	Back-fat thickness	70		Litter size (1st litters)	20
	Efficiency of food conversion	50	***Drosophila melanogaster***	Abdominal bristle number	50
	Weight gain per day	40		Body weight	40
	Litter size	5		Ovary size	30
				Egg production	20

Nature Reviews Genetics, 2008, 9: 255-266

FUNDAMENTAL CONCEPTS IN GENETICS

Heritability in the genomics era — concepts and misconceptions

Peter M. Visscher, William G. Hill‡ and Naomi R. Wray**

Abstract | Heritability allows a comparison of the relative importance of genes and environment to the variation of traits within and across populations. The concept of heritability and its definition as an estimable, dimensionless population parameter was introduced by Sewall Wright and Ronald Fisher nearly a century ago. Despite continuous misunderstandings and controversies over its use and application, heritability remains key to the response to selection in evolutionary biology and agriculture, and to the prediction of disease risk in medicine. Recent reports of substantial heritability for gene expression and new estimation methods using marker data highlight the relevance of heritability in the genomics era.

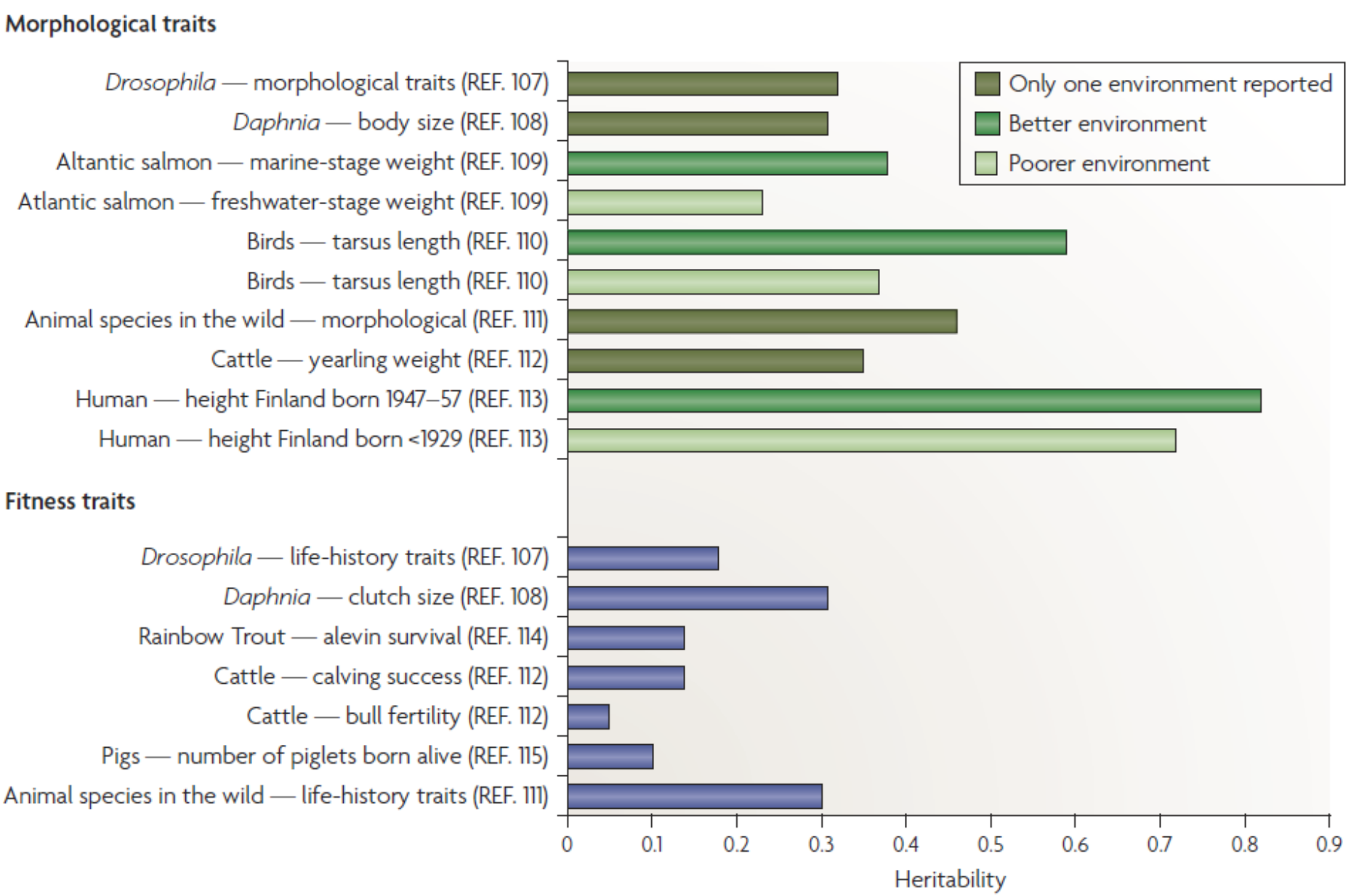


Figure 1 | **Examples of estimates of heritabilities of morphological and fitness traits.** Where possible, the estimates of heritability were taken from Reviews, and are the mean across a number of studies. The examples show that, on average, heritability estimates are larger for morphological traits than for fitness-related traits, and that heritability tends to be larger in better environments when compared with poorer environments.

2. Heritability is missing in GWAS. That's too sad and too bad!

What is missing Heritability?

- Wikipedia: The **"missing heritability" problem** can be defined as the fact that individual genes cannot account for much of a disease's heritability.
- Turkheimer (2011): **The missing heritability problem** refers to the gap between heritability estimates for complex human traits based on quantitative genetics and the small magnitude and unreliability of contemporary molecular genetics, especially genome wide association studies.
- E. Lander (2012, PNAS): Human genetics has been haunted by **the mystery of "missing heritability"** of common traits. Although studies have discovered >1,200 variants associated with common diseases and traits, these variants typically appear to explain only a minority of the heritability (20–30% in some well-studied cases and >50% in a few).

Evidence of the missing Heritability in GWAS

(adapted from Table 1, **Nature**, 2009, 461: 747-753)

Estimates of heritability and number of loci for several complex traits			
Disease	**No. of loci**	**Heritability explained**	**Heritability measure**
Age-related macular degeneration	5	**50%**	Sibling recurrence risk
Crohn's disease	32	**20%**	Genetic risk (liability)
Systemic lupus erythematosus	6	**15%**	Sibling recurrence risk
Type 2 diabetes	18	**6%**	Sibling recurrence risk
HDL cholesterol	7	**5.2%**	Residual* phenotypic variance
Height	40	**5%**	Phenotypic variance
Early onset myocardial infarction	9	**2.8%**	Phenotypic variance
Fasting glucose	4	**1.5%**	Phenotypic variance

* Residual is after adjustment for age, gender, diabetes.

Evidence of the missing Heritability in GWAS

(**Nature Genetics**, 2010, 42: 570-576)

Table 1 Estimated numbers of common susceptibility SNPs, and associated genetic variances explained, for three complex traits

	Estimated number of total loci (95% CI)	Total GV[a] explained by estimated loci (95% CI)	Observed range of effect sizes (% GV)
Height	201 (75, 494)	16.4 (10.6, 30.6)	0.04–1.13
Crohn's disease	142 (71, 244)	20.0 (15.7, 28.0)	0.07–1.96
BPC[b] cancers	67 (31, 173)	17.1 (11.6, 35.8)	0.14–1.82

All the projections were performed using a nonparametric method and are restricted to the range of observed effect sizes for known susceptibility SNPs (shown in the last column).

[a]All genetic variances (GV) are shown as a percentage of the total variance of the trait attributable to heritability. For Crohn's disease and BPC cancers, the variance due to heritability is computed from estimates of sibling relative risk using a log-normal model for risk[5]. [b]All estimates should be interpreted as averages over the three cancers.

Evidence of the missing Heritability in GWAS

(adapted from E. Lander's lecture at NIH, 20 May 2011)

Common variants: Heritability increasing		
Disease	**No. of loci**	**Heritability explained**
Type I diabetes	41	**60%**
Fetal Hemoglobin levels	3	**50%**
Age-related macular degeneration	3	**50%**
Crohn's disease	32	**25%**
Type 2 Diabetes	18	**25%**
HDL: LDL cholesterol	95	**25%**
Height	180	**12%**
Remainder:		

- *Common variants of smaller effects? Evidence for large contribution*
- *Rare variants? Some loci found so far, although contribution still small*
- *Something else?*

Who cares the missing heritability?

- **Nature vs. nurture (基因决定论对环境决定论) debate**
  - F. Galton: the evidence from twin studies **favored nature rather than nurture**. “It would be quite practicable to produce a highly-gifted of men by judicious marriages during several consecutive generations.”
    - 1860: F. Galton develops ideas of promoting “hereditary genius” through breeding
    - 1883: F. Galton coins term ‘eugenics’ to describe his movement
    - 1912: UK Mental Deficiencies Bill withdrawn after campaign by Josiah Wedgwood
    - 1927: US Supreme Court upholds compulsory sterilization laws in Buck v Bell case
    - 1933: 400,000 forced sterilization in Nazi Germany

Who cares the missing heritability?

- **Nature vs. nurture debate in human**
  - John B. Watson "Give me a dozen healthy infants, well-formed, and my own specified world to bring them up in and I'll guarantee to **take any one at random and train him to become any type of specialist** I might select – doctor, lawyer, artist, merchant-chief and, yes, even beggar-man and thief, regardless of his talents, penchants, tendencies, abilities, vocations, and race of his ancestors."
  - Behavioural genetics
    - Late 19th century: F. Galton studies heritable basis of behaviour
    - 1970s: Sociobiology movement suggests evolutionary influences on human behaviours
    - Late 20th century: Twin studies demonstrate heritable influence on multiple personality and behavioural traits

Who cares the missing heritability?

- **Nature vs. nurture debate in human**
  - Intelligence
    - Late 19th century: F. Galton studies heritable basis of intelligence
    - 1980s: Twin studies demonstrate heritable influence on multiple personality and behavioural traits
    - 1988: Discovery of possible link between IQ and IGF2R gene
  - Implications in law: In some cases, lawyers for violent offenders have begun to argue that an individual's genes, rather than their rational decision-making processes, can cause criminal activity.
    - 1991: Stephen Mobley robbed a branch of Domino's Pizza and shot the manager dead.
    - 1995: Stephen Mobley appeals murder conviction on grounds of his genetic profile. His lawyers argued, "Mobley's genes made him do it".
    - 2005: The appeal was thrown out, and Mobley was executed.
- **摩尔根遗传学与米丘林遗传学之争**