植物遗传资源学报
Journal of Plant Genetic Resources
ISSN 1672-1810,CN 11-4996/S

《植物遗传资源学报》网络首发论文

题目：两个甘薯野生近缘种的比较基因组分析
作者：肖世卓，许攀，王珧，戴习彬，周志林，曹清河
DOI：10.13430/j.cnki.jpgr.20241029003
收稿日期：2024-10-29
网络首发日期：2025-01-17
引用格式：肖世卓，许攀，王珧，戴习彬，周志林，曹清河．两个甘薯野生近缘种的比较基因组分析[J/OL]．植物遗传资源学报.
https://doi.org/10.13430/j.cnki.jpgr.20241029003

网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

网络首发时间：2025-01-17 13:48:08　网络首发地址：https://link.cnki.net/urlid/11.4996.s.20250117.1016.001

植物遗传资源学报
Journal of Plant Genetic Resources　DOI:10.13430/j.cnki.jpgr. 20241029003

两个甘薯野生近缘种的比较基因组分析

肖世卓[1],许攀[2]，王珧[1]，戴习彬[1]，周志林[1]，曹清河[1]

（[1]江苏徐淮地区徐州农业科学研究所/农业农村部甘薯生物学与遗传育种重点实验室, 徐州 221131；

[2]兰州大学草地农业科技学院/草种创新与草地农业生态系统全国重点实验室/

农业农村部草牧业创新重点实验室，兰州 730000）

摘要：甘薯是重要的农作物，其野生近缘种数量庞大，蕴含着丰富的基因变异。本研究利用公开发表的、基因组组装质量较高的两个甘薯野生种进行了比较基因组分析。发掘了覆盖全基因组的 1798184 个 SNP，同时还发掘了 2076 个易位和 84 个倒位，倒位中包含 2106 个基因，富集分析显示这些基因主要富集于次级代谢产物的生物合成等通路。其中 34 段倒位的断点位于基因内部，可能影响到基因的结构。两个基因组之间存在 10226 个插入型和 11411 个缺失型的结构变异（SV），对这些 SV 进行了注释，并对可能影响基因功能或表达的 SV 进行了基因富集分析，结果发现这些可能受影响的基因主要富集于次级代谢产物的通路，DNA 修复和复制等相关功能的基因也出现了富集。通过对两个野生种和其它近缘种共线性基因对的同义替换率（Ks）和四倍简并位点颠换率（4DTv）的统计，说明番薯属植物的共同祖先在发生了全基因组三倍化事件之后才分化成不同的种。以上研究将为甘薯野生近缘种优异变异的发掘和物种分化提供有力支撑。

关键词：甘薯；野生近缘种；比较基因组学

Comparative Genomic Analysis of Two Wild Relatives of Sweetpotato

XIAO Shizhuo[1], XU Pan[2], WANG Yao[1], DAI Xibin[1], ZHOU Zhilin[1], CAO Qinghe[1]

([1]*Xuzhou Institute of Agricultural Sciences in Jiangsu Xuhuai District/ Key Laboratory of Biology and Genetic Breeding of Sweetpotato, Ministry of Agriculture and Rural Affairs, Xuzhou 221131*; [2]*College of Pastoral Agriculture Science and Technology, Lanzhou University/ State Key Laboratory of Herbage Improvement and Grassland Agro-ecosystems/ Key Laboratory of Grassland Livestock Industry Innovation, Ministry of Agriculture and Rural Affairs*, *Lanzhou 730000*)

Abstract: Sweetpotato is an important crop, and it boasts numerous wild relatives that possess abundant genetic variations In this study, a comparative genomic analysis was conducted using two publicly available high-quality genome of sweetpotato wild relatives. A total of 1798184 single nucleotide polymorphisms (SNPs)

收稿日期:2024-10-29
第一作者研究方向为甘薯种质资源保存、鉴定以及优异基因的挖掘，E-mail：xiaoshizhuo@jaas.ac.cn
通信作者：曹清河，研究方向为甘薯种质资源，E-mail：caoqinghe@jaas.ac.cn
基金项目：江苏省自然科学基金（BK20221213）；国家甘薯产业技术体系（CARS-10-GW01）
Foundation projects: Natural Science Foundation of Jiangsu Province (BK20221213); China Agriculture Research System (CARS-10-GW01）

covering the whole genome were excavated. In additions, 2076 translocations and 84 inversions were detected. The inversions encompassed 2106 genes. Enrichment analysis showed that these genes are mainly enriched in pathways such as the biosynthesis of secondary metabolites. Among them, the breakpoints of 34 inversions are located inside the genes and may affect the structure of the genes. There are 10226 insertion-type and 11411 deletion-type structural variations (SVs) between the two genomes. We annotated these SVs and conducted gene enrichment analysis on those SVs that may affect gene function or expression. The results showed that these potentially affected genes were mainly enriched in pathways of secondary metabolites. Genes related to functions such as DNA repair and replication were enriched. Through the statistics of the synonymous substitution rate (*Ks*) and fourfold degenerate synonymous site (4DTv) of collinear gene pairs of two wild species and other related species, it was indicated that the common ancestor of the genus *Ipomoea* differentiated into different species after a whole-genome triplication event. The above research will provide strong support for the excavation of excellent variations in wild relatives of sweetpotato as well as for species differentiation.

Key words: sweetpotato; wild relatives; comparative genomics

作物的野生近缘种蕴含着丰富的、有价值的基因资源，不仅可以弥补栽培种长期驯化过程中丢失的重要性状，还可以为栽培作物的起源和驯化提供重要的遗传信息[1]。甘薯[*Ipomoea batatas* (L.) Lam.]是重要的粮食、饲料以及能源作物，其高产广适的特性为中国乃至世界的粮食安全做出了重要的贡献。甘薯在植物分类学上属于旋花科（Convolvulaceae）番薯属（*Ipomoea*）甘薯组（*batatas*），番薯属包含了约 800 种植物，是旋花科中最大的属，同时也是植物界最大的属之一，为植物分类学的研究提供了丰富的资源[2]。其中甘薯所在的甘薯组包含了约 15 个物种，这些种被认为是与甘薯亲缘关系最近的野生种，大部分为二倍体，也包含了四倍体和六倍体[2]。这些物种可能参与到了甘薯的起源或进化，其中三浅裂野牵牛（*I.trifida*）和毛果甘薯（*I. cordatotriloba*）就属于甘薯组中两个重要的野生近缘种。

三浅裂野牵牛（下文简称为 Itr）被认为是甘薯的二倍体祖先种之一，是甘薯驯化研究中最受关注的物种，其中包含二倍体和六倍体。因该种与甘薯杂交亲和，可以作为亲本与甘薯杂交，拓宽栽培种遗传背景。早在 1975 年，日本学者就利用六倍体的 Itr 与栽培种进行杂交，最终获得了高淀粉、高产、抗茎线虫病的品种“南丰”[3]。因其在甘薯起源中扮演了重要角色，二倍体的 Itr 是最早开展全基因组测序和组装的甘薯野生种。2015 年，日本学者分别组装了 Itr 自交系 Mx23Hm 和高度杂合系 0431-1 的基因组草图[4]。2018 年，由美国和中国等国家合作，共同绘制了首个染色体级别的 Itr 基因组[5]。2019 年，中国学者绘

制了一个具有膨大根系的 Itr 基因组[6]。但是上述 Itr 基因组的组装都是以二代测序技术为主，质量仍有待提高。2022 年，中日韩三国学者，利用 PacBio 测序技术，对自交系 Mx23Hm 进行了重新的测序、组装和注释[7]。获得了总长为 502.2 Mb 的高质量基因组，包含 34386 个基因。毛果甘薯（下文简称为 Ico）基因组的纯合度较高，自交亲和，结实率高。2024 年，由本团队主导，绘制了 Ico 的端粒到端粒（T2T，telomere-to-telomere）基因组，是旋花科番薯属中第一个真正意义上的 T2T 基因组。该基因组全长 454.8 Mb, 由无间隙的 15 条染色体组成，包含 40238 个蛋白编码基因[8]。从研究者构建的进化树中可以得知，在已完成基因组组装的番薯属植物中，Itr 和 Ico 是与甘薯的亲缘关系较近的野生种[8]。

植物比较基因组学是指通过比较不同植物物种的完整基因组序列，研究它们之间的基因组结构、功能和演化关系的科学。比较基因组不仅可以用来揭示植物之间的遗传相似性，还为理解植物的进化历史、适应机制以及功能基因的预测提供了重要工具。到 2024 年为止，已经公布了超过 1000 个植物基因组，涵盖大约 800 个物种[9]。借助测序技术和算法的进步，通过基因组之间的比较，能够在全基因组范围内识别 SNP 以及复杂的结构变异（SV）（通常>50 bp），包括拷贝数变异（CNVs）、存在/缺失变异（PAVs）、倒位和易位，用来进行物种进化的分析和解析复杂性状的遗传基础，从而服务于育种实践。研究人员利用全球 377 份栽培稻和野生稻的三代测序数据，构建了迄今为止最大规模的水稻群体水平倒位变异图谱，发现与逆境响应相关的基因在倒位内部或附近发生了明显富集，揭示了倒位变异对水稻抗逆性状的重要影响[10]。通过对 26 份大豆种质进行高质量的基因组组装，并与已公开的大豆基因组进行比较基因组分析，研究人员揭示了 SV 在调控关键农艺性状方面扮演的重要角色：例如在野生大豆与栽培大豆之间，查尔酮合成酶基因簇的结构变异是导致种皮颜色从黑色向黄色转变的主要驯化因素； *SoyZH13_14G179600* 基因的 SV 也造成了该基因在不同种质资源中表达模式的差异[11]。

甘薯基因组较为复杂，高质量组装较为困难，比较基因组研究还较少，主要在野生二倍体种中开展。通过比较基因组分析发现，研究人员发现贮藏蛋白基因在 *I. trifida*、*I. triloba* 以及马铃薯中发生了明显扩张，暗示贮藏蛋白在根茎类作物进化中扮演了重要角色[5]。但是上述比较基因组研究以二代测序为基础，无法对 SV 等大片段的变异进行发掘，对基因组间的变异认识受限。本研究拟通过比较甘薯组中组装质量较高的两个二倍体野生种基因组，来发掘甘薯组基因组上的变异，尤其是大的结构变异，并对变异进行系统地分析，为甘薯组植物的进化和分化，以及野生近缘种优异变异的育种利用奠定基础。

1 材料与方法

1.1 试验数据的获取

本研究中用到的基因组数据均来自公开数据库，具体下载地址如下所示：Itr (https://plantgarden.jp/en/list/t35884/genome/t35884.G002)、Ico (https://ngdc.cncb.ac.cn/gwh/Assembly/85948/show)、马鞍藤（*I. pes-caprae*）（下文简称 Ipes）(https://ngdc.cncb.ac.cn/gwh/Assembly/82945/show) 以及马铃薯 (*Solanum tuberosum*)（下文简称 Stu）(http://www.bioinformaticslab.cn/pubs/dm8/)。

1.2 基因组 SV 的发掘

本研究使用两种方法来发掘基因组的 SV。首先是基因组之间的比对，即使用 Smartie-sv[12]将 Ico 的基因组比对至 Itr 基因组，产生长度不小于 50bp 的结构变异，作为下一步研究的 SV。同时，使用 NGMLR 0.2.7[13]将 Ico 的 HiFi reads 比对至 Itr 基因组，然后使用 Sniffles[14]来获得 SV。两种方法产生的 SV 进行整合，重复检测到的 SV，即插入或者缺失类型相同、检出位置有重叠区域的 SV 被认为是高置信度的 SV。使用 ANNOVAR[15]对这些 SV 进行注释。使用 SyRI[16]来鉴定倒位和重复序列，最后使用 SyRI 将比对结果可视化。

1.3 *Ks* 和 4DTv 的分析

将 Itr、Ico、Ipes 以及 Stu 基因组的氨基酸序列进行比对，使用 MCScanX 101[17]鉴定共线性的基因，参数设为默认。使用 PAML[18]来评估共线性基因间发生同义替换的 SNP 数与同义替换位点数比值（*Ks*）。使用 KaKs_Calculator[19]计算同源基因对之间 *Ka/Ks* 的值。使用个人脚本来评估 4DTv 值。

1.4 KEGG 和 GO 分析

采用 InterProScan v5.52-86.0[20]对基因进行本体论（Gene Ontology，GO）注释。运用 EggNOG-mapper v2.1.4[21]进行基因表达通路（KEGG）和蛋白结构域（Pfam）注释。借助 topGO[22]软件包和 REVIGO[23]进行 GO 富集分析，使用 CirGo[24]和 ggplot2[25]软件包将结果可视化。通过 clusterProfiler[26]软件包进行 KEGG 富集分析。

2 结果与分析

2.1 基因组比较和 SNP 的发掘

Itr 和 Ico 的基因组长度分别为 502237654 bp 和 454814757 bp，contig N50 的长度分别为 3.7 Mb 和 30.9 Mb， BUSCO 评估基因组的完整度分别为 98.5%和 97.6%，预测的基因数量分别为 34386 个和 40238 个，Itr 的测序方法主要为 PacBio 和 Hi-C 技术，Ico 除了上述两

种测序技术之外，还使用了 Nanopore 技术（表 1）。

表 1 *I. trifida*（Itr）和 *I. cordatotriloba*（Ico）的基因组统计

Table 1 Statistics of *I. trifida*（Itr） and *I. cordatotriloba*（Ico） genome

参数 parameter	Itr	Ico
种质 Germplasm	Mx23Hm	*xiaoshu*
基因组长度 Length of genome (bp)	502237654	454814757
Contig N50 （Mb）	3.73	30.93
测序方法 Sequencing Method	PacBio, Hi-C	PacBio, Nanopore, Hi-C
BUSCO （%）	98.5	97.6
基因数量 Gene number	34386	40238

通过比较 Itr 和 Ico 的基因组发现，Itr 和 Ico 两个基因组之间存在 1798184 个 SNP 位点。除此之外，在重复序列中，相比于 Itr，Ico 多出了 112 个拷贝，总长 220663 bp，减少了 152 个拷贝，总长 559140 bp。两个基因组存在 50802 个高度分化的区域，这段区域在 Itr 中的长度是 228414336 bp，在 Ico 中则是 254217620 bp（表 2）。

表 2 Itr 和 Ico 基因组比较产生的变异

Table 2 The variations between Itr and Ico genome

变异类型 Variation type	数目 Count	长度 (bp) Length
SNPs	1798184	1798184/1798184
拷贝数增加 Copy-gains	112	-/220633
拷贝数减少 Copy-losses	152	559140/-
高度分化 Highly diverged	50802	228414336/254217620
串联重复 Tandem repeats	14	15681/25346

“/”符号之前为 Itr 基因组中的长度，“/”之后为 Ico 基因组中的长度

The number before the "/" symbol is the length in the Itr genome, and the number after the "/" symbol is the length in the Ico genome

2.2 基因组共线性分析

染色体倒位和易位会改变基因的顺序，影响基因的表达，同时会增加物种基因资源的遗传多样性。为了探究 Itr 和 Ico 之间可能存在的遗传多样性，对两个基因组进行了共线性的分析。结果显示两个基因组之间存在 1582 个共线性区域，这段区域在 Itr 基因组中的长度为 271201177 bp，占基因组全长的比例为 54.0%；在 Ico 基因组中的长度则为 304956210 bp，占基因组全长的比例为 67.0%。两个基因组之间存在 84 个倒位，倒位在 Itr 基因组中的长度为 41239691 bp，占比为 8.2%；在 Ico 基因组中的长度则为 32262135 bp，占比为 7.1%。两个基因组之间还存在 2076 个易位，易位的长度在 Itr 基因组中的长度为 13473827 bp，占比为 2.7%，在 Ico 基因组中的长度则为 13973042 bp，占比为 3.1%（表 3，图 1a）。

表 3 Itr 和 Ico 基因组的共线性分析

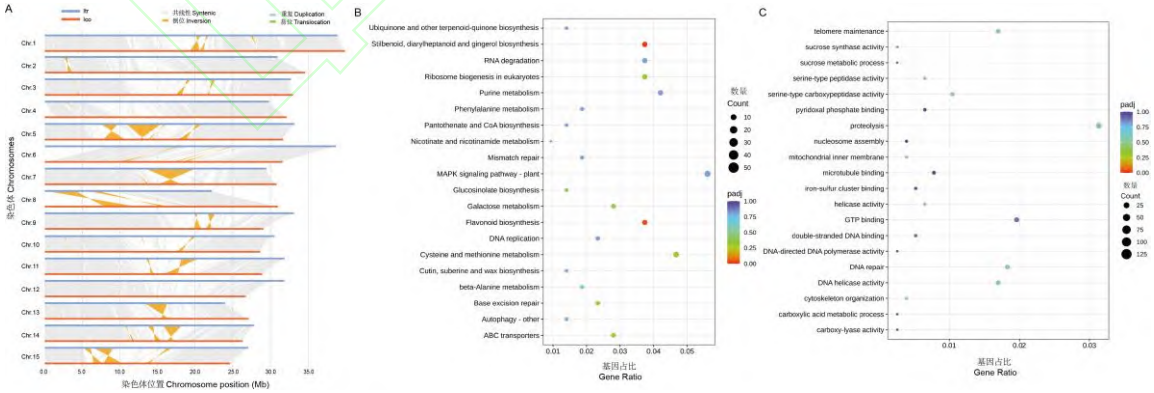
Table 3 Syntenic analysis between Itr and Ico genomes

变异类型 Variation type	数量 Count	Itr 中的长度（bp）和占比（%）[a] Length （bp）and proportion (%) on Itr genome	Ico 中的长度（bp）和占比（%） Length （bp）and proportion (%) on Ico genome
共线性区域 Syntenic regions	1582	271201177 (54.0)	304956210 (67.0)
倒位 Inversions	84	41239691 (8.2)	32262135 (7.1)
易位 Translocations	2076	13473827 (2.7)	13973042 (3.1)
Itr 中的重复 Duplications on Itr	918	4724592 (0.9)	-
Ico 中的重复 Duplications on Ico	2106	-	7285309 (1.6)
Itr 中未匹配序列 Not aligned on Itr	4131	135011642 (26.9)	-
Ico 中未匹配序列 Not aligned on Ico	5913	-	98197220 (21.6)

括号外数字表示长度，括号内数字表示占比

Numbers outside the brackets indicate length, and numbers inside the brackets indicate proportion

倒位可能会对基因的表达产生影响。通过对倒位中基因的研究发现，84 段倒位中共包含 2106 个基因，对这 2106 个基因分别进行了 KEGG 和 GO 富集分析。KEGG 分析显示基因富集较多的通路是二苯乙烯类、二芳基庚酸类和姜酚的生物合成（map00945）以及类黄酮的生物合成（map00941）等次级代谢产物的生物合成途径（图 1B）。除此之外，更值得关注的是，错配修复（map03430）、碱基切除修复（map03410）和 DNA 复制（map03030）等通路的基因也出现富集（图 1B）。GO 富集分析同样显示，端粒维持（GO:0000723）、DNA 解螺旋活性（GO:0003678）、双链 DNA 绑定（GO:0003690）、DNA 介导的 DNA 聚合酶活性（GO:0003887）以及 DNA 修复（GO:0006281）相关的基因出现显著的富集（图 1C）。这些基因跟倒位的形成和固定可能存在某种联系。



A：Itr 和 Ico 基因组之间的共线性；B：倒位中基因的 KEGG 富集分析；C：倒位中基因的 GO 富集分析

A: The synteny between Itr and Ico genomes; B: KEGG enrichment analysis of genes in inversion; C: G O enrichment analysis of genes in inversion

图 1 Itr 和 Ico 基因组之间的共线性以及倒位中的基因富集分析

Fig. 1 The synteny between Itr and Ico genomes and gene enrichment analysis in inversions

倒位会产生染色体断点，断点如果发生在基因内部会引起基因结构的变化从而影响其功能。经过统计发现，共有 34 段倒位的断点发生于 Itr 或者 Ico 中基因的内部，其中 1 段倒位的两端断点均位于基因内部（表 4）。34 段倒位共影响到 35 个基因，这些基因行使的功能包含转录因子、功能蛋白和酶等，参与到植物生长发育的不同阶段，没有明显规律，呈现出随机特点（表 4）。

2.3 插入缺失型结构变异的鉴定和分析

基因组上插入和缺失型的 SV 是引起表型变异的重要因素。本研究分别使用 Ico 三代测序得到的长序列以及组装完成的 T2T 基因组与 Itr 的基因组进行比较，得到全基因组的 SV。通过基因组之间的比较发现，相比于 Itr，Ico 基因组分别存在 45154 个插入型和 34638 个缺失型的 SV。通过测序长序列与基因组比较，发现相比于 Itr，Ico 基因组分别存在 60284 个插入型和 83558 个缺失型的 SV。由于 SV 检出存在较高的假阳性，为了降低假阳性率，将两种方法共同检测到的 SV 进行整合，保留重复检测到的 SV，最终获得了 10226 个插入和 11411 个缺失（表 5）。虽然这一过程可能会丢失一些真实存在的 SV，但是保留的 SV 被认为是具有高可信度的。通过序列注释发现，这些插入型的 SV 中，位于基因上游的有 1494 个，基因下游的有 1431 个，基因间的有 4162 个，内含子的有 2581 个，外显子的有 173 个，基因上游或者下游的有 380 个，位于剪接位点的有 5 个。缺失型的 SV 中，位于基因上游的有 1648 个，基因下游的有 1450 个，基因间区的有 5663 个，内含子的有 1998 个，外显子的有 348 个，基因上游或者下游的有 295 个，位于剪接位点的有 9 个（表 6）。

位于基因上游和外显子的 SV 可能会引起基因表达或者功能的变化，进而造成表型变异，本研究对这些基因（SV genes, SVGs）进行了富集分析。GO 富集分析显示，插入型的 SVG 主要富集于氧化胁迫应答（GO:0006979）、过氧化物酶活性（GO:0004601）、血红素结合（GO:0020037）等胁迫响应通路（图 2A）；缺失型的 SVG 主要富集于酸性磷酸酶活性（GO:0003993）、蛋白质水解作用（GO:0006508）和双链 DNA 结合（GO:0003690）等通路（图 2B）。KEGG 富集分析的结果显示，插入型 SVG 主要富集于角质，亚硫酸和蜡的生物合成（map00073）、N 糖的生物合成（map00510）和维生素 B6 的代谢（map00750）等途径（图 2C）；缺失型 SVG 主要富集于 N 糖的生物合成、苯丙烷类物质的生物合成（map00940）以及叶酸生物合成（map00790）（图 2D）。富集分析的结果说明，两个甘薯

野生近缘种之间的 SV，主要用来改变次级产物的变化，在保证了基本生存功能基因的保守性的同时，增加了两个物种对不同环境的适应性。

表 4 倒位断点处基因统计

Table 4 Statistics of genes covering inversion breakpoints

倒位编号 No. of inversions	基因组 Genomes	染色体 Chromosomes	倒位起止位置 The starting and ending position of the inversions	断点所在基因 The genes covering the breakpoints	基因起始位置 The starting and ending position of the genes	基因功能注释 Functional annotation of genes
1	Itr	Itr_chr01	11703012~11703394	Itr_chr01CG07580	11702887~11704047	-
2	Itr	Itr_chr03	19922914~20345663	Itr_chr03CG11900	20340885~20345769	poly(A) RNA polymerase GLD2
3	Itr	Itr_chr03	21722914~22704229	Itr_chr03CG12960	21714866~21724114	-
4	Itr	Itr_chr04	2752356~2873657	Itr_chr04CG04770	2870724~2887446	-
5	Itr	Itr_chr04	5793962~5795348	Itr_chr04CG08650	5784645~5795829	-
6	Itr	Itr_chr06	3927167~4015859	Itr_chr06CG06040	4013606~4016405	proton antiporter activity
7	Itr	Itr_chr06	21569168~21569727	Itr_chr06CG16690	21567864~21570685	light-harvesting complex II chlorophyll a/b binding protein 1
8	Itr	Itr_chr06	22213921~22728379	Itr_chr06CG17140	22712790~22738139	-
9	Itr	Itr_chr07	26827512~26828003	Itr_chr07CG19600	26821927~26828700	phosphorelay sensor kinase activity
10	Itr	Itr_chr08	27788218~27801091	Itr_chr08CG18580	27785342~27790952	lysosomal Pro-X carboxypeptidase
11	Itr	Itr_chr08	27816710~27817090	Itr_chr08CG18590	27814863~27820819	lysosomal Pro-X carboxypeptidase
12	Itr	Itr_chr09	911603~963422	Itr_chr09CG01660	961189~969086	sterol 24-C-methyltransferase
13	Itr	Itr_chr09	13287680~16314621	Itr_chr09CG15430	16314335~16315624	-
14	Itr	Itr_chr10	15750888~18193749	Itr_chr10CG14040	15742310~15753395	serine-type endopeptidase activity
15	Itr	Itr_chr13	11841296~11924724	Itr_chr13CG12110	11922947~11937966	AP-2 complex subunit mu-1
16	Itr	Itr_chr14	15978288~16031423	Itr_chr14CG11650	16021901~16031479	-
17	Itr	Itr_chr14	23223138~23349472	Itr_chr14CG15830	23221295~23223298	zinc ion binding
18	Itr	Itr_chr14	24546584~24707043	Itr_chr14CG16870	24543680~24547627	pectinesterase activity
18	Itr	Itr_chr14	24546584~24707043	Itr_chr14CG16990	24706321~24709541	-
19	Itr	Itr_chr14	24861187~25244280	Itr_chr14CG17470	25242071~25245346	-
20	Itr	Itr_chr15	106470~4084032	Itr_chr15CG05070	4080322~4084326	recognition of pollen
21	Itr	Itr_chr15	5682723~7141479	Itr_chr15CG06760	5682643~5684909	pectinesterase
22	Ico	Chr11	24496445~24496831	Icor_Chr11.01919	24494734~24497336	lysosomal Pro-X carboxypeptidase-like isoform X2
23	Ico	Chr12	12598615~12598997	Icor_Chr12.01502	12594644~12598979	B3 domain-containing protein At1g05920-like
24	Ico	Chr13	969575~1024868	Icor_Chr13.00176	968757~971843	cycloartenol-C-24-methyltransferase 1-like isoform X1
25	Ico	Chr13	15430378~16541871	Icor_Chr13.01604	15429529~15439711	katanin p80 WD40 repeat-containing subunit B1 homolog isoform X2
26	Ico	Chr14	25785754~25821651	Icor_Chr14.02284	25777482~25787636	vetispiradiene synthase 3-like
27	Ico	Chr15	11927803~12008190	Icor_Chr15.00998	11927513~11928117	-
28	Ico	Chr2	3042641~3683581	Icor_Chr02.00501	3682866~3684458	cysteine-rich repeat secretory protein 60
29	Ico	Chr3	4464551~4571946	Icor_Chr03.00679	4463814~4466804	cation/H(+) antiporter 15-like
30	Ico	Chr3	14014929~14036263	Icor_Chr03.01538	14008217~14016412	G-type lectin S-receptor-like serine/threonine-protein kinase
31	Ico	Chr5	16136141~16136714	Icor_Chr05.01181	16129497~16153598	alcohol dehydrogenase-like 2 isoform X2
32	Ico	Chr5	17695783~18808792	Icor_Chr05.01360	18805788~18810010	WAT1-related protein At1g43650-like
33	Ico	Chr6	2894889~2930955	Icor_Chr06.00381	2930162~~2947252	heavy metal-associated isoprenylated plant

34	Ico	Chr7	13546317~13782991	Icor_Chr07.01385	13780765~13783068	protein 39-like isoform X2 uncharacterized protein LOC109158392

表 5 Itr 和 Ico 基因组的插入和缺失型 SV 统计

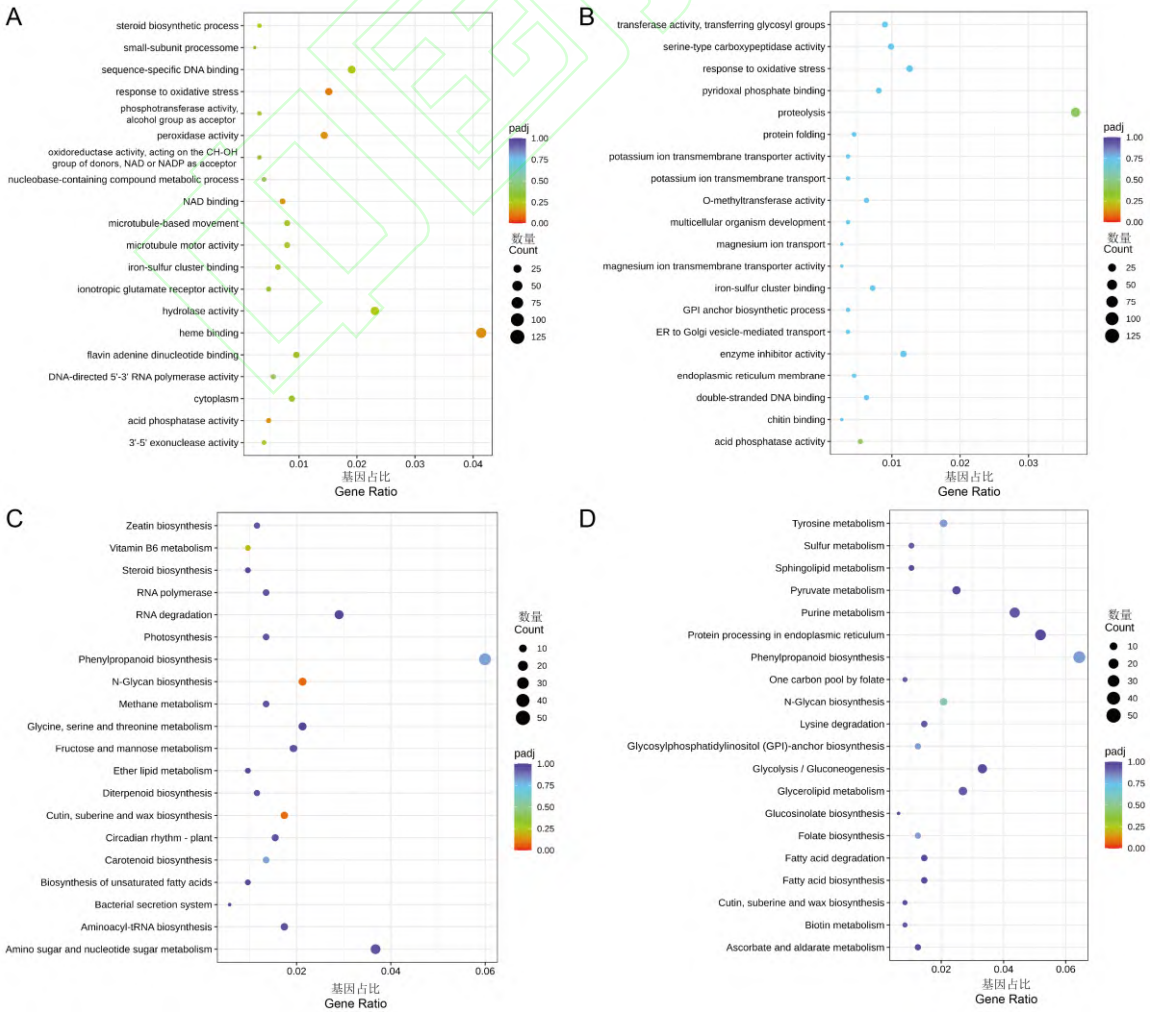
Table 5 Statistics of SVs by comparing Itr and Ico genomes

比对方法 Align methods	插入 Insertions	缺失 Deletions
基因组比对基因组 Genome to genome	45154	34638
读长比对基因组 Reads to genome	60284	83558
整合 Merge	10226	11411

表 6 SV 的位置注释

Table 6 Annotation of Positions of SVs

注释 Annotation	插入 Insertions	缺失 Deletions
上游 Upstream	1494	1648
下游 Downstream	1431	1450
基因间 Intergenic	4162	5663
内含子 Intronic	2581	1998
外显子 Exonic	173	348
上游或者下游 Upstream or downstream	380	295
剪接位点 Splicing	5	9



A：插入型 SVG 的 GO 富集分析；B：缺失型 SVG 的 GO 富集分析；C：插入型 SVG 的 KEGG 富集分析；D：缺失型 SVG 的 KEGG 富集分析

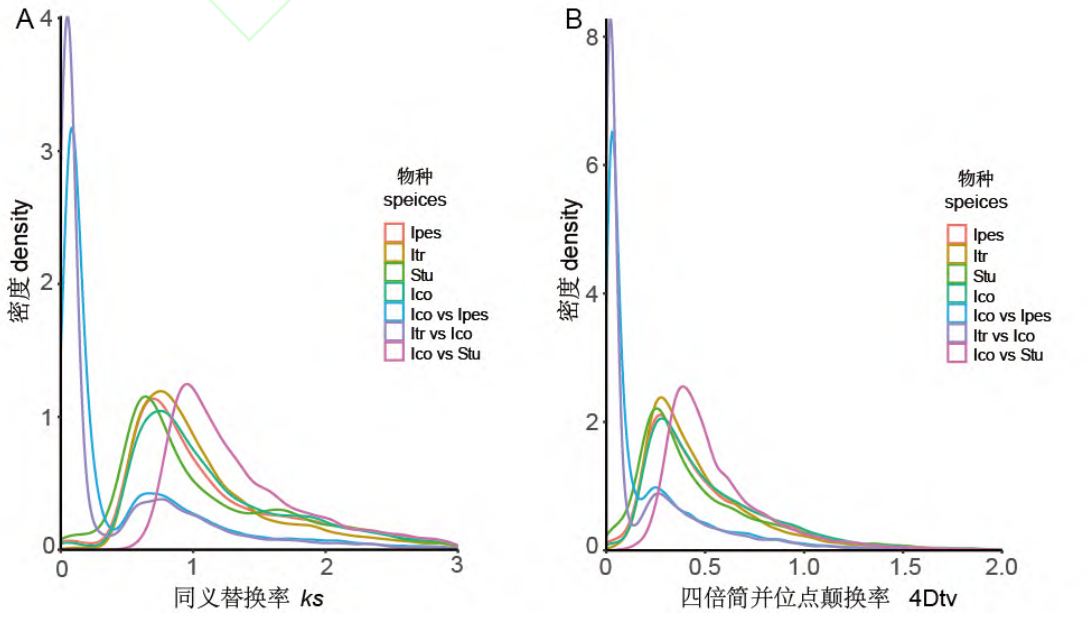
A: GO enrichment analysis of insertion SVGs; B: GO enrichment analysis of deletion SVGs; C: KEGG enrichment analysis of insertion SVGs; D: KEGG enrichment analysis of deletion SVGs

图 2 Itr 和 Ico 基因组间结构变异基因（SVGs）的富集分析

Fig. 2 Enrichment analysis of structural variant genes (SVGs) between Itr and Ico genomes

2.4 两个近缘野生种的进化和分化

为了探讨 Ico 和 Itr 在番薯属中的进化历程，本研究分析了 Itr、Ico 以及番薯属近缘种 Ipes 之间的共线性基因对的同义替换率（*Ks*），以马铃薯基因组[27]为参照。结果显示 Itr 的峰对应的 *Ks* 值（Ks^{Itr}）和 Ks^{Ico} 最为接近，其次是 Ks^{Ipes}，三者的峰都在 *Ks* = 0.75 左右（图 3A）。这说明番薯属这三个种在约 53.18 百万年前（Mya）发生过全基因组三倍化事件（WGT）[28]。进一步研究不同物种之间的 *Ks* 值分布，结果显示 Itr 和 Ico 之间主峰值对应的 *Ks* 值（$Ks^{Itr\ vs\ Ico}$）略小于 $Ks^{Ico\ vs\ Ipes}$，证实了 Itr 和 Ico 的分化时间略晚于 Ico 和 Ipes 的分化时间。Ks^{Itr}、Ks^{Ico} 和 Ks^{Ipes} 都大于 $Ks^{Itr\ vs\ Ico}$ 和 $Ks^{Ico\ vs\ Ipes}$，而小于 Ico 和马铃薯之间峰值对应的 *Ks* 值（$Ks^{Ico\ vs\ Stu}$）。这说明番薯属的 WGT 事件发生于属内不同种的分化之前，而在番薯属与马铃薯分化之后。除此之外，Itr 和 Ico 以及 Ico 和 Ipes 之间的 *Ks* 还存在一个较低的峰，相应的 *Ks* 值与三个物种各自的峰值 *Ks* 接近，大于主峰对应的 *Ks* 值，对应了 WGT 事件，也进一步证实了番薯属的 WGT 事件发生于属内不同种的分化之前。而马铃薯对应的峰 Ks^{Stu} 与番薯属较为接近，说明番薯属的祖先与马铃薯的祖先在相近时期分别独立发生了全基因组加倍事件。上述这些物种共线性碱基对之间的四倍简并位点颠换率（4DTv）统计得到了相同的结果（图 3B）。



A：Itr、Ico、Ipes 和马铃薯以及它们之间的 *Ks* 值分布；B：Itr、Ico、Ipes 和马铃薯以及它们之间的 4DTv 值分布。Stu 表示马铃薯

A: Distributions of *Ks* value of Itr, Ico, Ipes, potato and that between them. B: Distributions of 4DTv value of Itr, Ico, Ipes, potato and that between them. “Stu” means potato

图 3 Itr、Ico、Ipes 和马铃薯以及它们之间的 *Ks* 和 4DTv 值分布

Fig. 3 Distributions of *Ks* and 4DTv value of Itr, Ico, Ipes, potato and that between them

3 讨论

中国甘薯遗传基础较为狭窄，大部分推广品种具有南瑞苕或者胜利百号的血缘[29]。甘薯野生种质资源丰富，挖掘野生近缘种中的优异变异，应用于甘薯育种当中，有利于拓宽甘薯的遗传背景。同时，甘薯的起源尚不清晰，了解甘薯野生近缘种的分化可为甘薯的起源进化提供参考。本研究以甘薯野生近缘种中基因组组装质量较高的两个种为研究对象，进行了比较基因组的分析。挖掘了覆盖全基因组的 SNP，可以以此为基础，通过进一步的筛选开发分子标记，快速简便地对甘薯野生种进行分型和鉴定。

染色体倒位和易位是一种重要的遗传变异，可以迅速改变染色体的结构和基因组合，创造出新的基因连锁关系，并通过长时间的进化，在群体中维持一定的频率。新的基因组合可能赋予个体新的适应性特征，有助于物种适应环境的变化。同时，倒位还可能会影响基因的表达。Itr 和 Ico 之间存在 2076 个易位和 84 个倒位。人类染色体倒位和易位最主要的遗传效应是育性受损。天然的 Itr 自交育性较差，且基因组是杂合的，这些倒位和易位可能会存在于 Itr 的亚基因组之间，从而在减数分裂时期产生染色体异常的配子，最终导致自交亲和性较差。栽培甘薯的自交亲和性差可能也与此相关。这需要更高质量的单倍型基因组进行深入的研究。一些大型的倒位会抑制染色体的局部重组，产生生殖隔离进而推动物种分化[30]。Itr 和 Ico 的物种分化可能与本研究中发掘的 84 段倒位存在联系。但是倒位产生的原因仍然不清晰。Itr 和 Ico 基因组倒位断点处的基因有 35 个，这些基因功能注释多样，并未发现与倒位的发生存在明显的相关性。说明倒位的发生可能并不是由断点处基因起作用，可能与染色体的空间结构存在联系。

基于测序技术和算法的影响，以往对基因组变异的认识主要集中于 SNP，基因组变异与表型变异之间的关联也以 SNP 为主。但是随着技术的进步，越来越多的证据证明结构变异也是基因组变异的重要部分。利用短读长挖掘的 SV 假阳性较高，甚至达到了 89%[31]。本研究基于长读长测序挖掘插入和缺失型的 SV，并结合两种不同的比对方法进行筛选，最

终保留的 SV 不足四分之一，大大提高了 SV 的可靠性。SV 在物种分化和遗传多样性中扮演重要角色，一些影响物种生长发育的 SV 会导致物种本身无法生存，这些变异在进化中不会被保留，而相对轻微的变异，在自然选择和进化过程中可能会被保留下来，为物种的进化和适应提供了原材料。本研究通过比较 Itr 和 Ico 的基因组，发现 SVG 主要富集于次级代谢的通路，倒位中的基因也主要富集于次级代谢通路，次级代谢通路基因的改变赋予了甘薯野生种相应表型的变异以及对不同环境的适应性，丰富了甘薯野生种近缘种的遗传多样性。

全基因组加倍是物种进化和分化的重要驱动力，由于缺乏高质量参考基因组的原因，从全基因组加倍角度研究番薯属植物进化历程的研究还较少。甘薯及其野生近缘种在进化过程中经历了 WGT 事件，通过分析甘薯野生种与其它物种（马铃薯）共线性基因对的 *Ks* 和 4DTv，证实在漫长的进化历程中，番薯属植物先是与马铃薯所在的属分道扬镳，之后各自经历了 WGT 事件。番薯属的 WGT 与马铃薯的全基因组加倍发生的时间相近，约在 53.18 Mya。被子植物大规模全基因组加倍事件主要发生在三个地质历史时期，其中第二次大规模的加倍化事件发生在白垩纪-古近纪灭绝时期，距今约 65 Mya[32]。番薯属祖先和马铃薯祖先的全基因组加倍可能就发生在这个时期。全基因组加倍使其增强了对环境的适应性，在剧烈的环境变化中幸存了下来。随后番薯属植物分化成不同的种。 Itr 和 Ico（或者它们的祖先）先与 Ipes 分化，之后 Itr 和 Ico 发生了分化。这与前人对番薯属野生种的研究结果是相吻合的[5, 33]。

参考文献

[1] 乔卫华, 张宏斌, 郑晓明, 陈宝雄, 陈彦清, 李垚奎, 程云连, 张丽芳, 方沩, 孙玉芳, 杨庆文. 我国作物野生近缘植物保护工作近 20 年的成就与展望. 植物遗传资源学报, 2020, 21(6): 1329-1336

Qiao W H, Zhang H B, Zheng X M, Chen B X, Chen Y Q, Li Y K, Cheng Y L, Zhang L F, Fang W, Sun Y F, Yang Q W. Achievements of the conservation of wild relatives of crops in the past 20 years and the prospects in China. Journal of Plant Genetic Resources, 2020, 21(6): 1329-1336

[2] Muñoz-Rodríguez P, Wood J R I, Wells T, Carruthers T, Sumadijaya A, Scotland R W. The challenges of classifying big genera such as Ipomoea. Taxon, 2023, 72(6): 1201-1215

[3] 刘旭. 作物种质资源学教程. 北京: 高等教育出版社, 2024: 248

Liu X. Crop germplasmics. Beijing: Higher Education Press, 2024: 248

[4] Hirakawa H, Okada Y, Tabuchi H, Shirasawa K, Watanabe A, Tsuruoka H, Minami C, Nakayama S, Sasamoto S, Kohara M, Kishida Y, Fujishiro T, Kato M, Nanri K, Komaki A, Yoshinaga M, Takahata Y, Tanaka M, Tabata S, Isobe S N. Survey of genome sequences in a wild sweet potato, *Ipomoea trifida* (H. B. K.) G. Don. DNA Research, 2015, 22(2): 171-179

[5] Wu S, Lau K H, Cao Q, Hamilton J P, Sun H, Zhou C, Eserman L, Gemenet D C, Olukolu B A, Wang H, Crisovan E, Godden G T, Jiao C, Wang X, Kitavi M, Manrique-Carpintero N, Vaillancourt B, Wiegert-Rininger K, Yang X, Bao K, Schaff J, Kreuze J, Gruneberg W, Khan A, Ghislain M, Ma D, Jiang J, Mwanga R O M, Leebens-Mack J, Coin L J M, Yencho G C, Buell C R, Fei Z. Genome sequences of two diploid wild relatives of cultivated sweetpotato reveal targets for genetic

improvement. Nature Communications, 2018, 9(1): 4580

[6] Li M, Yang S, Xu W, Pu Z, Feng J, Wang Z, Zhang C, Peng M, Du C, Lin F, Wei C, Qiao S, Zou H, Zhang L, Li Y, Yang H, Liao A, Song W, Zhang Z, Li J, Wang K, Zhang Y, Lin H, Zhang J, Tan W. The wild sweetpotato (*Ipomoea trifida*) genome provides insights into storage root development. BMC Plant Biology, 2019, 19(1): 119

[7] Yoon U-H, Cao Q, Shirasawa K, Zhai H, Lee T-H, Tanaka M, Hirakawa H, Hahn J-H, Wang X, Kim H S, Tabuchi H, Zhang A, Kim T-H, Nagasaki H, Xiao S, Okada Y, Jeong J C, Nagano S, Shin Y, Lee H-U, Park S-U, Lee S J, Lee K, Yang J-W, Ahn B O, Ma D, Takahata Y, Kwak S-S, Liu Q, Isobe S. Haploid-resolved and chromosome-scale genome assembly in hexa-autoploid sweetpotato (*Ipomoea batatas* (L.) Lam). bioRxiv. (2022-12-25) [2024-4-01]. https://doi.org/10.1101/2022.12.25.521700

[8] Xiao S, Wang Y, Zhou Z, Zhao L, Zhao L, Gao B, Dai X, Xu P, Cao Q. *Xiaoshu*, a simple genetic model system for sweetpotato (*Ipomoea batatas* (L.) Lam.). Plant Biotechnology Journal, 2024, https://doi.org/10.1111/pbi.14528.

[9] Shi J, Tian Z, Lai J, Huang X. Plant pan-genomics and its applications. Molecular Plant, 2023, 16(1): 168-186

[10] He W, He H, Yuan Q, Zhang H, Li X, Wang T, Yang Y, Yang L, Yang Y, Liu X, Wei H, Zhang H, Zhang B, Guo M, Leng Y, Shi C, Lv Y, Chen W, Wang X, Zhang Z, Yu B, Zhang B, Xu Q, Qian H, Zhou Y, Wang S, Qian Q, Shang L. Widespread inversions shape the genetic and phenotypic diversity in rice. Science Bulletin, 2024, 69(5): 593-596

[11] Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou G A, Zhang H, Liu Z, Shi M, Huang X, Li Y, Zhang M, Wang Z, Zhu B, Han B, Liang C, Tian Z. Pan-Genome of Wild and Cultivated Soybeans. Cell, 2020, 182(1): 162-176 e113

[12] Kronenberg Z N, Fiddes I T, Gordon D, Murali S, Cantsilieris S, Meyerson O S, Underwood J G, Nelson B J, Chaisson M J P, Dougherty M L, Munson K M, Hastie A R, Diekhans M, Hormozdiari F, Lorusso N, Hoekzema K, Qiu R, Clark K, Raja A, Welch A E, Sorensen M, Baker C, Fulton R S, Armstrong J, Graves-Lindsay T A, Denli A M, Hoppe E R, Hsieh P, Hill C M, Pang A W C, Lee J, Lam E T, Dutcher S K, Gage F H, Warren W C, Shendure J, Haussler D, Schneider V A, Cao H, Ventura M, Wilson R K, Paten B, Pollen A, Eichler E E. High-resolution comparative analysis of great ape genomes. Science, 2018, 360(6393): eaar6343

[13] Sedlazeck F J, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz M C. Accurate detection of complex structural variations using single-molecule sequencing. Nature Methods, 2018, 15(6): 461-468

[14] Smolka M, Paulin L F, Grochowski C M, Horner D W, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D, Scholz S W, Carvalho C M B, Proukakis C, Sedlazeck F J. Detection of mosaic and population-level structural variants with Sniffles2. Nature Biotechnology, 2024, 42(10): 1571-1580

[15] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research, 2010, 38(16): e164-e164

[16] Goel M, Sun H, Jiao W B, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biology, 2019, 20(1): 277

[17] Wang Y, Tang H, DeBarry J D, Tan X, Li J, Wang X, Lee T-h, Jin H, Marler B, Guo H, Kissinger J C, Paterson A H. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Research, 2012, 40(7): e49

[18] Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution, 2007, 24(8): 1586-1591

[19] Zhang Z. KaKs_Calculator 3.0: Calculating selective pressure on coding and non-coding sequences. Genomics, Proteomics & Bioinformatics, 2022, 20(3): 536-540

[20] Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn A F, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. InterProScan 5: Genome-scale protein function classification. Bioinformatics, 2014, 30(9): 1236-1240

[21] Cantalapiedra C P, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Molecular Biology and Evolution, 2021, 38(12): 5825-5829

[22] Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics, 2006, 22(13): 1600-1607

[23] Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE, 2011, 6(7): e21800

[24] Kuznetsova I, Lugmayr A, Siira S J, Rackham O, Filipovska A. CirGO: An alternative circular way of visualising gene ontology terms. BMC Bioinformatics, 2019, 20(1): 84

[25] Ginestet C. ggplot2: Elegant graphics for data analysis. Journal of the Royal Statistical Society Series A: Statistics in Society, 2011, 174(1): 245-246

[26] Yu G, Wang L G, Han Y, He Q Y. clusterProfiler: An R package for comparing biological themes among gene clusters. Omics, 2012, 16(5): 284-287

[27] Yang X, Zhang L, Guo X, Xu J, Zhang K, Yang Y, Yang Y, Jian Y, Dong D, Huang S, Cheng F, Li G. The gap-free potato genome assembly reveals large tandem gene clusters of agronomical importance in highly repeated genomic regions. Molecular Plant, 2023, 16(2): 314-317

[28] Yang J, Moeinzadeh M H, Kuhl H, Helmuth J, Xiao P, Haas S, Liu G, Zheng J, Sun Z, Fan W, Deng G, Wang H, Hu F, Zhao S, Fernie A R, Boerno S, Timmermann B, Zhang P, Vingron M. Haplotype-resolved sweet potato genome traces back its hexaploidization history. Nature Plants, 2017, 3(9): 696-703

[29] LIU D, Ning Z, Hong Z, YU X, Qin J, WANG L, HE S, LIU Q. AFLP fingerprinting and genetic diversity of main sweetpotato varieties in China. Journal of Integrative Agriculture, 2012, 11(9): 1424-1433

[30] Hu H, Scheben A, Wang J, Li F, Li C, Edwards D, Zhao J. Unravelling inversions: Technological advances, challenges, and potential impact on crop breeding. Plant Biotechnology Journal, 2024, 22(3): 544-554

[31] Yuan Y, Bayer P E, Batley J, Edwards D. Current status of structural variation studies in plants. Plant Biotechnology Journal, 2021, 19(11): 2153-2163

[32] Fawcett J A, Maere S, Van de Peer Y. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proceedings of the National Academy of Sciences, 2009, 106(14): 5737-5742

[33] Jiang F, Wang S, Wang H, Wang A, Xu D, Liu H, Yang B, Yuan L, Lei L, Chen R, Li W, Fan W. A chromosome-level reference genome of a Convolvulaceae species *Ipomoea* cairica. G3 (Bethesda), 2022, 12(9): jkac187