遗传 Hereditas (Beijing) 2021 年 11 月, 43(11): 1023－1037
www.chinagene.cn

综 述

泛基因组：高质量参考基因组的新标准

边培培，张禹，姜雨

西北农林科技大学动物科技学院，杨凌 712100

摘要： 随着三代测序组装的高质量参考基因组的陆续发布，以及大规模重测序和群体遗传学分析的广泛进行，研究人员发现来自单一个体的参考基因组远不能涵盖整个物种的所有遗传序列，大量缺失序列导致群体遗传变异图谱不完整，而构建来自多个个体的泛基因组能很好地解决这一缺陷，其研究内容包括负责基本生物学功能及该物种主要表型特征的核心基因组以及与物种的遗传多样性和个体独特性相关的可变基因组。根据核心和可变基因组所占比例的不同，泛基因组存在开放型和闭合型两种类型。本文主要综述了细菌、真菌和动植物的泛基因组学研究进展，讨论了其在各生物类群中的特征，其中哺乳动物泛基因组是相对闭合的，而目前已知的微生物、被子植物和部分低等动物的泛基因组倾向于开放，通过泛基因组的构建可以完善现有参考基因组并获取整个物种的完整变异信息，将有助于深入研究遗传多样性和表型变异产生的分子机制。

关键词： 泛基因组；存在/缺失变异；核心基因组；可变基因组

Pan-genome: setting a new standard for high-quality reference genomes

Peipei Bian, Yu Zhang, Yu Jiang

College of Animal Science and Technology, Northwest A&F University, Yangling 712100, China

Abstract: With the release of high-quality reference genomes assembled by long reads from the third-generation sequencing technology, as well as extensive re-sequencing and population genetic analysis, researchers found that a single reference genome does not represent the diversity within a species. The missing sequences on the reference genome result in an incomplete population genetic polymorphism map. The emergence of pan-genome can well repair the deficiency of single reference genome, which include core genome (responsible for basic biological functions and the main phenotypic characteristics within a species) and the variable genome (related to the genetic diversity or biological characteristics). According to the core and variable genome proportion, the types of pan-genomes can be either open or closed. Here, we review the current exploring of pan-genome for a range of species, to discuss the characteristics of pan-genome in various biological groups. The pan-genome of mammals are more likely closed, while the pan-genomes of microbes, angiosperms, and some invertebrates are likely non-closed. It is possible to complete the reference genome and obtain complete variation

收稿日期： 2021−08−26; **修回日期：** 2021−10−28

基金项目： 国家自然科学基金项目(编号：31822052)资助[Supported by the National Natural Science Foundation of China(No. 31822052)]

作者简介： 边培培，在读博士研究生，专业方向：动物遗传。E-mail: bppisc@163.com

通讯作者： 姜雨，博士，教授，研究方向：动物遗传。E-mail: yu.jiang@nwafu.edu.cn

DOI: 10.16288/j.yczz.21-214

网络出版时间： 2021/10/29 16:32:29

URI: https://kns.cnki.net/kcms/detail/11.1913.R.20211029.0923.001.html

information through the pan-genomic study, which will contribute to the study of molecular mechanism for genetic diversity and phenotypic evolution.

Keywords: pan-genome; presence and absence variations; core genome; variable genome

随着功能基因组学对基因功能的研究越来越细致，一个物种是否拥有高质量的参考基因组成为了深入解析其遗传与表型关系的重要前提。然而在群体水平上，研究人员发现来自同一物种不同个体的基因组序列并不能完全与该物种的参考基因组一一对应。因此建立一个能够包含这个物种全部基因组序列和变异信息情况的完整集合对基因组学的研究变得极为重要。

2005 年，Tettelin 等[1]首次在细菌研究中提出泛基因组(pan-genome)的概念，指整个物种基因组序列的非冗余集合，其中包括存在于该物种几乎所有个体中的核心基因组(core genome)和仅在部分个体中存在的可变基因组(accessory/variable/dispensable genome)。相对于细菌来说，真核生物无法频繁的跨物种交换遗传物质，被认为存在相对较少的存在/缺失变异(presence and absence variations, PAVs)[2]。但是随着对动植物个体基因组之间的比较研究，研究者发现高等生物同样具有普遍的跨物种基因交流，也存在相当数量的 PAVs，且许多位于功能性区域，承担重要的生物学功能[3~5]。泛基因组现已在植物、真菌、动物基因组学研究中被广泛用于更为全面地评估物种内遗传多样性，探究跨物种的基因交流和驯化及改良过程。研究表明利用泛基因组可以获取更为准确全面的变异信息，通过与表型进行关联，筛查出可变基因组中的功能基因或功能序列，这将为物种的遗传改良提供宝贵的遗传资源[6~11]。在微生物方面，利用泛基因组还可以对菌种进化、适应性及群体结构进行研究分析[12]；同时可应用于菌株重要毒力因子的发现和疫苗的设计[13]。

本文综述了细菌、真菌和动植物的泛基因组学研究进展，讨论了其在各生物类群中的特征，并对其在完善参考基因组以及获取完整变异信息上的应用进行了分析和展望。

1 泛基因组的概念和特点

广义的泛基因组是一个捕获了物种全部遗传信息的集合。对于包含一定数量个体基因组信息的泛基因组来说，整个基因或序列集合可以被分为核心基因组和可变基因组(图 1A)，核心基因组(core，一般认为存在于超过 95%的个体基因组中)；可变基因组又可以被进一步分为壳基因组(shell，在所有个体基因组中存在比例大约为 5%~95%)和云基因组(cloud，仅存在约少于 5%的个体基因组中)，shell 和 cloud 作为可变基因组的子集，一般与生物对特定环境的适应或生物学特性有关。上述分类能够弥补在实际定义不同基因组类别时所面临的不确定性，核心基因组为 95%以上而不是 100%的存在比例，可以避免某个个体的低质量基因组序列或者是基因组缺陷而造成的分类错误，确保真实的核心基因组在注释和分类过程中不被遗漏；而 cloud 则可能是个别个体基因组意外获得的外源基因，或者是来自于该个体基因组异常装配或者是外源污染[14]。具体的分类比例并不固定，研究人员可以根据实际物种研究情况，进行合理定义。一些研究证明了泛基因组中基因频率呈不对称的“U”型分布(图 1B)，这说明大部分基因或以核基因组的方式存在于绝大多数个体中，或以云基因组的方式存在于个别个体中[4,14~16]。

根据泛基因组中核心基因组的比例，将泛基因组分为开放型和闭合型两种，具体状态取决于所分析的物种特征，如物种整合外源 DNA 的能力，以及物种的生活方式和环境[17]。与具有开放程度较小泛基因组的物种相比，具有大型开放泛基因组的物种可能占据更多样的生态位和具有更复杂的群落[18]以及更大的有效群体规模，多态性水平更高。一般认为完全闭合的泛基因组是不存在的，在构建泛基因组的时候随着个体数量的增加，无论是开放型还是闭合型的泛基因组，整个泛基因组的大小都是逐步增加的，而核心基因组的大小都是逐渐减少的(图 1: C, D)。对于一个既定的物种来说，除去云序列(仅存在于物种极少数个体中)以外的核心基因组和可变基因组是一个定值。对于闭合型的泛基因组，有限数量个体的增加，可以使核心基因组和整个泛基因组含量迅速到达平台期，趋近于真实的水平。而

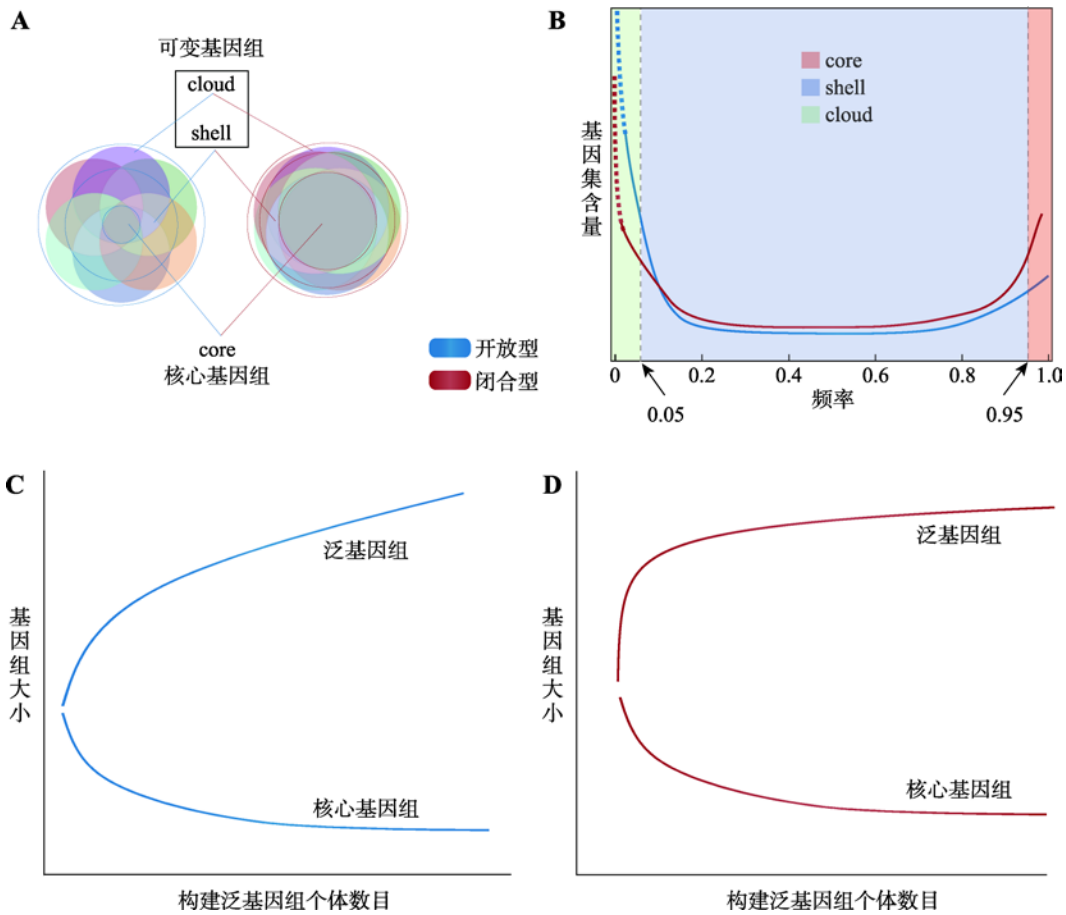

图 1　开放程度不同的泛基因组特征

Fig. 1　Pan-genome features with varying degrees of openness

A：泛基因组的组成。B：泛基因组中基因频率的不规则“U”型分布。C：开放型泛基因组：随着构建泛基因组个体的增加，整个泛基因组以及核心基因组大小的增长趋势。D：闭合型泛基因组：随着构建泛基因组个体的增加，整个泛基因组以及核心基因组大小的增长趋势。

开放程度高的泛基因组需要大量的个体才能获取这个真实值，在逐个增加研究个体时，到达平台期获得这个值的速度是缓慢的。基于以上差异，在进行闭合型泛基因组研究时，通过汇总有限数量个体的基因组序列，人们可以获取这个物种几乎全部的遗传信息。哺乳动物的泛基因组是比较典型的闭合型，其基因数量以及结构相对稳定，可变基因数量有限[5,19~20]，保证了高度复杂化的基因调节网络的稳定。而开放型泛基因组意味着，随着人们不断加入研究个体，其总是会有一定数量的新基因或者新序列的增长，也就是说通过一定数量的研究对象获取物种内全部遗传信息是不现实的，但是这种开放的模式为物种提供了丰富的遗传资源库，增加其功能多样性和复杂性，提高了其对动态环境的适应性。细菌、真菌和被子植物表现出开放型的特征，许多物种的核心基因比例小于 80%[2,21]。

2　群体中可变基因组的来源

当前泛基因组的研究主要是强调物种内部完整基因组序列的获取，所以更关注可变基因组，也就是在物种内部个体基因组之间一致性低的多态序列或者是产生了 PAVs 的序列集合。广义的泛基因组应该能够捕获该物种的全部遗传变异信息，但是当前的研究所构建的泛基因组大多体现不了那些小的插入缺失(insertions and deletions, indels)和单核苷酸多态性(single-nucleotide polymorphisms, SNPs)，以及不改变序列组成的易位(translocation)和倒位(inversion)

变异等，因此这种泛基因组可以被认为是狭义的泛基因组。

最初应用泛基因组概念的细菌，通常具有较小的基因组，其基因占据基因组序列的大部分，几乎没有基因间序列，而且数量差异很大，所以蛋白编码基因的含量是细菌等原核生物泛基因组研究的主要内容。原核基因组以不断变化的状态存在，通过水平基因转移，基因复制甚至可能以从头出现的方式而扩张，并通过基因丢失而收缩。在细菌中广泛的基因损失和水平基因转移(转化、接合和转导)是导致可变基因产生的两个主要进化过程[22]。不同模式真菌物种的泛基因组的研究表明真菌是通过菌株水平的创新来进化的，而不是大规模的水平基因转移。此外被子植物可通过全基因组复制(whole genome duplications, WGDs)、局部串联重复、转座因子(transposable elements, TEs)介导的重复、片段重复、近缘物种渗入、水平基因转移和从头基因诞生(*de novo* gene birth)获取新基因，同时也能通过染色体内重组和假基因化介导基因和序列的丢失[21]。虽然当前在动物上泛基因组的研究有限，但是众多的基因组学研究已经证明了在动物基因组上存在渗入、水平基因转移以及各种重复事件[23]。综上所述，正是通过序列的重复、近缘物种渗入、基因从头诞生或水平基因转移，以及后续的序列分歧/丢失或基因分裂/融合等多种过程，才产生了物种内广泛的 PAVs，形成了泛基因组。但是重复以及从头诞生的新基因一般很难在短时间内与原序列产生足够的分歧，因此在狭义泛基因组中难以被捕获。所以通常认为从狭义上来说，可变基因组的主要来源是基因和序列的丢失，渗入和水平基因转移(图 2)。

3 泛基因组的构建与呈现

目前构建泛基因组主要有基于迭代组装和基于从头组装两种方法(表 1)。

首先出现的是基于从头组装基因组构建泛基因组的方法[1]。这种方法分别对多个个体进行从头组装并注释，然后通过同一物种不同个体基因组间的相互比较，确定出核心基因组序列和可变基因组序列，最后将这些序列去冗余合并后构成一个包含该物种所有个体基因组序列的泛基因组[5]。这种方法的优势在于它能够检测到更多的结构变异(structural variations, SVs)，但对计算资源和样品的测序深度有较高的要求，不适用于基因组较大的物种和大规模群体的分析。迭代组装构建泛基因组方法的出现弥补了这些不足，其方式是由参考基因组起始，将每个样本的测序数据映射到参考基因组，提取未比对成功的序列进行组装，然后使用非冗余序列直接更新参考基因组，获得最终的扩展参考基因组即为该物种的泛基因组或者是对个体进行初步组装，从与参考基因组未比对上的 contigs 中移除冗余序列来构建代表性的非参考序列，结合参考基因组和代表性非参考基因组序列构建泛基因组。这种构建策略可以利用大规模的重测序数据，对测序深度要求很低，同时，因为只对未成功比对到参考基因组上的序列进行了组装，这种方法相对节省了计算资源，已在基因组较大的物种如小麦[24]以及大规模测序物种如水稻[10]中被应用。这种方法会在最终的泛基因组中产生大量的序列片段，并且无法检测每个个体的拷贝数变异(copy number variations, CNVs)，但对于基因的 PAVs 检测非常有效[25]。

这两种方法各有优缺点，目前均已被广泛应用于构建各种物种的泛基因组，研究人员通过将新发现的序列直接加入参考基因组的呈现形式产生了一系列的线性泛基因组，极大地丰富了人们对现有物种基因组的认识。然而，这种展示方式也带来了一些问题如：源于不同个体的变异信息被丢失，也几乎没有相应的程序和算法可以处理这种方式提供的

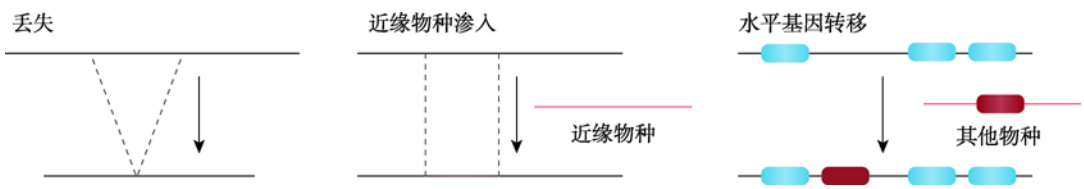

图 2 可变基因组的主要来源

Fig. 2 Origin of variable genome

表 1　泛基因组构建方法比较

Table 1　Comparison of pan-genome construction methods

构建方法	基于重测序 unmapping 序列的迭代组装	基于多个从头组装基因组的比较
(1)计算资源需求	低	高
(2)测序深度要求	低	高
(3)研究个体规模	大	小
(4)结构变异检测能力	低	高
(5)存在缺失变异检测能力	高	较低

变异信息。

获取可变基因组的序列组成和位置信息是展示和应用泛基因组的关键。但是线性泛基因组方式只呈现了可变基因组的序列组成，丢失了重要的染色体位置信息，因此在构建泛基因组的过程中，为防止重要信息的丢失，有两种方法：要么在线性泛基因组中标注序列位置信息，要么构建图结构的泛基因组。和线性基因组不同的是，图结构泛基因组是一个二维序列图谱，它以参考基因组为框架，以单个碱基作为图的节点，碱基间的前后关系作为图的边，存在序列差异的地方会自然形成不同的分支，呈现出一个图结构。这个图结构基因组可以依据新序列的加入不断扩展变化，最终它将会成为一个符合全物种的泛基因组图谱[26]。这种展示形式可以包含变异的嵌套，将同一位置的变异整合而不是单独占据一个区域，从而达到将所有变异精确纳入图谱的效果。这使得物种内大量复杂的变异可以紧凑的形式呈现。目前已有大量软件被开发用于这种图结构泛基因组的分析[27]，如 vg[28]、minigraph[26]、GraphType2[29]等，并且已在动植物基因组学研究中得到了初步应用[19,26,30~32]。

随着测序技术以及生物信息学工具的进步，包含全部序列变异信息的图结构泛基因组出现，尽管它受限于计算和存储当前只能应用于部分个体，但仍旧是向着广义泛基因组研究迈进的重要一步。未来技术的发展会让构建一个包含物种内全部遗传信息的泛基因组成为可能，实现精确处理大量基因组中的序列和变异信息，那时的基因组学研究才是真正在利用一个“参考”基因组。

4　泛基因组在不同物种中的研究进展

由于微生物基因组的可塑性和多样性，泛基因组的研究对其十分重要，同时，近年来测序和基因组组装成本的降低，研究人员在真核生物物种中发现了大规模的基因组变异，促使了泛基因组研究在真核生物中的扩展(图 3，表 2)。

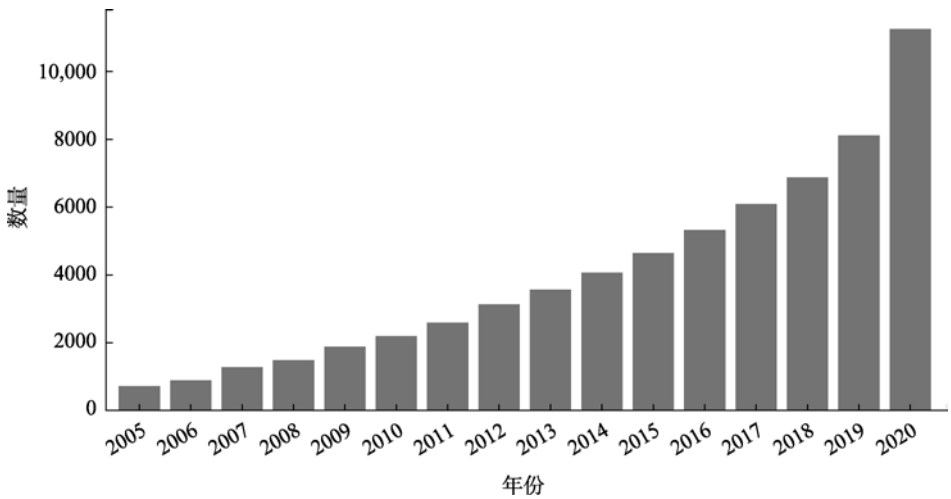

图 3　泛基因组相关研究数量的增长

Fig. 3　The growth of pan-genome publications

泛基因组的概念在 2005 年被首次提出之后，关键词“pangenome”或者“pan-genome”在 Europe Pubmed Central (https://europepmc.org/RestfulWebService)被搜索时的出现次数(检索日期：2021-08-17)。

表 2 泛基因组代表性研究

Table 2 Representative pangenome studies

年份	物种	基因组大小	个体数目	构建策略	主要新进展	参考文献
2005	无乳链球菌(*Streptococcus agalactiae*)	~2 Mb	8	基于多个从头组装的基因组的比较	泛基因组概念的引入	[1]
2007	N/A	N/A	N/A	N/A	综述文章，第一次在植物中应用泛基因组这个术语	[45]
2010	人(*Homo sapiens*)	~3.2 Gb	3	基于多个从头组装的基因组的比较	估计一个完整的泛基因组可能包含 19~40 Mb 在当前参考基因组中不存在的新序列，鉴定了额外 86 个新基因	[5]
2014	大豆(*Glycine soya*)	~0.9 Gb	7	基于多个从头组装的基因组的比较	第一个植物泛基因组文章，测序和重新组装了野生大豆个体的基因组，将注释基因聚类到基因家族，核心基因簇的比例为 49%	[46]
	玉米(*Zea mays*)	~2.4 Gb	503	基于多个从头组装的转录组的比较	获得了约 8600 个有代表性的在参考基因组中不存在的转录本，其中的 16.4%在所有品系中表达，82.7%在部分品系中表达	[50]
2016	甘蓝(*Brassica oleracea*)	~650 Mb	10	reads 映射到参考基因组；unmapping reads 的组装；通过新组装的 contigs 更新旧序列来产生新的参考序列。(将从每个基因组获得的 reads 映射到不断增长的泛基因组)	核心基因簇比例占泛基因组总数的 81%，近 20%的基因受到存在/缺失变异的影响	[53]
2018	水稻(*Oryza sativa*)	~400 Mb	3010	对个体测序数据进行组装，通过从与参考基因组 unaligned 的 contigs 中移除冗余序列来构建具有代表性的非参考序列，结合参考基因组构建泛基因组	鉴定了超过 10,000 个新的全长蛋白编码基因和大量的存在–缺失变异，核心基因簇比例占泛基因组总数的 54%~62%	[10]
2019	人(*Homo sapiens*)	~3.2Gb	910	reads 映射到参考基因组，组装 unmapping 的 reads，保留新组装的长度大于 1 kb 的非参考序列的 contigs 用于构建泛基因组	利用非洲血统的人类群体基因组构建泛基因组，获取了参考基因组中 296 Mb 不存在的序列	[20]
	番茄(*Solanum lycopersicum*)	~810Mb	725	对个体测序数据进行组装，通过从与参考基因组 unaligned 的 contigs 中移除冗余序列来构建具有代表性的非参考序列，结合参考基因组构建泛基因组	鉴定出一个约 4 kb 与风味相关的基因 *TomLoxC* 的启动子的存在缺失变异，表明泛基因组研究可以帮助物种恢复驯化或者改良过程中丢失的理想性状	[4]
	猪(*Sus scrofa*)	~2.7 Gb	12	基于从头组装的基因组之间的相互比较	第一个家养动物的线性泛基因组，获得了额外的 72.5 Mb 序列	[3]
	山羊(*Capra hircus*)	~2.9 Gb	10	基于从头比较来自近缘物种的基因组	第一个跨物种比对的泛基因组，从参考基因组中寻找缺失序列的有效且可靠的策略，获得了 38.3 Mb 序列	[70]
2020	大豆(*Glycine soja* 和 *Glycine max*)	~1 Gb	29	基于从头组装的基因组之间的相互比较，图结构泛基因组	鉴定了大的结构变异和基因融合事件，将结构变异与基因表达和农艺性状联系起来	[30]

续表

年份	物种	基因组大小	个体数目	构建策略	主要新进展	参考文献
	牛(*Bos taurus*)	~2.6 Gb	300	集成了线性参考基因组坐标和预先选择的变异(<50 bp)，图结构泛基因组	第一个家养动物的图结构泛基因组，在人类以外的大基因组动物上对图结构泛基因组的首个尝试	[32]
	贻贝(*Mytilus galloprovincialis*)	~1.28 Gb	16	测序 reads 被映射到贻贝参考基因组上，收集未映射的 reads 从头组装。新组装的 contigs 被添加到参考基因组中，构建了一个贻贝泛基因组(将从每个基因组获得的 reads 映射到不断增长的泛基因组)。	开放型的动物泛基因组，高比例的可变基因组(45%)，展示了动物泛基因组的潜能	[71]
2021	水稻(*Oryza sativa* 和 *Oryza glaberrima*)	~400 Mb	33	基于从头组装的基因组之间的相互比较，图结构泛基因组	共鉴定了 171,072 个 SVs 和 25,549 个 gCNVs，可以用于全基因组关联研究	[31]
	牛(*Bos taurus*)	~2.6 Gb	6	基于从头组装的基因组之间的相互比较，图结构泛基因组	70 Mb 的非参考基因组等位序列，提供了一个构建图结构泛基因组的框架，适合于多种物种	[19]

4.1　细菌泛基因组

首个细菌泛基因组由无乳链球菌(*Streptococcus agalactiae*)构建，每个菌种的核心基因组约占任何单个基因组的 80%[1]，这说明有一定数量的可变基因组仅在部分或者个别菌种中存在，很明显单个基因组序列不能反映细菌物种内的整个遗传变异性。细菌栖息在千差万别的生态位中，并具有大量相应的调节机制，以适应多变的环境[33]，核心基因的比例可以从 5%至 98%。除了使基因组垂直向下传给后代外，细菌还具有通过水平转移从环境中获取遗传物质的能力[34]，在获得基因的同时，为了维持细菌基因组小而紧凑的结构特征，基因还经常复制或丢失[35]。垂直传播和水平转移的混合作用使细菌基因组的系统发育分析复杂化[22]。在同一种细菌内，在基因组水平上也可能存在很大程度的个体差异。如在大肠杆菌(*Escherichia coli*)泛基因组中，任何一种大肠杆菌的基因组核心基因的比例都少于泛基因总数的 10%，即使在转录因子水平上，大肠杆菌基因组之间也存在巨大差异[36]。考虑到这种高水平的遗传变异，重建细菌的系统发育和种群历史，泛基因组研究是有必要的，并且可以作为细菌分类的重要依据[37]。Freschi 等[38]基于 1311 个铜绿假单胞菌的高质量基因组进行了泛基因组分析，研究了水平基因转移在人类病原体铜绿假单胞菌的抗菌素耐药性和毒力机制中的贡献，基于核心基因组的系统发育为其种群结构提供了强有力的证据。同样分枝杆菌泛基因组学研究证明了水平基因转移在进化过程中对其适应新环境和宿主中有重要作用[39]。随着测序成本降低以及数据库中可用细菌基因组的快速增加促进了泛基因组软件工具的开发[40]，一些在线软件例如 PanX[41]等，只要遵循特定步骤，即可生成泛基因组分析结果，加速了细菌泛基因组的研究进展。

在细菌泛基因组研究中发现一些可变基因在不断变化的环境中具有适应性优势[42]，另一些则和菌株的致病性和耐药性相关[18,43]。细菌泛基因组的研究在临床微生物学中有许多应用。它可以揭示细菌的致病潜力和抵抗抗菌素的能力，鉴定特定序列并预测抗原表位，从而可以设计分子或血清学检测方法和疫苗[40]。

4.2　植物泛基因组

从不同植物中获得的数据向人们展示了植物基因组的可塑性[44]，单个基因组已无法表征全部的遗传多样性，促使在基因组学研究中引入了植物泛基因组的概念[45]，这有助于深入了解植物产生遗传多样性和表型变异的过程。

首个植物泛基因组在 2014 年被报道，其基于

对 7 份代表性野生大豆(*Glycine soja*)全基因组的组装比较，发现了与生物抗性、种子组成、开花和成熟时间等重要农艺性状有关的可变基因[46]。泛基因组分析使人们能够追踪驯化和育种过程中基因的保留和丢失，开发将基因重新引入现代品种的潜力，恢复物种失去的遗传多样性。Gao 等[4]使用了具有广泛品种和地理代表性的 725 个番茄(*Solanum lycopersicum*)个体，揭示了参考基因组中不存在的 4873 个基因，PAVs 分析表明，在驯化和改良过程中有大量的基因丢失以及基因和启动子的负选择，并且丢失或者受到负选择的基因具有重要功能，尤其是与抗病性相关。此外，该研究还鉴定出在驯化阶段受选择的 *TomLoxC* 启动子上与番茄风味有关的稀有等位基因，利用其杂合子优势，可直接应用于生产中的性状改良。目前对泛基因组的研究并不局限于基因本身，基因以外的基因组区域也解释了作物表型变异的很大一部分，许多重要的农艺性状可能是由基因调控的变化而不是基因的存在/缺失变异决定的[21]。由于 SVs 的大小能够造成更多的核苷酸序列差异，因此可能会表现出不成比例的大表型效应[47]，已被确定为许多罕见和常见疾病的致病因素，并且通常被认为它们是通过影响基因的表达来起作用的。多个植物泛基因组研究也发现，SVs 导致基因组变异的同时，能够引起表型变异[48]。2020 年对番茄 PanSV 基因组的深入研究揭示了这一点，几乎一半的 SVs 与基因或调控序列重叠，并且半数影响编码序列的 SVs 与基因差异表达有关[49]。

泛基因组对于揭示物种内完整的遗传变异信息至关重要，尤其是近年来图结构泛基因组的发展，其构建及应用策略越来越稳定和完善，包含的功能元素和序列空间越来越充足，能够作为分析其他个体的参考，提高了研究人员对许多个体和物种基因组复杂性的理解。2020 年，有研究将 26 个大豆株系从头组装的基因组和 3 个先前报道的基因组构建了一个基于图形的泛基因组，结合 2898 个不同株系的重测序数据，揭示了众多仅用单个参考基因组无法检测到的变异，为大豆的进化和功能基因组学研究提供了更加完整的基因组图谱，并且通过对全基因组复制区域及 SVs 的研究，表明基因组复制是 SVs 进化的重要驱动力[30]。同样基于多个参考基因组水平的高质量组装基因组，2021 年，Qin 等[31]构建了高质量的水稻(*Oryza sativa* 和 *Oryza glaberrima*)图结构泛基因组。其研究提供了水稻基因组变异和驯化的遗传资源，并推断了整个水稻种群中 SVs 的派生状态，分析了 SVs 的分布并评估了 SVs 形成的机制以及 SVs 对基因表达的影响。此研究提供了 SVs 和基因的拷贝数变异(gene copy number variations, gCNVs)如何直接影响环境适应性和农艺性状的示例，展示了高质量基因组组装和图结构泛基因组在植物基因组学和功能基因组学中的重要作用。

迄今为止，已经有 10 余种植物建立了泛基因组包括玉米(*Zea mays*)[50]、大豆[30,46]、二穗短柄草(*Brachypodium distachyon*)[14]、辣椒(*Capsicum*)[51]、小麦(*Triticum* spp.)[24,52]、甘蓝(*Brassica oleracea*)[53]、水稻[10,31,54]、番茄[4,49]、狗尾草(*Setaria viridis*)[55]、向日葵(*Helianthus annuus* L.)[56]、大麦(*Hordeum vulgare* ssp.)[57,58]、桃子(*Prunus persica*)[6]，高粱(*Sorghum bicolor*)[59,60]等，除了重要农作物还包括驯化作物的野生和杂草近缘种，在每个被研究的物种中都有一个可观的可变基因含量(10%~60%)。可变基因经常被注释为与生物和非生物胁迫耐受性相关，同时这些基因相较于核心基因具有较低的进化约束和表达水平。通过泛基因组研究可以获取更多准确和大片段的 SVs，其中一些涉及改变基因剂量和表达水平的 SVs 影响了许多重要的农艺性状，包括水果的味道、大小和产量。这些发现强调了泛基因组研究在作物改良中的重要性和效用。

4.3 真菌泛基因组

研究人员使用长 reads 组装了驯化酵母及其野生近缘种的 12 个端到端的基因组，核基因组的大小从 11.73 到 12.14 Mb 不等，通过多个参考质量的基因组序列的比较，在驯化和野生个体之间观察到的许多差异可能反映了人类活动对基因组结构进化的影响[8]。接着通过对 1011 个酿酒酵母分离株的泛基因组构建，结合表型分析工作，提供了酿酒酵母变异的详细信息，为其全基因组关联分析(genome-wide association study, GWAS)奠定了基础，并为基因型–表型关系提供了新的见解，在规模上提供了与其他模式生物体相匹配的群体基因组资源[61]。2019 年报道了四种模式真菌的泛基因组：酿酒酵母(*Saccharomyces cerevisiae*)、白色念珠菌(*Candida albi-*

cans)、新型隐球菌(*Cryptococcus neoformans*)和烟曲霉(*Aspergillus fumigatus*)。研究发现，在这些物种中，每个菌株的所有基因中 80%~90%属于核心基因[62]，其余的可变基因可能与发病机制和抗菌素耐药有关。对物种祖先核心基因组和可变基因组的分析表明：基因复制等过程可能是影响真菌全基因组进化的主要因素，水平基因转移的作用有限。真菌病原体反复击败农作物抗性，变得对农药耐受，威胁着全球粮食生产，种群内的遗传变异多样性常常助长了这种进化过程[63]。小麦叶枯菌(*Zymoseptoria tritici*)会导致小麦枯萎病，2019 年其泛基因组的研究仅鉴定出了 58%的核心基因，其余的可变基因为其适应性进化提供了基础[64]。此外，有研究人员组装了来自六大洲的小麦叶枯菌的 19 个完整基因组，构建了小麦真菌病原体的高质量泛基因组，表明了染色体重排是广泛的基因存在/缺失变异的基础，同时发现可变基因组中富含与发病机制相关的功能基因[65]。

与细菌相似，真菌生物在基因含量上也显示出种内变异性。真菌泛基因组可用于获取大量菌株完整的变异信息，有助于真菌的驯化以及基因型-表型的关联研究。同时研究表明可变基因通常在致病性中起重要作用，通过泛基因组研究可以追踪确定参与感染和宿主反应的新基因的来源，也将有助于解决与作物–病原体共同进化相关的问题。

4.4　动物泛基因组

目前，相对于微生物和植物来说，动物泛基因组的研究范围还很有限，主要集中在人类(*Homo sapiens*)和家养动物。2010 年，Li 等[5]整合了亚洲人和非洲人新组装的基因组以及当时的人类参考基因组，构建了人类的第一个泛基因组。该研究在每个新组装基因组中获取了~5 Mb 在参考基因组中不存在的新序列，推断完整的人类泛基因组将包含现有参考基因组中不存在的 19~40 Mb 新序列。跨物种保守性分析表明这些新序列中包含的某些基因在哺乳动物基因组之间是保守的，很可能具有生物学功能。此研究证实了单个人类基因组序列中存在大量未证明的遗传区域，并且可以通过非常深的测序和从头组装来鉴定。对来自冰岛的 15,219 个人进行测序，仅关注非重复，非参考基因组序列，该研究共发现了 3719 个约 0.33 Mb 的新序列[66]。2019 年构建的汉人泛基因组发现了~29.5 Mb 的新序列，还鉴定了 188 个新的蛋白质编码基因[67]，而对 1000 个瑞典基因组的分析发现了~46 Mb 的新序列，大部分为重复序列(56%)[68]。Sherman 等[20]利用 910 个非洲后裔个体组成的深度测序数据集，构建的泛基因组比当前参考基因组多近 300Mb 的新序列，这是迄今为止报道的找到最多新序列的人类泛基因组。这些研究说明，单一参考基因组不足以进行基于群体的人类遗传学研究，更好的方法可能是为不同的人类群体创建参考基因组。

猪、牛和羊在畜牧业中都占据重要地位，猪也是重要的生物医学模型[69]，构建猪、牛和羊的泛基因组对优质种质资源的保护和利用，解析人类驯化动物的历史及作为模式动物探究生命奥秘有重要意义。Tian 等[3]使用了来自欧亚大陆的 12 个基因组构建了猪的泛基因组，相较于参考基因组(Sscrofa11.1)共获取了 72.5 Mb 的非冗余的新序列，且发现了脂肪分解的必要调节基因 *TIG3* 在猪群中显示为 PAVs，并且可能导致不同猪种之间的生理差异。山羊泛基因组研究利用其他 9 个从头组装的 Caprini 物种基因组共鉴定出了 38.3 Mb 山羊参考基因组上不存在的新序列，通过山羊全基因组重测序和转录组数据进一步验证了它们在山羊中的存在，证明了对亲缘关系近的物种基因组进行比较是一种基于参考基因组寻找缺失序列的有效且可靠的策略，这种方法也可能适用于其他物种[70]。这两项研究都表明使用泛基因组作为参考可产生更高质量的变异集合和更准确的基因表达量化，改善广泛的基因组分析。2020 年，研究人员使用来自约 300 头牛的变异信息(<50 bp)构建了家养动物的第一个图结构泛基因组[32]，提高了序列比对和基因分型的准确性，这是在人类以外的大基因组动物上对图结构泛基因组的首个尝试，为其他动物的研究提供了重要参考。稍后研究人员利用 6 只牛的基因组构建了图结构泛基因组，发现了参考基因组中缺失的功能序列[19]，其中包括参与免疫反应和免疫调节的基因，此研究提供了用于建立和利用更多样化的参考基因组的方法和框架。

除了上述哺乳动物以外，研究人员还报道了地中海贻贝的开放型泛基因组[71]。贻贝是具有生态和经济意义的食用双壳类生物，对生物和非生物应激源具有高度的侵袭性和复原力，其泛基因组具有

15,000 个可变基因，占全部泛基因组数量的 25%，平均出现的时间晚，表达水平低并且容易受到 PAVs 的影响，开放阅读框较短，基因结构复杂性低，并且参与了与防御和生存相关的功能，对生物适应性具有重要价值。此外，泛基因组也在昆虫基因组学的研究中得到了应用。蜱虫(Acari: Ixodidae)是传播最多样化的人类和动物病原体，对其泛基因组的研究揭示了不同蜱种的遗传结构和病原体组成主要受生态和地理因素的影响，并进一步确定了与不同宿主范围、生命周期和分布相关的物种特异性决定因素[72]，这也将为蜱虫生物学、病媒–病原体相互作用、疾病传播和控制策略的研究开辟新途径。熊蜂(Hymenoptera: Apidae)的泛基因组研究表明在系统发育框架中对多个基因组进行比较分析，大大提高了进化分析的精度和灵敏度，并可以提供识别基因组稳定和动态特征的可靠结果[73]。此研究也将助力于功能基因定位和克隆，以及重测序和群体基因组学研究，为熊蜂在农业中的使用提供基础的遗传信息。

上述研究表明，目前的动物单一参考基因组对于具有高适应能力，高杂合度，高水平重复元素以及复杂群体历史的物种还远远不够完整，并且强调了参考基因组缺失的基因对于临床和农业应用的潜在影响。后续研究应集中于动物高质量泛基因组的构建，获取完整的泛基因组序列，以及构建可用的图结构泛基因组，寻找更多可应用于经济动物选育和改良的遗传信息。

5 结语与展望

基因组时代的前期，研究人员采取的主要策略就是为目标物种提供一个单一的“参考”基因组，该基因组成为各种遗传分析(包括研究物种内部和物种之间的变异)的基础[25]。随着测序新技术的发展，测序质量进一步提高，同时成本大大降低，成千上万的新基因组被测序，物种间大量变异被获取，人们开始意识到单一参考基因组不足以代表一个物种全部的遗传信息。泛基因组分析提供了一个平台，可通过收集物种的整个基因组信息库来获取其全部的遗传多样性，在细菌、真菌以及动植物中已经得到了广泛的应用。

在目前泛基因组的研究中仍存在一些问题亟待解决：各种生物的基因组组装还不完整，尽管长 reads 测序被证明已经能够解析基因组中一些具有挑战性的区域，检测以前无法获取的 SVs[74~76]，但是为物种中每一个个体实现完整、无间隙的装配是不现实的；此外，基因组的测序、组装，泛基因组的构建策略，序列注释，判断 PAVs 等一系列方法并没有标准化的流程，导致不同研究获取的泛基因组序列不能直接比较，汇集所有数据建立一个完整的泛基因组将是一个巨大的挑战。

微生物和被子植物相比于哺乳动物，基因组可塑性更高，物种内的遗传多样性更为丰富，因而有了相对广泛的研究。哺乳动物基因组相对保守，通常只有基因间或片段化的基因区域参与基因组序列的增减，但是这并不意味着动物泛基因组的重要性降低。从对贻贝的研究[71]中可以看到动物泛基因组的潜力，随着泛基因组研究扩展到更多的物种，才能真正准确地评估一个生物类群的多样性水平。近年来泛基因组学研究为植物多样性研究和改良提供了新的思路[21,44]，但在除人类以外的动物中泛基因组学研究有限，在其他动物泛基因组的研究上还需要努力，以期为动物遗传相关研究打下坚实基础。

当前泛基因组研究的核心是用更丰富的数据结构取代传统的线性参考基因组[27]，相对于传统的单一线性参考基因组，泛基因组作为参考基因组能更加全面地呈现群体基因组信息，同时更有益于变异信息的获取和利用。随着图结构泛基因组的构建方式和分析策略的逐步完善，利用泛基因组将会更加高效地辅助解决功能基因组学研究的难题，从而彻底改变基因组学的研究。

参考文献(References):

[1] Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Ros IMY, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou LW, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O’Connor KJB, Smith S, Utterback TR, White

O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci USA*, 2005, 102(39): 13950–13955. [DOI]

[2] Golicz AA, Bayer PE, Bhalla PL, Batley J, Edwards D. Pangenomics comes of age: from bacteria to plant and animal applications. *Trends Genet*, 2020, 36(2): 132–145. [DOI]

[3] Tian XM, Li R, Fu WW, Li Y, Wang XH, Li M, Du D, Tang QZ, Cai YD, Long YM, Zhao Y, Li MZ, Jiang Y. Building a sequence map of the pig pan-genome from multiple *de novo* assemblies and Hi-C data. *Sci China Life Sci*, 2019, 63(5): 750–763. [DOI]

[4] Gao L, Gonda I, Sun HE, Ma QY, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, Thannhauser TW, Foolad MR, Diez MJ, Blanca J, Canizares J, Xu YM, van der Knaap E, Huang SW, Klee HJ, Giovannoni JJ, Fei ZJ. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet*, 2019, 51(6): 1044–1051. [DOI]

[5] Li RQ, Li YR, Zheng HC, Luo RB, Zhu HM, Li QB, Qian WB, Ren YY, Tian G, Li JX, Zhou GY, Zhu X, Wu HL, Qin JJ, Jin X, Li DF, Cao HZ, Hu XD, Blanche H, Cann H, Zhang XQ, Li SG, Bolund L, Kristiansen K, Yang HM, Wang J, Wang J. Building the sequence map of the human pan-genome. *Nat Biotechnol*, 2010, 28(1): 57–63. [DOI]

[6] Cao K, Peng Z, Zhao X, Li Y, Liu KZ, Arus P, Zhu GR, Deng SH, Fang WC, Chen CW, Wang XW, Wu JL, Fei ZJ, Wang LR. Pan-genome analyses of peach and its wild relatives provide insights into the genetics of disease resistance and species adaptation. *BioRxiv*, 2020, doi: 10.1101/2020.07.13.200204. [DOI]

[7] Schreiber M, Stein N, Mascher M. Genomic approaches for studying crop evolution. *Genome Biol*, 2018, 19(1): 140. [DOI]

[8] Yue JX, Li J, Aigrain L, Hallin J, Persson K, Oliver K, Bergström A, Coupland P, Warringer J, Lagomarsino MC, Fischer G, Durbin R, Liti G. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat Genet*, 2017, 49(6): 913–924. [DOI]

[9] Kaas RS, Friis C, Ussery DW, Aarestrup FM. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics*, 2012, 13(1): 577. [DOI]

[10] Wang WS, Mauleon R, Hu ZQ, Chebotarov D, Tai SS, Wu ZC, Li M, Zheng TQ, Fuentes RR, Zhang F, Mansueto L, Copetti D, Sanciangco M, Palis KC, Xu JL, Sun C, Fu BY, Zhang HL, Gao YM, Zhao XQ, Shen F, Cui X, Yu H, Li ZC, Chen ML, Detras J, Zhou YL, Zhang XY, Zhao Y, Kudrna D, Wang CC, Li R, Jia B, Lu JY, He XC, Dong ZT, Xu JB, Li YH, Wang M, Shi JX, Li J, Zhang DB, Lee S, Hu WS, Poliakov A, Dubchak I, Ulat VJ, Borja FN, Mendoza JR, Ali J, Li J, Gao Q, Niu YC, Yue Z, Naredo MEB, Talag J, Wang XQ, Li JJ, Fang XD, Yin Y, Glaszmann JC, Zhang JW, Li JY, Hamilton RS, Wing RA, Ruan J, Zhang GY, Wei CC, Alexandrov N, McNally KL, Li ZK, Leung H. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, 2018, 557(7703): 43–49. [DOI]

[11] Song JM, Guan ZL, Hu JL, Guo CC, Yang ZQ, Wang S, Liu DX, Wang B, Lu SP, Zhou R, Xie WZ, Cheng YF, Zhang YT, Liu K, Yang QY, Chen LL, Guo L. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat Plants*, 2020, 6(1): 34–45. [DOI]

[12] Zou YQ, Xue WB, Luo GW, Deng ZQ, Qin PP, Guo RJ, Sun HP, Xia Y, Liang SS, Dai Y, Wan DW, Jiang RR, Su LL, Feng Q, Jie ZY, Guo TK, Xia ZK, Liu C, Yu JH, Lin YX, Tang SM, Huo GC, Xu X, Hou Y, Liu X, Wang J, Yang HM, Kristiansen K, Li JH, Jia HJ, Xiao L. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat Biotechnol*, 2019, 37(2): 179–185. [DOI]

[13] Naz K, Naz A, Ashraf ST, Rizwan M, Ahmad J, Baumbach J, Ali A. PanRV: Pangenome-reverse vaccinology approach for identifications of potential vaccine candidates in microbial pangenome. *BMC Bioinformatics*, 2019, 20(1): 123. [DOI]

[14] Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu SQ, Stritt C, Roulin AC, Schackwitz W, Tyler L, Martin J, Lipzen A, Dochy N, Phillips J, Barry K, Geuten K, Budak H, Juenger TE, Amasino R, Caicedo AL, Goodstein D, Davidson P, Mur LAJ, Figueroa M, Freeling M, Catalan P, Vogel JP. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat Commun*, 2017, 8(1): 2184. [DOI]

[15] Brito PH, Chevreux B, Serra CR, Schyns G, Henriques AO, Pereira-Leal JB. Genetic competence drives genome diversity in *Bacillus subtilis*. *Genome Biol Evol*, 2018, 10(1): 108–124. [DOI]

[16] Vincent AT, Schiettekatte O, Goarant C, Neela VK, Bernet

E, Thibeaux R, Ismail N, Khalid MKNM, Amran F, Masuzawa T, Nakao R, Korba AA, Bourhy P, Veyrier FJ, Picardeau M. Revisiting the taxonomy and evolution of pathogenicity of the genus *Leptospira* through the prism of genomics. *PLoS Negl Trop Dis*, 2019, 13(5): e0007270. [DOI]

[17] Lefébure T, Pavinski Bitar PD, Suzuki H, Stanhope MJ. Evolutionary dynamics of complete campylobacter pan-genomes and the bacterial species concept. *Genome Biol Evol*, 2010, 2: 646–655. [DOI]

[18] Rouli L, Merhej V, Fournier PE, Raoult D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect*, 2015, 7: 72–85. [DOI]

[19] Crysnanto D, Leonard AS, Fang ZH, Pausch H. Novel functional sequences uncovered through a bovine multiassembly graph. *Proc Natl Acad Sci USA*, 2021, 118(20): e2101056118. [DOI]

[20] Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, Boorgula MP, Chavan S, Vergara C, Ortega VE, Levin AM, Eng C, Yazdanbakhsh M, Wilson JG, Marrugo J, Lange LA, Williams LK, Watson H, Ware LB, Olopade CO, Olopade O, Oliveira RR, Ober C, Nicolae DL, Meyers DA, Mayorga A, Knight-Madden J, Hartert T, Hansel NN, Foreman MG, Ford JG, Faruque MU, Dunston GM, Caraballo L, Burchard EG, Bleecker ER, Araujo MI, Herrera-Paz EF, Campbell M, Foster C, Taub MA, Beaty TH, Ruczinski I, Mathias RA, Barnes KC, Salzberg SL. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet*, 2019, 51(1): 30–35. [DOI]

[21] Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new reference. *Nat Plants*, 2020, 6(8): 914–920. [DOI]

[22] Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol*, 2014, 12(1): 66. [DOI]

[23] Richard GF. Eukaryotic pangenomes. The Pangenome, Springer International Publishing, 2020, 253–291. [DOI]

[24] Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan CKK, Visendi P, Lai KT, Doležel J, Batley J, Edwards D. The pangenome of hexaploid bread wheat. *Plant J*, 2017, 90(5): 1007–1013. [DOI]

[25] Sherman RM, Salzberg SL. Pan-genomics in the human genome era. *Nat Rev Genet*, 2020, 21(4): 243–254. [DOI]

[26] Li H, Feng XW, Chu C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol*, 2020, 21(1): 265. [DOI]

[27] Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, Rautiainen M, Garg S, Paten B, Marschall T, Sirén J, Garrison E. Pangenome graphs. *Annu Rev Genom Hum G*, 2020, 21(1): 139–162. [DOI]

[28] Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten B, Durbin R. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol*, 2018, 36(9): 875–879. [DOI]

[29] Eggertsson HP, Kristmundsdottir S, Beyter D, Jonsson H, Skuladottir A, Hardarson MT, Gudbjartsson DF, Stefansson K, Halldorsson BV, Melsted P. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat Commun*, 2019, 10(1): 5402. [DOI]

[30] Liu YC, Du HL, Li PC, Shen YT, Peng H, Liu SL, Zhou GA, Zhang HK, Liu Z, Shi M, Huang XH, Li Y, Zhang M, Wang Z, Zhu BG, Han B, Liang CZ, Tian ZX. Pan-genome of wild and cultivated soybeans. *Cell*, 2020, 182(1): 162–176.e13. [DOI]

[31] Qin P, Lu HW, Du HL, Wang H, Chen WL, Chen Z, He Q, Ou SJ, Zhang HY, Li XZ, Li XX, Li Y, Liao Y, Gao Q, Tu B, Yuan H, Ma BT, Wang YP, Qian YW, Fan SJ, Li WT, Wang J, He M, Yin JJ, Li T, Jiang N, Chen XW, Liang CZ, Li SG. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, 2021, 184(13): 3542–3558.e16. [DOI]

[32] Crysnanto D, Pausch H. Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. *Genome Biol*, 2020, 21(1): 184. [DOI]

[33] Boutte CC, Crosson S. Bacterial lifestyle shapes stringent response activation. *Trends Microbiol*, 2013, 21(4): 174–180. [DOI]

[34] Soucy SM, Huang JL, Gogarten JP. Horizontal gene transfer: building the web of life. *Nat Rev Genet*, 2015, 16(8): 472–482. [DOI]

[35] Lefébure T, Stanhope MJ. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol*, 2007, 8(5): R71. [DOI]

[36] Cook H, Ussery DW. Sigma factors in a thousand *E. coli* genomes. *Environ Microbiol*, 2013, 15(12): 3121–3129. [DOI]

[37] O'Callaghan A, Bottacini F, O'Connell Motherway M, van Sinderen D. Pangenome analysis of *Bifidobacterium*

longum and site-directed mutagenesis through by-pass of restriction-modification systems. *BMC Genomics*, 2015, 16(1): 832. [DOI]

[38] Freschi L, Vincent AT, Jeukens J, Emond-Rheault JG, Kukavica-Ibrulj I, Dupont MJ, Charette SJ, Boyle B, Levesque RC. The *pseudomonas aeruginosa* pan-genome provides new insights on its population structure, horizontal gene transfer, and pathogenicity. *Genome Biol Evol*, 2019, 11(1): 109–120. [DOI]

[39] Dumas E, Christina Boritsch E, Vandenbogaert M, de la Vega RCR, Thiberge JM, Caro V, Gaillard JL, Heym B, Girard-Misguich F, Brosch R, Sapriel G. Mycobacterial pan-genome analysis suggests important role of plasmids in the radiation of type VII secretion systems. *Genome Biol Evol*, 2016, 8(2): 387–402. [DOI]

[40] Anani H, Zgheib R, Hasni I, Raoult D, Fournier PE. Interest of bacterial pangenome analyses in clinical microbiology. *Microb Pathog*, 2020, 149: 104275. [DOI]

[41] Ding W, Baumdicker F, Neher RA. panX: pan-genome analysis and exploration. *Nucleic Acids Res*, 2018, 46(1): e5. [DOI]

[42] Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. *Curr Opin Microbiol*, 2015, 23: 148–154. [DOI]

[43] Fu J, Qin QW. Pan-genomics analysis of 30 *Escherichia coli* genomes. *Hereditas(Beijing)*, 2012, 34(6): 765–772
付静, 秦启伟. 30 株大肠杆菌的泛基因组学特征分析. 遗传, 2012, 34(6): 765–772. [DOI]

[44] Golicz AA, Batley J, Edwards D. Towards plant pangenomics. *Plant Biotechnol J*, 2016, 14(4): 1099–1105. [DOI]

[45] Morgante M, De Paoli E, Radovic S. Transposable elements and the plant pan-genomes. *Curr Opin Plant Biol*, 2007, 10(2): 149–155. [DOI]

[46] Li YH, Zhou GY, Ma JX, Jiang WK, Jin LG, Zhang ZH, Guo Y, Zhang JB, Sui Y, Zheng LT, Zhang SS, Zuo QY, Shi XH, Li YF, Zhang WK, Hu YY, Kong GY, Hong HL, Tan B, Song J, Liu ZX, Wang YS, Ruan H, Yeung CKL, Liu J, Wang HL, Zhang LJ, Guan RX, Wang KJ, Li WB, Chen SY, Chang RZ, Jiang Z, Jackson SA, Li RQ, Qiu LJ. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol*, 2014, 32(10): 1045–1052. [DOI]

[47] Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang YJ, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, MacArthur DG, MacDonald JR, Onyiah I, Pang AWC, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME. Origins and functional impact of copy number variation in the human genome. *Nature*, 2010, 464(7289): 704–712. [DOI]

[48] Liu YC, Tian ZX. From one linear genome to a graph-based pan-genome: a new era for genomics. *Sci China Life Sci*, 2020, 63(12): 1938–1941. [DOI]

[49] Alonge M, Wang XA, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, Levy Y, Harel TH, Shalev-Schlosser G, Amsellem Z, Razifard H, Caicedo AL, Tieman DM, Klee H, Kirsche M, Aganezov S, Ranallo-Benavidez TR, Lemmon ZH, Kim J, Robitaille G, Kramer M, Goodwin S, McCombie WR, Hutton S, Van Eck J, Gillis J, Eshed Y, Sedlazeck FJ, van der Knaap E, Schatz MC, Lippman ZB. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, 2020, 182(1): 145–161.e23. [DOI]

[50] Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K, de Leon N, Kaeppler SM, Buell CR. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*, 2014, 26(1): 121–135. [DOI]

[51] Ou LJ, Li D, Lv JH, Chen WC, Zhang ZQ, Li XF, Yang BZ, Zhou SD, Yang S, Li WG, Gao HZ, Zeng Q, Yu HY, Ouyang B, Li F, Liu F, Zheng JY, Liu YH, Wang J, Wang BB, Dai XZ, Ma YQ, Zou XX. Pan-genome of cultivated pepper (*Capsicum*) and its use in gene presence-absence variation analyses. *New Phytol*, 2018, 220(2): 360–363. [DOI]

[52] Walkowiak S, Gao LL, Monat C, Haberer G, Kassa MT, Brinton J, Ramirez-Gonzalez RH, Kolodziej MC, Delorean E, Thambugala D, Klymiuk V, Byrns B, Gundlach H, Bandi V, Siri JN, Nilsen K, Aquino C, Himmelbach A, Copetti D, Ban T, Venturini L, Bevan M, Clavijo B, Koo DH, Ens J, Wiebe K, N'Diaye A, Fritz AK, Gutwin C, Fiebig A, Fosker C, Fu BX, Accinelli GG, Gardner KA, Fradgley N, Gutierrez-Gonzalez J, Halstead-Nussloch G, Hatakeyama M, Koh CS, Deek J, Costamagna AC, Fobert P, Heavens D, Kanamori H, Kawaura K, Kobayashi F, Krasileva K, Kuo T, McKenzie N, Murata K, Nabeka Y, Paape T, Padmarasu S, Percival-Alwyn L, Kagale S, Scholz U, Sese J, Juliana P, Singh R, Shimizu-Inatsugi R, Swarbreck D, Cockram J, Budak H, Tameshige T, Tanaka T, Tsuji H, Wright J, Wu JZ, Steuernagel B, Small I, Cloutier S, Keeble-Gagnère G, Muehlbauer G, Tibbets J, Nasuda S, Melonek J, Hucl PJ,

Sharpe AG, Clark M, Legg E, Bharti A, Langridge P, Hall A, Uauy C, Mascher M, Krattinger SG, Handa H, Shimizu KK, Distelfeld A, Chalmers K, Keller B, Mayer KFX, Poland J, Stein N, McCartney CA, Spannagl M, Wicker T, Pozniak CJ. Multiple wheat genomes reveal global variation in modern breeding. *Nature*, 2020, 588(7837): 277–283. [DOI]

[53] Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CKK, Severn-Ellis A, McCombie WR, Parkin IAP, Paterson AH, Pires JC, Sharpe AG, Tang HB, Teakle GR, Town CD, Batley J, Edwards D. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat Commun*, 2016, 7(1): 13390. [DOI]

[54] Zhao Q, Feng Q, Lu HY, Li Y, Wang AH, Tian QL, Zhan QL, Lu YQ, Zhang L, Huang T, Wang YC, Fan DL, Zhao Y, Wang ZQ, Zhou CC, Chen JY, Zhu CR, Li WJ, Weng QJ, Xu Q, Wang ZX, Wei XH, Han B, Huang XH. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat Genet*, 2018, 50(2): 278–284. [DOI]

[55] Mamidi S, Healey A, Huang P, Grimwood J, Jenkins J, Barry K, Sreedasyam A, Shu SQ, Lovell JT, Feldman M, Wu JX, Yu YQ, Chen C, Johnson J, Sakakibara H, Kiba T, Sakurai T, Tavares R, Nusinow DA, Baxter I, Schmutz J, Brutnell TP, Kellogg EA. A genome resource for green millet *Setaria viridis* enables discovery of agronomically valuable loci. *Nat Biotechnol*, 2020, 38(10): 1203–1210. [DOI]

[56] Hübner S, Bercovich N, Todesco M, Mandel JR, Odenheimer J, Ziegler E, Lee JS, Baute GJ, Owens GL, Grassa CJ, Ebert DP, Ostevik KL, Moyers BT, Yakimowski S, Masalia RR, Gao LX, Ćalić I, Bowers JE, Kane NC, Swanevelder DZH, Kubach T, Muños S, Langlade NB, Burke JM, Rieseberg LH. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat Plants*, 2019, 5(1): 54–62. [DOI]

[57] Ma YL, Liu M, Stiller J, Liu CJ. A pan-transcriptome analysis shows that disease resistance genes have undergone more selection pressure during barley domestication. *BMC Genomics*, 2019, 20(1): 12. [DOI]

[58] Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D, Himmelbach A, Ens J, Zhang XQ, Angessa TT, Zhou GF, Tan C, Hill C, Wang PH, Schreiber M, Boston LB, Plott C, Jenkins J, Guo Y, Fiebig A, Budak H, Xu DD, Zhang J, Wang CC, Grimwood J, Schmutz J, Guo GG, Zhang GP, Mochida K, Hirayama T, Sato K, Chalmers KJ, Langridge P, Waugh R, Pozniak CJ, Scholz U, Mayer KFX, Spannagl M, Li CD, Mascher M, Stein N. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, 2020, 588(7837): 284–289. [DOI]

[59] Tao YF, Luo H, Xu JB, Cruickshank A, Zhao XR, Teng F, Hathorn A, Wu XY, Liu YM, Shatte T, Jordan D, Jing HC, Mace E. Extensive variation within the pan-genome of cultivated and wild sorghum. *Nat Plants*, 2021, 7(6): 766–773. [DOI]

[60] Wang B, Jiao YP, Chougule K, Olson A, Huang J, Llaca V, Fengler K, Wei XH, Wang LY, Wang XF, Regulski M, Drenkow J, Gingeras T, Hayes C, Armstrong JS, Huang YH, Xin ZG, Ware D. Pan-genome analysis in sorghum highlights the extent of genomic variation and sugarcane aphid resistance genes. *BioRxiv*, 2021, doi: 10.1101/2021.01.03.424980. [DOI]

[61] Peter J, De Chiara M, Friedrich A, Yue JX, Pflieger D, Bergström A, Sigwalt A, Barre B, Freel K, Llored A, Cruaud C, Labadie K, Aury JM, Istace B, Lebrigand K, Barbry P, Engelen S, Lemainque A, Wincker P, Liti G, Schacherer J. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature*, 2018, 556(7701): 339–344. [DOI]

[62] McCarthy CGP, Fitzpatrick DA. Pan-genome analyses of model fungal species. *Microb Genom*, 2019, 5(2): e000243. [DOI]

[63] Badet T, Croll D. The rise and fall of genes: origins and functions of plant pathogen pangenomes. *Curr Opin Plant Biol*, 2020, 56: 65–73. [DOI]

[64] Plissonneau C, Hartmann FE, Croll D. Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biol*, 2018, 16(1): 5. [DOI]

[65] Badet T, Oggenfuss U, Abraham L, McDonald BA, Croll D. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen *Zymoseptoria tritici*. *BMC Biol*, 2020, 18(1): 12. [DOI]

[66] Kehr B, Helgadottir A, Melsted P, Jonsson H, Helgason H, Jonasdottir A, Jonasdottir A, Sigurdsson A, Gylfason A, Halldorsson GH, Kristmundsdottir S, Thorgeirsson G, Olafsson I, Holm H, Thorsteinsdottir U, Sulem P, Helgason A, Gudbjartsson DF, Halldorsson BV, Stefansson K. Diversity in non-repetitive human sequences not found in the reference genome. *Nat Genet*, 2017, 49(4): 588–593. [DOI]

[67] Duan ZQ, Qiao YY, Lu JY, Lu HM, Zhang WM, Yan FZ,

Sun C, Hu ZQ, Zhang Z, Li GC, Chen HZ, Xiang Z, Zhu ZG, Zhao HY, Yu YY, Wei CC. HUPAN: a pan-genome analysis pipeline for human genomes. *Genome Biol*, 2019, 20(1): 149. [DOI]

[68] Eisfeldt J, Mårtensson G, Ameur A, Nilsson D, Lindstrand A. Discovery of novel sequences in 1,000 swedish genomes. *Mol Biol Evol*, 2020, 37(1): 18–30. [DOI]

[69] Lunney JK. Advances in swine biomedical model genomics. *Int J Biol Sci*, 2007, 3(3): 179–184. [DOI]

[70] Li R, Fu WW, Su R, Tian XM, Du D, Zhao Y, Zheng ZQ, Chen QM, Gao S, Cai YD, Wang XH, Li JQ, Jiang Y. Towards the complete goat pan-genome by recovering missing genomic segments from the reference genome. *Front Genet*, 2019, 10: 1169. [DOI]

[71] Gerdol M, Moreira R, Cruz F, Gómez-Garrido J, Vlasova A, Rosani U, Venier P, Naranjo-Ortiz MA, Murgarella M, Greco S, Balseiro P, Corvelo A, Frias L, Gut M, Gabaldón T, Pallavicini A, Canchaya C, Novoa B, Alioto TS, Posada D, Figueras A. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. *Genome Biol*, 2020, 21(1): 275. [DOI]

[72] Jia N, Wang JF, Shi WQ, Du LF, Sun Y, Zhan W, Jiang JF, Wang Q, Zhang B, Ji PF, Bell-Sakyi L, Cui XM, Yuan TT, Jiang BG, Yang WF, Lam TTY, Chang QC, Ding SJ, Wang XJ, Zhu JG, Ruan XD, Zhao L, Wei JT, Ye RZ, Que TC, Du CH, Zhou Y-H, Cheng JX, Dai PF, Guo WB, Han XH, Huang EJ, Li LF, Wei W, Gao YC, Liu JZ, Shao HZ, Wang X, Wang CC, Yang TC, Huo QB, Li W, Chen HY, Chen SE, Zhou LG, Ni XB, Tian JH, Sheng Y, Liu T, Pan YS, Xia LY, Li J, Tick Genome and Microbiome Consortium (TIGMIC), Zhao FQ, Cao WC. Large-scale comparative analyses of tick genomes elucidate their genetic diversity and vector capacities. *Cell*, 2020, 182(5): 1328–1340.e13. [DOI]

[73] Sun C, Huang JX, Wang Y, Zhao XM, Su L, Thomas GWC, Zhao MY, Zhang XT, Jungreis I, Kellis M, Vicario S, Sharakhov IV, Bondarenko SM, Hasselmann M, Kim CN, Paten B, Penso-Dolfin L, Wang L, Chang YX, Gao Q, Ma L, Ma LN, Zhang Z, Zhang HB, Zhang HH, Ruzzante L, Robertson HM, Zhu YH, Liu YJ, Yang HP, Ding LL, Wang QG, Ma DN, Xu WL, Liang C, Itgen MW, Mee L, Cao G, Zhang Z, Sadd BM, Hahn MW, Schaack S, Barribeau SM, Williams PH, Waterhouse RM, Mueller RL. Genus-wide characterization of bumblebee genomes provides insights into their evolution and variation in ecological and behavioral traits. *Mol Biol Evol*, 2021, 38(2): 486–501. [DOI]

[74] Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet*, 2018, 19(6): 329–346. [DOI]

[75] Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet*, 2020, 21(10): 597–614. [DOI]

[76] Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 2011, 12(5): 363–376. [DOI]

(责任编委：李海鹏)