



评述

基于基因组结构变异解析动物复杂性状的研究进展

汪亚军, 王晓萍*, 于黎*

云南大学, 生命科学学院, 省部共建云南生物资源保护与利用国家重点实验室, 昆明 650091

* 联系人, E-mail: wangxp@ynu.edu.cn; yuli@ynu.edu.cn

收稿日期: 2023-09-14; 接受日期: 2023-11-15; 网络版发表日期: 2024-01-09

国家自然科学基金(批准号: 32230013, 31925006, 32360120)和云南省科技厅应用基础研究计划面上项目(批准号: 202001BB050058)资助

摘要 动物为了适应环境而演化出多种多样的表型性状。揭示物种复杂性状的分子机制对理解物种适应性演化至关重要。基因组结构变异是动物复杂性状重要分子机制之一, 它反映了基因组不同位置和序列的可变性, 包括大片段的平衡变异倒位、易位, 以及非平衡变异插入、缺失和重复等多种多样变异类型。相对于单碱基变异和小片段插入缺失, 基因组结构变异在基因组上影响的碱基范围更广, 对动物复杂性状的影响更大。本文通过对基因组结构变异的类型和结构变异的检测鉴定技术的研究历史进行概述, 并从动物毛色性状、极端环境适应性性状和驯养动物经济性状等方面综述基因组结构变异与复杂性状关联的研究进展, 为从基因组结构变异的角度解析动物复杂性状产生的分子机制提供重要参考。

关键词 基因组结构变异, 复杂性状, 动物, 适应性进化

在适应性演化的驱动下, 生物演化出多种复杂性状。生物复杂性状的产生是物种生存和持续存在的重要策略, 是物种适应性演化的结果和物种多样性存在的基础^[1], 因此, 揭示物种复杂性状的分子机制对理解物种适应性演化至关重要^[2,3]。在性状演化的分子机制中, 基因组结构变异(structural variation, SV)是影响物种复杂性状形成的重要驱动力之一, 它能够反映出基因组不同位置和序列的可变性, 如大片段的平衡变异倒位、易位, 以及非平衡变异插入、缺失和重复等多种多样变异类型。基因组结构变异不仅可以改变编码序列顺序或拷贝数, 还可以影响如启动子(promoter)和

增强子(enhancer)等顺式调控元件(cis-regulatory elements, CREs)的位置或功能, 导致相关基因区域的三维空间结构和染色质开放区域等的改变, 通过改变调控元件与靶基因间的相互作用, 影响靶基因的转录表达和翻译过程, 最终导致物种的性状发生改变^[4-7]。传统意义上, 由结构变异导致的表型通常都是有害的, 尤其是在人类疾病方面, 如参与瘦素和胰岛素信号传导的衔接蛋白*Sh2b1*基因编码区发生的缺失与人类早发性肥胖相关^[8]; 非编码区的增强子重复或缺失会导致其靶基因(如*Myc*, *Ar*, *Sox9*或*Klf5*)的错误调节, 导致癌症和性发育障碍^[9-11]。然而, 近年来, 越来越多的研究

引用格式: 汪亚军, 王晓萍, 于黎. 基于基因组结构变异解析动物复杂性状的研究进展. 中国科学: 生命科学, 2024, 54: 283-299
Wang Y J, Wang X P, Yu L. Research progresses in revealing complex traits of animals based on genomic structural variations (in Chinese). Sci Sin Vitae, 2024, 54: 283-299, doi: 10.1360/SSV-2023-0216

报道基因组结构变异在物种适应性演化过程中扮演着重要的角色^[12,13]。从基因组结构变异的角度去解析物种的复杂性状有助于理解物种适应性演化的本质,而且也具有广泛的应用前景,如濒危物种的保护以及驯养动植物经济性状的改善等^[14]。

植物中的基因组结构变异与复杂性状关联研究起步较早,目前已有大量研究对基因组结构变异与植物驯化经济性状的关联报道,尤其是在作物育种方面^[15,16],而针对动物复杂性状与基因组结构变异的关联研究开始较晚^[17,18]。近年来,越来越多的研究报道,基因组结构变异与动物的毛色性状、极端环境的适应、驯养动物经济性状等复杂性状的产生紧密相关,基因组结构变异在物种的这些复杂性状形成过程中发挥着重要的作用。本研究将对基因组结构变异类型及其检测鉴定技术的研究历史,以及解析动物复杂性状的研究进展进行综述,为理解动物复杂性状形成的分子机制提供重要参考。

1 基因组结构变异概述

1.1 基因组结构变异

遗传变异是导致基因组差异的主要因素,变异类型中的基因组结构变异是区别于单碱基变异(SNP/ SNV)和小片段插入缺失(Indel≤50 bp)的特殊变异类型。基因组结构变异通常被定义为不同个体或物种间在基因组DNA序列的组成、位置和顺序的差异,差异碱基长度大于50 bp^[14,19]。相对于单碱基变异和小片段插入缺失,基因组结构变异在基因组上影响的碱基范围更广,对物种复杂性状带来的影响更大^[20-22]。

基因组结构变异的类型多种多样(图1),有简单的结构变异和复杂的结构变异。其中大部分简单的结构变异模式按照基因组DNA序列的稳定性可分为平衡的(DNA序列含量不变)和不平衡的(DNA序列含量发生增减)结构变异。前者包含基因片段的倒位(inversion, INV)、染色体内或者染色体间DNA片段的易位(translocation, TRA)等类型;后者包括插入(insertion, INS)、缺失(deletion, DEL)和重复(duplication, DUP)等类型,其中插入变异类型根据插入序列可分为一般核苷酸片段插入和转座元件的插入,重复变异类型根据重复片段(segmental duplications, SD)在基因组上的位置分布情况可分为串联重复(tandem duplications,

TD)和散在重复(dispersed duplications, DD)^[23,24]。基因组结构变异类型中的重复和缺失也被称为拷贝数变异(copy number variation, CNV)^[25]。CNV是由基因组发生重排而导致的,通常是指1 kb以上的基因组发生大范围拷贝数的增加或者减少,CNV也可以叫做大规模的结构变异,是基因组结构变异的重要组成部分之一^[26-28]。此外,结构变异中还包含少部分复杂变异,这类结构变异是指由多个序列位点断裂-融合形成新的链接(junction)而组合成复杂的序列,通常还会伴随着CNV的发生^[29]。

基因组结构变异的形成有多种机制,主要发生在细胞分裂的过程中^[31-33],但不分裂的细胞(如神经元细胞等)也可以在DNA修复时发生结构变异^[34]。最常见的SV形成机制包括以下五种:(i) 非等位基因同源重组(non-allelic homologous recombination, NAHR),这种类型是在有丝分裂或减数分裂期间高度相同的序列错位而发生的,指两个非等位基因的同源染色体片段发生交叉不对等互换,引起染色体序列片段的增加、减少以及顺序的改变。该机制的发生可能导致插入、缺失和倒位^[35];(ii) 非同源末端连接(non-homologous end joining, NHEJ),是DNA双链断裂修复时发生错误,经常参与DNA双链断裂修复的机制,在完成对DNA断裂位点的修复后会留下一个“痕迹”,该修复机制主要导致缺失^[35];(iii) 微同源介导末端链接(microhomology-mediated end joining, MMEJ),是DNA双链断裂修复中的备用机制,MMEJ利用DNA双链断裂两侧5~25 bp的区域进行DNA修复,通常发生在NHEJ和同源重组修复机制受到抑制时,该机制的发生与易位和插入相关^[32];(iv) 复制叉停与模板转换(fork stalling and template switching, FoSTeS),是DNA复制过程中的模板链转换导致的错误,是单链断裂的修复机制,即当DNA复制过程受阻时,滞后链与聚合酶发生解离并与基因组DNA另一序列区域相关联的过程,通常会产生大片段结构变异。该机制的发生可能与倒位、易位和重复相关^[32]。此外,FoSTeS的复制分叉断裂也可以引起微同源介导断裂诱导复制(microhomology-mediated break-induced replication, MMBIR)来修复,MMBIR通常会导致插入和重复^[36];(v) 移动元件(mobile genetic elements, MEs),又被称为转座子或转座元件,其中Alu,长散在核元件1(long interspersed nuclear elements, LINE1)和SVA等是在人类基因组中比较活跃的移动元

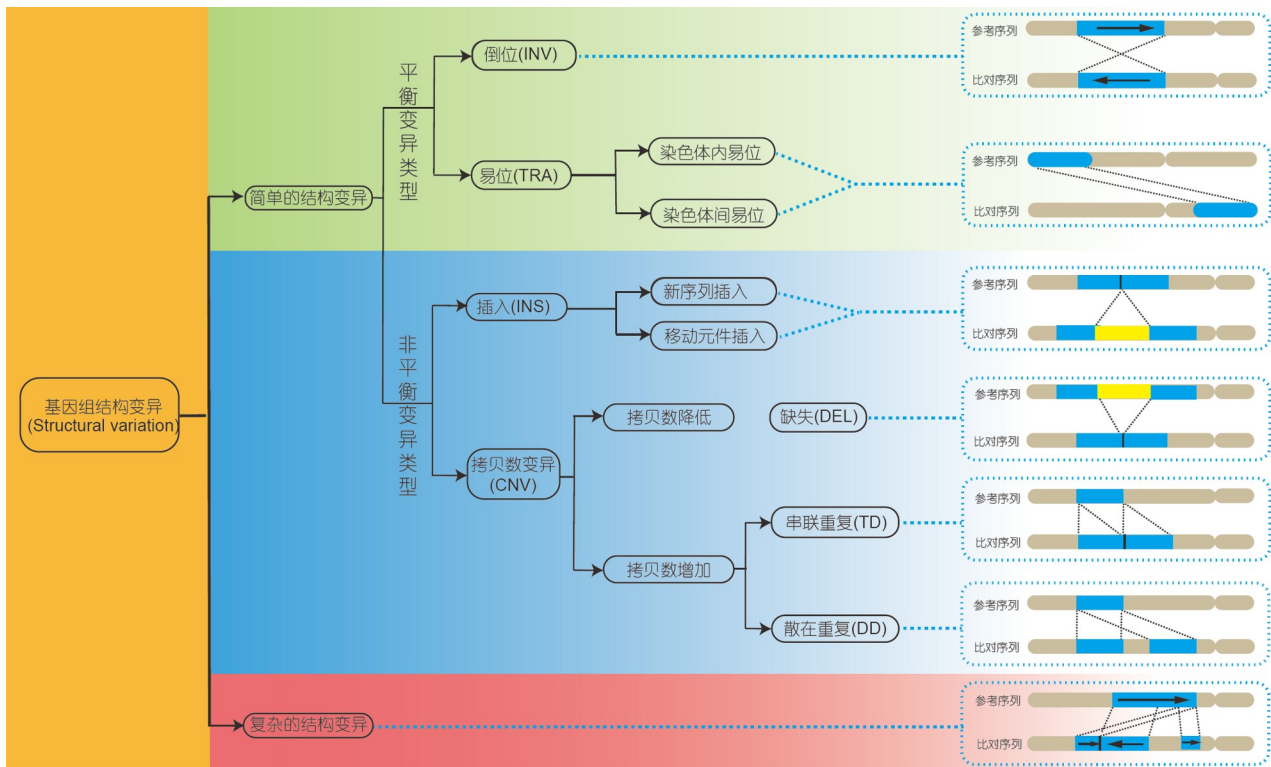


图 1 基因组结构变异的类型及分类(修改自文献[30](开放获取))

Figure 1 Types and Classifications of structural variations (Modified from ref.[30](Open Access))

件家族, 并参与结构变异的形成^[5,37]. MEs可以通过转座作用插入到基因组的其他位置^[38,39]. 该机制通常会与结构变异类型中插入和缺失相关^[32]. 在这四种结构变异发生机制当中, NAHR和MEs是SV最常见的发生原因^[39,40].

在基因组SV的平衡变异类型中, 倒位是指基因组DNA序列发生反向颠倒, 导致DNA序列的顺序也发生反向改变, 但倒位后的碱基数目并不改变. NAHR和FoSTeS是倒位形成的主要机制^[33]. 染色体倒位是与物种分化相关的最常见的结构变异^[41-44]. 易位是保持基因组DNA序列平衡的另外一种变异类型, 是指将基因组DNA序列从一个位置移除, 重新插入到另一位置. 根据该DNA序列在染色体中的相对位置, 可将易位分为染色体间易位(inter-chromosomal translocation)和染色体内部易位(intra-chromosomal translocation). 易位发生的机制较多, 主要有NAHR, NHEJ和FoSTeS等. 发生染色体易位的精子或卵子受精后, 胚胎的染色体出现增减的概率会大大提高, 从而影响物种的生育和流产^[45,46].

在基因组SV的不平衡变异类型中, 插入是指将一段DNA序列(如转座元件)插入到基因组序列中, 导致基因组DNA碱基数目增加. 转座元件是一类可移动的核苷酸重复序列, 其中非长末端重复序列(non long terminal repeat, non-LTR)逆转录转座子(retrotransposons)在人类基因组中是唯一具有自主转座能力的转座子^[47]. 插入发生的可能机制主要是NAHR, MMEJ, MMBIR和MEs^[32]. 缺失是最常见和最简单的一种基因组结构变异类型, 指染色体上的DNA序列片段的丢失导致基因组DNA碱基数目减少, 缺失的产生的机制主要有NAHR和NHEJ等^[32]. DNA序列不平衡变异类型除了缺失和插入, 重复也是一种较常见的基因组结构变异, 是指一段DNA序列在基因组中重复出现, 重复的产生机制主要包括NAHR, FoSTeS和MMBIR等^[32,35]. 其中, 基于串联重复形成的基因家族成员通常紧密排列在同一条染色体上, 形成序列相似、功能相近的基因簇, 而散在重复导致重复的基因距离较远, 甚至位于不同的染色体. 这些基因组结构变异被报道会导致动物复杂性状的改变, 包括毛色变化(如北美雪鞋兔的

白色毛色)、动物生理适应相关性状(如牦牛对高原的生理适应)以及驯养动物经济性性状(如牛肌肉的发育)等^[48-50]。

1.2 基因组结构变异的研究历史

基因组结构变异这一机制,很早就出现在科学家的视野,如在20世纪初,科学家就发现果蝇存在染色体倒位和基因序列的重复等现象^[51,52]。此外,也有研究通过核型分析发现染色体非整倍性引起的各种疾病,如克氏综合征(Klinefelter syndrome)^[53]和唐氏综合征(Down syndrome)^[54]等。因此,结构变异很早就被关注。但是,在20世纪70年代,分子生物学的兴起和快速发展,研究主要集中在单核苷酸多态性、扩增片段长度多态性(amplified fragment length polymorphism, AFLP)和微卫星序列(simple sequence repeat, SSR)等遗传标记方面,而结构变异的研究并没有被广泛普及,直到高通量测序技术的出现,对于基因组结构变异的检测和鉴定才慢慢再次被关注^[13,55]。

随着测序分析技术的不断改进和完善,结构变异检测技术方法也多种多样(表1)。早期对于结构变异的检测,最常用的是基于细胞核型和染色体倍性检测的传统分子技术。结构变异类型主要是基于细胞分裂中期,染色体长臂和短臂相对于着丝粒的长度变化来判断,该技术的局限性主要只适用于发生大范围DNA序列变化的结构变异(≥ 10 Mb)^[56]。直到21世纪初期,微阵列比较基因组杂交(array comparative genomic hybridization, aCGH)技术的发展,使得高分辨率CNV的发现成为了可能。aCGH是对两个不同染色体DNA与相同

靶标寡核苷酸杂交的强度进行比较分析的方法。相比于染色体核型分析,此方法相对提高了CNV检测的分辨率(分辨率 ≥ 2 Mb),并大量应用于临床,如产前医学、智力障碍、孤独症谱系障碍以及遗传咨询等领域^[57,58]。但是,aCGH无法检测到平衡的结构变异(倒位和易位)以及无法提供序列的具体拷贝数量。因此,在该技术的基础上,SNP芯片技术(SNP array)逐渐发展起来。SNP-array是指将荧光标记的DNA序列作为探针放置在玻片上,随后将待测序列与探针杂交,通过分析杂交信号的结果来检测CNV变异^[59]。aCGH和SNP芯片技术的共同点是都可以用于检测基因组拷贝数变异,但二者在检测CNV的分辨率高低方面存在差异^[60]。相较于aCGH方法,SNP芯片技术不仅提高了检测CNV的分辨率(3~5 kb),而且还能够检测出SNP。尽管如此,SNP芯片技术分辨率和可靠性还是比较低^[61]。

继array-CGH和SNP-array技术之后,随着高通量测序技术的迅猛发展,全基因组测序技术(whole genome sequencing, WGS)的出现使得全面探索和研究SV成为可能,其中全基因组测序数据主要包括二代测序(next-generation sequencing, NGS)和三代测序(third-generation sequencing, TGS)。NGS与之前的SV检测技术相比较,具有分辨率高、可灵敏的检测到一些平衡变异如倒位和易位以及一些新序列插入等基因组SV类型的优势^[62]。然而,由于二代测序的读长较短,而基因组上存在许多重复序列,导致短的读长不能完全跨过重复序列的区域,使得检测的SV出现了低质量比对和假阳性较高的问题^[25,63,64]。随着PacBio公司的单分子实时测序技术(single molecule real time, SMRT)和

表1 基因组结构变异检测技术的发展历史

Table 1 The development history of genomic structural variation detection technologies

检测技术	可检测SV类型	技术特点	数据输出类型
染色体核型显微观察	大片段序列增加或减少	有效检测染色体规模的SV, 分辨率很低(≥ 10 Mb), 无法检测基因组DNA序列的SV	染色体核型图
微阵列比较基因组杂交(aCGH)	拷贝数变异	相对较高分辨率检测CNV的可靠方法(分辨率 ≥ 2 Mb), 但无法检测平衡变异和更小的CNV变异	基因组核苷酸芯片
SNP芯片技术	拷贝数变异以及SNP	除检测CNV变异外(分辨率在3~5 kb), 还可以有效检测SNP变异, 但可靠性较低	基因组核苷酸芯片
二代测序技术(NGS)	插入、缺失、重复(拷贝数变异)、倒位以及易位	分辨率较高, 高通量, 成本较低, 但测序读长比较短	二代短读长测序数据
三代测序技术(TGS)	插入、缺失、重复(拷贝数变异)、倒位以及易位	超长读长(10 kb)、避免PCR扩增引入错误、无GC偏好性且变异检测灵敏度和准确度高, 但单读长的错误率偏高, 需二代测序数据进行纠错, 成本较高	三代长读长测序数据

牛津纳米孔公司(oxford nanopore technologies, ONT)的纳米孔单分子测序技术(nanopore sequencing)为代表的第三代测序技术(TGS)应运而生, 基于其单分子测序且测序长读长(>10 kb)、以及不需要PCR扩增、无GC偏好性等优势, TGS能够检测出长片段序列及复杂序列的结构变异, 如插入、缺失、倒位和易位等, 弥补了NGS对大片段结构变异检测的不足^[23,65]. 有研究显示, 在50~1000 bp范围内, 利用三代测序检测的结构变异数目要比二代测序多3~4倍^[66]. 三代测序技术的长读长优势已被证明可以有效地检测和解析二代短读长不能检测的结构变异, 尤其是在检测基因组重复序列区域的SV类型方面. 如在一项分析乳腺癌基因组结构变异的研究中, 研究者们利用三代长读长测序平台ONT和SMRT对乳腺癌患者的组织样本进行测序分析发现, 使用三代长读长测序方法在患者基因组中发现了*Egfr*基因的3'UTR处存在大约250 bp的插入, 进一步分析发现该区域靠近7号染色体的着丝粒位置, 此位置含有许多TA重复序列, 而使用二代短读长测序技术无法检测到这个区域的结构变异类型^[67]. 此外, 已有研究通过三代测序技术分析了藏族和汉族人群SV的变异图谱, 揭示了SVs在藏族人适应高原环境中发挥的重要作用, 通过比较发现, 三代长读长技术检测的大约有一半的结构变异数量是用NGS数据检测不到的, 进一步分析结构变异位点周围的序列结构发现, 这一半用二代短读长测序技术未检测到的SV类型大多位于基因组重复区域^[68].

测序技术的发展极大的促进了基因组结构变异的检测效率^[25]. 测序技术的不同也意味着结构变异检测原理和方法的多样化. 目前经典的基于NGS数据的结构变异检测原理主要有三种: (i) 读对方法(read-pair, RP), 即根据双末端测序(paired-endsequencing, PE)获得两端成对reads的距离和方向等信息, 然后将测序的reads与参考基因组进行比对, 分析两个reads在参考基因组上的距离和方向信息, 寻找比对前后距离和方向与参考基因组不一致的reads, 以此特征为原则去判断SVs的类型, 可以判断出缺失、插入和倒位等变异类型^[69]. 该方法的局限性在于无法检测大片段以及重复区域的结构变异, 同时对小于200 bp的DEL检测的精确度低; (ii) 读段深度(read-depth, RD), 是将高通量测序获得的样本基因组的reads比对到参考基因组上, 然后分析测序深度信号, 通过测序深度在基因组某些

区域的差异变化来确定结构变异的类型, 该方法可以确定重复和缺失变异类型. 读段深度的方法在检测结构变异类型中的重复和缺失时的效果非常显著, 即发生重复变异区域的测序深度明显增加, 而缺失变异区域的测序深度则明显降低. 但是, 该方法无法检测其他类型的结构变异和无法区分串联重复和散在重复, 并且只能判断片段中是否存在结构变异, 而不能判断出结构变异的准确位置; (iii) 分裂片段方法(split-read, SR), 是在未比对到参考基因组的reads上设置一个断点位置信息, 然后从该断点处将reads分裂成两个小片段, 再将小片段比对到参考基因组上, 通过观察两个小片段比对到参考基因组中的情况, 从而判断结构变异的类型. 这种方法的优势在于能够精确到插入和缺失变异的单个碱基的断裂位点, 但是也和大多数的RP方法一样, 无法解决大片段和复杂的SVs. 对于SR来说, reads越长, 检测效果越好. 当前对于二代数据的常用比对软件有BWA(<https://github.com/lh3/bwa>)和Bowtie2(<https://github.com/BenLangmead/bowtie2>)等, 而结构变异检测软件大部分是基于上述三种原理中的一种或几种联合开发的, 如DELLY^[70]软件是基于RP和SR的方法开发, 该软件提高了断点预测的准确性, 并允许检测较小的缺失, 而LUMPY^[71]软件则同时使用RP, SR和RD进行联合分析等^[72,73]; 基于TGS数据检测结构变异的原理和二代数据是一致的, 即将测序数据比对到已经组装完成的物种参考基因组, 主要的三代数据比对软件有Minimap2(<https://github.com/lh3/mini-map2>), PBmm2(<https://github.com/PacificBiosciences/pbmm2>)和NGMLR(<https://github.com/philres/ngmlr>)等, 然后再通过变异检测软件去检测出相关的结构变异类型. 基于TGS数据的结构变异检测软件也非常成熟, 如PBSV(<https://github.com/PacificBiosciences/pbsv>)和Sniffles(<https://github.com/fritzsedlazeck/Sniffles>)等. 三代测序的长读长有利于识别跨越高度重复区域或其他复杂区域的结构变异, 但也存在检测SV结果的假阳性较高以及研究成本增加的缺点; 另外, 基于二代和三代测序数据的基因组从头组装方法(*de novo assembly*)也流行了起来, 该方法是通过将不同个体的reads从头组装成基因组, 然后与参考基因组进行比对, 从而清楚地判断从头组装基因组的结构变异类型. 相较于前两种方法, 基因组从头组装策略采取了截然不同的非reads比对方法, 直接克服了reads的长度问题, 该方法

是最经典, 最直接有效鉴定所有结构变异类型的办法。但是从头组装方法针对高度重复序列区域的拼装依赖于高质量的较长的测序片段和精密的组装算法, 组装高质量物种基因组需要的成本将更大。基于从头组装的方法是目前近缘物种间结构变异检测最流行的方法, 主要用到的软件有Assemblytics^[74], SMARTie-SV (<https://github.com/zeceev/smartie-sv>)和MUM&Co^[75]等(表2)。

2 基因组结构变异与动物复杂性状的关联研究

2.1 基因组结构变异与动物毛色性状的研究

动物毛色在自然界中的表型多种多样, 通常与动物的适应性捕食、防御和伪装相关, 其分子机制备受关注^[82]。目前, 关于动物毛色性状的遗传学研究主要集中在以SNP作为遗传标记来展开, 而基于基因组结构变异的毛色多态性分子机制研究相对较少^[83]。已有的研究通过比较基因组学分析发现特定环境下的物种类群相比于其他群体发生了特异性的基因组结构变异, 进而影响与毛色基因功能或调控相关的蛋白结构

或基因表达, 从而产生毛色的适应性改变。比如, 通过对北美雪鞋兔白色种群和棕色种群的群体基因组学研究发现, 该物种不同毛发颜色的转换与Agouti信号蛋白基因(*Asip*)的顺式调控元件相关。在白色雪鞋兔(*Lepus americanus*)基因组中, *Asip*基因的5'UTR区存在大约1000 bp左右的碱基插入, 使得雪鞋兔在冬季其毛色由棕色转变为白色, 从而有利于适应冬季环境以达到躲避天敌的伪装目的^[50,84]。这个基因在其他物种中也存在相似的调控效果。白水牛(*Bubalus bubalis*)的*Asip*基因上游存在2809 bp的LINE-1插入, 使得该基因出现新的强启动子功能, 导致*Asip*基因的表达量显著提高, 而*Asip*基因高表达可以抑制黑色素细胞的发育, 从而产生白色表型^[85]。此外, 貂(*Nyctereutes procyonoides*)的被毛颜色也有黄色、棕色等, 有研究将黄色貂与野生貂(棕色或黑色)的基因组进行比较分析, 发现促黑色素细胞激素受体基因*Mclr*起始密码子上游5'UTR中有152 bp的缺失, 与野生型貂相比, 黄色貂的*Mclr*蛋白水平明显降低, 表明褐黑素(一种红色或黄色色素)合成增加, 从而导致黄色貂表型的产生^[86]。另外, 缺失变异与鸡的羽毛呈现深棕色密切相关, 在*Sox10*基因的上游14 kb处存在一个约8.3 kb的较大缺失, 研究进一步证

表 2 基因组结构变异检测方法

Table 2 Methods of detecting genomic structural variations

方法	原理	方法优缺点	可检测SV类型	常用序列比对软件	常用结构变异检测软件
基于NGS数据的方法	读对方法(RP): 根据双端reads与参考基因组比对后的距离和方向等信息, 寻找比对前后不一致的reads, 以此特征为原则去判断SVs的类型	此方法检测SV类型比较全面, 但无法检测大片段和复杂的结构变异, 以及难以精确检测较小缺失(<200 bp)	缺失、插入、倒位、易位、重复	BWA ^[76] Bowtie2 ^[77]	Delly ^[70] Lumpy ^[71]
	读段深度(RD): 将样本基因组的reads 比对到参考基因组上, 通过测序深度在基因组某些区域的差异变化来确定SVs的类型	检测重复和缺失效果显著, 但无法确定其他结构变异类型, 并且较小变异难以检测	缺失、重复	BWA Bowtie2	Lumpy
	分裂片段方法(SR): 在样本基因组reads上设置一个断点位置信息, 然后从该断点处将reads分裂成两个小片段, 再将小片段比对到参考基因组上, 通过观察两个小片段比对到参考基因组中的位置及顺序等信息判断SVs的类型	能够精确检测SV断点, 但无法解决大片段和复杂变异, 并且测序reads越长, 检测效果越好	缺失、插入、倒位、重复	BWA, Bowtie2	Delly Lumpy SVseq2 ^[78]
基于TGS数据的方法	将长读长测序数据比对到参考基因组上, 通过检测软件鉴定SVs的类型	可测类型多, 可检测大片段和复杂结构变异, 但相较于短reads, 长reads的精确度较低, 并且成本较高	缺失、插入、倒位、易位、重复	Minimap2 ^[79] PBmm2 ^[80] NGMLR	PBSV Sniffles ^[72]
基因组从头组装的方法	基于二代数据和三代数据组装物种染色体水平的基因组, 通过基因组间比对后识别SVs的类型	结构变异检测类型多, 可检测大片段和复杂结构变异, 但依赖于基因组组装质量, 并且成本较高	缺失、插入、倒位、易位、重复	Mummer ^[81] Minimap2	Assemblytics ^[74] SMARTie-sv MUM&Co ^[75]

实该缺失位点影响*Sox10*的表达,导致鸡羽毛出现深棕色^[87]。牛(*Bos taurus*)的侧腹颜色是一种引人注目的毛色表型(有黑色、蓝色或白色斑点的侧腹颜色),研究发现其侧腹皮毛颜色的多样性与基因组6号染色体片段上的*Kit*基因的492 kb序列易位到29号染色体相关^[88]。此外,也有研究发现染色体上发生的倒位与白喉带鹀(*Zonotrichia albicollis*)的眉羽颜色有关,即白喉带鹀基因组的2号染色体上存在大范围的倒位(超过100 Mb),该倒位的发生涉及到的*Lyst*基因可能与白喉带鹀的白色和棕褐色两种眉羽颜色相关^[89]。

2.2 基因组结构变异与动物极端环境适应性状的研究

极端环境(extreme conditions)是指自然环境中普通生物无法生存的特殊环境,比如低渗透压、低温低氧、剧毒以及炎热干旱等极端环境^[90]。动物在极端环境下的适应性演化分子机制一直是研究的热点问题。基于基因组结构变异解析动物极端环境适应的分子机制的研究层出不穷。最典型的例子是鱼类适应淡水低渗环境的分子机制研究。有研究以刀鲚(*Coilia nasus*)的洄游种群及其淡水定居种群为研究对象,发现在刀鲚淡水定居种群的染色体LG6和LG22上存在两个较大的染色体倒位,两条染色体倒位区域超过50 Mb,而在洄游种群中倒位发生频率较低或者没有发生,两个染色体倒位区域富集了与代谢过程和渗透压调节等多个生物过程相关的约1800个基因,提示该染色体倒位可能与刀鲚两种生态位种群间在适应性演化过程中形态上的分化,如椎骨的数量、眼睛的大小、肝脏的形状和大小等,和生理的区别,如渗透调节,以及行为上的差异比如运动量等有关,提示基因组结构变异对刀鲚淡水种群在适应淡水的过程中起到重要作用^[91]。

高海拔(低氧、低温、强紫外线辐射等)环境的适应性演化研究也是极端环境适应的重要模型。牦牛(*Bos mutus*)作为青藏高原的动物代表,它的栖息地具有昼夜温差大,氧气稀薄以及营养条件缺乏等特征,通过长期的自然选择,牦牛演化出多种适应性性状,如毛发变厚、消化功能提高以及肺孔洞变多,来应对高原环境。Gao等人^[48]对生活在高原环境的野牦牛和家牦牛种群与低海拔牛种群进行比较基因组学分析,发现高原牦牛的基因组中有6733个基因中存在显著的基因组结构变异,其中127个基因在普通牛和牦牛的肺中差

异表达,如*Arnt*, *Hoxb5*和*Klf5*等。通过对受到选择的结构变异和差异表达基因等数据进行整合分析发现高原牦牛肺脏中内皮细胞群表达含结构变异的差异基因最多,表明结构变异可能与牦牛低氧适应有关,作者进一步通过对牦牛肺组织进行染色观察发现,牦牛肺组织中血管中膜相较于普通牛是比较厚的,并且还存在着较多的弹性纤维,这些性状的改变可以增强牦牛肺的伸缩能力,有利于牦牛对高原的低氧环境的适应。除了高原极端环境以外,也有研究表明内洛尔牛(*Bos indicus*)适应闷热潮湿环境的机制与结构变异相关。在闷热潮湿的环境中,内洛尔牛的头、颈、驼峰、膝盖等部位一般都表现出深色(darkness of hair coat, DHC)。研究对内洛尔牛群体基因组数据和澳大利亚婆罗门母牛(与内洛尔牛有共同祖先)群体基因组分析,鉴定出与DHC有关的许多结构变异,其中位于13号染色体的*Asip*基因附近存在1155 bp的缺失,该缺失可能影响了*Asip*非编码外显子的招募,从而影响基因的表达^[92],特别是内洛尔牛的头、颈、驼峰和膝盖部位的表达。在不同的环境中,一些植物可能会存在着毒性,从而对植食性动物造成一定的威胁,因此有报道在牛(*Bos taurus*)基因组中发现*Tas2r46*基因中存在约17 kb的缺失,该基因与味觉适应相关,该基因的突变可能有助于牛识别环境中的有毒植物^[93,94]。

人类工业的发展伴随着对动物生活环境的污染。桦尺蠖(*Biston betularia*)的工业黑化体现了这个物种适应工业环境带来的负面影响(工厂煤烟使得环境变为黑褐色),研究发现桦尺蠖表型的黑化与转座元件的插入有关,即转座元件插入到桦尺蠖*Cortex*基因(调控多种鳞翅目体色模式)的内含子上,导致*Cortex*基因发生高表达,调节了细胞周期,进而引起桦尺蠖黑化性状的出现,适应环境而不被天敌捕捉^[95]。另外,在高毒环境下物种也会产生一些适应性改变,如在美国大西洋海岸,生存着有被重工业混合物严重污染的鳉鱼(*Fundulus heteroclitus*)。研究通过对污染区域耐受群体和敏感群体进行比较分析,发现耐受鳉鱼基因组中的芳香烃受体AHR的多种同源物(如*Ahr1a*和*Ahr2a*)发生不同程度的拷贝数缺失变异,使鳉鱼耐受群体能够适应污染区域^[96]。此外,也有研究表明染色体倒位在物种适应极端环境中扮演着重要角色,如生活在空气干燥温度偏高的非洲地区的冈比亚按蚊(*Anopheles gambiae*)。研究发现冈比亚按蚊基因组的2号染色体上

发生2La(22 Mb)和2Rb(7 Mb)两个片段的倒位, 倒位片段存在着约149个有相关功能报道的基因, 通过进一步分析发现该倒位变异涉及的功能基因可能与冈比亚按蚊适应干燥炎热环境相关^[97]。

2.3 基因组结构变异与驯养动物复杂经济性状的研究

动物的驯化是指在人工选择条件下为了提高生活质量而筛选出与人类互利互惠的复杂经济性状。驯养动物性状产生的选择压力主要与动物适应农场环境和提高生产力有关, 如动物的性情、生长状态和繁殖特征等性状的改变^[98], 这种性状的产生往往与基因组结构变异密切相关, 在家犬驯化适应的研究中^[99], 对家犬(*Canis lupus familiaris*)、灰狼(*Canis lupus*)和豺狗(*Cuon alpinus*)三个犬科物种的基因组进行比较分析发现, 在家犬基因组的14号染色体上存在一个1355 bp的逆转录转座子插入, 该插入片段包含一个新基因(ENSCAFG00000002440), 该基因是*Akr1b1*基因的新拷贝, 在小肠和肝脏中高表达, 参与脂肪的从头合成和降低高淀粉饮食带来的毒性等功能, 有利于家犬适应高淀粉饮食。此外, 有研究解析了牛(*Bos taurus*)基因组中的*App12*基因的第二个内含子上发生了168 bp的Bov-tA2转座子插入, 认为这种插入可能与牛的复杂经济性状相关, 如肌肉增多以及免疫力增强等^[100,101]。同时, 在野猪(*Sus scrofa*)的驯化过程中, 通过联合二代和三代测序数据分析, 对中国本土猪基因组和欧洲本土猪基因组比较分析发现了大量的基因组结构变异, 并在猪基因组中*Myl4*基因的内含子上检测出281 bp的缺失变异, 研究发现该缺失可能促进了皮下脂肪沉积从而提高猪肉产量^[49]。肌肉雄激素的主要作用靶点是脂肪组织, 在另一研究中发现猪基因组上的*Cyp3a29*基因的3'-UTR区存在一个长241 bp的SINE元件, 该元件的插入或者缺失可能调节*Cyp3a29*基因的表达, 进而调控猪的肌肉雄激素水平, 提高猪肉的质量^[102]。此外, 也有研究发现一些高频SVs可能与梅山猪高的繁殖性能相关, 如在与精子发生和生育相关的*Tex11*基因(睾丸表达基因家族TEX的家族成员)中发现了一个内含子缺失, 这个缺失影响该基因的表达^[103]; 对绵羊(*Ovis aries*)驯化遗传机制的研究中发现位于绵羊的26号染色体上*Bag4*基因的第二个内含子发生了1600 bp的拷贝数缺失, 与绵羊的肩高、体长等经济性状有关^[104]。

此外, 最近的一项研究显示插入变异与绵羊的长尾性状相关, 即在长尾绵羊*Hoxb13*基因的5'UTR区有一个168 bp的插入, 而其他短尾绵羊中没有, 该研究还报道了9个结构变异(6个插入和3个缺失)附近的*Bmp2*和*Pdgfd*基因可能与绵羊肥尾性状相关^[105](图2)。

2.4 基因组结构变异的作用模式

从以上研究可以看出, 基因组结构变异不仅可以影响基因的编码区(如外显子), 还可以影响基因内调控区(非外显子区域)以及基因间区(图3)^[106]。结构变异一方面通过改变基因的外显子, 进而破坏原有基因的蛋白结构从而产生新的功能, 另一方面通过改变顺式调控区的序列, 导致结合不同的转录因子, 从而调控靶基因的表达。此外, 对于大片段结构变异的倒位或易位则可以直接通过改变基因的位置来影响基因功能, 形成新的性状。

3 小结与展望

随着高通量测序以及三代测序技术的不断进步和升级, 比较基因组学的研究如火如荼, 其中基因组结构变异是物种基因组进化的潜在动力, 基因组结构变异分析已经是比较基因组学的一个重要内容之一, 为生物复杂性状的系统解析提供了一个新的角度。

对结构变异的全面精准检测是基因组SV研究的核心环节。虽然测序技术发展迅速, 参考基因组质量越来越高, 对于结构变异的检测与之前相比也更全面, 但由于SV高发区域序列重复度高、存在大量未知复杂结构变异类型, 因此, 并非所有类型的结构变异都能完全检测到。例如有小部分结构变异是由多种插入、散在重复以及多等位基因CNV等组成的复杂结构重排, 这种复杂变异类型是一般检测软件算法所不能完全检测到的^[107]。因此, 利用超长片段测序技术和光学图谱定位技术组装高质量基因组来检测复杂或大片段结构变异显得尤为重要, 如单细胞模板链测序(single-cell template strand sequencing, Strand-seq)方法能分别对单细胞的双亲DNA模板链进行测序, 获得高分辨率的姐妹染色体交换图谱, 从而检测到更复杂的基因重排^[108,109]。此外, BAC-long(>150 kb)技术可以让测序长度更长, 更容易跨过端粒、着丝粒等含有复杂重复序列的区域, 在基因组组装中的表现更优越^[110]。因此,

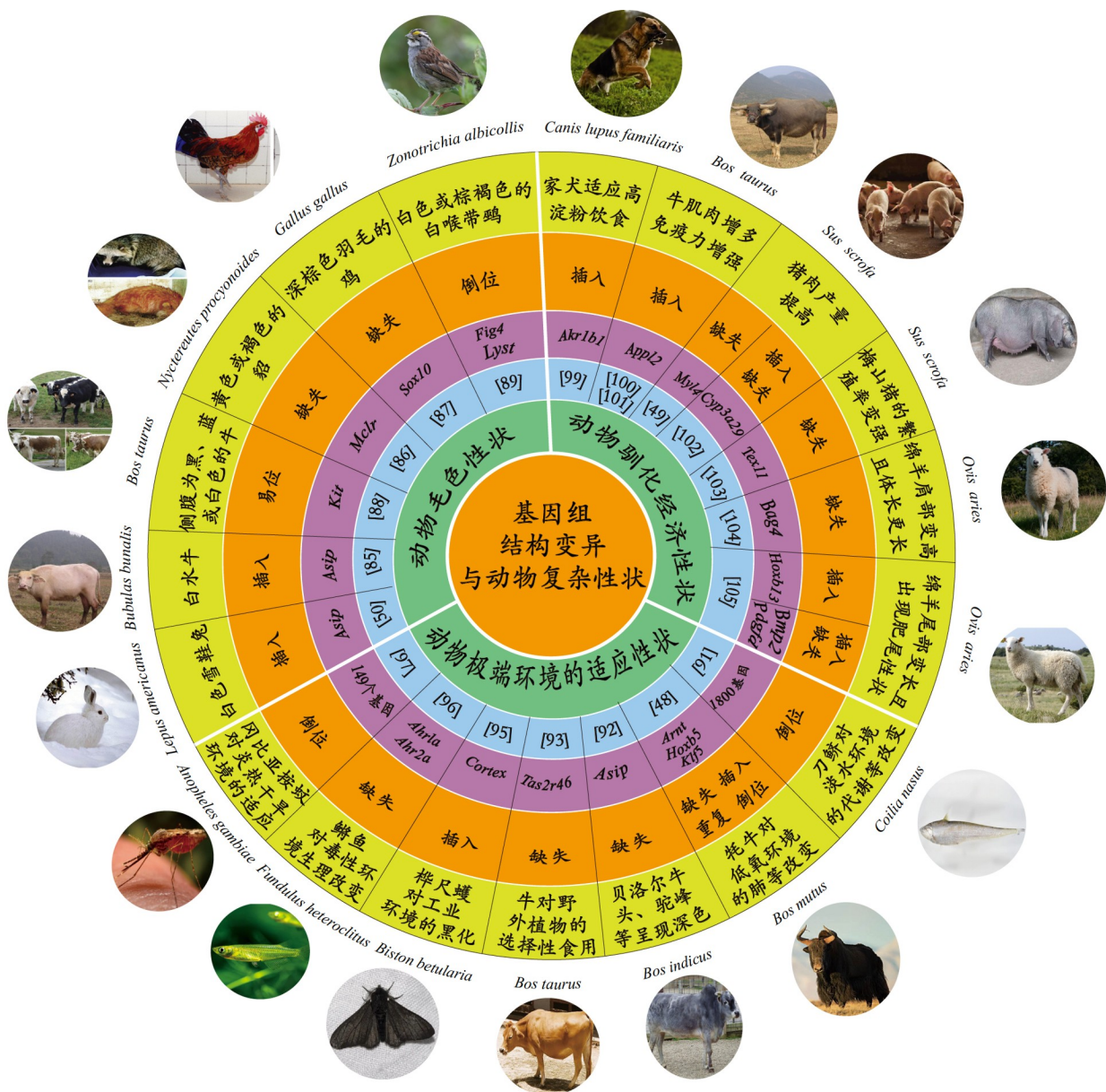


图 2 基因组结构变异和动物复杂性状总结

Figure 2 Summary of genome structural variations and animal complex traits

将Strand-seq, BAC-long等技术与三代测序技术结合起来则更有可能获得复杂的重复序列区域信息. 同时, 单分子光学图谱技术(optical mapping)可以对Nanopore或PacBio测序拼接的Contig序列进行排序和定位, 并且在基因组组装中还可以帮助识别并纠正错误, 可以大大提高基因组的组装质量^{[111], [112]}. 另外, 在基因组结构变异分析中, 针对传统建模策略的算法中存在许多错检、漏检等缺点. 近年来已有研究针对基因组重复

区域背景噪声高、复杂结构变异类型未知且建模难等问题, 通过将基因组结构变异检测从二维序列转换为三维空间图像, 实现了复杂类型结构变异检测和准确表征^[113]. 然而, 三维空间图像层面也面临着技术革新与流程复杂的问题, 将来能否广泛应用于基因组结构变异的检测还有待进一步考量^[114]. 此外, 基因组结构变异的全面检测与基因组组装的完整度高低有关. 而近年来, 自人类T2T(telomere-to-telomere)基因组公布

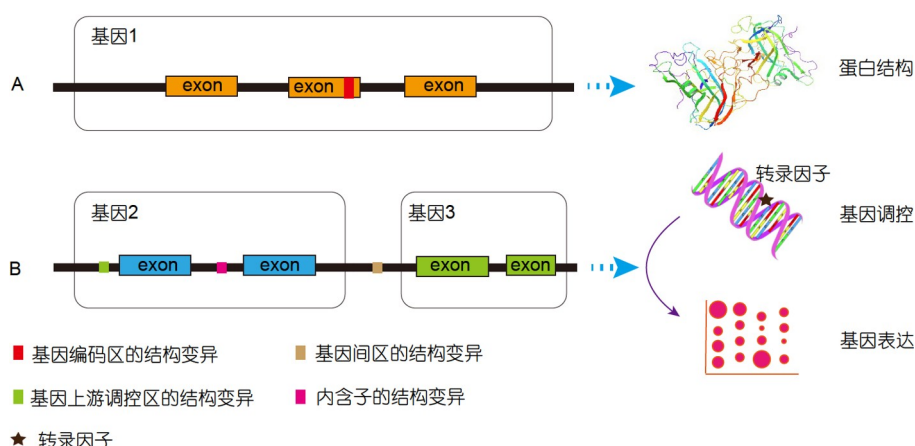


图 3 基因组结构变异的作用模式。A: 结构变异位于基因1的外显子上, 直接影响该基因的蛋白结构而产生新的功能; B: 结构变异位于基因内上游调控区或内含子或基因间区, 通过结合转录因子而调控基因的表达

Figure 3 The action mode of structural variations. A: The structural variations are located on the exon of gene1 and directly affect the protein structure of the genes to create new functions; B: The structural variations are located in the upstream regulatory regions or introns or intergenic regions within the genes and regulate gene expression by binding to transcription factors

后, 掀起了T2T组装的热潮。T2T基因组是指基于Pac-Bio HiFi, ONT Ultra-long和Hi-C等多种测序技术相结合, 能够实现一条或多条染色体端粒到端粒水平组装的零gap的基因组, 对基因组中高重复、高杂合的复杂区域实现全覆盖, 填补了参考基因组中最后一块“空白”区域^[24,115], 如最新公布的人类T2T基因组(T2T-CHM13)与基于三代测序技术获得的基因组GRCh38.p14(GRCh38)相比较发现^[116], T2T-CHM13揭示了以前没有被发现的大约8%的生物学重要区域, 比如新发现的*Klrc2*基因的重复和缺失可能与人类脑、免疫等疾病息息相关^[117]。目前, 随着越来越多T2T动物基因组被公开, 基因组质量不断提高, 基于比较基因组学分析检测和鉴定SV的流程也亟待更新。

目前参考基因组的组装大部分都是依赖一个物种的单个个体或者几个个体, 尽管测序技术大大提高基因组组装质量, 但是得到的参考基因组始终是比较单一的, 基因组还是不能完全代表着一个物种所有个体的遗传信息, 导致在检测结构变异的过程中就会遗漏一些可能有着重要表型调控的变异类型。因此, 在此问题基础上新的策略应运而生, 泛基因组(pan-genome)是指一个包含物种全部遗传信息的集合, 能够解决单一参考基因组对物种基因组代表性不足的问题, 有助于全面鉴定并解析物种基因组结构变异^[118,119]。Jiang等人^[103]构建的11个品种猪的研究中也体现泛基因组的优势, 利用构建的泛基因组揭示了206 Mb的新

序列, 并检测到183352个结构变异, 其中有63%是之前研究中未报道的新变异。参考基因组的组装质量直接决定基因组SV类型的检测效率, 不可避免的是结果也会受参考基因组质量情况的影响。因此使用无参考基因组的比对方式去解析结构变异有着天然的优势。无参比对是基于两两物种间的序列比对方式, 而不仅仅局限于比对到参考基因组上, 例如无参比对软件(Cactus, <https://github.com/ComparativeGenomicsToolkit/cactus>)解决了现有基因组间比对软件的弊端(即依赖参考基因组), 并且多基因组同时进行相互比对, 可以极大地提高了跨物种的比对效率。有研究团队运用这套方法完整描绘出鸟类基因组系统演化图谱, 为后续无参比对的研究提供参考^[120,121]。但目前还没有基于无参比对的方式去进行结构变异的研究, 尽管如此, 无参比对模式仍然有着巨大的研究前景。

目前基因组结构变异的检测研究大多集中在种内不同个体或群体间的研究, 并且大部分基于组装基因组间结构变异检测的软件也都是针对近缘物种, SVs的检测方法也多与物种亲缘关系远近相关^[122]。近年来, 远缘物种间的结构变异与性状的关联分析也开始被关注和受到重视。比如有研究通过对人属(*Homo*)中的智人(*Homo sapiens*)、猩猩属(*Pongo*)的苏门答腊猩猩(*Pongo abelii*)以及黑猩猩(*Pan*)属的黑猩猩(*Pan troglodytes*)的比较基因组分析发现17789个智人的特异性结构变异, 还发现了智人特异性结构变异所涉及的

Glg1, *St3gal2*等基因是人与黑猩猩之间大脑表达差异相关^[123]。另外,最近的一项研究报道了灵长目不同科之间的结构变异检测。Zhou等人^[124]对长臂猿科(Hylobatidae)中的白颊长臂猿(*Nomascus leucogenys*)和人科(Hominidae)中的智人(*Homo sapiens*)、黑猩猩(*Pan troglodytes*)、大猩猩(*Gorilla gorilla*)和红毛猩猩(*Pan paniscus*)以及猴科(Cercopithecidae)中的猕猴(*Macaca mulatta*)进行全基因组比较分析,获得15885个人科物种特异的结构变异(great-ape-specific structural variants, GSSVs),进一步分析找到4个发生在编码区域的结构变异所对应的基因可能与人科物种独特表型相关,包含与骨骼发育相关的*Acan*基因12号外显子上的60 bp的缺失可能与人科物种体型变大相关以及位于*Cmya5*基因外显子的264 bp的插入可能在人科物种的大脑进化中起作用。

通过比较基因组学的优化检测,越来越多的研究报告,基因组结构变异不仅发生在编码蛋白区域,而且还发生在与非编码调控元件区域或基因组结构染色质特征等相关的区域。因此,结构变异不仅可以影响基因功能,还可以通过改变染色体的高级结构来影响基因组的三维空间构象,从而影响离变异区域很远的功能基因表达^[125]。为了解析基因组结构变异是否调控下游复杂性状相关基因表达的分子机制,已有研究发现基因组结构的稳定性(染色体的三维空间结构),即拓扑结构域(topologically associated domain, TAD)与调控基因表达过程密切相关^[126]。比如,软骨鱼类的双鳍鱼小鲷鱼(*Leucoraja erinacea*)具有增大的翼状胸鳍,研究人员通过结合RNA测序、高通量染色质开放区域测序(assay for transposase-accessible chromatin with high throughput sequencing, ATAC-seq)技术以及染色质三维构象数据发现小鲷鱼特有的基因组重排改变了参与平面细胞极性(planar Cell Polarity, PCP)途径基因的三维调控景观(TAD),对PCP信号的功能性抑制导致了鲷鱼前鳍尺寸的变小,同时研究还发现一个增强子能够与几个*Hoxa*基因互作,这可能与小鲷鱼胸鳍的发育相关。该研究显示了PCP途径和*Hox*基因调控是鲷鱼鳍形态发育的关键贡献者,同时体现了染色体三维空间结构对基因表达调控的作用^[127]。近年来,随着新兴前沿技术的开发和应用,通过联合高通量染色体构象捕获(high-throughput chromosome conformation capture, Hi-C)技术和ATAC-seq测序技术,将有助于解析

存在SV的基因组是如何改变染色体的三维空间结构TAD内部基因互作强度,破坏TAD边界位点,打破序列对应的TAD稳定性,形成新的TAD,导致新的TAD内部调控元件-复杂性状或表型相关基因的异常互作,从而产生新的性状或表型^[128~132]。因此,从基因组结构变异角度研究动物复杂适应性性状不能仅仅停留在从编码基因去解析的范式,更需要关注非编码调控区域的影响,尤其是导致染色质三维空间构象改变的结构变异类型。

此外,目前大部分研究对于动物表型性状与基因组结构变异的关联研究仅仅展示了与适应性状相关的候选基因或调控元件,而对于该候选基因或调控该基因所对应的某种基因组结构变异类型是否与该性状相关,是否会导致表型改变,仍然缺乏功能实验验证。已报道的研究中,只有非常少量的研究对动物表型是否与SV位点相关进行过初步的功能实验验证,如白水牛*Asip*基因上游存在2809 bp碱基的插入,该研究通过实时荧光定量PCR(quantitative real-time PCR, QRT-PCR)以及转录组表达数据对该结构变异进行验证,结果表明,该插入片段会导致*Asip*基因表达上调,从而影响到水牛的白色表型^[85]。因此,今后的研究有必要从体外细胞水平和体内个体水平对基因组结构变异进行更全面的功能验证,例如,通过体外的QRT-PCR,双荧光素酶报告(dual-luciferase reporter)实验及慢病毒转染等细胞水平的实验验证,检测结构变异所涉及的功能基因或者调控元件活性是否发生改变,然后再结合前沿的CRISPR/Cas9基因编辑技术进行个体水平的功能验证,从而证实基因组结构变异与动物复杂性状的关联^[133]。

综上,本文从多方面总结展望了基因组结构变异的研究瓶颈和未来解决的方向。第一,随着测序技术的不断升级,越来越多的高质量动物基因组数据被更新和公布,检测SV的方法因基因组数据类型不同(如泛基因组、T2T基因组等)而不同;第二,针对SV检测和鉴定方法,未来对于近缘或远缘物种间的结构变异的检测和鉴定,尤其是远缘物种间的结构变异的检测和鉴定方法急需开发和优化完善;第三,为了检测SV的作用模式及功能,结合多组学技术检测SV如何直接影响或间接调控基因的功能,如通过改变染色体的高级结构来影响基因组的三维空间构象,进而影响新的表型性状的产生,将有助于解析SV影响新性状产

生的作用模式; 第四, 功能实验验证是分子机制研究中的重要一环, 对基因组结构变异的候选基因或候选调控元件做体内体外功能实验验证, 以进一步验证结果. 总之, 随着测序技术不断成熟和发展, 如转录组学、表

观组学和空间组学等, 以及大量非模式物种的高质量基因组被测序报道, 将为全面研究基因组结构变异提供了前所未有的机会, 为最终系统、全面解析动物复杂性状产生的分子机制提供新视角和新线索.

参考文献

- 1 Hu Y, Wang X, Xu Y, et al. Molecular mechanisms of adaptive evolution in wild animals and plants. *Sci China Life Sci*, 2023, 66: 453–495
- 2 Goddard M E, Hayes B J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet*, 2009, 10: 381–391
- 3 Stranger B E, Stahl E A, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 2011, 187: 367–383
- 4 Chiang C, Scott A J, Davis J R, et al. The impact of structural variation on human gene expression. *Nat Genet*, 2017, 49: 692–699
- 5 Sudmant P H, Rausch T, Gardner E J, et al. An integrated map of structural variation in 2504 human genomes. *Nature*, 2015, 526: 75–81
- 6 Wang X, Xu J, Zhang B, et al. Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nat Methods*, 2021, 18: 661–668
- 7 Zichner T, Garfield D A, Rausch T, et al. Impact of genomic structural variation in *Drosophila melanogaster* based on population-scale sequencing. *Genome Res*, 2013, 23: 568–579
- 8 Bochukova E G, Huang N, Keogh J, et al. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, 2010, 463: 666–670
- 9 Croft B, Ohnesorg T, Hewitt J, et al. Human sex reversal is caused by duplication or deletion of core enhancers upstream of SOX9. *Nat Commun*, 2018, 9: 5319
- 10 Zhang X, Choi P S, Francis J M, et al. Somatic superenhancer duplications and hotspot mutations lead to oncogenic activation of the KLF5 transcription factor. *Cancer Discov*, 2018, 8: 108–125
- 11 Zhang X, Choi P S, Francis J M, et al. Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat Genet*, 2016, 48: 176–182
- 12 Hamala T, Wafula E K, Guiltinan M J, et al. Genomic structural variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the chocolate tree. *Proc Natl Acad Sci USA*, 2021, 118: e2102914118
- 13 Wellenreuther M, Mérot C, Berdan E, et al. Going beyond SNPs: the role of structural genomic variants in adaptive evolution and species diversification. *Mol Ecol*, 2019, 28: 1203–1209
- 14 Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. *Front Bioeng Biotechnol*, 2015, 3: 92
- 15 Gabur I, Chawla H S, Snowdon R J, et al. Connecting genome structural variation with complex traits in crop plants. *Theor Appl Genet*, 2019, 132: 733–750
- 16 Song B, Ning W, Wei D, et al. Plant genome resequencing and population genomics: current status and future prospects. *Mol Plant*, 2023, 16: 1252–1268
- 17 Lye Z N, Purugganan M D. Copy number variation in domestication. *Trends Plant Sci*, 2019, 24: 352–365
- 18 Voss-Fels K, Snowdon R J. Understanding and utilizing crop genome diversity via high-resolution genotyping. *Plant Biotechnol J*, 2016, 14: 1086–1094
- 19 Soto D C, Uribe-Salazar J M, Shew C J, et al. Genomic structural variation: a complex but important driver of human evolution. *Am J Biol Anthropol*, 2023, 181: 118–144
- 20 Feuk L, Carson A R, Scherer S W. Structural variation in the human genome. *Nat Rev Genet*, 2006, 7: 85–97
- 21 Maydan J S, Lorch A, Edgley M L, et al. Copy number variation in the genomes of twelve natural isolates of *Caenorhabditis elegans*. *BMC Genomics*, 2010, 11: 62
- 22 Gazave E, Darré F, Morcillo-Suarez C, et al. Copy number variation analysis in the great apes reveals species-specific patterns of structural

- variation. *Genome Res*, 2011, 21: 1626–1639
- 23 Ho S S, Urban A E, Mills R E. Structural variation in the sequencing era. *Nat Rev Genet*, 2020, 21: 171–189
 - 24 Vollger M R, Guitart X, Dishuck P C, et al. Segmental duplications and their variation in a complete human genome. *Science*, 2022, 376: 55
 - 25 Mahmoud M, Gobet N, Cruz-Dávalos D I, et al. Structural variant calling: the long and the short of it. *Genome Biol*, 2019, 20: 246
 - 26 Redon R, Ishikawa S, Fitch K R, et al. Global variation in copy number in the human genome. *Nature*, 2006, 444: 444–454
 - 27 Sebat J, Lakshmi B, Troge J, et al. Large-scale copy number polymorphism in the human genome. *Science*, 2004, 305: 525–528
 - 28 Stankiewicz P, Lupski J R. Structural variation in the human genome and its role in disease. *Annu Rev Med*, 2010, 61: 437–455
 - 29 Hadi K, Yao X, Behr J M, et al. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. *Cell*, 2020, 183: 197–210.e32
 - 30 van Belzen I A E M, Schönhuth A, Kemmeren P, et al. Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. *NPJ Precis Oncol*, 2021, 5: 15
 - 31 Ottaviani D, LeCain M, Sheer D. The role of microhomology in genomic structural variation. *Trends Genet*, 2014, 30: 85–94
 - 32 Wang W J, Li L Y, Cui J W. Chromosome structural variation in tumorigenesis: mechanisms of formation and carcinogenesis. *Epigenet Chromatin*, 2020, 13: 49
 - 33 Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. *Brief Funct Genomics*, 2015, 14: 305–314
 - 34 Dileep V, Boix C A, Mathys H, et al. Neuronal DNA double-strand breaks lead to genome structural variations and 3D genome disruption in neurodegeneration. *Cell*, 2023, 186: 4404–4421.e20
 - 35 Carvalho C M B, Lupski J R. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet*, 2016, 17: 224–238
 - 36 Burssed B, Zamariolli M, Bellucco F T, et al. Mechanisms of structural chromosomal rearrangement formation. *Mol Cytogenet*, 2022, 15: 23
 - 37 Ebert P, Audano P A, Zhu Q, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 2021, 372: eabf7117
 - 38 Niu Y, Teng X, Zhou H, et al. Characterizing mobile element insertions in 5675 genomes. *Nucleic Acids Res*, 2022, 50: 2493–2508
 - 39 Kojima S, Koyama S, Ka M, et al. Mobile element variation contributes to population-specific genome diversification, gene regulation and disease risk. *Nat Genet*, 2023, 55: 939–951
 - 40 Fujimoto A, Wong J H, Yoshii Y, et al. Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. *Genome Med*, 2021, 13: 65
 - 41 Merot C. Karyotype-environment associations support a role for a chromosomal inversion in local adaptation among North American populations of the seaweed fly *Coelopa frigida*. *bioRxiv*, 2018, 2018: 278317
 - 42 Nadeau N J. Genes controlling mimetic colour pattern variation in butterflies. *Curr Opin Insect Sci*, 2016, 17: 24–31
 - 43 Wellenreuther M, Bernatchez L. Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol Evol*, 2018, 33: 427–440
 - 44 Ayala D, Guerrero R F, Kirkpatrick M. Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito. *Evolution*, 2013, 67: 946–958
 - 45 Trunca C, Mendell N R, Schilit S L P. Reproductive risk estimation calculator for balanced translocation carriers. *Curr Protocols*, 2022, 2: e633
 - 46 Tunc E, Ilgaz S. Robertsonian translocation (13;14) and its clinical manifestations: a literature review. *Reprod Biomed Online*, 2022, 45: 563–573
 - 47 Liang Z B, Liang C, Geng Y Q, et al. The mechanisms of host restriction on LINE-1 element. *Acta Agronomica Sin*, 2013, 40: 711
 - 48 Gao X, Wang S, Wang Y F, et al. Long read genome assemblies complemented by single cell RNA-sequencing reveal genetic and cellular mechanisms underlying the adaptive evolution of yak. *Nat Commun*, 2022, 13: 4887
 - 49 Ma H, Jiang J, He J, et al. Long-read assembly of the Chinese indigenous Ningxiang pig genome and identification of genetic variations in fat metabolism among different breeds. *Mol Ecol Resour*, 2022, 22: 1508–1520
 - 50 Jones M R, Mills L S, Alves P C, et al. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science*, 2018, 360: 1355–1358
 - 51 Bridges C B. The bar “gene” a duplication. *Science*, 1936, 83: 210–211
 - 52 Sturtevant A H. The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J Exp Zool*, 1913, 14: 43–59

- 53 Jacobs P. The somatic chromosomes in mongolism. *Lancet*, 1959, 273: 710
- 54 Jacobs P A, Strong J A. A case of human intersexuality having a possible *XXY* sex-determining mechanism. *Nature*, 1959, 183: 302–303
- 55 Merot C, Oomen R A, Tigano A, et al. A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends Ecol Evol*, 2020, 35: 561–572
- 56 Shah M S, Cinnioglu C, Maisenbacher M, et al. Comparison of cytogenetics and molecular karyotyping for chromosome testing of miscarriage specimens. *Fertil Steril*, 2017, 107: 1028–1033
- 57 Coughlin C R, Scharer G H, Shaikh T H. Clinical impact of copy number variation analysis using high-resolution microarray technologies: advantages, limitations and concerns. *Genome Med*, 2012, 4: 80
- 58 Keren B. The advantages of SNP arrays over CGH arrays. *Mol Cytogenet*, 2014, 7: 131
- 59 Sander C, Wallenborn M, Brandt V P, et al. Central neurocytoma: SNP array analyses, subtel FISH, and review of the literature. *Pathol Res Pract*, 2019, 215: 152397
- 60 Hollox E J, Zuccherato L W, Tucci S. Genome structural variation in human evolution. *Trends Genet*, 2022, 38: 45–58
- 61 Pinto D, Darvishi K, Shi X, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol*, 2011, 29: 512–520
- 62 Hu T, Chitnis N, Monos D, et al. Next-generation sequencing technologies: an overview. *Hum Immunol*, 2021, 82: 801–811
- 63 Chaisson M J P, Sanders A D, Zhao X, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun*, 2019, 10: 1784
- 64 Sedlazeck F J, Lee H, Darby C A, et al. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet*, 2018, 19: 329–346
- 65 Xiao T, Zhou W. The third generation sequencing: the advanced approach to genetic diseases. *Transl Pediatr*, 2020, 9: 163–173
- 66 Audano P A, Sulovari A, Graves-Lindsay T A, et al. Characterizing the major structural variant alleles of the human genome. *Cell*, 2019, 176: 663–675.e19
- 67 Hu T, Li J, Long M, et al. Detection of structural variations and fusion genes in breast cancer samples using third-generation sequencing. *Front Cell Dev Biol*, 2022, 10: 854640
- 68 Quan C, Li Y, Liu X, et al. Characterization of structural variation in tibetans reveals new evidence of high-altitude adaptation and introgression. *Genome Biol*, 2021, 22: 159
- 69 Alkan C, Coe B P, Eichler E E. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 2011, 12: 363–376
- 70 Rausch T, Zichner T, Schlattl A, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 2012, 28: i333–i339
- 71 Layer R M, Chiang C, Quinlan A R, et al. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*, 2014, 15: R84
- 72 Sedlazeck F J, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*, 2018, 15: 461–468
- 73 Kosugi S, Momozawa Y, Liu X, et al. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol*, 2019, 20: 117
- 74 Nattestad M, Schatz M C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics*, 2016, 32: 3021–3023
- 75 O'donnell S, and Fischer G. MUM&Co: accurate detection of all SV types through whole-genome alignment. *Bioinformatics*, 2020, 36: 3242–3243
- 76 Jung Y, and Han D. BWA-MEME: BWA-MEM emulated with a machine learning approach. *Bioinformatics*, 2022, 38: 2404–2413
- 77 Langdon W B. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *Biodata Min*, 2015, 8: 1
- 78 Zhang J, Wang J, Wu Y. An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. *BMC Bioinf*, 2012, 13: S6
- 79 Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 2018, 34: 3094–3100
- 80 Luth T, Schaake S, Grünwald A, et al. Benchmarking low-frequency variant calling with long-read data on mitochondrial DNA. *Front Genet*, 2022, 13: 887644

- 81 Delcher A L, Salzberg S L, Phillippy A M. Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinformatics*, 2003, doi: 10.1002/0471250953.bi1003s00
- 82 Hubbard J K, Uy J A C, Hauber M E, et al. Vertebrate pigmentation: from underlying genes to adaptive function. *Trends Genet*, 2010, 26: 231–239
- 83 Radke D W, Lee C. Adaptive potential of genomic structural variation in human and mammalian evolution. *Brief Funct Genomics*, 2015, 14: 358–368
- 84 Mills L S, Bragina E V, Kumar A V, et al. Winter color polymorphisms identify global hot spots for evolutionary rescue from climate change. *Science*, 2018, 359: 1033–1036
- 85 Liang D, Zhao P, Si J, et al. Genomic analysis revealed a convergent evolution of line-1 in coat color: a case study in water buffaloes (*Bubalus bubalis*). *Mol Biol Evol*, 2021, 38: 1122–1136
- 86 Han J I, Yang H, Jeung E B, et al. Altered expression of melanocortin-1 receptor (MC1R) in a yellow-coloured wild raccoon dog (*Nyctereutes procyonoides*). *Vet Dermatol*, 2012, 23: 187
- 87 Gunnarsson U, Kerje S, Bed'hom B, et al. The Dark brown plumage color in chickens is caused by an 8.3-kb deletion upstream of *SOX10*. *Pigment Cell Melanoma Res*, 2011, 24: 268–274
- 88 Durkin K, Coppieters W, Drögemüller C, et al. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature*, 2012, 482: 81–84
- 89 Tuttle E M, Bergland A O, Korody M L, et al. Divergence and functional degradation of a sex chromosome-like supergene. *Curr Biol*, 2016, 26: 344–350
- 90 Shu W S, Huang L N. Microbial diversity in extreme environments. *Nat Rev Microbiol*, 2022, 20: 219–235
- 91 Zong S B, Li Y L, and Liu J X. Genomic architecture of rapid parallel adaptation to fresh water in a wild fish. *Mol Biol Evol*, 2021, 38: 1317–1329
- 92 Trigo B B, Utsunomiya A T H, Fortunato A A A D, et al. Variants at the *ASIP* locus contribute to coat color darkening in Nellore cattle. *Genet Sel Evol*, 2021, 53: 40
- 93 Leonard A S, Crysnanto D, Fang Z H, et al. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nat Commun*, 2022, 13: 3012
- 94 Dong D, Jones G, Zhang S. Dynamic evolution of bitter taste receptor genes in vertebrates. *BMC Evol Biol*, 2009, 9: 12
- 95 Hof A E V T, Campagne P, Rigden D J, et al. The industrial melanism mutation in British peppered moths is a transposable element. *Nature*, 2016, 534: 102–105
- 96 Reid N M, Proestou D A, Clark B W, et al. The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science*, 2016, 354: 1305–1308
- 97 Ayala D, Zhang S, Chateau M, et al. Association mapping desiccation resistance within chromosomal inversions in the African malaria vector *Anopheles gambiae*. *Mol Ecol*, 2019, 28: 1333–1342
- 98 Larson G, Piperno D R, Allaby R G, et al. Current perspectives and the future of domestication studies. *Proc Natl Acad Sci USA*, 2014, 111: 6139–6146
- 99 Wang G D, Shao X J, Bai B, et al. Structural variation during dog domestication: insights from gray wolf and dhole genomes. *Natl Sci Rev*, 2019, 6: 110–122
- 100 Zhou Y, Yang L, Han X, et al. Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and evolutionary history. *Genome Res*, 2022, 32: 1585–1601
- 101 Cheng K K Y, Zhu W, Chen B, et al. The Adaptor protein APPL2 inhibits insulin-stimulated glucose uptake by interacting with TBC1D1 in skeletal muscle. *Diabetes*, 2014, 63: 3748–3758
- 102 Liang X F, Ran X Q, Niu X, et al. Association analysis between structural variation of *CYP3A29* gene 3'-UTR and rogen level of muscle tissue in Xiang pigs(in Chinese). *China Animal Husbandry & Vet Med*, 2023, 50: 3199–3209 [梁小芬, 冉雪琴, 牛熙, 等. 香猪 *CYP3A29* 基因 3'-UTR 结构变异及其与肌肉雄激素水平的关联分析. *中国畜牧兽医*, 2023, 50: 3199–3209]
- 103 Jiang Y F, Wang S, Wang C L, et al. Pangenome obtained by long-read sequencing of 11 genomes reveal hidden functional structural variants in pigs. *iScience*, 2023, 26: 106119
- 104 Yang Z, Cao X, Ma Y, et al. Novel copy number variation of the *BAG4* gene is associated with growth traits in three Chinese sheep populations.

- anim Biotechnol, 2021, 32: 461–469
- 105 Li R, Gong M, Zhang X, et al. A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes. [Genome Res](#), 2023, 33: 463–477
- 106 Alonge M, Wang X, Benoit M, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. [Cell](#), 2020, 182: 145–161.e23
- 107 Liu Z, Roberts R, Mercer T R, et al. Towards accurate and reliable resolution of structural variants for clinical diagnosis. [Genome Biol](#), 2022, 23: 68
- 108 Akbari V, Hanlon V C T, O'Neill K, et al. Parent-of-origin detection and chromosome-scale haplotyping using long-read DNA methylation sequencing and Strand-seq. [Cell Genomics](#), 2023, 3: 100233
- 109 Sanders A D, Hills M, Porubský D, et al. Characterizing polymorphic inversions in human genomes by single-cell sequencing. [Genome Res](#), 2016, 26: 1575–1587
- 110 Nuttle X, Giannuzzi G, Duyzend M H, et al. Emergence of a *Homo sapiens*-specific gene family and chromosome 16p11.2 CNV susceptibility. [Nature](#), 2016, 536: 205–209
- 111 Jain M, Olsen H E, Paten B, et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. [Genome Biol](#), 2016, 17: 239
- 112 Chan E K F, Cameron D L, Petersen D C, et al. Optical mapping reveals a higher level of genomic architecture of chained fusions in cancer. [Genome Res](#), 2018, 28: 726–738
- 113 Levy-Sakin M, Pastor S, Mostovoy Y, et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. [Nat Commun](#), 2019, 10: 1025
- 114 Lin J, Wang S, Audano P A, et al. SVision: a deep learning approach to resolve complex structural variants. [Nat Methods](#), 2022, 19: 1230–1233
- 115 Attwaters M. The final pieces of the human genome. [Nat Rev Genet](#), 2022, 23: 321
- 116 Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. [Science](#), 2022, 376: 44–53
- 117 Yang X, Wang X, Zou Y, et al. Characterization of large-scale genomic differences in the first complete human genome. [Genome Biol](#), 2023, 24: 157
- 118 Gong Y, Li Y, Liu X, et al. A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals? [J Anim Sci Biotechnol](#), 2023, 14: 73
- 119 Sherman R M, Salzberg S L. Pan-genomics in the human genome era. [Nat Rev Genet](#), 2020, 21: 243–254
- 120 Armstrong J, Hickey G, Diekhans M, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. [Nature](#), 2020, 587: 246–251
- 121 Feng S, Stiller J, Deng Y, et al. Dense sampling of bird diversity increases power of comparative genomics. [Nature](#), 2020, 587: 252–257
- 122 Derezanin L, Blažytė A, Dobrynin P, et al. Multiple types of genomic variation contribute to adaptive traits in the mustelid subfamily Guloninae. [Mol Ecol](#), 2022, 31: 2898–2919
- 123 Kronenberg Z N, Fiddes I T, Gordon D, et al. High-resolution comparative analysis of great ape genomes. [Science](#), 2018, 360: eaar6343
- 124 Zhou B, He Y, Chen Y, et al. Comparative genomic analysis identifies great-ape-specific structural variants and their evolutionary relevance. [Mol Biol Evol](#), 2023, 40: msad184
- 125 Spielmann M, Lupiáñez D G, Mundlos S. Structural variation in the 3D genome. [Nat Rev Genet](#), 2018, 19: 453–467
- 126 Dixon J R, Selvaraj S, Yue F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. [Nature](#), 2012, 485: 376–380
- 127 Marletaz F, de la Calle-Mustienes E, Acemel R D, et al. The little skate genome and the evolutionary emergence of wing-like fins. [Nature](#), 2023, 616: 495–503
- 128 Barutcu A R, Lajoie B R, McCord R P, et al. Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. [Genome Biol](#), 2015, 16: 214
- 129 Dixon J R, Xu J, Dileep V, et al. Integrative detection and analysis of structural variation in cancer genomes. [Nat Genet](#), 2018, 50: 1388–1398
- 130 Du Y, Gu Z, Li Z, et al. Dynamic interplay between structural variations and 3D genome organization in pancreatic cancer. [Adv Sci](#), 2022, 9: e2200818
- 131 Lupianez D G, Kraft K, Heinrich V, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer

interactions. [Cell](#), 2015, 161: 1012–1025

132 Valton A L, Dekker J. TAD disruption as oncogenic driver. [Curr Opin Genet Dev](#), 2016, 36: 34–40

133 Sapozhnikov D M, Szyf M. Unraveling the functional role of DNA demethylation at specific promoters by targeted steric blockage of DNA methyltransferase with CRISPR/dCas9. [Nat Commun](#), 2021, 12: 5711

Research progresses in revealing complex traits of animals based on genomic structural variations

WANG YaJun, WANG XiaoPing & YU Li

State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, School of Life Sciences, Yunnan University, Kunming 650091, China

Animals have evolved a wide variety of phenotypic traits to adapt to their environment. Among the molecular mechanisms of trait evolution, genomic structural variation is one of the important driving forces affecting complex traits in animals, which reflects variability at different locations and sequences in the genome, including a wide range of variant types such as large segment balanced variant inversions and translocations, and unbalanced variant insertions, deletions and duplications. Compared with single-base variants and small insertion deletions, genomic structural variants affect a wider range of DNA sequences in the genome and have a greater impact on complex traits in species. In this article, we will summarize the types of genomic structural variants and the development history of techniques for their detections and identifications, and review the research progress on the association between genomic structural variants and complex traits in terms of animal coat color traits, extreme environmental adaptations, and economic traits of domesticated animals, to provide important references to analyze the molecular mechanisms of complex traits from the perspective of genomic structural variants.

genomic structural variation, complex traits, animals, adaptive evolution

doi: [10.1360/SSV-2023-0216](https://doi.org/10.1360/SSV-2023-0216)