

壳斗目植物多倍化数据管理系统研究与实现

李英杰¹, 王 瑶¹, 王希胤^{1,2*}

¹华北理工大学理学院, 河北 唐山

²河北省数据科学与应用重点实验室, 河北 唐山

收稿日期: 2025年3月28日; 录用日期: 2025年4月21日; 发布日期: 2025年4月28日

摘 要

壳斗目植物生态价值与经济价值较高, 部分壳斗目植物逐渐完成测序, 并已有研究分析了多个影响壳斗目物种进化与功能创新的关键因素, 多倍化在其中起着重要作用, 对于理解壳斗目植物在生物学进化历史中的位置具有重要意义。本研究通过对壳斗目4个科12个物种的基因组数据进行分析, 构建了一个在线数据库PCGB。在对壳斗目植物多倍化过程的研究中, 使用数据库发现壳斗目植物除了被子植物共有的一次全基因组三倍事件之外, 只有胡桃科的植物单独进行过一次全基因组加倍事件, 其余科的物种并未产生多倍化事件。在对壳斗目植物进行同源结构分析、构建多基因组联合比对列表与绘制多基因组联合比对谱图后, 发现该目植物在进化过程中产生了少量且不连续的基因丢失现象, 并且个别科的物种发生了染色体重构。本研究为进一步探索壳斗目植物进化历史提供了指导, 同时为壳斗目植物的进化历程研究提供了平台。

关键词

壳斗目, 多倍化, 共线性, 同源基因

Research and Implementation of a Polyploidy Data Management System for Fagales

Yingjie Li¹, Yao Wang¹, Xiyin Wang^{1,2*}

¹College of Science, North China University of Science and Technology, Tangshan Hebei

²Hebei Key Laboratory of Data Science and Application, Tangshan Hebei

Received: Mar. 28th, 2025; accepted: Apr. 21st, 2025; published: Apr. 28th, 2025

Abstract

Fagales plants possess high ecological and economic values. The genomes of some Fagales plants

*通讯作者。

have been gradually sequenced, and previous studies have analyzed several key factors influencing the evolution and functional innovation of Fagales species, among which polyploidization plays a significant role and is crucial for understanding the position of Fagales plants in the history of biological evolution. In this study, we analyzed the genomic data of 12 species from four families of Fagales and constructed an online database named PCGB. In the study of the polyploidization process of Fagales plants, the database revealed that, apart from the whole-genome triplication event common to angiosperms, only plants in the Juglandaceae family have undergone a separate whole-genome duplication event, while species in the remaining families have not experienced any polyploidization events. After conducting homologous structure analysis, constructing a multi-genome joint alignment list, and drawing a multi-genome joint alignment spectrum for Fagales plants, it was found that there were a small number of discontinuous gene loss events during the evolutionary process of these plants, and chromosomal rearrangements occurred in some species of individual families. This study provides guidance for further exploration of the evolutionary history of Fagales plants and offers a platform for the research on their evolutionary journey.

Keywords

Fagales, Polyploidy, Collinearity, Homologous Gene

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

壳斗目(Fagales)是双子叶植物下的一个分支,壳斗目植物大多为乔木和灌木[1][2],具有单叶、互生的叶片,该目包含 33 个属约 1055 种,主要分支为胡桃科(Juglandaceae)、桦木科(Betulaceae)、壳斗科(Fagaceae)和杨梅科(Myricaceae)。壳斗目植物具有极高的经济价值和研究价值,胡桃科的胡桃和山核桃等,果实富含营养,是人们较多食用的坚果类食物,桦木科与壳斗科植物提供的木材坚硬、耐保存,是重要的建筑与家具原材料。此外,壳斗科植物的化石记录丰富,能为研究植物的演化和古气候提供重要线索。

全基因组加倍事件(Whole Genome Duplication, WGD),即多倍化(Polyploidy),是植物进化过程中一个重要的现象[3]-[5]。它指的是整个基因组的复制,使得一个物种在其基因组中拥有多个相同的基因副本。基因组加倍通常会导致物种具有更强的适应能力和更高的遗传多样性。许多植物种类,尤其是被子植物,在其进化过程中经历了多次基因组加倍事件。在植物谱系中,多倍化事件持续发生,从而形成了现今复杂且广泛的植物类群。多倍化不仅能促进新基因的生成和功能的多样化,还能为植物提供更多的表型选择,从而在面对环境变化时具有更强的适应力。

使用生物信息学方法对 12 个壳斗目植物进行鉴定,并分析了其序列、同源结构、系统发育、全基因组加倍事件,以研究壳斗目植物的同源结构和加倍事件。然后将研究结果分析汇总后构建壳斗目植物基因组多倍化数据库,为进一步研究壳斗目植物进化历程提供帮助。

2. 材料与方法

2.1. 植物基因组数据收集

根据植物基因组数据库 PlaBiPD (<https://www.plabipd.de/>)和 NCBI (<https://www.ncbi.nlm.nih.gov/>)和收录的测序植物的基因组信息,获取了 12 个组装在染色体水平上的壳斗目植物的基因组数据,详细信息如

下表所示(表 1)。

Table 1. Genomic information of representative species in the Fagales
表 1. 壳斗目代表性物种基因组信息

科	中文名	拉丁名	缩写	染色体数量
胡桃科	美国山核桃	<i>Carya illinoensis</i>	Cil	16
	胡桃	<i>Juglans regia</i>	Jre	16
桦木科	白桦	<i>Betula platyphylla</i>	Bpl	14
	川黔千金榆	<i>Carpinus fangiana</i>	Cfa	8
	榛	<i>Corylus heterophylla</i>	Che	11
	欧洲栎木	<i>Alnus glutinosa</i>	Agl	14
	欧洲水青冈	<i>Fagus sylvatica</i>	Fsy	12
壳斗科	栎	<i>Quercus lobata</i>	Qlo	12
	木姜叶柯	<i>Lithocarpus litseifolius</i>	Lli	12
	红槲栎	<i>Quercus rubra</i>	Qru	12
	夏栎	<i>Quercus robur</i>	Qro	12
杨梅科	杨梅	<i>Morella rubra</i>	Mru	8

2.2. 序列比对与同源结构分析

使用序列比对工具 BLAST+在两个不同植物基因组之间寻找相似的基因序列，期望值为 $1e-5$ ，序列匹配分数 ≥ 100 。使用共线性比对分析工具 WGDI 构建植物基因组内和基因组间的同源结构图，识别并提取基因组内保留相对顺序的共线性片段。

图 1 展示了使用 BLAST+比对美国山核桃与胡桃后的结果，第一列为基准物种美国山核桃的基因，第二列为目标物种胡桃的基因，这二者使用了数据归一化处理，每个基因的命名方式为物种名缩写、染色体号与基因编号。第三列为序列比对的一致性百分比，第四列为符合比对的比对区域的长度，第五列为比对区域的错配数，第六列为比对区域的间隔数目，第七列为比对区域在基准物种上的起始位点，第八列为比对区域在基准物种上的终止位点，第九和第十列为比对区域在目标物种上的起始与终止位点，第十一列为比对结果的期望值，最后一列为比对结果的打分值。

2.3. 基于不同事件的同源共线性片段区分

使用同源共线性软件 WGDI，搜索壳斗目植物基因组内和基因组间的同源共线性基因。要求片段内同源基因对的 E-value 小于 1×10^{-5} ，同源片段打分值大于 50，且同源片段的基因含量大于 5，从而得到壳斗目物种基因组内和基因组间比较准确的同源共线性片段[6]。

为了区分不同加倍事件或物种分化形成产生的共线基因块，使用 WGDI 软件的 Ks 模块计算了同义核酸取代率(Ks)，用于表示同源基因之间的差异水平。通过 bi 模块整合共线性和 Ks 值结果，然后根据已知的全基因组加倍事件，使用 c 模块对同源共线性片段进行过滤，保留经历多倍化事件的共线性片段。最终绘制了使用 Ks 值标记了不同进化事件的同源结构图(图 2(A))、共线性区块上 Ks 的拟合分布图(图 2(B))和共线性区块的核密度估计图(图 2(C))。

Cil1g0002	Jre7g0001	94.889	763	39	0	1	763	1	763	0.0	1436	
Cil1g0002	Jre12g0004	86.387	764	103	1	1	763	1	764	0.0	1333	
Cil1g0002	Jre12g0005	60.406	394	148	2	102	495	86	471	1.41e-164		489
Cil1g0002	Jre10g0588	39.556	541	306	12	125	655	104	633	1.34e-107		343
Cil1g0002	Jre1g0869	39.114	542	307	13	125	655	116	645	5.30e-104		334
Cil1g0002	Jre12g1316	38.674	362	213	5	129	487	104	459	1.21e-74		250
Cil1g0002	Jre4g1498	38.933	375	210	7	128	487	163	533	1.14e-72		248
Cil1g0002	Jre7g2125	38.122	362	215	5	129	487	108	463	1.68e-72		245
Cil1g0002	Jre3g0255	40.947	359	183	8	149	487	183	532	1.67e-71		245
Cil1g0002	Jre1g0350	38.889	360	204	6	124	477	130	479	1.05e-70		245
Cil1g0002	Jre14g1657	38.564	376	211	7	128	487	163	534	4.44e-70		241
Cil1g0002	Jre10g0243	37.598	383	222	7	130	506	174	545	7.58e-70		243
Cil1g0002	Jre1g2295	38.904	365	200	9	128	476	134	491	1.50e-68		237
Cil1g0002	Jre9g0309	38.056	360	214	5	131	487	248	601	6.57e-68		236
Cil1g0002	Jre2g1027	36.744	430	224	13	130	536	122	526	9.42e-68		233
Cil1g0002	Jre5g0482	42.136	337	180	5	150	476	123	454	1.22e-67		231
Cil1g0002	Jre7g0929	35.641	390	201	4	130	477	507	888	7.49e-67		240
Cil1g0002	Jre16g1079	39.353	371	182	9	131	472	303	659	5.45e-64		227
Cil1g0002	Jre15g0244	36.612	366	211	8	129	488	45	395	1.37e-62		215
Cil1g0002	Jre6g0290	36.612	366	211	8	129	488	45	395	7.73e-62		213

Figure 1. Display of sequence alignment results

图 1. 序列比对结果展示

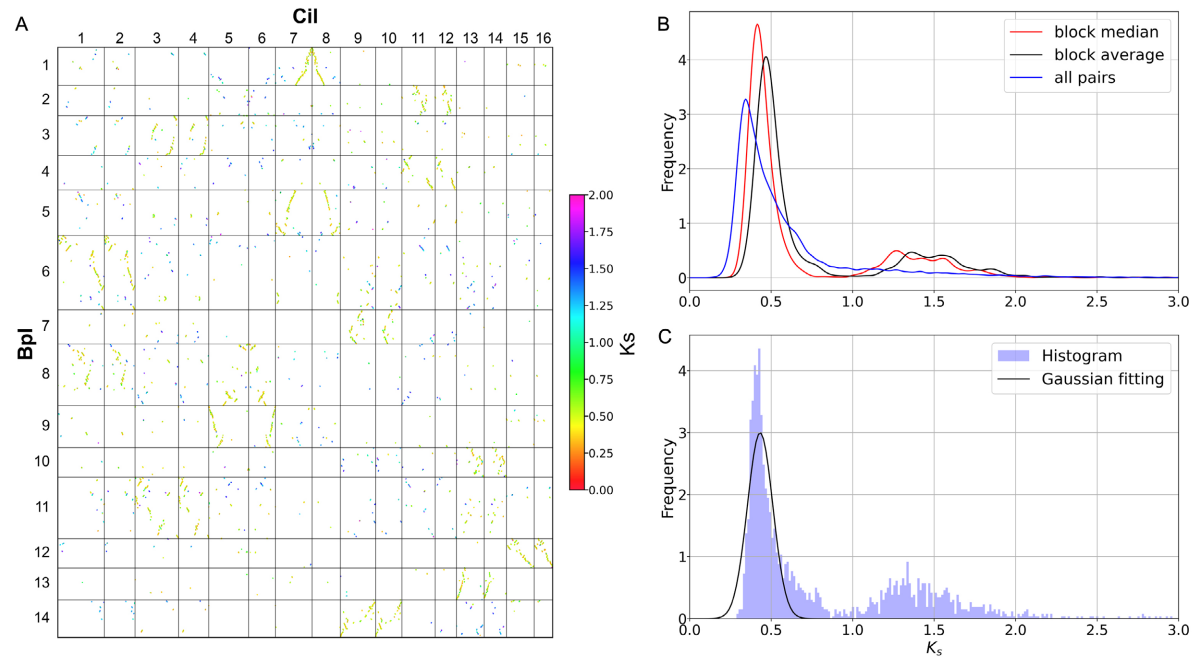


Figure 2. Ks distribution diagram

图 2. Ks 分布图

2.4. 客户端架构设计

客户端采用分层架构和模块化设计，分为用户界面层(UI Layer)、业务逻辑层(Business Logic Layer)和数据访问层(Data Access Layer)。用户界面层基于 Vue.js 构建，采用响应式设计；业务逻辑层通过块管理器协调各功能模块，执行数据处理和业务逻辑操作；数据访问层负责与本地数据库和外部 API 交互，

获取和存储数据。客户端结合 Vuex 进行状态管理，使用 D3.js 和 ECharts 实现数据可视化，并通过 Axios 与服务端通信。整体设计注重可扩展性和可维护性，为用户提供高效、直观的交互体验。

2.5. 服务端架构设计

服务端采用 Spring Boot 框架构建，提供高效、稳定的数据支持和业务处理能力。数据访问层负责与 MySQL 数据库交互，执行数据的增删改查操作，并通过 Redis 缓存机制提升性能。业务逻辑层封装了数据归一化、序列比对、共线性分析等复杂计算任务，支持异步处理以优化响应速度。接口层以 RESTful API 形式对外提供服务，支持多种数据格式交互，并通过身份验证和授权机制保障数据安全。此外，集成日志管理和监控系统，实时监控性能和异常，确保系统的高可用性和可维护性。

3. 结果与分析

3.1. 数据库概述

PCGB 在 NCBI 等平台收集原始数据，对数据进行标准归一化处理，进行序列比对分析，使用共线性分析程序构架多重比对列表，并对其进行可视化展示。共设计了 6 大模块：首页、物种信息模块、同源结构分析模块、多倍化分析模块、多倍化联合比对图谱模块以及下载模块。本研究整合上述模块的功能需求，最终构建了整体的系统架构(图 3)。

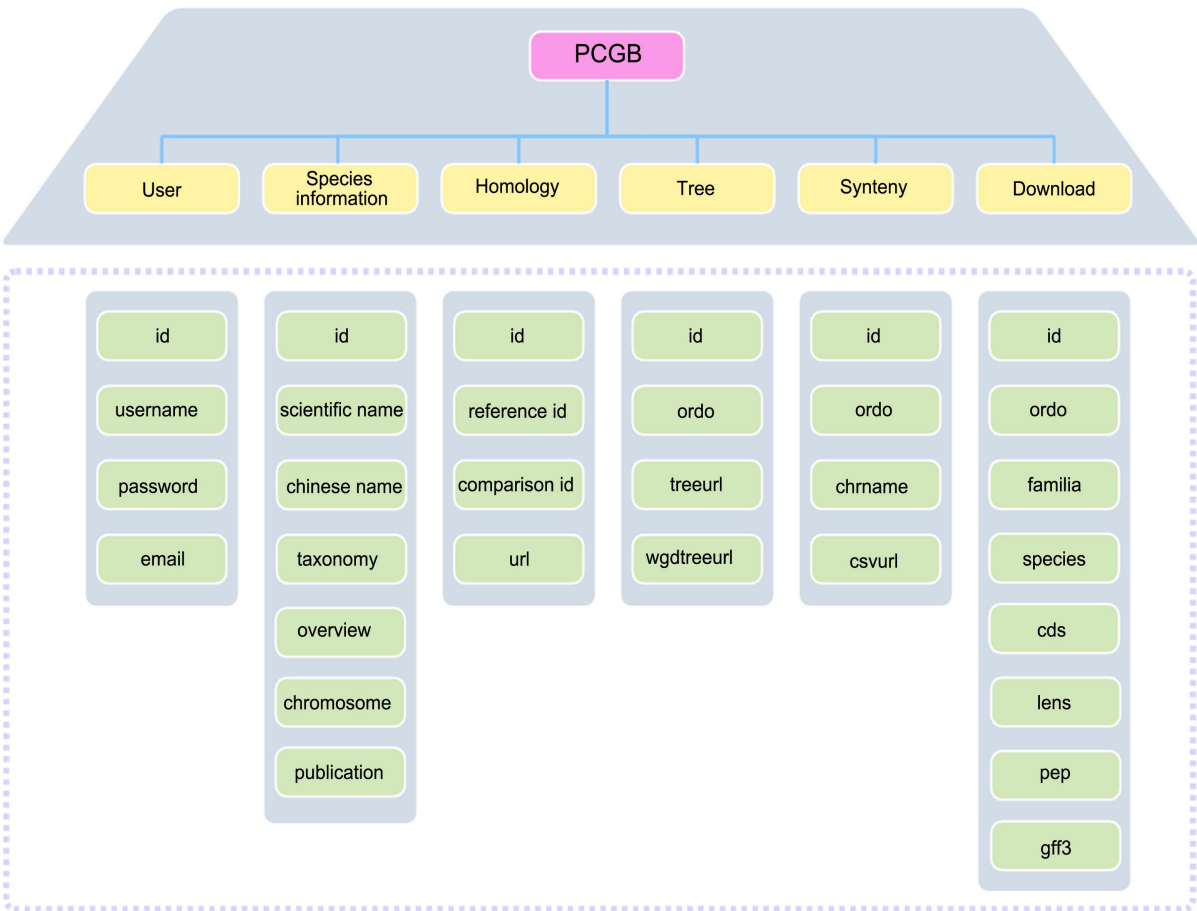


Figure 3. Database system architecture diagram
图 3. 数据库系统架构图

3.2. 首页

PCGB 提供了全面且用户友好的支持,用于分析植物基因组之间以及基因组内部的多重比对关系,通过推断每个基因组内以及基因组之间的共线性基因,下图展示了数据管理系统的首页(图 4)。

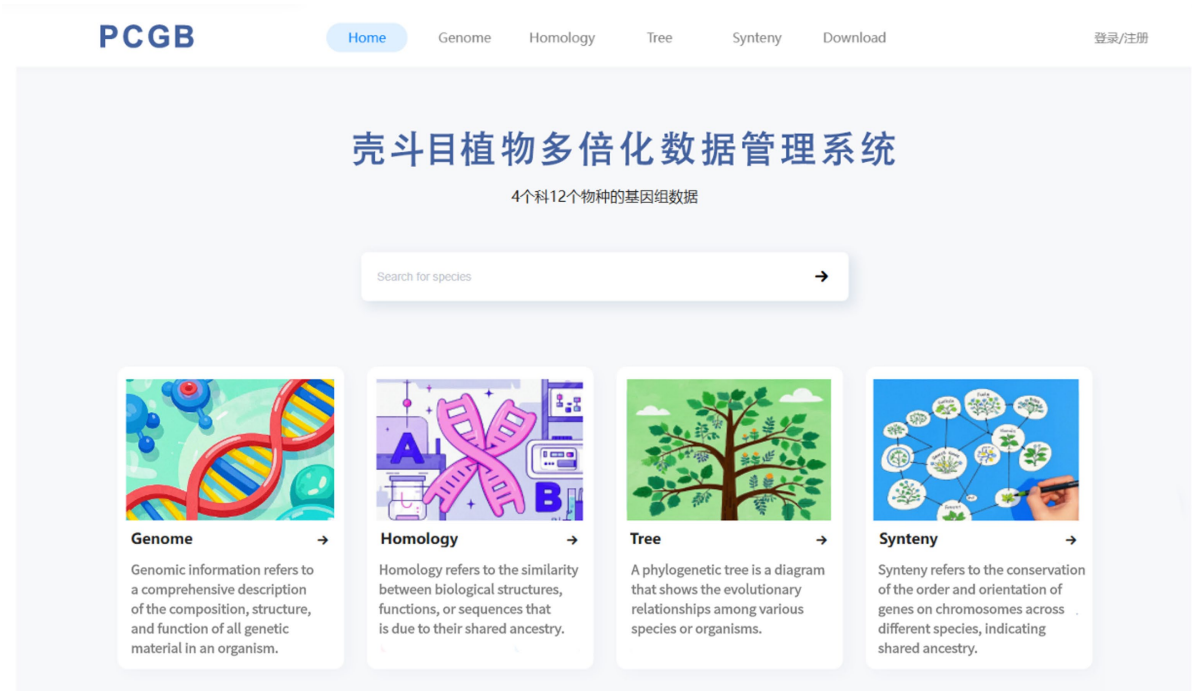


Figure 4. Management system home page
图 4. 管理系统首页

3.3. 物种信息页

对于每个物种,我们提供一个网页,展示基本信息,包括其拉丁学名、常见名称、中文名称、图片、分类学分类以及诸如地理分布、生物学特征和栖息地偏好的详细资料。该页面还包括染色体数目、出版参考文献以及其他相关数据,下图展示了壳斗目每个物种包含的染色体数量和基因数量(图 5)。

3.4. 同源结构分析模块

同源基因点阵图可以使用推断的共线性基因信息绘制。通过显示这些共线性基因之间的 K_s 值(即同义核苷酸位点上的同义核苷酸替代估算值),可以方便地区分由不同进化事件产生的同源基因块,如共享的古老多倍体化、特定的多倍体化影响某一植物或植物亚群体,以及物种形成。

在这个模块中,用户可以选择一个参考物种,然后选择一个比较物种,以查看它们之间的比对结果。参考物种通常是正在研究的关键植物,它可能是另一个物种的外群,或者是经历较少多倍体化事件的物种,因此基因组结构更为简单。科学地选择参考物种将有助于探索所研究的基因组是如何构成、形成、(重新)塑造和进化的。

下图展示了白桦(*Betula platyphylla*)与榛(*Corylus heterophylla*)的同源基因点阵图(图 6),该图显示:白桦与榛的加倍对应关系为 1:1,并且大多数染色体对应完整,只有白桦的 9 号染色体进行了重组,分别对应上了榛的 1 号与 4 号染色体,其他染色体的重组与基因丢失现象发生不明显,说明这两个物种进化关系较为亲近。

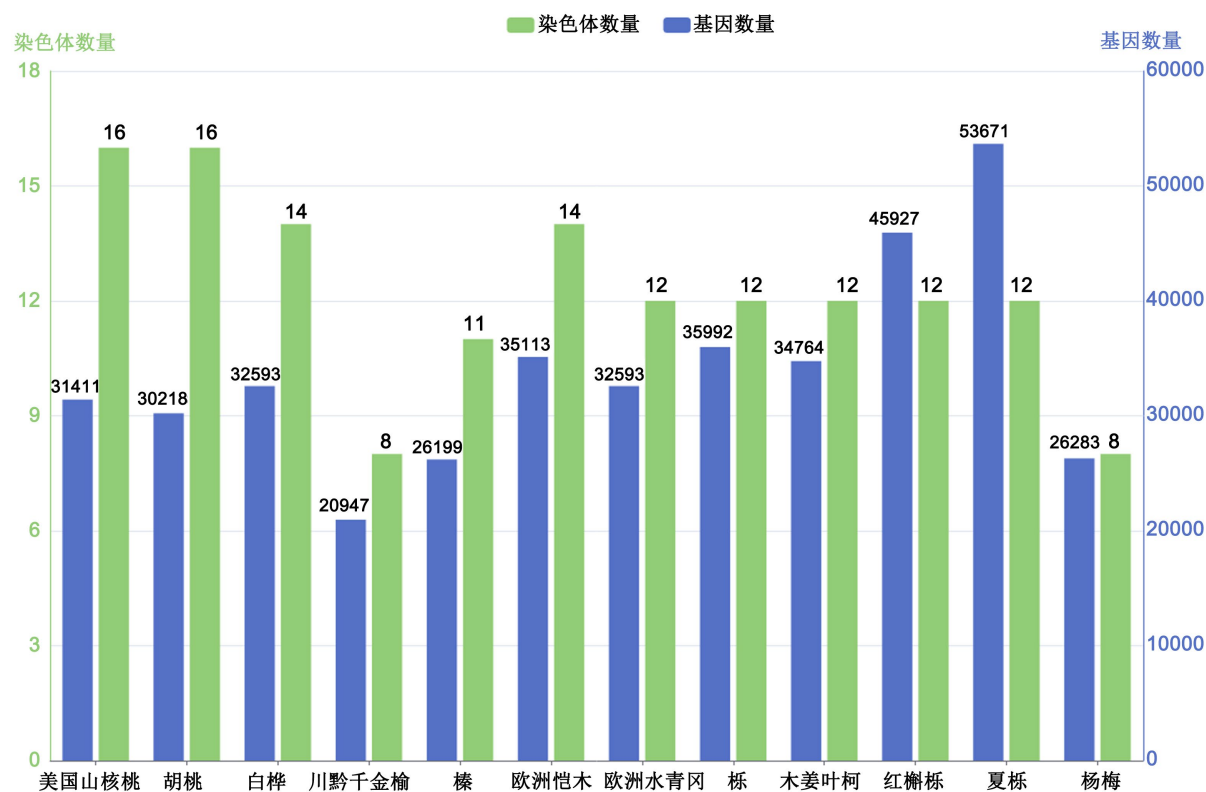


Figure 5. Fagales plant genome information diagram
图 5. 壳斗目植物基因组信息图

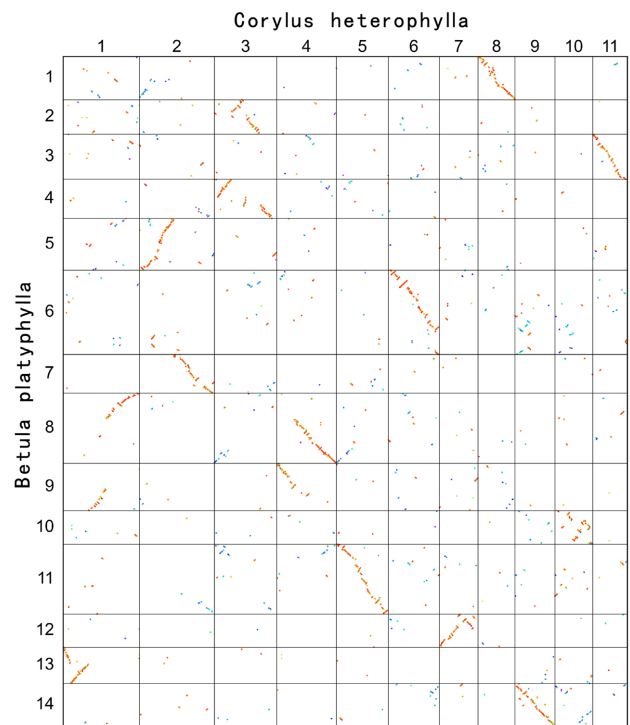


Figure 6. Dotplot of Betula platyphylla and corylus heterophylla homologous genes
图 6. 白桦与榛同源基因点阵图

3.5. 多倍化分析模块

该模块的第一个功能构建了壳斗目系统发育树(图 7)，展示了所考虑物种在分类群内的关系，以及同源基因之间的关系[7]。当用户选择一个分类群时，模块会生成树状图：显示所考虑物种之间的关系，并展示影响该分类群部分植物的多倍体化事件。

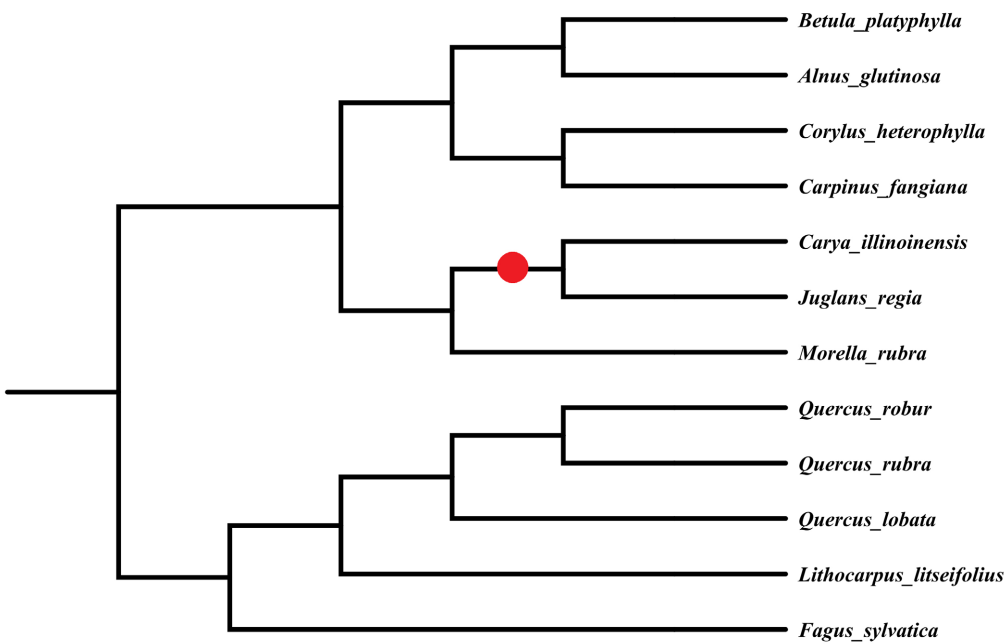


Figure 7. Phylogenetic dendrogram of the Fagales
图 7. 壳斗目系统发育树图

第二个功能生成了研究分类群的基因共线性列表(表 2)。用户可以搜索相应的同源基因，并展示染色体级别的相关基因共线性。基因共线性列表展示了每个基因的原始 ID (数据源中使用的 ID)、修改后的 ID (格式化以便进行跨基因组分析)以及基因在染色体上的物理位置。此外，我们还提供了下载 CSV 文件的功能，方便用户在分类群中进行自己的基因共线性分析。

Table 2. Fagales partial genome alignment list
表 2. 壳斗目部分基因组联合比对列表

Bpl	Jre	Jre	Cil	Cil	Cfa	Che	Mru	Agf	Fsy
Bpl1g0001									
Bpl1g0002		Jre12g0225							
Bpl1g0003				Cil2g0230					
Bpl1g0004		Jre12g0226							
Bpl1g0005	Jre7g0411		Cil1g0452						
Bpl1g0006					Cfa5g1054	Che8g0051	Mru8g3077	AGL6g0064	
Bpl1g0007	Jre7g0412		Cil1g0454		Cfa5g1057	Che8g0053	Mru8g3078	AGL6g0068	Fsy4g2550
Bpl1g0008			Cil1g0455					AGL6g0069	
Bpl1g0009	Jre7g0413								

3.6. 多倍化联合比对图谱模块

多倍化联合比对图谱模块提供了一种方法来检查一个分类群内的古老多倍体化[8]。它为观察物种间基因共线性关系提供了直观的方式。用户可以查看分类群内不同物种染色体之间的基因共线性。图表数据展示了基于基因共线性的对齐染色体，这使得用户能够无缝访问跨基因组比对的详细信息，其中包括事件相关的同源基因，涵盖旁系同源基因和直系同源基因，从而帮助用户研究他们感兴趣的特定功能基因。

下图展示了壳斗目全基因组多重比对联合图谱(图 8)，图中的每层圈代表一个植物的基因组，由内至外共 14 圈，解释了 12 种植物的基因组，其中胡桃(*Juglans regia*, 缩写 Jre)和美国山核桃(*Carya illinoensis*, 缩写 Cil)由于均经历过一次全基因组加倍事件，因此这两个物种均有两圈，其余物种没有经历过全基因组加倍事件。图中显示：这 12 个物种的全基因组对应较为完整，没有发现大片段的基因丢失，但每个物种的基因组均出现了小片段且不连续的基因组空缺现象，这说明壳斗目这一类群在进化过程中虽然经历过功能进化但没有经历根源结构上的改变。

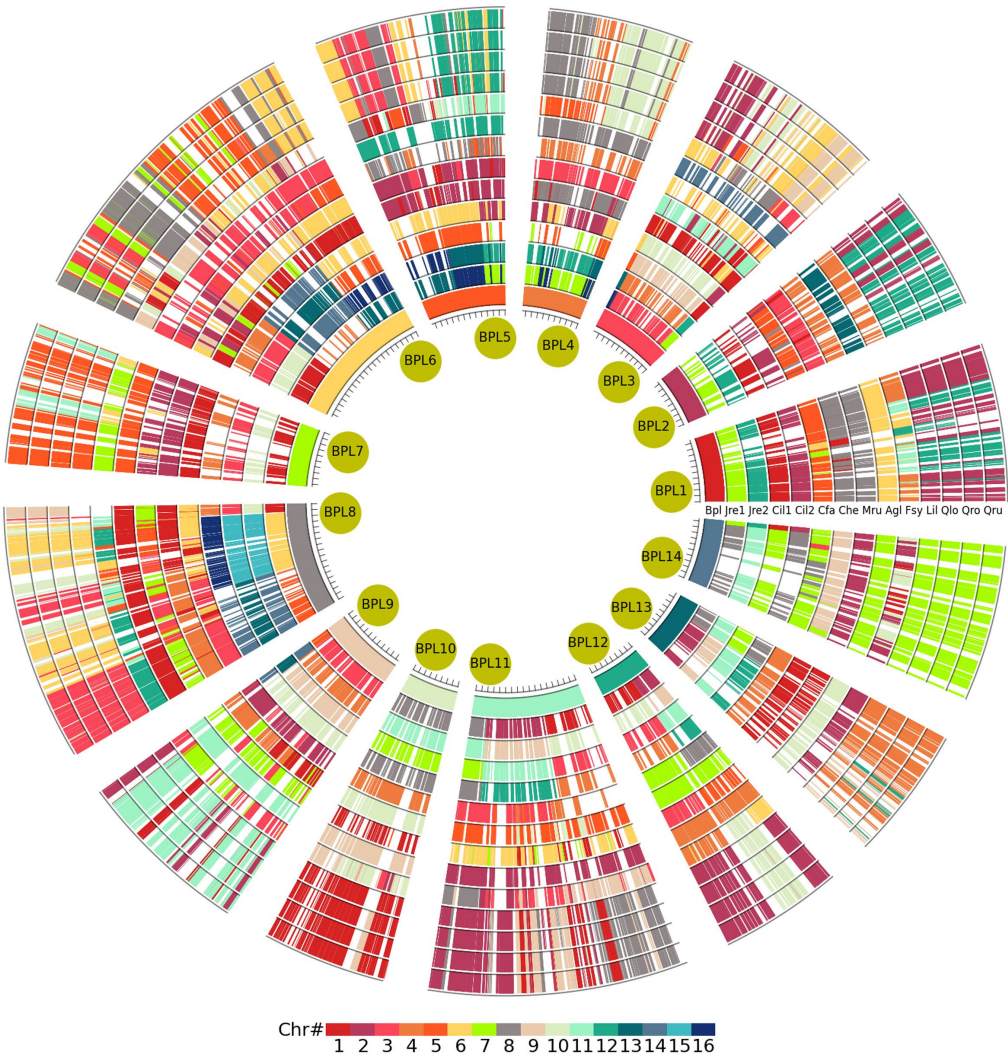


Figure 8. Genome-wide joint alignment map of Fagales
图 8. 壳斗目植物全基因组联合比对图谱

3.7. 下载模块

在下载页面，我们提供了涉及 PCGB 数据库的 12 种植物的原始数据，具体为三类文件：基因组注释文件、基因组序列文件和蛋白质序列文件，用户可以根据需求下载数据并进行后续分析。

4. 讨论

随着高通量测序技术以及大数据技术的发展，越来越多的植物基因组测序不断完成，为科研人员研究植物的进化历史提供了数据支持。多倍化在植物进化过程中起着至关重要的作用，经历多倍化后的植物基因组，会出现染色体断裂、异常重组、基因位移和丢失等现象，为植物进化创新和物种多样性提供了新的土壤。

壳斗目植物作为双子叶植物的一个分支，是森林生态系统中的建群种和优势种，为森林生态系统的可持续发展提供了支撑，包括土壤保持、水源涵养以及生物多样性的维持。此外，壳斗目植物作为重要的木材来源，在建筑与家具行业中具有极高的经济价值，其木材质地坚硬、纹理美观，是建筑用材和家具制造的优质原料，为木材产业提供了丰富的物质基础，使其成为建筑与家具行业可持续发展的关键资源。在此，收集了 4 个科 12 种壳斗目植物，并对其进行同源结构分析和多倍化事件判断，最后构建联合比对图谱，构建了壳斗目植物多倍化数据管理平台 PCGB，为研究这些植物的直系同源与旁系同源关系提供了方法。

我们的数据库提供了直观的工具，帮助用户在全局和局部尺度上进行多基因组联合比对与可视化检索，并将多倍化事件绑定到对应的系统发育树上，为深入研究基因的进化和功能创新提供了基础。

参考文献

- [1] Siniscalchi, C.M., Correa-Narvaez, J., Kates, H.R., *et al.* (2023) Fagalean Phylogeny in a Nutshell: Chronicling the Diversification History of Fagales.
- [2] Folk, R.A., Siniscalchi, C.M., Doby, J., *et al.* (2023) Spatial Phylogenetics of Fagales: Investigating the History of Temperate Forests.
- [3] Yang, Y.Y., Stull, G.W., Qu, X.J., *et al.* (2023) Genome Duplications, Genomic Conflict, and Rapid Phenotypic Evolution Characterize the Cretaceous Radiation of Fagales.
- [4] Ding, Y., Pang, X., Cao, Y., Zhang, W., Renner, S.S., Zhang, D., *et al.* (2023) Genome Structure-Based Juglandaceae Phylogenies Contradict Alignment-Based Phylogenies and Substitution Rates Vary with DNA Repair Genes. *Nature Communications*, **14**, Article No. 617. <https://doi.org/10.1038/s41467-023-36247-z>
- [5] She, R., Chu, J.S., Uyar, B., Wang, J., Wang, K. and Chen, N. (2011) Genblastg: Using BLAST Searches to Build Homologous Gene Models. *Bioinformatics*, **27**, 2141-2143. <https://doi.org/10.1093/bioinformatics/btr342>
- [6] Wray, G.A. (2007) Evolutionary Dissociations between Homologous Genes and Homologous Structures. In: *Novartis Foundation Symposium 222-Homology: Homology: Novartis Foundation Symposium 222*, John Wiley & Sons, 189-206.
- [7] Dufayard, J., Duret, L., Penel, S., Gouy, M., Reichenmann, F. and Perriere, G. (2005) Tree Pattern Matching in Phylogenetic Trees: Automatic Search for Orthologs or Paralogs in Homologous Gene Sequence Databases. *Bioinformatics*, **21**, 2596-2603. <https://doi.org/10.1093/bioinformatics/bti325>
- [8] Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., *et al.* (2012) Mcscanx: A Toolkit for Detection and Evolutionary Analysis of Gene Synteny and Collinearity. *Nucleic Acids Research*, **40**, e49. <https://doi.org/10.1093/nar/gkr1293>