DOI: 10.13376/j.cbls/2019090

文章编号：1004-0374(2019)07-0739-09

动植物高质量基因组的获取及其主要应用

霍冬赦[1]，王跃斌[2]*，陈庆富[1]*

(1 贵州师范大学荞麦产业技术研究中心，贵阳 550001；2 华中农业大学植物科学技术学院，武汉 430070)

摘　要：自国际千人基因组计划实施以来，伴随着测序技术的发展和成本的下降，几乎所有重要的动植物都拥有了参考基因组以及全基因组重测序数据。针对二代和三代测序技术产生的海量数据，准确和高效的组装是获得高质量基因组的关键。对于重复序列较多、杂合性较高的复杂基因组的组装尤其具有挑战性，基因组从头组装算法不断被更新，联合组装策略正在发挥强大优势。高质量的基因组不仅能提高精细定位效率，还能提高全基因组关联分析的准确性和精度，为动植物复杂性状的遗传机制解析奠定基础。同时，高质量的基因组对于比较基因组以及泛基因组的研究都具有重要的推动作用。

关键词：高质量基因组；组装；变异分析；复杂性状

中图分类号：Q3-3；R3　　**文献标志码**：A

Acquisition of high quality genomes of animals and plants and their main applications

HUO Dong-Ao[1], WANG Yue-Bin[2]*, CHEN Qing-Fu[1]*

(1 Research Center of Buckwheat Industry Technology, Guizhou Normal University, Guiyang 550001, China;
2 College of Plant Science & Technology, Huazhong Agriculture University, Wuhan 430070, China)

Abstract: Since the implementation of the International Thousand Human Genome Project, almost all important plants and animals have attained reference genomes and genome-wide resequencing data with the development of sequencing technology and the decrease in sequencing costs. Accurate and efficient assembly of the vast amounts of data generated by second- and third-generation sequencing technologies is the key to acquire high-quality genomes. The assembly of complex genomes with more repeats and higher heterozygosity rate is particularly challenging. The genome *de novo* assembly algorithm is constantly being updated, and the joint assembly strategy is showing strong advantages. High-quality genomes not only improve the efficiency of fine mapping, but also improve the accuracy and precision of genome-wide association analysis, thus laying the foundation for the genetic dissection of complex traits of plants and animals. Meanwhile, high-quality genomes would promote the study of comparative genomics and pan-genomics.

Key words: high quality genome; assembly; variation analysis; complex trait

高质量的参考基因组在动植物遗传学和基因组学研究中有着极为重要的作用。解析复杂性状的遗传机制，首先需要在全基因组范围挖掘与该性状紧密相关的关键变异，在尚未实现群体基因组组装的物种中，变异的检测都是基于该物种的参考基因组。因此，参考基因组质量的高低决定了变异基因型鉴

收稿日期：2019-01-24；**修回日期**：2019-03-05

基金项目：国家自然科学基金项目 (31471562，31860408，U1812401)；黔科合 LH 字 [2017]7356；国家燕麦荞麦现代农业产业技术体系专项资金 (CARS-07-A-5)；贵州省高层次创新型人才培养对象十百千计划 [2015] 4020；贵州省科技支撑计划 ([2017] 2505, [2018] 2320)

***通信作者**：E-mail: chenqf1966@126.com(陈庆富)；624547075@qq.com(王跃斌)

定的准确性，进而影响对复杂性状遗传机制的解析。高质量参考基因组的获得需要高深度的测序结果以及合适的组装算法。相应地，使用准确率高的二代高通量测序数据和读长超长的三代测序数据进行联合组装，可提升基因组组装的完整性和重复序列组装的准确性。获得高质量参考基因组后，针对种内的其他个体，只需进行低深度测序就能准确鉴定相对于参考基因组的变异类型，为数量遗传学研究提供大量分子标记，并可提高性状变异位点的定位精度，同时提高了鉴定基因组上结构变异的准确性。目前针对二代测序技术的重测序数据检测基因组结构变异的算法不断被开发出来，通过收集测序片段比对回参考基因组上的错配和不恰当比对的信息，就能在一定程度上判断测序目标相对于参考基因组的插入和缺失的结构变异。随着基因组学的发展，获得一个高质量的基因组难度不断降低，直接在多个基因组上进行比对，发现关键变异尤其是集中分布的关键变异的方法开始被更多地使用。在此过程中，基因组学的研究范畴不断延伸。泛基因组成为真正意义上能够代表整个物种遗传物质多样性的“参考基因组”，因此基于群体水平的泛基因组也被越来越多地关注。本文回顾了基因组组装算法的发展，重点讨论了对于重复序列较多、杂合度较高的基因组组装新算法，分享了近 5 年来高质量基因组在动植物遗传机制解析中的成果，以及在比较基因组和泛基因组研究中的应用。

1 基因组的从头组装(*de novo*)算法

随着测序技术的发展，不论从技术还是成本上获得高质量的基因组测序数据都变得具有较高的可行性，对于富集了大量重复序列和高度杂合的玉米 [1] 和小麦 [2] 基因组而言，对基因组从头组装算法的研究早已是新的挑战和研究焦点。

针对最早的 Sanger 测序数据，早期开发的是 OLC (overlap-layout-consensus) 算法，即寻找两条 Sanger 序列之间的重叠区域并进行拼接。随着高通量测序技术的出现，这样的算法则不再适用，原因是二代测序通量更高但片段更短 (通常双端测序片段长度只有 150 bp)，不能记录下所有重叠区域的信息，而且对于复杂基因组而言，寻找短测序片段 (read) 之间的重叠区域也更困难和不可靠 [3]。在这样的背景下，针对二代测序数据进行基因组组装的 DBG graph(de-bruijn-graph) 算法应运而生 [4]。DBG 算法的核心是 k-mer，即将所有的短 reads 打断为更短的长度为 k 的序列，两个相邻的 k-mer 相差一个碱基，将一组相邻 k-mer 的最后一个碱基 (edges) 相连即成为一条组装结果 (图 1)。K-mer 组装解决了基因组测序覆盖深度带来的数据冗余问题，并且从理论上，要获得组装结果只需要记录 k-mer 之间相连的信息 (read path)。DBG 算法相较于 OLC 不仅能更充分地利用高通量二代测序结果，还极大地减少了运算时间和成本。

Soap denovo[5-6] 是利用 DBG graph 算法开发的典型基因组组装软件，其组装过程分为以下四步。(1) 二代测序数据自纠错，相较于三代数据高达 15%~40% 的错误率，二代数据的准确性仍然具有强大优势 [7]。虽然二代数据的错误率只有 1%~2%，但是对于大型基因组组装而言，为了达到更准确的组装效果，正式组装前需要进行自纠错。(2) 选取合适的 k-mer，组装 contig。K-mer 的选取非常重要，k-mer 的值过小，不利于构建更长的 contig 以及利用 reads 本身的长度跨越一些小片段的重复区域，但如果 k-mer 的值过大，则会造成运算时间和消耗内存指数级别的上升。同时，需要指出的是，为了避免回文序列造成的组装错误，一般不选取偶数 k-mer。初步组装完成的 DBG graph 非常粗糙，包含了大量的错误和不确定信息，首先要去除一些连接着两个独立 contig 的过短或者覆盖度过低的中间序列，其次由于基因组本身高度杂合的特性产生的一些相似度很高的中间序列 (bubble)，将由 soap denovo 选择覆盖度更高的一条作为代表序列。(3) 组装 scaffold。scaffold 的实质是一条更长的，连续的 contig，要达到这样的组装级别，只有二代数据

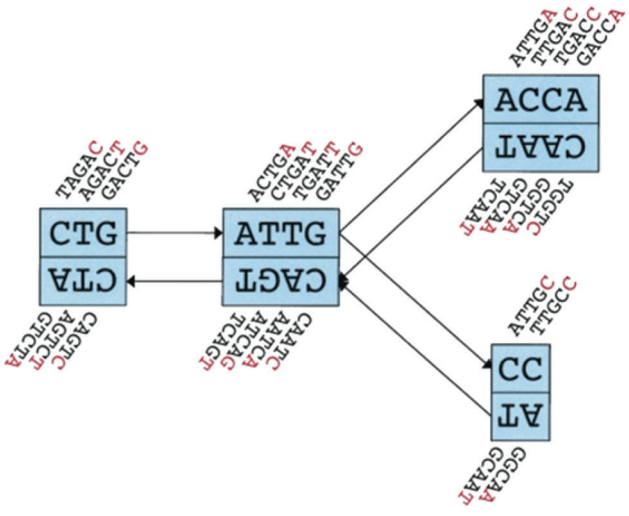


图1 DBG算法原理示意图[4]

是不够的，还需要插入片段大小 (insert size) 更大的 mate pair reads，以提供锚定的信息，将在物理距离上相隔较远的数条 contig 连接在一起，mate pair read 的 insert size 越大，最后能得到的组装效果也越好。(4) 缩小组装漏洞。在 scaffold 组装完全后，程序会收集一端落在 scaffold 内部，另一端没有被锚定的 reads，作为修补组装漏洞的数据。从原则上来讲，这样的操作仍然是在延伸 scaffold 的长度，以期达到更好的组装效果。

由于 Soap denovo 在对亚洲和非洲人基因组组装中的优异表现，这款软件的核心算法 DBG 及其运算流程已经成为了组装大型基因组的代表。除此之外，还有专门针对高杂合基因组组装的软件 Platnus[8]，其算法重点考虑了杂合基因组本身对基因组组装带来的影响，在构建 contig 的过程中不仅保留了所有相似度高的连接两个 contig 的中间序列 (bubble)，还将这些 bubble 重新锚定回 scaffold 上，保留与 bubble 相连的 contig 信息作为杂合区段。

从这些软件的算法和运算流程我们可以看出，完善在复杂基因组中占有相当比例的重复序列的组装几乎是所有算法面临的问题和挑战 [9]。而三代测序技术 (single-molecule, real-time sequencing) 的超长读长，旨在从源头上解决复杂基因组组装面对的各种难题。平均 10 k 以上的读长所带来的显著优势是原始 reads 可以轻易跨越一些中等长度的重复序列区段，甚至不需要组装即可获得完整的基因组序列 (这一点已经在转录组转录本的测序中得到了证明)。也正因为如此，适用于二代测序数据的 DBG 算法不再适合于三代测序数据，因此 OLC 算法也再次回到人们的视野中。早期针对三代组装开发的软件致力于寻找长 reads 之间的重叠区段，从而将两条 reads 相连。但对于三代测序超长读长，并且本身携带了较高测序错误的 reads 而言，OLC 算法的使用面临两个挑战：一是用于组装的 reads 自纠错，即使是像 HGAP 这样成功的三代组装软件也不能回避 reads 纠错所带来的运算时间和内存消耗问题 [10]；二是由于寻找重叠区段必须进行多重比对，三代测序的超长读长带来的运算负担相较于一代数据而言早已呈指数级别的增长。为解决上述问题，出现了 DBG 和 OLC 算法联合组装的策略 [11]。同时对复杂基因组进行二代和三代测序，不仅能在 OLC 正式组装前，利用二代数据对三代数据进行高效率的校正，还能在由计算三代 reads 最佳重叠区域而得的组装骨架基础上，联合二代数据进行 scaffold 的延伸和补洞。

联合组装的策略由于兼具了二代数据的准确性和三代数据的读长优势，已经被广泛运用到各类复杂的基因组，尤其是高度重复序列的基因组组装案例中 [1,12]。即使各类算法和软件开发飞速发展，仍然没有一个大型基因组是完全没有瑕疵的。虽然现有的高通量测序技术极大地推动了人类对基因组的研究与理解，但对基因组复杂区域的组装与研究仍困难重重。

2　高质量参考基因组在寻找功能基因和重要变异位点中的重要作用

由于测序技术和组装算法的改进，出现了越来越多高质量的参考基因组。一些长久以来因为基因区段或者序列的复杂性而没有办法解释的现象逐步得到了理解，如抗病基因簇 [13-14] 或者某些具有功能的转座子 [15-16]。一个高质量的参考基因组不仅是了解自然群体变异形式的开端，更是解析功能基因和重要变异位点的前提。

2.1　高质量参考基因组提高传统基因克隆手段——精细定位的效率

精细定位是克隆基因的传统遗传学手段之一，通过设计定位标记，筛选重组，将功能基因锁定在基因组的某一个区段内。在这个过程中，如果能同时得到定位群体两个亲本的高质量基因组，则能在很大程度上缩短定位的年限并加深对功能基因变异形式的理解。例如在对玉米单向杂交不亲和基因的精细定位中 [17]，借助参考基因组和组装另一亲本相应区段的 BAC 序列，确定了功能基因在其中一个亲本中发生了提前中止；而此前由于定位区段在两个亲本基因组中发生了重大变异，其中一个基因组在该区段完全未知而导致无法进一步缩小区段 [18-19]。

2.2　高质量参考基因组对GWAS(genome-wide association study)结果的影响

近十年来，得益于 GWAS 方法的迅猛发展，传统的寻找功能基因或者功能变异的遗传学方法中所体现出的缺点，如耗时长、工作量大及无法充分挖掘自然群体中的等位变异等都得到了明显的改善，但很少有人强调高质量基因组在获得准确的 GWAS 结果中所起到的作用。一方面，早期的 GWAS 多产生于芯片测序的结果；另一方面，人们对稀有变异的理解也没有今时今日那么深刻 [20]。

GWAS 虽然成功地在人类 [21]、玉米 [22-25]、棉花 [26-27]、大豆 [28-29]、水稻 [30-31] 等具有复杂基因组的

物种中解析了多种多样的复杂性状，但其局限性也很明显，即受限于群体大小和芯片测序质量，很多关键性的变异位点不一定能被检测到，这也是为什么当使用一般芯片的数据时，GWAS 的效果往往不够好 [20]。由于不能够精确和全面地覆盖群体变异，GWAS 的优势逐渐丧失。

解决这一问题的有效方法是提高基因组质量。基于一个高质量的参考基因组，只需要对群体进行中等甚至低覆盖率的重测序 (1~30×) 就可以实现在群体水平上对全基因组的 SNP 进行分析和鉴定。

二代测序技术的高通量优势是实现大批量样本全基因组测序的基础。在获得低倍覆盖短 reads 的基础上，将 reads 回帖到序列准确和注释完整的参考基因组上，不仅可以挖掘更多稀有变异，更能宏观且准确地了解这些稀有变异在全基因组上所处的物理位置及其附近的基因信息。

2018 年，Du 等 [32] 基于一个 *de novo* 组装的高质量棉花 A 基因组和 243 份平均测序深度为 6× 的棉花材料，共获得了 17 883 108 个 SNP 和 2 470 515 个 InDel，成功构建了棉花高密度遗传图谱，基于该图谱鉴定了 98 个与 11 个农艺性状紧密关联的位点。Huang 等 [33] 基于 517 份水稻 landrace 种质资源，在只进行了一倍测序的前提下，与高质量参考基因组比对后鉴定到了 3 625 200 个非同义突变 SNP，且在已被报道证实的和农艺性状相关基因附近至少发现了 6 个显著 SNP。综上，我们不难得出高密度遗传图谱的构建是 GWAS 成功的关键因素，而高质量的参考基因组则是在群体水平上鉴定准确基因型的首要条件。

2.3 基于高参考质量基因组鉴定影响复杂性状的基因组结构变异

随着国际千人基因组计划的实现，人们对基因组变异的了解也越来越深入，一些从前未能被发现和认识到的大型变异开始逐渐被人们研究 [34]，基因组结构变异 (structure variation) 一般是指大小超过 1 kb 的插入、缺失或者倒位。鉴定结构变异不同于鉴定一般的 SNP，主要原因是由于在二代测序技术被用于群体水平高通量测序时，其 read 读长过短，不足以跨越或者组装出基因组上一些较大的结构变异。因此，越来越多针对检测大型结构变异，包括拷贝数变异 (copy number variation) 的算法和实验平台被开发出来，通过将实验对象的测序 reads 直接比对回参考基因组上，保留下比对结果中错配与不恰当比对的结果，鉴定不同个体中不同于参考基因组的结构变异。基于这样的原理，一个高质量的参考基因组几乎成为了所有检测结构变异算法的基础，参考基因组只有在保证组装正确的前提下才能正确鉴定出其他材料或者样本的结构变异。在此，我们讨论几种常见的鉴定结构变异 (structure variation) 的方法 [35]，虽然这些算法基于不同的原理发展而来，但都必须依赖于一个高质量的参考基因组。

2.3.1 利用二代双端reads鉴定基因组结构变异

将全基因组打断为固定长度的短片段进行建库是二代正式建库测序的步骤，因此一对双端 reads 之间的距离往往是固定且可控的。如果在双端 reads 比对回参考基因组时相隔距离过大，则可能是由于研究对象相对于参考基因组插入了一段序列；如果相隔距离过小，则可能是由于研究对象相对于参考基因组丢失了一段序列。依据这样的算法原理，理论上可以检测所有的插入和缺失变异，且不受变异大小的影响 [36-37]。

2.3.2 利用测序reads比对参考基因组鉴定结构变异断点

对于一些需要明确知道断点的变异而言，如果测序覆盖度够高，总能够找到位于变异附近，且一部分能够比对到参考基因组上另一部分为未知序列的 reads。一些较小的插入和缺失甚至能直接被长 reads 跨越。这样在比对中产生的不能完整比对到参考基因组上的 reads 被称为 split reads，split reads 在检测基因组结构变异中最明显的优势就是能够提供明确的断点信息 [38-39]。

2.3.3 利用测序reads覆盖基因组深度的差异鉴定结构变异

为避免扩增不均匀，PCR-free 方法被开发出来，全基因组的每一个区段都能保证被均匀覆盖，得到测序深度基本一致的结果。因此，大多数测序深度陡增或者陡降的基因组区域都有结构变异的可能性，尤其是基因拷贝数变异 (copy number variation)。人们最早认识到拷贝数变异的重要性是源于一些与人类疾病显著相关的区段只包含很少的 SNP。研究表明，每个人都携带着 8% 左右的大小超过 500 kb 的拷贝数变异 [35]，拷贝数变异通过剂量效应直接造成基因表达量的变化，一些严重的疾病，如艾滋病、肾炎等都和基因拷贝数变异紧密相关 [40]。但要检测这样的变异则不容易，因为拷贝数变异从本质上来说并不是基因组序列的变化，不存在测序 reads 错配或者不恰当匹配的问题。可行的方法是通过将测序 reads 比对到参考基因组上发现覆盖深度变异的

部分，判断拷贝数变异的发生。如果比对深度相对于参考基因组的其他区段有明显上升，则可能是拷贝数增加造成的；如果比对深度相对于参考基因组的其他区段有显著下降，则可能是基因拷贝数减少造成的。

综合以上所有在群体水平上鉴定基因型和结构变异的方法，不难看出参考基因组在其中所起的作用几乎是决定性的。在基因组尚未实现的物种中，其变异都是相对于参考基因组而言的，一个高质量的参考基因组是建立群体遗传学的基础，更是挖掘功能基因变异形式，寻找功能位点的前提。

3　基因组组装质量在比较基因组学研究中所起的作用

基因组学发展至今，人们对变异的挖掘和理解已经不再局限于一个基因或一种性状。从全基因组的角度出发，探究某一个物种在整个进化历史上所处的位置或某一类影响重要性状的基因在多个物种中的作用，更有利于我们理解每一个生命个体的由来和进化。基于这样的理念，比较基因组学应运而生。

比较基因组学是基因组学发展到一定程度的产物，在比较基因组学的应用中，并不存在参考基因组概念。所进行比较的物种都有自身完整的基因组序列，通过最直接的序列比对，理论上可以检测到存在于多个基因组上所有的变异信息。因此，保证多个基因组组装的正确性往往决定了比较结果的可靠性。坚持使用同一套组装标准，让所有基因组重测序数据基于同样的参数进行组装则是组装质量保持一致的前提。

比较基因组学基于多个基因组之间的相互比较，最后讨论的问题往往离不开基因组之间一致且保守的区域以及各自特有的部分。利用这样的结论不仅可以构建更清晰的物种进化树，还可以发现那些在进化中至关重要的基因。

Zhang 等 [41] 收集了 48 套已公布的鸟类基因组重测序数据，并且进行了统一的基因组组装与注释，随后在它们之间进行了相互比较，构建了一个清晰完整的鸟类进化树。同时，他们还发现鸟类虽然是最古老的哺乳动物之一，但相较于其他哺乳动物，鸟类的基因组大小却在相当程度上缩减了。通过进一步研究比较结果之后，作者发现鸟类，尤其是现代鸟类，与其他哺乳动物 (海龟、鳄鱼) 相比，经历了更多的染色体小片段丢失事件，但这样的丢失却没有对鸟类的生存造成重大的影响，其中关键的原因是由于这些经历了丢失的片段内所包含的基因大部分在基因组上都有同源基因，可以在一定程度上对丢失的基因进行功能互补。Stein 等 [14] 通过选取 13 个具有代表性的水稻品种，同时组装全基因组序列，并且在比较后发现，虽然水稻各品种间的差异已经很小，但是仍然有一些染色体重排事件只存在于某些品种内，这导致了转座子和一些新的非编码区序列的诞生。同时，作者还比较了 13 个基因组之间抗病基因家族的一致性，发现虽然抗病基因由于偏向于形成基因簇而很难被研究清楚，但是基因的排布却有一定的规律可循，两个相邻的抗病基因更倾向于首对首 (head-to-head) 地分布，这可能是为了更好地形成抗病复合体。

比较基因组学着眼于全基因组，在一定程度上为人们解释更宏观的科学问题提供了方法。不同基因组之间的比较往往适用于解释不同的问题。自 2017 年 PGA 会议发展“重测序项目”以来，比较基因组学由于信息来源的广泛性和几乎覆盖所有重要动植物基因组的众多重测序项目而进入了一个黄金时代，但基因组和基因组之间参差不齐的质量仍然值得关注和改善。在比较基因组学探究具体的生物学话题和意义之前，保证基因组的质量和正确性，甚至保证由组装误差造成的错误都尽可能地一致，将所有基因组放在同样的水平上进行比较是比较的前提。

4　基于高质量参考基因组获得涵盖物种内更多变异信息的泛基因组

随着基因组学的发展，更多的参考基因组被组装出来。在芯片测序和短序列比对的过程中发现了大量变异。人们开始思考，一个参考基因组是否真的能代表整个物种？一个基因组上的一种变异是否足够解释物种内所有表型变异？虽然有很多方法可以挖掘物种间变异，例如前文介绍的鉴定结构变异 (structure variation) 的算法等，但是对于较为复杂的基因组而言，从头组装一个高质量基因组是最直接和简便的研究全新变异的办法。

在微生物中首次提出泛基因组的概念。随着二代测序的普及，已经可以针对某个物种分别组装单个基因组，即使是在具有复杂基因组的真核生物中也能实现 [42-43]。泛基因组包含了两个部分，一个是必需基因组，即指在该物种的所有个体中都存在的基因组片段；另一个是非必需基因组，指仅在个别

个体基因组上出现的基因组片段[44]。鉴定非必需基因组是所有泛基因组研究的初衷，因为这类片段代表了物种内基因资源的多样性。非必需基因组是基因组结构变异的一部分，但又不同于一般的结构变异，因为它必需在群体水平上展示极端的有和无的分布差异。

虽然泛基因组最终是以一个基因组的形式出现，但这其实是一个群体的概念。首先，一段非必须基因组片段的鉴定需要以物种内多个个体基因组组装为前提条件；其次，每一个个体的组装质量都有可能成为鉴定非必需基因组的变量，群体水平上单个基因组的组装不一定要上升到染色体级别，但覆盖度和尽量高的完整性是非常关键的因素；之后，单个基因组分别与参考基因组比对，鉴定出不属于参考基因组的片段作为候选的新基因组片段，此时的新基因组片段不一定是非必需基因组，有可能是因为污染而错误引入的其他物种基因组片段。因此，接下来最重要的部分是比对回 NCBI 的非冗余数据库中，排除所有其他微生物、动植物等基因组污染的影响。早期的人类泛基因组研究中，对非必需基因组的要求是只需要在一个个体中出现即可，后来更正为两个或两个以上。之后还需要再经历测序深度、局部组装以及侧翼序列比对验证的检验，最终才能确定一段没有出现在参考基因组中的新基因组片段能否成为代表该物种非必需基因组的片段[44-45]。在这些鉴定过程中，泛基因组借鉴了比较基因组的方法与思路，被比对的参考基因组必须保证相当的质量和尽量少的组装 gap，才能准确鉴定出真实存在于物种间某些个体上的新基因组片段。事实上，在一些非必需基因组鉴定的例子中，确实是因为个体基因组组装片段比对到了参考基因组的 gap 区而“不得已”归类为了非必需基因组[46]。

泛基因组的发展让人们得以在群体水平上认识基因组真实的变异和包含在这些变异中的基因。Maretty 等[47] 通过组装 150 个丹麦人的高质量基因组序列，成功鉴定了 7 个与组织相容性相关的单倍型，发现了两个此前从未被发现的 kb 级别的变异位点。同时，联合千人计划基因组与丹麦人泛基因组重新矫正了一组用于鉴定青少年肥胖的基因型数据，发现了 5 个与该性状紧密连锁的新结构变异。

泛基因组来源于对同一物种内不同个体基因组的组装与比较。由于人们对复杂基因组的认知有限，还没有任何一个大型物种基因组能够做到完全没有组装漏洞，也没有任何一个物种泛基因组的组装能涵盖所有的变异类型。个体基因组的组装为了最大程度地保留变异信息，往往采取从头组装的方式。参考基因组是泛基因组提供非必需基因组的“底线”，只有参考基因组的组装漏洞较少才能准确鉴定不存在于参考基因组上的新序列；新组装个体基因组则是在此基础上建成的“大厦”，由一系列全新变异的“砖石”堆叠而成。在检测新序列的算法中，侧翼序列比对参考基因组的准确和质量也将成为新序列存在的重要证据之一，所以要尽可能地保证其连续与完整。

5 基于高质量基因组解析生物复杂性状的遗传机制(case study)

高质量的基因组究竟能对研究结果产生多大的影响？野生的葫芦科植物具有强烈的苦味 (葫芦素)，在自然界中可以保护植物。虽然葫芦素可以提高人体免疫力以及抑制癌细胞生长，但就适口度而言却是一种不利性状[48]。黄瓜作为一种被人类驯化的葫芦科植物，在基因组上仍存在着两个控制葫芦素合成的位点，其中 Bi 导致整个植株都带有苦味[49]，而 Bt 只让植株果实带有苦味[50]。Shang 等[51] 通过对 155 个黄瓜重测序构建了一个高密度的黄瓜遗传图谱，检测到一个位于 6 号染色体上与苦味显著相关的位点，并且在该位点附近发现由于携带了一个非同义突变而导致葫芦素不能合成的基因，即为 Bi 基因。同时，研究人员对一个带有苦味的黄瓜品种 (XY-2) 和一个不带苦味的黄瓜 (XY-3) 进行全基因组重测序后，比较了两个基因组上携带的变异位点，发现一个位于 loop-helix 转录因子上的变异可以显著影响 Bi 的表达量；并在蛋白质互作实验中验证了 Bi 与该转录因子结合的真实性，并推断这个在黄瓜叶片中特异表达的转录因子通过与 Bi 基因的结合间接地影响葫芦素的含量。进一步分析 GWAS 显著位点附近的基因在两个基因组上的差异，研究人员惊奇地发现 Bi 附近、1 号以及 3 号染色体上分别存在 4 个、1 个和 3 个被注释为酰基转移酶的基因与它享有相同的表达谱，且在两个重测序的黄瓜基因组上也呈现一致的表达趋势。而酰基转移酶是在葫芦素合成途径下游中起到关键氧化和乙酰化作用的酶，进一步的 RNAi 实验显示在降低了任何一个酰基转移酶表达量之后，葫芦素的含量都有明显下降。因此，研究人员大胆推测，8 个酰基转移酶与 Bi 基因共同作用，调控黄瓜中葫芦素的合成 (图 2)。

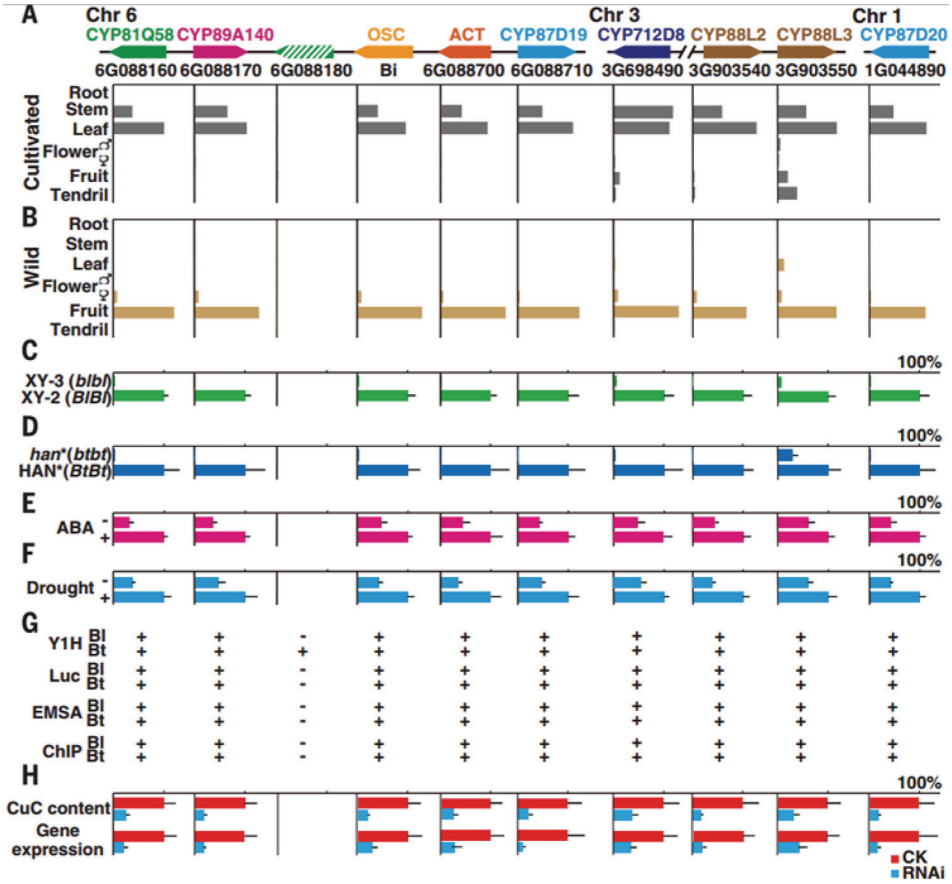


图2　黄瓜基因组上9个表达模式一致且共同负责调控葫芦素合成的酰基转移酶基因[51]

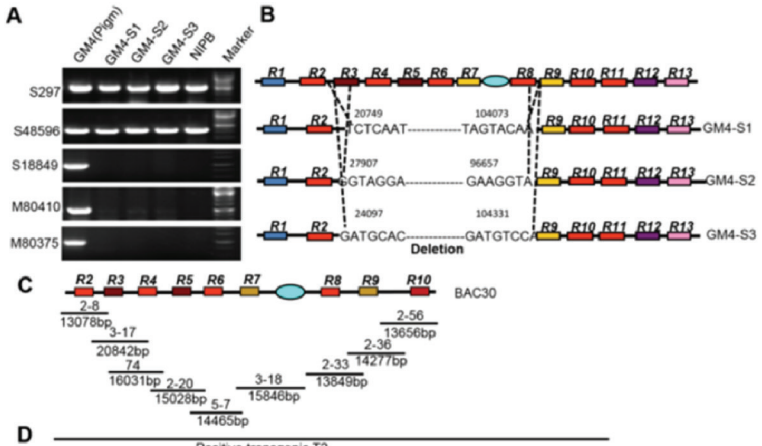
在对控制黄瓜苦味的葫芦素基因的研究中，两个高质量的黄瓜基因组无疑起到了关键性的作用。基因簇的形成往往伴随着重复片段的富集，很容易造成组装的漏洞和组装错误，如果缺少真实和正确的组装结果就可能造成对基因结构变异 (例如 copy number variation) 的误判。正是由于基因组学的发展，使得在该研究中，不仅可以完整地收集所有在基因组上被注释为酰基转移酶的基因，还可以比较出它们在不同品种间的基因结构变异与表达变异，从而完善研究结论。

基因簇的组装因为基因位置的集中和基因序列的高相似性不仅考验基因组组装流程，更考验注释流程。这也是抗病基因解析较为复杂的主要原因。抗病基因相互之间的高相似性使得通过重测序某一区段分离单个抗病基因变得几乎不可能，因此得到一段覆盖抗病基因簇的基因组序列是最可行和最有效的克隆抗病基因的方法。

Deng 等 [13] 正是通过对不同遗传背景下、性状分离的水稻品种进行 BAC 序列的筛选和测序，得到了一系列不同的控制稻瘟病基因组片段，随后通过两两比对，最终在分布着众多抗病基因的复杂区域内确定了唯一一个提供抗病效应的 *Pigm* 基因 (图 3)。

6　前景展望

基因组学的发展不仅依赖于生物信息学的发展，更得益于测序技术的发展。自二代高通量测序平台投入使用以来，几乎所有重要的动植物基因组都有了重测序数据。本文虽然只着重于回顾基因组组装的方法及组装质量对解析动植物复杂性状所起到的作用，但高质量基因组的内涵并不仅止于此，还包括了基因结构与功能注释、转座子注释，甚至于染色体交互信息注释。但这一切都需要基于基因组序列的正确性。迄今为止，还没有一个大型基因组可以完全做到没有组装漏洞。而在这其中复杂区域、转座子和重复序列的作用与影响也许远超过我们的想象，很多尚未得到解释的生物学现象也许就隐藏在我们无法获得的基因组漏洞背后，要获得这部分序列的正确组装结果仍然任重道远。



D

Clone	*NLR* gene	Positive transgenic T2 lines	Resistant	Susceptible	Reaction
2-8	*R2*	56	0	56	S
3-17	*R3*	48	0	48	S
74	*R4*	46	0	46	S
2-20	*R5*	53	0	53	S
5-7	*R6/PigmR*	62	62	0	R
3-18	*R7*	44	0	44	S
2-33	*R8/PigmS*	34	0	34	S
2-36	*R9*	38	0	38	S
2-56	*R10*	46	0	46	S

图3　鉴定在水稻中具有抗稻瘟病功能的*Pigm*基因[13]

致谢：感谢北京市农林科学院王夏青博士在本文修改过程中提供的帮助。

[参　考　文　献]

[1] Jiao Y, Peluso P, Shi J, et al. Improved maize reference genome with single-molecule technologies. Nature, 2017, 546: 524-7

[2] The International Wheat Genome Sequencing Consortium (IWGSC), Appels R, Eversole K, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science, 2018, 361: eaar7191

[3] Chu TC, Lu CH, Liu T, et al. Assembler for *de novo* assembly of large genomes. Proc Natl Acad Sci USA, 2013, 110: E3417-24

[4] Zerbino D, Birney E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res, 2008, 18: 821-9

[5] Li R, Zhu H, Ruan J, et al. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res, 2010, 20: 265-72

[6] Xie Y, Wu G, Tang J, et al. SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. Bioinformatics, 2014, 30: 1660-6

[7] Ye C, Ma ZS. Sparc: a sparsity-based consensus algorithm for long erroneous sequencing reads. Peer J, 2016, 4: e2016

[8] Kajitani R, Toshimoto K, Noguchi H, et al. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res, 2014, 24: 1384-95

[9] Gnerre S, Maccallum I, Przybylski D, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA, 2011, 108: 1513-8

[10] Chin CS, Alexander DH, Marks P, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods, 2013, 10: 563-9

[11] Ye C, Hill CM, Wu S, et al. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci Rep, 2016, 6: 31900

[12] Wang M, Tu L, Yuan D, et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. Nat Genet, 2019, 51: 224-9

[13] Deng Y, Zhai K, Xie Z, et al. Epigenetic regulation of antagonistic receptors confers rice blast resistance with yield balance. Science, 2017, 355: 962-5

[14] Stein JC, Yu Y, Copetti D, et al. Publisher correction: genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. Nat Genet, 2018, 50: 285-96

[15] Yang Q, Li Z, Li W, et al. CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. Proc Natl Acad Sci USA, 2013, 110: 16969-74

[16] Huang C, Sun H, Xu D, et al. ZmCCT9 enhances maize adaptation to higher latitudes. Proc Natl Acad Sci USA, 2018, 115: E334-41

[17] Zhang Z, Zhang B, Chen Z, et al. A *PECTIN METHYLESTERASE* gene at the maize *Ga1* locus confers male function in unilateral cross-incompatibility. Nat Commun, 2018, 9: 3678

[18] Zhang H, Liu X, Zhang Y, et al. Genetic analysis and fine mapping of the *Ga1-S* gene region conferring cross-incompatibility in maize. Theor Appl Genet, 2012, 124: 459-65

[19] Lauter ANM, Muszynski MG, Huffman RD, et al. A pectin methylesterase *ZmPme3* is expressed in *gametophyte*

factor1-s (*Ga1-s*) silks and maps to that locus in maize (*Zea mays* L.). Front Plant Sci, 2017, 8: 1926
[20] Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet, 2017, 101: 5-22
[21] Huyghe JR, Bien SA, Harrison TA, et al. Discovery of common and rare genetic risk variants for colorectal cancer. Nat Genet, 2019, 51: 76-87
[22] Li H, Peng Z, Yang X, et al. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet, 2013, 45: 43-50
[23] Xue Y, Warburton ML, Sawkins M, et al. Genome-wide association analysis for nine agronomic traits in maize under well-watered and water-stressed conditions. Theor Appl Genet, 2013, 126: 2587-96
[24] Yang N, Lu Y, Yang X, et al. Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. PLoS Genet, 2014, 10: e1004573
[25] Chen Q, Han Y, Liu H, et al. Genome-wide association analyses reveal the importance of alternative splicing in diversifying gene function and regulating phenotypic variation in maize. Plant Cell, 2018, 30: 1404-23
[26] Fang L, Wang Q, Hu Y, et al. Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. Nat Genet, 2017, 49: 1089-98
[27] Liu R, Gong J, Xiao X, et al. GWAS analysis and QTL identification of fiber quality traits and yield components in upland cotton using enriched high-density SNP markers. Front Plant Sci, 2018, 9: 1067
[28] Fang C, Ma Y, Wu S, et al. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol, 2017, 18: 161
[29] Wang J, Zhao X, Wang W, et al. Genome-wide association study of inflorescence length of cultivated soybean based on the high-throughout single-nucleotide markers. Mol Genet Genomics, 2019, 294: 607-20
[30] Chen W, Gao Y, Xie W, et al. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat Genet, 2014, 46: 714-21
[31] Chen W, Wang W, Peng M, et al. Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. Nat Commun, 2016, 7: 12767
[32] Du X, Huang G, He S, et al. Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. Nat Genet, 2018, 50: 796-802
[33] Huang X, Sang T, Zhao Q, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet, 2010, 42: 961-7
[34] Moraes F, Góes A. A decade of human genome project conclusion: scientific diffusion about our genome knowledge. Biochem Mol Biol Educ, 2016, 44: 215-23
[35] Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet, 2011, 12: 363-76
[36] Tuzun E, Sharp AJ, Bailey JA, et al. Fine-scale structural variation of the human genome. Nat Genet, 2005, 37: 727-32
[37] Kidd JM, Cooper GM, Donahue WF, et al. Mapping and sequencing of structural variation from eight human genomes. Nature, 2008, 453: 56-64
[38] Mills RE, Luttig CT, Larkins CE, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res, 2006, 16: 1182-90
[39] Ye K, Schulz MH, Long Q, et al. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics, 2009, 25: 2865-71
[40] McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. Nat Genet, 2007, 39: S37-S42
[41] Zhang G, Li C, Li Q, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science, 2014, 346: 1311-20
[42] Sun C, Hu Z, Zheng T, et al. RPAN: rice pan-genome browser for ~3000 rice genomes. Nucleic Acids Res, 2016, 45: 597-605
[43] Sherman RM, Forman J, Antonescu V, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet, 2018, 51: 30-5
[44] Yao W, Li G, Zhao H, et al. Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol, 2015, 16: 187
[45] Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One, 2011, 6: e17288
[46] Li R, Li Y, Zheng H, et al. Building the sequence map of the human pan-genome. Nat Biotechnol, 2010, 28: 57-63
[47] Maretty L, Jensen JM, Petersen B, et al. Sequencing and *de novo* assembly of 150 genomes from Denmark as a population reference. Nature, 2017, 548: 87-91
[48] Ford R, Schwartz L, Dancey J, et al. Lessons learned from independent central review. Eur J Cancer, 2009, 45: 268-74
[49] Da Costa CP, Jones CM. Cucumber beetle resistance and mite susceptibility controlled by the bitter gene in *Cucumis sativus* L. Science, 1971, 172: 1145-6
[50] Qi J, Liu X, Shen D, et al. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat Genet, 2013, 45: 1510-5
[51] Shang Y, Ma Y, Zhou Y, et al. Plant science. Biosynthesis, regulation, and domestication of bitterness in cucumber. Science, 2014, 346: 1084-8