瓦维诺夫拟态中人类选择的基因组证据

翻译自：

Ye C Y, Tang W, Wu D, et al. Genomic evidence of human selection on Vavilovian mimicry[J]. Nature ecology & evolution, 2019, 3(10): 1474-1482.

译者：

浙江大学农业与生物技术学院应用生物科学专业（1701）

彭影彤 唐瑜悦 施杨琪 古爱媛 邓博文 许焕一 叶永澎 李超峰 刘畅 刘臣涛 陈斌焕 陈之豪

修改：

叶楚玉副教授

*该译文来自《生物信息学》课程作业（樊龙江主讲）

【摘要】

瓦维诺夫拟态是指杂草在形态上逐渐进化成与驯化作物相似的一种植物拟态现象，被认为是人类无意识选择的结果。解析瓦维诺夫拟态的分子机制将扩展我们对拟态这一进化现象的认知，同时有助于我们理解农业杂草的起源和进化，而这也是作物学领域的重要研究内容之一。为此，我们收集了长江流域的拟态和非拟态稗草（*Echinochloa crus-galli*）材料，进行了表型观察以及基因组重测序。稻田中的拟态稗草逐渐进化出小分蘖角从而使得苗期稗草在形态上与栽培水稻相似。进一步分析表明，拟态稗草是在约1000年前从非拟态稗草群体中演化化而来的，并经历遗传瓶颈效应。稗草基因组在拟态过程中受到选择的区域包含87个株型相关基因（这其中包括控制植物分蘖角度的关键基因 *LAZY1*）。该项研究为瓦维诺夫拟态中的人类选择提供了基因组水平的证据。

瓦维诺夫拟态是以前苏联著名植物学家和遗传学家 Nikolai Vavilov的名字命名的，瓦维诺夫拟态（或称作物拟态)描述的是一种进化适应性现象，指杂草通过进化使得其在特定时期（如稗草苗期）的形态与驯化作物相似，从而避免被清除。在早期农业生产

中，人类依赖肉眼辨别作物与杂草来进行杂草清除，而 瓦维诺夫拟态使得农民难以区分二者，从而让杂草可以逃避拔除。与大多数其他拟态系统相似，瓦维诺夫拟态涉及三个参与者：（1）被模仿者，被模仿的作物；（2）模仿者，即模仿作物的杂草；（3）受欺骗者，即需要识别并清除杂草的农民。

阐述得最清楚的拟态是在动物中，包括贝氏拟态和缪勒拟态。这两种类型拟态的分子机制已经分别在凤蝶和袖蝶中进行了解析。植物拟态描述得较少。如在兰花中花的拟态较为常见，在一些植物中也有叶片拟态的报道。瓦维诺夫拟态是一种独特的植物拟态，发生在农业生态系统中，是人类无意识选择的结果。瓦维诺夫拟态的一个著名的例子就是稻田里的稗属杂草在形态上类似于栽培水稻。稗属(Poaceae)植物属于禾本亚科，多为杂草。其中，稗草*E.crus-galli* 在稻田中为优势种，形态多样，与水稻拟态和非拟态的均有。Barrett[1]比较了拟态和非拟态稗草与栽培水稻的15个形态和生长特征，结果表明，栽培水稻和一种四倍体拟态稗草（*E.crusu-galli* var.*oryzicola*，也称为*E.oryzicola*或*E.phyllopogon*）在这些表型特征上相似（聚集在一起），而非拟态的六倍体稗草*E. crus-galli* var. *crus-galli*明显不同。拟态稗草具有小的分蘖角和叶片夹角，而非拟态稗草具有松散的、下垂的叶片，株型通常呈匍匐生长形态。Barrett 实验室还比较了几种稗草的生活史特征，包括开花时间、干重分配、繁殖力和种子特性。例如稗草*E.crus-galli* var.*oryzicola*的种子是 *E.crus-galli* var.*crus-galli*的两到三倍重。基于同工酶变异分析发现，相比于拟态的稻田稗草*E.oryzoides*（六倍体）和*E.phyllopogon*（四倍体），稗草*E.crus-galli* var.*crus-galli*具有更丰富的遗传多态性。

杂草中也存在种子拟态——一个典型的例子是一种一年生杂草——*Camelina sativa*，它的种子表现出类似亚麻种子的风选特性，使得两者在扬场时很难被分开。作物拟态也存在于其他一些物种中，包括黑麦(*Secale* spp.)、燕麦(*Avena* spp.)和毒麦(*Lolium temulentum*)。由于这些拟态杂草与驯化物种有许多共同的适应性特征，Fuller和Stevens也把它们称为寄生作物（parasitic domesticoids）。

拟态是一种有趣的现象，了解其机制在进化生物学领域有重要意义。而瓦维诺夫拟态发生在有人类活动参与的农业生态系统中，因此揭示其分子机制将有助于我们更进一步理解拟态这一适应性进化现象，以及农田杂草的起源和进化。虽然已经有些文献描述了稗草 *E.crus-galli* 拟态水稻的现象，但这种拟态现象背后的分子机制以及拟态过程中基因组的变化在很大程度上仍是未知的。我们课题组之前已经报导了稗草 *E.crus-galli* 全基因组序列[29]，为揭示这一独特拟态现象背后的基因组机制提供了机会。稗草拟态水稻被认为是由于不断的农民除草引起的，因此我们认为：**在稗草拟态过程中，来自人类的选择压是可以在稗草基因组区域中检测到的。**为了验证这一假设，我们对典型水稻产区的中国长江流域的 328 个稗草 *E.crus-galli* 进行了全基因组重测序。我们的分析结果揭示了

稗草 *E.crus-galli* 与“人为选择下的 Vavilovian 拟态”有关的基因组足迹。

【结果】

【稗草（*E. crus-galli*）在苗期的拟态】

在苗期时，稻田稗草 *E.crus-galli* 和栽培水稻形态相似，使得农民难以区分二者，从而难以被清除。这种苗期拟态包括了多种表型性状。基于先前的研究结果[1]以及我们自己的观察发现，拟态稗草和栽培稻在苗期的相似表型主要包括：分蘖角小、茎节直、茎基部绿色和叶片紧凑（即叶夹角小）。相比之下，非拟态稗草通常呈现出松散或匍匐的株型，常伴随膝状节、红色或紫色的茎基部以及披散的叶片（即叶片夹角大）等表型性状(图 1 a, b 和补充图.1)。因此，我们利用这四个代表性的形态特征来定义稗草苗期的拟态。将这四个性状赋不同数值，以它们的和来评估拟态水平，我们称之为拟态指数（MI）(详见方法和补充图 2)。MI 范围为 10 到 0，代表拟态水平从拟态变为非拟态。根据观察，将 MI≥8 的稗草定义为典型的拟态类型，MI≤2 的稗草定义为非拟态类型。

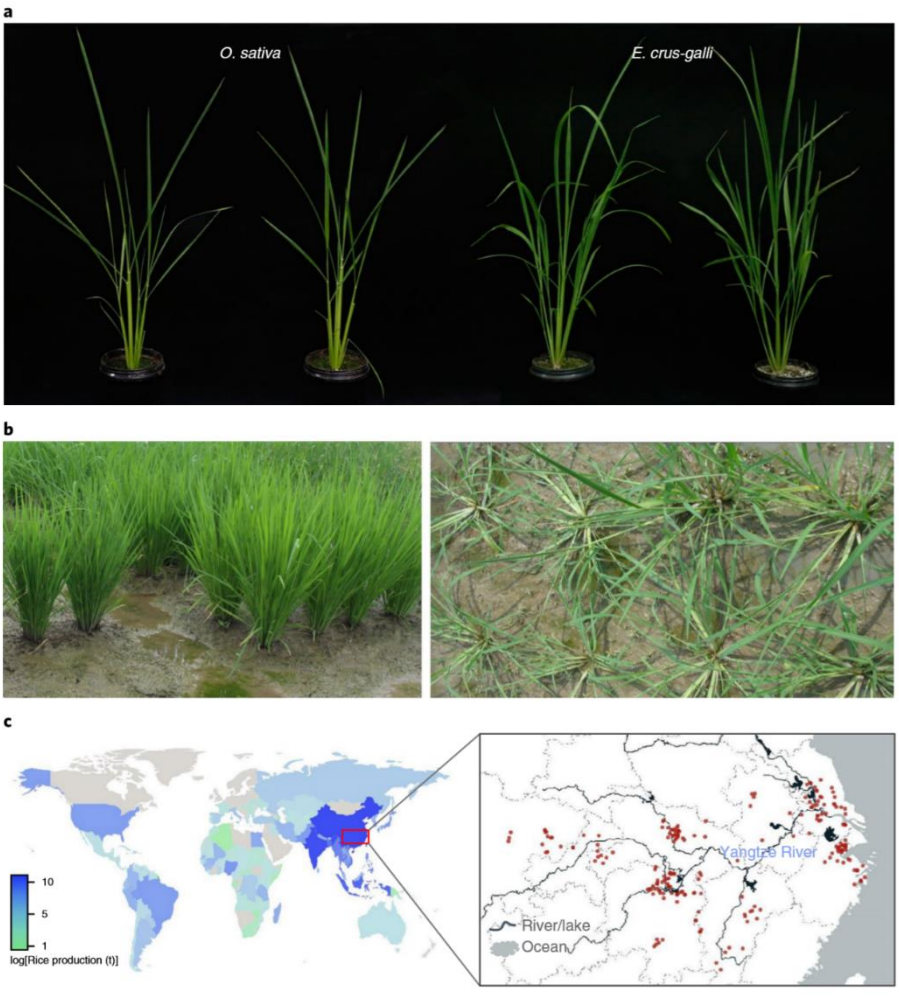


图 1. 稗草（*E. crus-galli*）拟态表型性状以及稗草的取样。a, 在温室中种植的栽培水稻（左）和拟态稗草（右）；b, 田间种植的典型的拟态（左）和非拟态（右）稗草；c, 本研究中的稗草取样，左图为稗草采样位置（方框标记）。地图上颜色深浅表示水稻产量，右图为取样稗草的地理分布（红点；n=328）。

【稗草（*E. crus-galli*）的取样、表型鉴定及测序】

我们从典型的水稻种植区域，我国长江流域不同地点的稻田及稻田周边区域收集了328 份稗草（*E. crus-galli*） 材料(图 1c)。将这些材料在田间进行种植，并进行为期两年（2017、2018）的苗期表型测定。根据拟态指数 MI，这些样品中的 150 份（占 45.7%）归为典型的拟态类型，121 份（36.9%）是非拟态类型，两年表型一致。同时，在田间观测到的拟态和非拟态稗草间的一些重要的表型差异，我们在温室两种不同的光周期条件下也观察到了一致的表型（图 1 b 和补充图 4）。

我们对这328份稗草进行了全基因组重测序，总共获取了7.17Tb的数据，平均覆盖深度15×（补充表1）。质控后的测序读段比对到稗草*E. crus-galli*（STB08）参考基因组上。每个样本大约有96.27%的读序可以覆盖超过92%的参考基因组,这个结果连同基因组大小估计(补充表1)表明该研究中所有样本均属于物种*E.crus-galli*（注：稗属植物较难从形态上鉴别具体物种）。我们获得了903万个全基因组高质量的单核苷酸多态位点（SNPs）和240万个小的插入/缺失（indels），平均每千碱基有9.07个变异。预计共有13101个变异位点对功能产生重要影响，这些SNP导致启动密码子丢失、转录延长或密码子终止（补充表2）。

【拟态和非拟态稗草（*E. crus-galli*）之间的基因组分化】

我们基于全基因组 SNP 构建了系统发生树，可以将这 328 份稗草材料分为三个主要的进化分枝（图 2a）。主成分分析（PCA）得到一致结果，也可以将这些材料分为三组（图 2b）。根据表型，我们将其中两个组分别定义为拟态（MIC）和非拟态（NMC）组，在这两个组中的大多数材料都是拟态或非非拟态的（图 2c 和补充图 5）。第三组被定义为混合组（MIX），其拟态指数 MI 均一分布，平均 MI 为 5.4（图 2c）。 拟态组（n=131）的平均 MI 值为 7.7，有 97 份（74.1%）材料表现出典型的拟态表型（即 MI≥8）。在非拟态组中（n=88），平均 MI 为 1.9，有 68（77.3%）份材料的 MI≤2（图 2c）。不同组的稗草材料其分蘖角度同样有显著差异，拟态组、混合组和非拟态组的平均分蘖角度分别为 27°、42°和 75°（图 2d）。后续分析集中在具有极端 MI 值的材料上，也就是来自拟态组的拟态材料（n = 97）和来自非拟态组的非拟态材料（n=68）。

从逻辑上讲，拟态稗草应该是在稻田出现之后从非拟态稗草中演变而来的。为了检验该假设，我们使用四倍体稗草 *E.oryzicola*（六倍体稗草 *E. crus-galli* 的父本）构建了系统进化树。 可以看出，非拟态组位于树的基部位置，更靠近四倍体稗草 *E. oryzicola*，拟态群体来自非拟态材料，支持了我们的假设（补充图 6）。

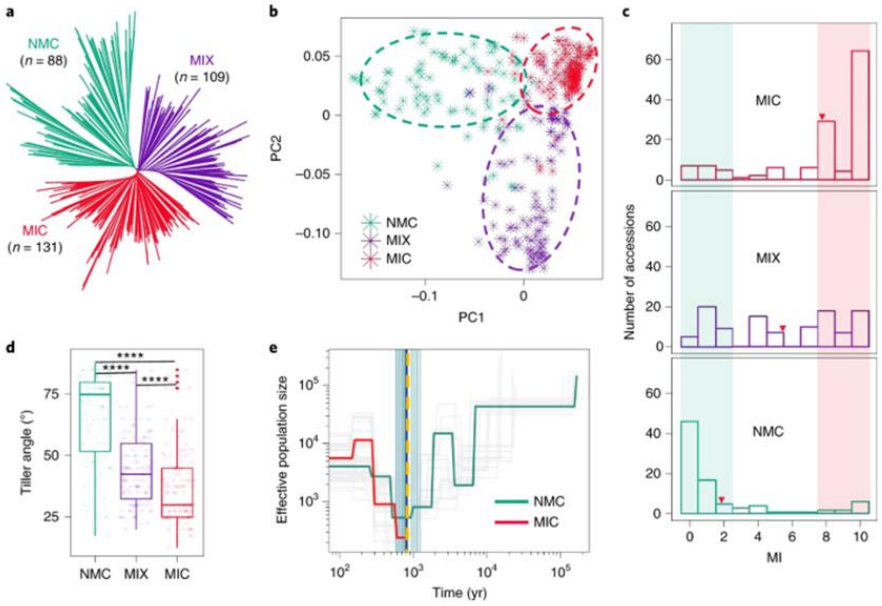


图 2 长江流域稗草（*E.crus-galli*）的系统发生和表型的分化。a，328 份稗草材料的最大似然树。b，基于全基因组 SNP 的 PCA 图。PC1 和 PC2 分别占遗传变异的 18.9％和 15.3％。红色（MIC），绿色（NMC）和紫色（MIX）虚线框分别表示三组。c，用拟态指数 MI 定量表示的三组之间的拟态表型分化。三组的平均 MI 用红色三角形标记。红色和绿色阴影分别表示典型的拟态和非拟态 MI 值。d，三组之间的分蘖角差异。****P<0.0001。e，利用 SMC ++计算的拟态和非拟态稗草群体之间的分化时间。中间虚线表示分化时间（大约 1000 年前）。

我们还发现，拟态群体中 92.5％的变异（SNPs 和 indels）是非拟态群体所具有的变异的子集，而非拟态群体包含更多变异，其中 17.1％是拟态材料不具有的。这些结果进一步支持了拟态群体由非拟态材料演变而来的假设。我们利用 SMC ++估算了分化时间，结果表明大约 1000 年前拟态群体从非拟态群体中演化而来（图 2e），这处于中国宋朝（公元 960-1279 年）时期。PSMC 分析显示，在 1 万年以前，拟态和非拟态的种群历史是一致的，这也支持了这两组是近期才分离（补充图 7）。

我们采用固定指数（FST）衡量了拟态和非拟态组之间的基因组分化情况。结果表明，在全基因组水平上，该值相对较低（FST = 0.062），这与拟态和非拟态组之间较近的分离时间的估计是一致的。我们检测到 255 个大于 100kb 的基因组区域其 FST 值大于 0.147（阈值最高 5%），包括七个 Mb 级的区域（补充表 3）。这表明，拟态和非拟态稗草在基因组的某些局部区域存在显著分化。

【稗草拟态过程中的正向选择信号】

我们利用遗传多态性和 Tajima's D 来检测稗草拟态过程中基因组区域是否受到选择。相对于非拟态稗草，拟态稗草在全基因组水平上遗传多态性降低（π），下降 1.324 倍（ROD=πNMC/πMIC），这表明稗草在拟态过程中经历了遗传瓶颈效应（图 3a）。检测中性位点、同义突变位点和基因区域的遗传多态性表明，全基因组遗传多态性较中性位点（ROD=1.113，P <0.0001）和同义突变位点（ROD=1.150，P<0.0001）其降低值更为明显，但基因区域的遗传多态性降低最为明显（ROD=1.650，P<0.0001）。这些结果表明，尽管有遗传漂移和其他因素的影响（补充图 8），正向选择在拟态进化过程中导致遗传多态性降低中也起着重要作用。此外，非拟态群体中 21.0%的中频 SNP（0.4<MAF≤0.5）（n=130,213）都在拟态群体中被固定（MAF≤0.2 或≥0.8），这也表明了拟态过程中的基因组选择（补充图 9）。在拟态群体中，有 280 个长于 100kb 的基因组区域其遗传多态性显著降低（ROD<2.504；阈值最高 5%），这表明这些区域在拟态进化过程中可能受到选择（补充表 4）。拟态（Tajima's D=0.456）和非拟态（Tajima's D=0.401；图 3b）群体的平均 Tajima's D 值相似。但是在拟态群体中，我们发现了总长度为 39.9 Mb 的基因组区域（包含 127 个长于 100 kb 的片段）具有显著 Tajima's D 负值（小于-1.774；置信度 95%；（图 3b 和补充表 5））。相反，在非拟态组中，只有 15.6Mb 区域的 Tajima's D 值小于−1.784（置信度为 95%），而且这部分区域具有显著 Tajima's D 负值的部分原因可能是非拟态群体中具有稀有等位基因比拟态群体多所导致的。综上所述，稗草基因组某些基因组区域在拟态过程中受到了选择。以两个长的基因组区域为例，scaffold10 和 scaffold16 中的两个基因组区域（分别为 2.21Mb 和 1.59Mb 的长度），其遗传多态性降低，拟态群体具有显著 Tajima's D 负值（补充图 10）。

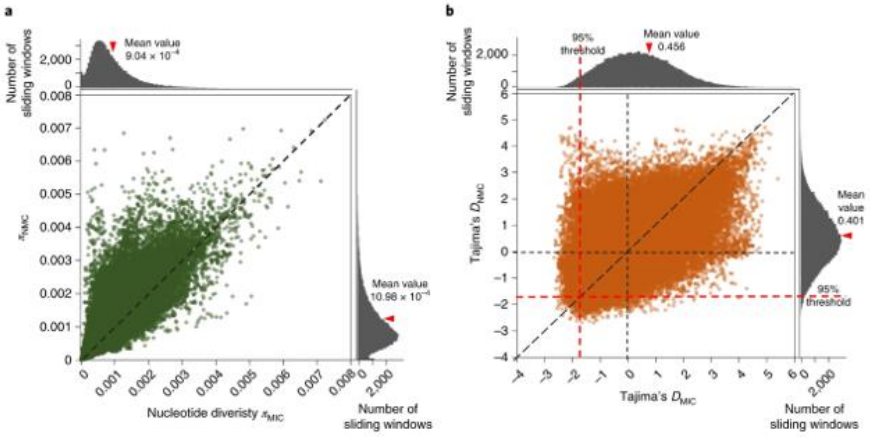


图 3 稗草拟态过程中基因组受到选择。a，20kb 步长、50kb 滑动窗口基因组核苷酸多态性的分布，表明非拟态稗草群体的 π 值高于拟态群体。b，20 kb 长的滑动窗口 Tajima's D 在整个基因组中分布的情况，表明拟态稗草基因组区域具有显著 Tjajima's D 负值。

【稗草拟态过程受选择的基因中株型相关基因显著富集】

根据上述基因组分化、遗传多态性和中性测验的结果，在基因组可能受选择区域中发现了总共 7596 个基因，具有 114652 个变异（93063 个 SNP 和 21589 个 indel）。进一步基于拟态和非拟态群体之间的等位基因频率的显著差异（$P < 1\times10^{-20}$）以及变异的影响情况（不包括同义突变）进行了过滤，共有 1986 个基因（占全基因组注释基因的 1.83%）在拟态过程中受到选择，包括了 8373 个变异（7167 个 SNP 和 1206 个 indel）位点。在这些基因中，有 87 个是水稻已知的株型相关基因的同源基因，这些基因可能是与稗草拟态性状相关的高可信度基因（补充表 6）。这 87 个基因中共有 455 个变异，包括 398 个 SNP 和 57 个 indel（图 4 和补充表 7）在拟态和非拟态群体中显著分化。此外，在所有 1986 个候选基因中，与稗草全基因组的基因相比，87 个株型相关基因是显著富集的（$P<0.001$）。

同时我们发现，与重力响应相关基因显著富集（$P<0.05$；图 4 和补充图 11）。之前研究表明，重力作用通过生长素的不对称分布和生长素信号应答途径导致分蘖的近轴和远轴面之间的细胞伸长差异，从而在控制植物分蘖角中发挥着关键的作用。在拟态性状相关的高可信的 87 个基因集里，发现了至少 1 个与植物重力感知有关的同源基因、与生长素极性分布有关的 14 个同源基因和与生长素信号响应有关的 7 个同源基因（补充图 11 和补充表 7）。此外，受选择基因中包括油菜素内酯信号传导途径相关的 5 个基因

的 22 个同源基因，油菜素内酯途径在叶夹角的调节中发挥重要作用（补充图 12 和补充表 7）。

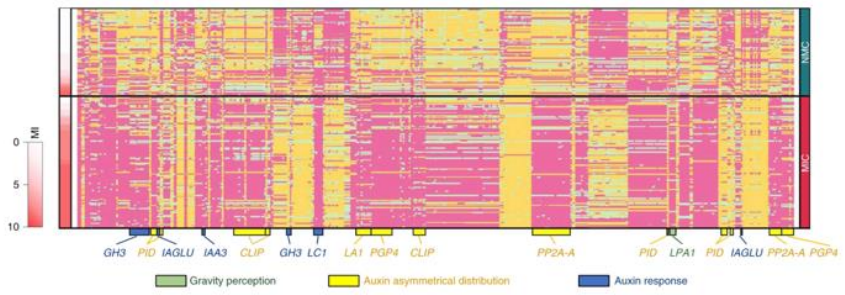
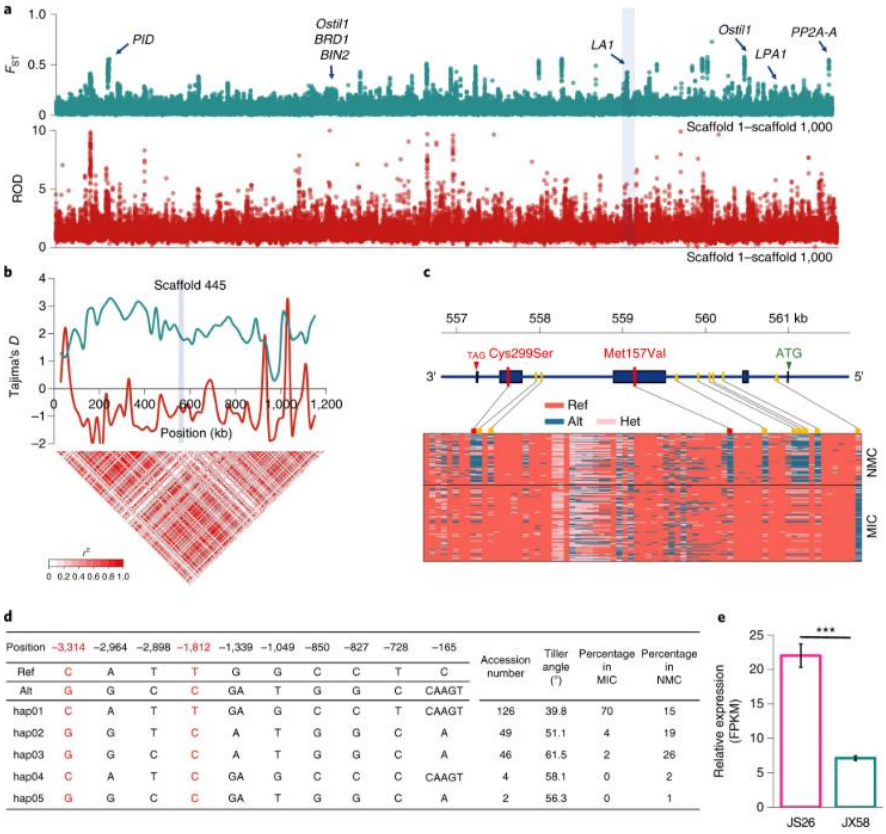


图 4 稗草拟态过程中 87 个植物株型相关基因的 455 个变异的单倍型。粉色，黄色和绿色方块分别代表与参考基因组位点相同的纯合单倍型、与参考基因组位点不同的纯合单倍型以及杂合位点。拟态指数 MI 标注在左侧。重力响应相关基因标记在底部，不同颜色表示不同的途径。

LAZY1（LA1）是迄今为止鉴定出的控制植物分蘖的最重要的基因之一。它通过重力响应参与植物生长素的重新分配，其功能在水稻、玉米和拟南芥中均保守。在稗草拟态过程中，LA1（EC_v6.g079654）的基因组区域显著分化（FST = 0.476），其遗传多态性显著降低（ROD = 4.956;图 5a 和补充表 7）。中性测验表明，在非拟态群体中该基因区域具有显著的 Tajima’s D 正值，而在拟态群体中则急剧下降至负值（图 5b 和补充表 1）。较高程度的连锁不平衡也暗示该基因是拟态过程中潜在的选择区域（图 5b）。由两个导致氨基酸变异和八个内含子变异组成的 LA1 的 5 个纯合单倍型（hap01 至 hap05）在拟态和非拟态群体中分布明显不同（图 5c，d 和补充表 9）。hap01 在拟态群体所有单倍型中占 70%，但在非拟态中仅占 15%。相应地，与 hap02（平均分蘖角：51.1°，P <0.0001）和 hap03（平均分蘖角：61.5°，P <0.0001）相比，具有单倍型 hap01 稗草材料的分蘖角（平均角：39.8°）要小得多（图 5d）。以两个导致氨基酸变异的位点来看，拟态群体中 76%的材料具有 CT 单倍型，而非拟态群体中 57%的稗草材料具有 GC 单倍型（补充图 13）。此外，在苗期，拟态稗草（JS26）LA1 的基因表达水平远高于非拟态稗草（JX58；图 5e）。综上，这些结果表明在稗草拟态过程中，LA1 受到了人为驱动的选择，产生分化。实际上在水稻（Oryza sativa ssp.japonica）驯化过程中，LA1 也受到了选择，具有显著的 Tajima’s D 负值、遗传多态性显著降低（Tajima’s $D_{cultivar}$=−1.871； ROD=6.333；补充图 14a）。单倍型分析显示，水稻 LA1 中的几个位点在野生稻和栽培稻之间有高度分化，并且野生水稻中的稀有位点在栽培稻上被固定（补充图 14b）。这表明，LA1 基因在稗草拟态进化过程中和水稻驯化过程中经历了平行选择。



Position	−3,314	−2,964	−2,898	−1,812	−1,339	−1,049	−850	−827	−728	−165	Accession number	Tiller angle (°)	Percentage in MIC	Percentage in NMC
Ref	C	A	T	T	G	G	C	C	T	C				
Alt	G	G	C	C	GA	T	G	G	C	CAAGT				
hap01	C	A	T	T	GA	G	C	C	T	CAAGT	126	39.8	70	15
hap02	G	G	T	C	A	T	G	G	C	A	49	51.1	4	19
hap03	G	G	C	C	A	T	G	G	C	A	46	61.5	2	26
hap04	C	A	T	C	GA	G	C	C	C	CAAGT	4	58.1	0	2
hap05	G	G	C	C	GA	T	G	G	C	A	2	56.3	0	1

图 5 LA1 基因在稗草拟态进化过程中受到正向选择作用。a，FST（上图）和 ROD（下图）分布情况，一些已知的植物株型相关基因用箭头标记， LA1 的基因组区域（scaffold445）用灰色阴影突出显示。b，Tajima'D 分布（上）和连锁不平衡热图。c，LA1 的基因结构、变异和单倍型。上图：稗草 LA1 的基因结构。下图：LA1 单倍型。d，LA1 在 328 份稗草材料中的单倍型分析。两个非同义的 SNP 用红色突出显示。e，在拟态稗草（JS26）和非拟态稗草（JX58）中 LA1 的基因表达情况。 *** P ＜0.001。

【讨论】

在该项研究中，我们提供了瓦维诺夫拟态中人类选择的基因组学证据（也即提供了瓦维诺夫拟态的基因组学证据）。我们发现，拟态和非拟态稗草基因组的局部区域，存在着显著的基因组分化和正向选择信号。研究结果扩展了我们关于拟态的认知，也提升了我们对稻田杂草起源和进化的理解。

我们的分析显示：拟态稗草的一些基因组区域具有选择信号，其遗传多态性显著降

低、具有 Tajima's D 显著负值。我们发现 87 个与植物株型相关的同源基因在拟态群体中受到选择，在稗草的拟态进化和水稻驯化过程中观察到平行选择现象，这些结果似乎表明稗草拟态化的遗传机制与水稻驯化过程中的遗传机制部分类似。我们的结果支持一开始的假设，即瓦维诺夫拟态是由人类的无意识选择导致的，尽管这种选择与农作物中的人工选择不同，后者是一种有意识、定向的选择。

拟态分子机制解析得较为清楚的通常涉及的是简单的表型特征，例如颜色（蝴蝶翅膀），而瓦维诺夫的拟态特征复杂。瓦维诺夫拟态可能是由多个基因控制的，分子机制更为复杂。我们在该项研究中就发现了数百个与稗草拟态可能相关的基因组区域以及数十个具有选择信号的高可信拟态特征相关基因。

在本研究中，我们估算出稗草开始出现拟态的时间大约是 1000 年前。根据中国的历史记载，在宋代，经济中心由黄河流域转变为长江流域，人口迅速增长，并且稻米取代了小麦成为主粮。正是由于该时期人口的快速增长而需要更多的稻米，更强的人类选择压导致了拟态稗草的形成。我们的系统发生树和 PCA 结果支持拟态稗草来源于非拟态类型，并且长江流域的拟态稗草具有单一起源（图 2 和补充图 6）。因为如果该地区稗草拟态发生了多次起源，那么在系统发生树上应该能够观察到多个类群，每个类群均包含非拟态稗草及其演化的拟态稗草，而不是现在进化树上观察到的两个独立的分枝。至于在进化树上拟态和非拟态组中具有异常表型的少量材料，一个合理的解释是基因渗入。但是，我们不能排除，在更多的区域，稗草有多次拟态起源的可能性。需要更多其他地区的材料来解决此问题。混合组中包括了许多处于拟态和非拟态表型之间的材料。两组（拟态和非拟态）之间的杂交可能会导致这种现象。另外，这些材料可能是处于拟态进化过程中的中间阶段（半拟态），或是模仿了尚未改良为理想（紧凑型）株型的较早期的水稻地方品种。总之该项工作为进一步揭示瓦维诺夫拟态这一进化过程提供了重要基础。

【研究方法】

【植物材料采样和基因组重测序】

从长江流域不同地点的稻田和及邻近稻田周边区域的地区收集了 328 份稗草样品（补充表 1），每份样品的种子都发芽并种植在中国富阳中国水稻研究所的稻田中，每份材料种植 8 株。在 2017 年，我们开始记录它们的表型，包括分蘖角大小，茎节形状（直或弯曲），茎基部的颜色（绿色或红色/紫色）和叶型（紧凑或披散，补充图 1）。每份稗草样品的分蘖角通过三株个体的平均值来衡量。在 2018 年再次记录表型，结果与 2017

年保持一致。所有表型的记录时间是将三叶期幼苗移入稻田后 3 周的分蘖期。用 MI 值来表示拟态水平，四个特征值的总和≥8 或≤2 分别表示典型的拟态和非拟态类型：分蘖角<30°，30–50°和>50°分别被赋值为 5、3 和 0；直茎节和膝状茎节分别赋值为 3 和 0；茎基部为绿色和茎基部为红色/紫色分别赋值为 1 和 0；叶片紧凑和叶片披散分别赋值为 1 和 0。我们还从拟态组（CQ1，JS26，HB17 和 SC22）和非拟态组（HB10，ZJ102，JX67 和 JX58）中随机选择了四份样品，在两个光周期（14h/10h 和 16h/8h）下种植于温室中用来观察表型。温室中表现出的表型与在稻田中观察到的一致。使用常规方案从绿叶中提取 DNA。利用 Illumina HiSeq 4000 生成了总计超过 7.7 Tb 的配对末端序列数据，每个样品的平均覆盖深度约为 15X。图 1b 中有关世界大米产量的信息是从粮食及农业组织（http://www.fao.org/faostat/）下载的，并用 R 语言重绘图表(https://madlogos.github.io/recharts/)进行说明。

【识别变异】

首先使用 NGS QC 工具包 v2.3.3 整理原始的末端配对读数，其标准是 Phred 质量得分大于 20，且与给定序列的匹配度大于 70%。使用 BOWTIE2 v2.2.1（默认设置）将每份稗草的整理后末端配对读数映射到 *E.crus-galli* 参考基因组（STB08）上。一个综合性计算流程，主要使用 SAMtools v0.1.19 和 GATK v2.3 来检测全基因组变异（SNPs 和插入/缺失）。为了满足识别变异的标准，使用自定义脚本根据以下参数过滤了 SNP 和插入/缺失：QUAL <30，DP <5，QD <2，MQ <20，FS> 60，HaplotypeScore> 13，ReadPosRankSum <-8，MAF> 0.01，整合率> 0.8。这些变异由 SnpEff v3.6 注释，并由自定义脚本进行汇总。

【系统发生分析】

系统发生分析共使用了 219 万个全基因组 SNPs，其中 MAF 的过滤标准大于 0.05，完整性比率大于 0.9，杂合位点比率小于 0.1。使用 FastTree v2.1 构建了所有 328 个样本的极大似然系统树，其中 bootstrap 值设为 1000。为了测试拟态群体是否由非拟态群体演变，我们使用四倍体父本 *E.oryzicola*（ZJU2）作为外类群构建系统发生树。我们使用 iTOL（itol.embl.de）来显示和修改已构建的系统发生树，并使用 EIGENSOFT（v6.1.3）的 smartPCA 脚本（默认设置）进行 PCA 分析。

【分化时间估计】

由非拟态类群开始向拟态类群分化的时间使用 SMC ++ v1.13.1 估计，它可以根据全基因组序列数据推断有效的种群规模历史，并且对于在较短的时间跨度内恢复历史非常有效。由于 SMC ++分离模型要求在两个种群分离演化后没有基因流动，因此仅分别选择拟态类群和非拟态类群中具有典型拟态和非拟态类型的样本。该分析排除了可能由基

因流动引起的每个群体中那些具有异常表型的个体。然后分别随机在拟态类群中选择七个世系，在非拟态类群选择六个世系作为显著个体，以改进对有效种群规模的估计和分化时间的推断，这些分别是由 SMC++估计和划分的。PSMC 分析（v0.6.5-r67）通过选择具有较高测序深度（>20X）的个体，也用于拟态类群和非拟态类群的群体历史推断。共有序列由 SAMtools 50 生成。为了提高准确性，将读取深度小于 10 或大于 50 的位点过滤掉。参数-p（指定原子时间间隔）被两次设置为“ 4 + 30×2 + 4 + 6 + 10”和“ 4 + 25×2 + 4 + 6”，其他参数被设置为默认值。假定突变率为 $\mu=6.5\times10^{-9}$ 突变×bp^{-1}（每个碱基对）×$generation^{-1}$（每一代），而且和水稻一样一年种一代，SMC ++和 PSMC 分析均以此调整。

【检测选择信号】

使用 VCFtools v0.1.15 来计算整个基因组的遗传统计数据，其中，全基因组的滑动窗口为 50 kb，π 和 F_{ST} 的步长为 20 kb ，Tajima’s D 的滑动窗口为 20 kb 且基于完整性比率>0.8 的 SNP 数据集。除去少于 10 个 SNP 的窗口。然后使用自定义脚本计算 ROD。选择 5%的经验阈值以找到 ROD 和 F_{ST} 值的异常窗口。Tajima’s D 值小于相应样本规模的 95%置信限被认为是显著的。如果两组之间的窗口的 F_{ST} 值非常低（小于 5%的经验阈值），窗口在拟态类群和非拟态类群中都显示出显著的负的 Tajima’s D 值，则被排除掉。间隔长度小于 50 kb 的异常值窗口被合并后用于检测。在模拟过程中，只有在拟态类群和非拟态类群之间的差异（$P<1\times10^{-20}$，Fisher 精确检验）和潜在影响（SnpEff 注释结果为 MODIFIER, MODERATE 或 HIGH）中具有显著性差异的基因才被视为筛选条件下的候选基因。稗草 *E.crus-galli* 中株型相关基因的同源基因通过 BLASTP V2.6.0 与已知的株型相关基因相联系并进行注释，临界值小于 1×10^{-20}（补充表 6）。

【*LA1* 单倍型分析】

使用 Beagle v5.0（默认设置）对包含 *LA1* 遗传区域的 scaffold445 中的 SNPs 和插入/缺失进行划分。通过 R 语言 PopGenome v2.2.4 计算单基因遗传学数据；通过 Plink v1.9 为窗口大小为 5 Mb 的每对 SNP 计算 r^2 值，并通过 R 语言 LDheatmap v0.99-5 进行了说明。为了分析水稻中 *LA1* 的选择模式，从国家生物技术信息中心下载了 100 份 *Oryza sativa ssp.japonica* 和 43 份 *O.rufipogon* 的全基因组重测序数据（补充表 10）。将整理后的读数定位到水稻参考基因组（MSU6.1）。然后进行 SNP 调用，扫描选择信号和 *LA1* 单倍型分析。

【RNA-seq 分析】

分别选择具有典型拟态和非拟态特征的样品 JS26（拟态类群）和 JX58（非拟态类群）进行转录组测序。从收集的分蘖基中提取 RNA（长度~1cm，采样时间为分蘖期，

即将三叶期幼苗移入稻田后 3 周），并按 Illumina HiSeq 4000 的常规规程进行测序。接头序列，低质量序列和空标签在映射读取到稗草 *E.crus-galli* 的参考基因组之前被移除。通过 TopHat v2.1.1 将读数映射到参考基因组（STB08）。将映射到基因的标签数量标准化为 FPKM 值，并通过 Cufflinks v2.2.1 进行差异表达的基因分析。根据调整后的错误发现率 0.05，鉴定出存在表达差异的基因，这些基因必须展现出最小为 2 的折叠变化。