Genetic and Geographic Patterns of Woody Plant Biosynthesis

Muhamad Fahmi[1*], Kojiro Takanashi[2,3], Yusuke Kakei[4], Yasuhiro Kubota[5,6] and Keiichiro Kanemoto[1,7,8*]

[1]Research Institute for Humanity and Nature, 457-4 Motoyama, Kamigamo, Kita-ku, Kyoto, 603-8047, Kyoto, Japan.
[2]Department of Science, Graduate School of Science and Technology, Shinshu University, 3-1-1 Asahi, Matsumoto, 390-8621, Nagano, Japan.
[3]Department of Biology, Faculty of Science, Shinshu University, 3-1-1 Asahi, Matsumoto, 390-8621, Nagano, Japan.
[4]The Institute of Vegetable and Floriculture Science, National Agriculture and Food Research Organization, 3-1-1 Kannondai, Tsukuba, 305-8517, Ibaraki, Japan.
[5]Faculty of Science, University of the Ryukyus, 1, Sembaru, Nishihara, 903-0213, Okinawa, Japan.
[6] Think Nature Inc., 1, Sembaru, Nishihara, 903-0213, Okinawa, Japan
[7]Graduate School of Environmental Studies, Tohoku University, Aoba, 468-1, Aramaki, Aoba-ku, Sendai, 980-8572, Miyagi, Japan.
[8] Research Center for Social Systems, Shinshu University, Matsumoto 390-8621, Japan

*Corresponding author(s). E-mail(s): fahmi@chikyu.ac.jp, keiichiro.kanemoto@gmail.com;
Contributing authors: fahmi@chikyu.ac.jp; takanashi@shinshu-u.ac.jp; kakeiy197@affrc.go.jp; kubota.yasuhiro@gmail.com;

Abstract

The conservation of biodiversity is important for a number of reasons, one of which being drug discovery. Such conservation would benefit from better understanding of the geographical and genetic patterns of natural product (NP) biosynthesis in

plants. Here, we integrate publicly accessible next-generation sequencing and species distribution data of 1,612 woody plant species and their 166 enzymes and establish critical metrics to highlight specific enzyme-rich regions. Our results show that species with patulone-forming prenyltransferase, which catalyzes patulone, a potential bioactive compounds in treating disease mediated by platelet-activating factor, are mainly distributed in the South Central region of the United States, while those with enzymes associated with shikonin derivatives biosynthesis, which are associated with potent anti-cancer activity, are mainly found in certain regions of Asia. Those species distribution patterns are different from the species richness. Moreover, we analyse the complex interplay between numerous enzymes in biosynthetic pathways using shikonin and benzylisoquinoline alkaloids biosynthesis as examples. We detect contrasting geographical distribution patterns, with shikonin derivatives biosynthesis confined to specific regions and benzylisoquinoline alkaloids biosynthesis more globally distributed. These findings suggest policymakers adopt biodiversity conservation policies based on not only the species richness and threat level but also the genetic resources embedded in species. The maps presented in this study will help understand and set such policies.

Keywords: Biodiversity, Woody Plants, Natural Products, Genetic Diversity

1 Introduction

Natural products (NPs), especially those derived from plants, have served as the cornerstone of human medicinal practices for millennia [1]. The earliest evidence of such use dates back to Mesopotamia around 2600 BCE [2, 3]. NPs are valuable to the pharmaceutical and biotechnology industries, who have produced numerous modern medicines from them [4]. Indeed, by the middle of the 1970s, NPs were the main source of new therapies in the pharmaceutical drug pipeline. Between the 1980s and the 2010s, unmodified NPs (5%), NP analogues (28%), and NP pharmacophores (35%) accounted for two-thirds of all modern medications [5].

Currently, the pharmaceutical landscape leans more towards semi-synthetic and fully synthetic drugs, reducing their reliance on NPs [2]. Despite this shift, the genetic resources integral to NP biosynthesis continue to hold substantial value for drug discovery. These genetic resources, consisting of enzymes and biosynthetic pathways involved in the formation of NPs, are key tools that allow us to explore, recreate, and improve natural chemical structures. Even for synthetic drugs, knowledge derived from the study of NP biosynthesis can guide the design of new molecules and the development of more efficient production methods [6].

A multitude of genetic resources play an indispensable role in shaping the future of healthcare and advancing drug discovery. For instance, the cytochrome P450 monooxygenase family (CYP enzymes) stands out for its remarkable versatility and breadth. These heme-containing proteins catalyze a wide range of oxidation reactions during the post-assembly phase of NPs [7, 8]. The importance of CYP enzymes is exemplified by CYP725A4, which is involved in the biosynthesis of taxol, a potent anti-cancer agent derived from the Pacific Yew tree (*Taxus brevifolia*) [9]. Similarly, CYP71AV1 contributes to the biosynthesis of artemisinin, an effective anti-malarial drug from the sweet wormwood plant (*Artemisia annua*) [10]. Beyond their crucial roles in natural biosynthesis, these enzymes have been harnessed for metabolic engineering, utilizing heterologous microorganisms for the renewable production of drugs [11]. Importantly, these examples reiterate the importance of protecting biodiversity, as it harbors genetic resources for future drug discovery.

The amount and quality of genetic resources for drug discovery depend heavily on functioning natural ecosystems [12]. Three decades of research have demonstrated that biodiversity loss can impede such functioning, and the current dominance of humans over global ecosystems has resulted in a significant rise in the rate of extinction [12]. Consistently, assessments by the International Union for Conservation of Nature (IUCN) Red List revealed over 41,000 species are currently at risk of extinction [13]. Moreover, according to one study, between 2000 and 2010, 5.96% of worldwide forest cover was lost [14]. Deforestation was particularly prevalent in tropical regions, with South American rainforests experiencing nearly half of the loss. Another study found that 27% of the loss was due to commodity-driven deforestation, with steady rates shifting geographically from Brazil to other tropical forests in Latin America and Southeast Asia [15].

Given this considerable threat and limited conservation resources, urgent and strategic conservation efforts are vital. However, before such actions are taken, an understanding of the present state of biodiversity, the likely response by biodiversity to global environmental change, and the best conservation practices per region are needed. Many studies have already investigated species richness from local to global dimensions [16–20]. Recently, a metric called Species Threat Abatement and Recovery (STAR) was created to evaluate the potential contribution of threat mitigation and habitat restoration in a given location towards global biodiversity objectives [21]. By enhancing conservation strategies, genetic diversity assessments, mostly using mitochondrial markers, have been combined with phylogenetic and functional diversity to better understand the adaptability and evolutionary potential of species and ecosystems [22–26]. While these studies and metrics share the ultimate aim of conserving ecosystem services in general, none have focused on drug discovery. In this study, we address this gap by uncovering the global pattern of genetic resources to gain insights on targeted conservation strategies.

2 Results

To gain insight into regions with a high potential of genetic resources for drug discovery, we first mapped species richness, genetic data coverage, and the potential yield of genetic resources for woody plants (Fig. 1). These metrics leveraged the genetic and distribution data of approximately 1,612 and 82,974 distinct species, respectively, from publicly available databases. The results show a general distribution of biodiversity; biodiversity hotspots are located near the equator, particularly in the South American Amazon, East Africa, and Southeast Asia (Fig. 1A). However, areas with abundant species richness lack adequate genetic data representation (Fig. 1B). Moreover, the average number of enzymes linked to drug discovery shows a surprisingly even distribution (Fig. 1C). The overall drug discovery potential, therefore, corresponds almost exactly to the species richness (Fig. 1D). This finding suggests complex interplay between enzymes in a biosynthetic pathway and implies a limitation in the sensitivity of the metric for this study. As such, it becomes apparent that aggregate averages of enzyme markers may not provide a detailed representation of NP diversity, necessitating the development of refined metrics to elucidate the intricate nature of NP biosynthesis.

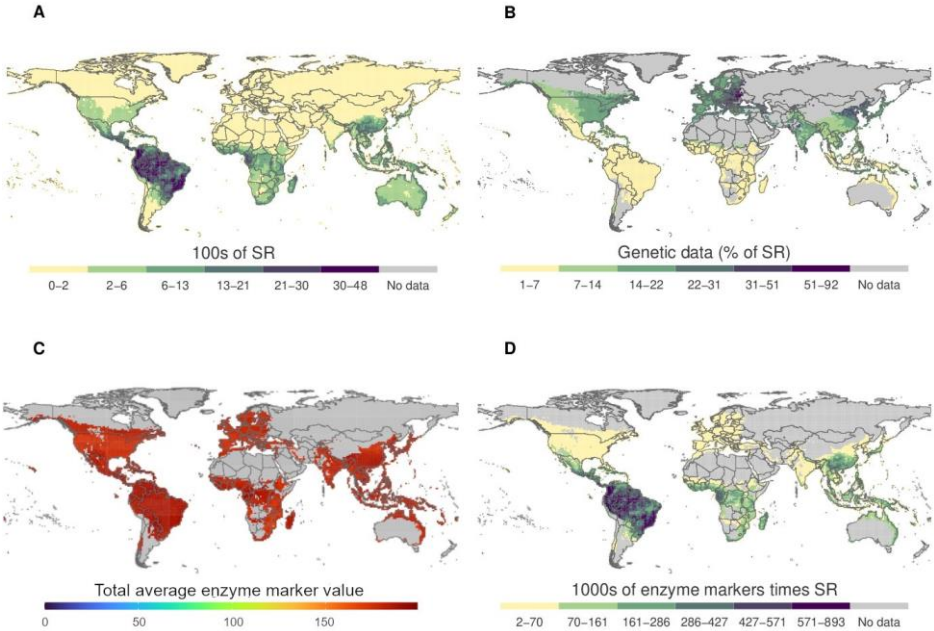

Fig. 1 Global Pattern of Woody Plant Species Richness and Genetic Diversity. (A) The predicted species richness of woody plants from 89,826 species. (B) The percentage of species possessing genetic data in relation to species richness (SR). (C) The aggregate average of enzyme markers for evaluating natural product diversity. (D) The product of average enzyme markers and species richness.

Given the observed uniform distribution of aggregate enzyme marker averages, we decided to delve deeper into the distribution patterns of individual markers across the 166 enzymes (Supplementary table 1). The rationale behind this approach was to pinpoint any key enzyme that might show an uneven distribution across diverse geographical regions, since medicinal uses vary from enzyme to enzyme. Such patterns may suggest unique biosynthetic capabilities, the potential diversity of NPs, and guide regional conservation policies within the areas. Notably, several enzyme markers, including UDP glycosyltransferase 82A1 (UGT82A1), 2-oxoglutarate-dependent dioxygenaselike protein (DOXC20), CYP702A5, 8-prenyl-1,3,6,7-tetrahydroxyxanthone 8-prenyltransferase (PTpat), UGT83A1, and CYP710A1 exhibit high geographical heterogeneity scores, indicating a disparate distribution (Fig. 2), which is unlike the even distribution shown in Fig. 1C. The remarkable geographical heterogeneity observed in the distribution of specific enzymes may hold profound implications for the spatial patterns of the corresponding NPs they synthesize.

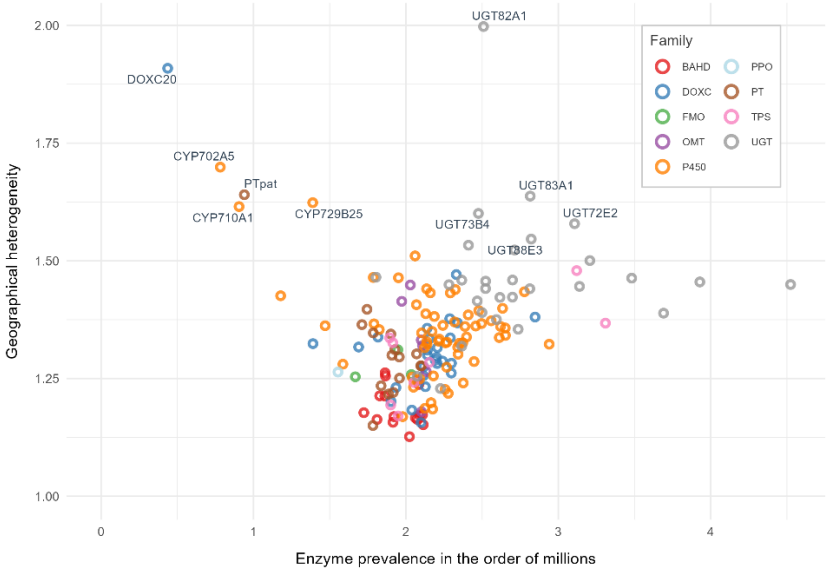

Fig. 2 Geographical Heterogeneity and Prevalence of Enzyme Markers. The heterogeneity of each enzyme across different geographical regions was ranked by calculating the ratio of the maximum to the mean of the average enzyme count across all grids. Concurrently, the prevalence of the enzymes was estimated by the product of the total number of enzymes and the species count across all grid regions. The enzyme families are colour-coded, with each scatter point colour representing a different family.

While the functional roles of UGT82A1 (Fig. 3A), CYP702A5, and DOXC20 still warrant investigation, the biological attributes of UGT83A1, PTpat, and

CYP710A1 (Fig. 3B, C, and D) have been explored in previous research. PTpat is involved in the biosynthesis of patulone, a compound noted for its anti-inflammatory properties [27]. UGT83A1 has been implicated in the biosynthesis of sinapoyl anthocyanin, a compound with potent antioxidant properties [28, 29]. In addition, CYP710A1 is involved in the synthesis of stigmasterol from β-sitosterol via C22 desaturation. Stigmasterol is suggested to possess anticancer and anti-inflammatory properties [30, 31].

Conversely, we can also find several enzymes that are linked to potential lead compound biosynthesis and show even geographical distribution. These include, for example, CYP719A23 and *Taxus cuspidata* Baccatin III:3-amino-3-phenylpropanoyltransferase (TcBAPT) (Fig. 3E and F). The CYP719 family has a distinct ability to produce a methylenedioxy ring, a characteristic chemical structure in etoposide, which is used to treat various cancers [32, 33]. The methylenedioxy rings catalyzed by the CYP719 family are also used in other bioactive compounds including piperine and berberine [34, 35]. Meanwhile, TcBAPT is involved in paclitaxel biosynthesis, a widely-used chemotherapy medication in cancer treatment.

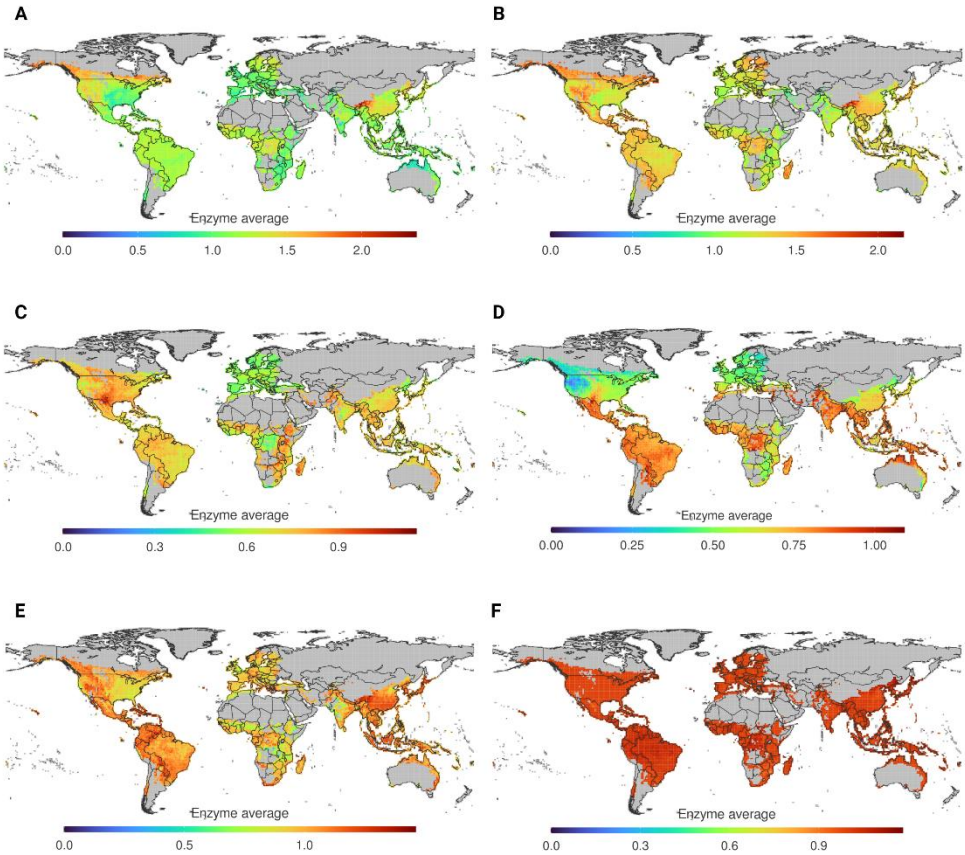

Fig. 3 Global Patterns of Selected Enzymes Showing Geographical Heterogeneity and Even Distribution, including (A) UGT82A1, (B) UGT83A1, (C) PTpat, (D) CYP710A1 (showing heterogeneity), (E) CYP719A23, and (F) TcBAPT (showing even distribution).

The limitations observed in our initial strategy led us to adopt a more targeted approach for subsequent analyses. By focusing on well-known, specific groups of enzymes responsible for the biosynthesis of two distinct NPs, we aimed to distinguish subtler details that may be overlooked when analysing aggregate enzyme markers and individual enzyme averages. In this analysis, shikonin and benzylisoquinoline alkaloids were chosen due to their high pharmacological significance and distinctive biosynthetic pathways (Supplementary table 2) [36, 37]. Assessing the average richness of enzymes involved in the biosynthesis of these particular compounds provides a finer comprehension of the potential for specific NPs in a given region. Additionally, this method may assist in identifying regions that, although not necessarily characterized by high species diversity, possess a high potential for drug discovery.

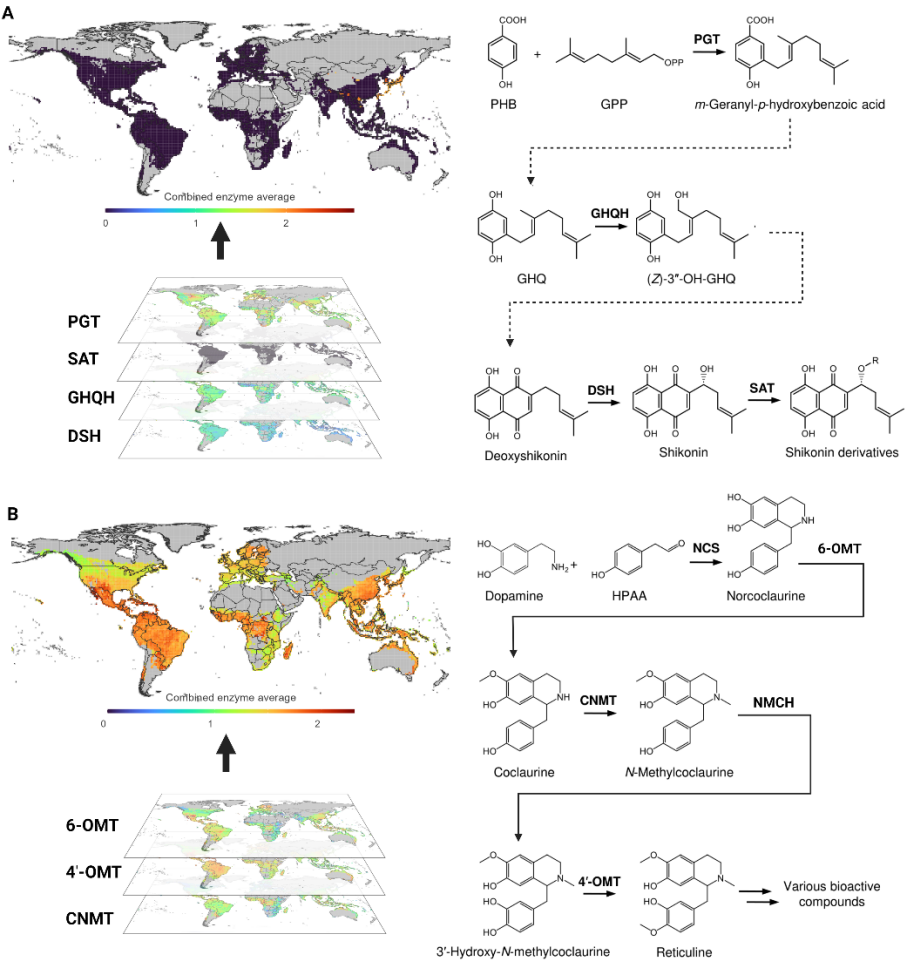

Fig. 4 Global Pattern of Enzymes Associated with Shikonin Derivatives and Benzylisoquinoline Alkaloids Biosynthetic Pathways. (A) Geographic distribution of enzymes linked to shikonin derivatives biosynthesis (left) and the corresponding biosynthetic pathway (right). The distributions are primarily concentrated in Japan, South Korea, Taiwan, Northern Vietnam, and areas around the Himalayas, suggesting region-specific ecological or evolutionary influences. The map was developed by overlaying the distributions of PGT, SAT, GHQH, and DSH. Dashed arrows in the biosynthetic pathway indicate reaction steps for which enzymes have not yet been identified. (B) Geographic distribution of enzymes associated with benzylisoquinoline alkaloids biosynthesis (left) and the corresponding biosynthetic pathway (right). The distribution was more widespread than that of shikonin. The map was developed by overlaying the distributions of 6-OMT, CNMT, and 4′-OMT. PHB, p-hydroxybenzoic acid; GPP, geranyl diphosphate; GHQ, geranylhydroquinone; (Z)-3′′′′-OH-GHQ, (Z)-3′′′′-hydroxy-geranylhydroquinone; PGT, PHB geranyltransferase; GHQH, geranylhydroquinone hydroxylase; DSH, deoxyshikonin hydroxylase; SAT, shikonin O-acyltransferase; HPPA, 4-hydroxyphenylacetaldehyde; NCS, norcoclaurine synthase; 6-OMT, norcoclaurine 6-O-

methyltransferase; CNMT, coclaurine *N*-methyltransferase; NMCH, *N*-methylcoclaurine 3′′-hydroxylase; 4′-OMT, 3′-hydroxy-*N*-methylcoclaurine 4′′-O-methyltransferase.

Our analyses showed that employing a targeted strategy allows for a more refined comprehension of the NP biosynthesis. For instance, the distribution of shikonin derivatives, a potent natural compound, presents an intriguing geographical constraint, with shikonin biosynthesis being concentrated predominantly in specific regions (Fig. 4A). These regions encompass Japan, South Korea, Taiwan, northern Vietnam, and areas along the Himalayan range. This pattern may reflect distinct evolutionary adaptations or specific ecological niches inhabited by plant species in these regions that promote the biosynthesis of shikonin derivatives. The confinement of shikonin biosynthesis to these locales emphasizes the crucial role of biodiversity conservation in these specific regions.

In contrast, the distribution of benzylisoquinoline alkaloids do not exhibit such pronounced heterogeneity (Fig. 4B). Instead, its biosynthesis is more evenly distributed across different regions globally. This broad geographic distribution may be indicative of a more generalized ecological adaptation or a wider range of habitats conducive to the biosynthesis of benzylisoquinoline alkaloids. The ubiquity of benzylisoquinoline alkaloids biosynthesis and the constrained distribution of shikonin derivatives underscore the complexity and diversity of NP biosynthesis pathways and how they can be differentially influenced by the environmental, ecological, and evolutionary contexts of different regions. This finding further illustrates the importance of considering both the species and enzyme richness in a region while assessing the potential for drug discovery in nature.

3 Discussion

The global environmental crisis in modern times threatens every aspect of biodiversity, ranging from individual genes to entire ecosystems [38, 39]. With limited conservation resources, a thorough biodiversity evaluation becomes crucial for directing targeted conservation efforts effectively. Yet genomic contribution to biodiversity evaluation and conservation efforts is underexplored, even though genetic diversity is recognized as fundamental at all levels of biology [40]. Here we show that integrating the genetic (genomic and transcriptomic) and distributional data of woody plants enables a nuanced geospatial understanding of biodiversity which is essential for developing targeted conservation strategies for future drug discovery. We examined 166 enzymes associated with NP biosynthesis and observed distinct distribution patterns. Certain enzymes were primarily located in specific regions, implying that these areas are the hotspots for particular NPs. On the other hand, numerous enzymes were widespread, indicating a larger-scale availability of certain NPs.

We observed a similar pattern when investigating well-known enzyme groups responsible for synthesizing shikonin derivatives and benzylisoquinoline alkaloids (Fig. 4). This pattern of distribution offers a window into the endemic nature of certain NPs. As such, the concept of endemism extends beyond species, affecting these NPs and, by extension, impacting our potential to discover new drugs. These spatial patterns of NPs may offer significant guidance for policymakers to develop targeted conservation measures for sustainable nature-derived drug discovery in the future.

The distribution of numerous NPs, which are highly localized, links them to potential risks from regional environmental changes. These risks can undermine their multifaceted roles. For instance, we found that the combined spatial patterns of enzymes associated with the biosynthesis of shikonin derivatives are predominantly localized to East Asian regions and areas along the Himalayan range. These compounds are known for their therapeutic potential in treating cancer and inflammation, as well as their traditional role in wound healing [41]. Remarkably, shikonin derivatives, which accumulate exclusively within the roots of plants, also demonstrate potent antibacterial activity against soil-borne microorganisms [42]. This particular characteristic, functioning as a chemical barrier protecting root tissues from biological stressors, may reflect unique regional evolutionary adaptations or ecological niches favouring shikonin derivative biosynthesis, a hypothesis warranting further investigation. The presence of these versatile NPs in specific regions underscores the urgency of targeted conservation strategies. Understanding these spatial patterns can provide valuable insights into how ongoing adverse impacts on the global environment could influence the availability of certain NPs. By safeguarding these areas, we ensure the continued availability of these valuable NPs, supporting both pharmacological advancements and ecological balance.

The growing popularity of certain plants for use as herbal medicines or as sources of bioactive compounds for pharmaceutical drugs often leads to a surge in global market demand [43]. This increase can result in overharvesting from natural populations, putting the species at risk of extinction [43, 44]. It is estimated that between 60% and 80% of medicinal plants traded worldwide are gathered from the wild, representing a trade value that surpasses $3 billion [43]. Additionally, the distribution of medicinal plants throughout the world is unequal, with China and India leading the way in terms of utilization with 11,146 and 7,500 species, respectively [45]. Understanding the spatial patterns of NPs may open avenues for influencing market dynamics. Regions rich in endemic NPs can be preserved and utilized responsibly via targeted conservation and management. Additionally, if overharvesting is a concern in regions that need conservation, recommendations for alternative medicinal plants with similar NP profiles can be made. These approaches have the potential to alleviate pressure

on endangered or over-harvested regions by shifting the demand in the market towards more sustainable sources.

The confined distribution of specific NPs in certain regions presents challenges in extraction and sustainability, especially when demand for these compounds is high. By identifying and analysing the spatial patterns of these unique NPs, scientists can pinpoint which compounds would benefit most from genetic engineering. The use of metabolic engineering, either in plants or microorganisms, can produce selected NPs in large quantities without the need for extracting them from their native sources [46–48]. This approach not only alleviates the pressure on ecosystems where these NPs are naturally found but also allows for more controlled and efficient production.

Beyond their therapeutic applications, NPs are also integral to numerous ecological and evolutionary processes, extending across complex ecological networks [49–54]. Plants, unable to flee from predators or rely on an immune system, have evolved NPs for survival strategies, a trait developed early in their evolutionary history about 140 million years ago [54]. These compounds serve multiple functions, including protection against viruses, bacteria, and herbivores, as well as the attraction of pollinators and seed dispersers through fragrance or colour [42, 49, 51, 54]. Some compounds even function dually, like anthocyanins, which can attract insects while also having insecticidal properties [55]. The multifaceted role of NPs underscores their irreplaceable value not only to individual plant species but also to entire ecosystems. Conservation strategies that prioritize the protection and sustainable utilization of NPs are therefore not just a matter of preserving nature-derived drugs but of maintaining ecological stability and resilience.

Additionally, this study revealed the stark contrast between species richness and the availability of corresponding genetic data (genomic and transcriptomic), particularly in biodiversity hotspots like the Amazon Basin and Andean region. This observation indicates a significant gap in our understanding of the genetic diversity within these regions and underscores the urgency for the targeted collection and analysis of genetic data from these underrepresented areas. Such action will ensure a comprehensive understanding of the biosynthetic potential of NPs and contribute to ongoing efforts for biodiversity conservation. In general, economically developed and low-diversity regions are more likely to have a substantial proportion of genetic data.

While our study provides significant insights into the spatial patterns of enzymes relevant to drug development, several limitations should be acknowledged. Our study primarily focused on woody plants, which leaves out a significant portion of plant biodiversity, such as herbaceous plants, grasses, and aquatic plants. This limited scope may consequently exclude potential drug candidates found within these groups. Additionally, the quality and completeness of the genetic and distributional data can affect the accuracy of our results.

Incomplete or biased data could potentially lead to the misinterpretation or oversight of significant patterns. While we found significant geospatial heterogeneity in numerous enzymes, these enzymes alone do not directly translate into drug compounds. Finally, our approach might not capture the complexities of plant biochemical pathways. Future research that addresses these gaps will broaden the application of our findings to the real world.

4 Methods

4.1 Data Collection

We utilized the woody angiosperm species list outlined by [56] as a basis for determining a list of woody plant species. Their dataset is based on information in the national floras, including tree distribution checklists for each country, botanical literature, and various databases. We then collected species genetic data from publicly accessible databases based on next-generation sequencing (NGS). The NGS data incorporated assembled genomes, assembled transcriptomes, and raw transcriptomes. For each species, a single data type was selected based on availability, with a preference hierarchy of assembled genomes (most preferred), followed by assembled transcriptomes, and raw transcriptomes (least preferred). The assembled genome data was collected from EnsemblePlants (release 52) and RefSeq, whereas the assembled transcriptomes were obtained from OneKP and Transcriptome Shotgun Assembly [57]. For species lacking assembled genetic data, we collected their raw transcriptome data from the Sequence Reads Archive. To reduce the complexity in the assembling of raw transcriptome data, third-generation sequencing data was omitted and Illumina-derived data was preferred over data acquired from other second-generation sequencing instruments due to the abundance of the data. Additionally, only a single raw RNA-seq dataset from each experiment accession was collected for each species, and data with the biggest size was preferred. In total, we collected 52 assembled genomes, 349 assembled transcriptomes, and 1211 raw transcriptome data for 1612 species (Supplementary Table 3).

4.2 De Novo Assembly Simulation

We downloaded the coding sequence (CDS) and protein sequence of selected plants (Supplementary table 4) from EnsemblePlants (release 52) as a dataset. The CDSs were later simulated to produce RNA-seq reads using Polyester v.1.36.0 with 100 reads per transcript [58]. We performed de novo assembly by initially trimming the detected adapters, bases under 20 Phred score, and PolyG artifacts with a minimum length of 10 and by discarding reads with a read length equal to or below 36 after trimming using Fastp v.0.21.0 [59]. Then, multiQC v.1.10.1 was used to look at the results of the quality control of the RNA-seq data with similar

read lengths [60]. We removed the ribosomal RNA fragments using SortMeRNA v.4.3.3 tool with SMR v4.3 default as the reference [61]. We then performed de novo assembly into scaffold levels using SOAP denovo-Trans v.1.03 with the parameters set as follows: avg ins = 200, reverse seq = 0, and asm flags = 3, and multiple k-mer lengths, including 21, 31, 35, and 41 (Assembly 1-4) [62]. For RNA-seq data with paired-end reads, before the de novo assembly process at the scaffold level, we first estimated the average insert size with the following steps. Contigs were assembled from RNAseq reads using SOAP denovo-Trans with parameters set as follows: avg ins = 200, reverse seq = 0, and asm flags = 3. Then, the first 400.000 RNA-seq reads were sampled and mapped to contigs with BWA-MEM, and the average insert size from the alignment was calculated. We later assembled RNA-seq contigs into scaffold levels using SOAP denovo-Trans v.1.03 with the estimated average insert size and the specified multiple k-mer lengths as above. We discarded scaffolds or contigs shorter than 200 bp and developed merged assemblies from the merging of assemblies 1-4 by cd-hit-est v4.6.1 with similarity thresholds of 93% (Assembly 5). The pipeline of this de novo assembly was developed using Nextflow [63]. The performance of each assembly was assessed with simulated benchmarks and common assembly statistics, which include mean, median, longest, and N50 sizes of scaffolds or contigs, and mean coverage, which was calculated from the mapping of the assembly with its reads using BWA-MEM v.0.7.17. The simulated benchmark scores included the correct-to-incorrect ratio, accuracy, and ratio of the detected reference (Supplementary table 4). These metrics were retrieved based on the Blastx v.2.12.0 search of each scaffold or contig in an assembly to its original protein transcripts as a reference. From the Blastx alignment, we assumed a scaffold or contig to be misassembled if its alignment result had no match to the reference transcript with an identity and sequence length threshold of 95% and 50%, respectively. Otherwise, we assumed correct assembly; the total amount of correctly assembled and misassembled was considered a true positive (TP) and true negative (TN), respectively. Based on the alignment, we also calculated the amount of protein reference that was found and missing; the amount of missing reference was considered a false negative (FN). The ratio of correct to incorrect predictions was calculated by dividing the number of TP by the number of TN. Accuracy, on the other hand, was calculated by dividing the number of TP by the sum of TP, TN, and FN.

4.3 De Novo Assembly for Raw Transcriptome Data

We performed a de novo assembly process following the steps described in 4.2. Steps for assembly 5 was selected as this resulted to the best benchmark (Supplementary table 4). Further, we removed low complexity regions of the assembly using DustMasker v.1.0.0 and identified candidate ORFs using TransDecoder v.5.7 (https://github.com/TransDecoder/TransDecoder/) [64].

4.4 Assembly Completeness and Enzyme Annotation

The completeness of the assembly was investigated using Benchmarking Universal Single-Copy Orthologs (BUSCO) v2.0 with embryophyte (odb10) as the reference [65]. A completeness investigation was also performed for the assembled genome and transcriptome data. Genetic data in which more than 42.5% of BUSCO sequences were missing was removed [66], resulting in a refined dataset comprising 1,434 species (Supplementary table 3). Further, we performed a focused search of specific enzyme families using BLAST v.2.12.0. These families included O-methyltransferases (OMT), prenyltransferase (PT), DOXC-class 2-oxoglutarate-dependent dioxygenase (DOXC), flavin-containing monooxygenase (FMO), terpenoid synthase (TPS), BAHD acyltransferase (BAHD), polyphenol oxidase (PPO), CYP and UGT. We also performed a BLAST search for well-known specific groups of enzymes responsible for the biosynthesis of shikonin (PGT, SAT, GHQH, and DSH) and benzylisoquinoline alkaloids (6-OMT, CNMT, and 4′-OMT). The specified identity threshold for each enzyme is shown in Supplementary table 1 and 2, with a bit score threshold set at 50.

4.5 Spatial Pattern of Species Richness and Enzyme Markers

We used the dataset of occurrence of woody plants compiled by [56], which includes 89,826 species based on The Plant list (http://www.theplantlist.org/). Focusing on species with more than five occurrence records, we estimated species distribution of our list of species with NGS data. We specifically predicted species distributions at a scale of 1 degree (approximately 100 km x 100 km) per grid cell. A species distribution model was developed using Maxent version 3.4.1 [67] with 70 environmental variables, which included climatic (bio1–bio19), soil, geological, topographical, and geographical conditions, as the predictor variables. The relative suitability was computed within the range of countries where each species has a native distribution; the tree distribution checklist of each country was used to exclude non-native distribution occurrence records. The Maxent output (maximum training sensitivity plus specificity Cloglog threshold) is a continuous map with shading intensities ranging from 0 to 1, allowing fine distinctions to be made between the modelled suitability of different areas. We used the threshold criteria of the sensitivity-specificity sum maximizer (MST) to generate a binary (presence/absence) prediction map for each species [68] and confirmed the accuracy of each model using the area under the receiver operating characteristic curve. The relative suitability below the MST threshold value was converted to zero. We superimposed the distributions of the 1612 species with NGS data and summed the number of species per grid cell. We finally created a map of the geographical patterns of species richness for drug species overall. Further, to calculate the proportion of species with available NGS data in each grid cell, we divided the number of species possessing such data by the total

species richness in each one-degree square grid. This ratio was then expressed as a percentage. For each grid cell, the aggregate average of enzyme markers was calculated by summing the number of enzyme markers identified from BLAST matches for all species present within the grid cell. This total was then divided by the total number of species with the NGS data within the cell, providing an average enzyme marker count per species for each grid. In addition to this aggregate average of all enzyme markers, we also calculated the average for each individual enzyme and the overlap of enzymes involved in the biosynthesis of shikonin and benzylisoquinoline alkaloids. We also calculated the product of average enzyme markers and species richness by multiplying the species richness with the average enzyme markers in each grid cell.

Supplementary information.

- **Supplementary table 1**. Comprehensive enzyme information with emphasis on functions, sequences, and BLAST identity thresholds.
- **Supplementary Table 2**. Comprehensive enzyme information associated to shikonin and benzylisoquinoline alkaloids biosynthesis with emphasis on functions, sequences, and BLAST identity thresholds.
- **Supplementary Table 3**. Species genetic data information and enzyme presence.
- **Supplementary Table 4**. De novo assembly simulation.
- **Supplementary Figures**. Spatial patterns of 166 enzymes associated with natural products.

Acknowledgments. The authors would like to thank Nguyen Tien Hoang for his helpful comments and the members of supply chain project in RIHN for their supports.

Declarations

- Funding: This work was supported by the Japan Society for the Promotion of Science through the Grant-in-Aid for Grant-in-Aid for Scientific Research (A) 20H00651.
- Competing interests: The authors declare that they have no competing interests.
- Availability of data and materials: The data that support the findings of this study are available from the corresponding author, Muhamad Fahmi, upon reasonable request.
- Authors' contributions: All authors conceived and designed the analysis. MF, KT, and YK collected the data. MF and YK performed the data analysis. All authors contributed to the article and approved the submitted version.

References

[1] Bernardini, S., Tiezzi, A., Masci, V.L., Ovidi, E.: Natural products for human health: an historical overview of the drug discovery approaches. Natural Product Research **32**(16), 1926–1950 (2018) https://doi.org/10.1080/14786419.2017.1356838. https://doi.org/10.1080/14786419.2017.1356838. PMID: 28748726

[2] Cragg, G.M., Newman, D.J.: Natural products: A continuing source of novel drug leads. Biochimica et Biophysica Acta (BBA) - General Subjects **1830**(6), 3670–3695 (2013). https://doi.org/10.1016/j.bbagen.2013. 02.008

[3] De Pasquale, A.: Pharmacognosy: The oldest modern science. Journal of Ethnopharmacology **11**(1), 1–16 (1984). https://doi.org/10.1016/ 0378-8741(84)90092-8

[4] Yao, H., Liu, J., Xu, S., Zhu, Z., Xu, J.: The structural modification of natural products for novel drug discovery. Expert Opinion on Drug Discovery **12**(2), 121–140 (2017) https://doi.org/10.1080/17460441.2016.1272757. https://doi.org/10.1080/17460441.2016.1272757. PMID: 28006993

[5] Newman, D.J., Cragg, G.M.: Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. Journal of Natural Products **83**(3), 770–803 (2020) https://doi.org/10.1021/acs.jnatprod.9b01285. https://doi.org/10.1021/acs.jnatprod.9b01285. PMID: 32162523

[6] Karageorgis, G., Foley, D.J., Laraia, L., Waldmann, H.: Principle and design of pseudo-natural products. Nature chemistry **12**(3), 227–235 (2020). https://doi.org/10.1038/s41557-019-0411-x

[7] Fischbach, M.A., Walsh, C.T.: Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms.

Chemical Reviews **106**(8), 3468–3496 (2006). https://doi.org/10.1021/cr0503097

[8] Zhang, X., Li, S.: Expansion of chemical space for natural products by uncommon p450 reactions. Nat. Prod. Rep. **34**, 1061–1089 (2017). https://doi.org/10.1039/C7NP00028F

[9] Jennewein, S., Long, R.M., Williams, R.M., Croteau, R.: Cytochrome p450 taxadiene 5alpha-hydroxylase, a mechanistically unusual monooxygenase catalyzing the first oxygenation step of taxol biosynthesis. Chemistry 'I&' Biology **11**(3), 379–387 (2004). https://doi.org/10.1016/j.chembiol.2004.02.022

[10] Teoh, K.H., Polichuk, D.R., Reed, D.W., Nowak, G., Covello, P.S.: *Artemisia annua* L. (Asteraceae) trichome-specific cDNAs reveal CYP71AV1, a cytochrome P450 with a key role in the biosynthesis of the antimalarial sesquiterpene lactone artemisinin. FEBS Letters **580**(5), 1411–1416 (2006) https://febs.onlinelibrary.wiley.com/doi/pdf/10.1016/j.febslet.2006.01.065. https://doi.org/10.1016/j.febslet.2006.01.065

[11] Chemler, J.A., Koffas, M.A.: Metabolic engineering for plant natural product biosynthesis in microbes. Current Opinion in Biotechnology **19**(6), 597–605 (2008). https://doi.org/10.1016/j.copbio.2008.10.011. Chemical biotechnology / Pharmaceutical biotechnology

[12] Cardinale, B.J., Duffy, J.E., Gonzalez, A., Hooper, D.U., Perrings, C., Venail, P., Narwani, A., Mace, G.M., Tilman, D., Wardle, D.A., Kinzig, A.P., Daily, G.C., Loreau, M., Grace, J.B., Larigauderie, A., Srivastava, D.S., Naeem, S.: Biodiversity loss and its impact on humanity. Nature **486**(7401), 59–67 (2012). https://doi.org/10.1038/nature11148

[13] IUCN: The IUCN Red List of Threatened Species. Version 2022-2. https://www.iucnredlist.org/ Accessed 2022-10-30

[14] Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G.: High-resolution global maps of 21stcentury forest cover change. Science **342**(6160), 850–853 (2013) https://www.science.org/doi/pdf/10.1126/science.1244693. https://doi.org/10.1126/science.1244693

[15] Curtis, P.G., Slay, C.M., Harris, N.L., Tyukavina, A., Hansen, M.C.: Classifying drivers of global forest loss. Science **361**(6407), 1108–1111 (2018) https://www.science.org/doi/pdf/10.1126/science.aau3445. https://doi.org/10.1126/science.aau3445

[16] Gaston, K.J.: Global patterns in biodiversity. Nature **405**(6783), 220–227 (2000). https://doi.org/10.1038/35012228

[17] Willig, M.R., Kaufman, D.M., Stevens, R.D.: Latitudinal gradients of biodiversity: Pattern, process, scale, and synthesis. Annual Review of Ecology, Evolution, and Systematics **34**(1), 273–309 (2003) https://doi.org/10.1146/annurev.ecolsys.34.012103.144032. https: //doi.org/10.1146/annurev.ecolsys.34.012103.144032

[18] Pressey, R.L., Cowling, R.M., Rouget, M.: Formulating conservation targets for biodiversity pattern and process in the cape floristic region, south africa. Biological Conservation **112**(1), 99–127 (2003). https://doi.org/10.1016/S0006-3207(02)00424-X. Conservation Planning in the Cape Floristic Region

[19] Gray, J.S.: Marine biodiversity: patterns, threats and conservation needs. Biodiversity & Conservation **6**(1), 153–175 (1997). https://doi.org/10.1023/A:1018335901847

[20] Irigoien, X., Huisman, J., Harris, R.P.: Global biodiversity patterns of marine phytoplankton and zooplankton. Nature **429**(6994), 863–867 (2004). https://doi.org/10.1038/nature02593

[21] Mair, L., Bennun, L.A., Brooks, T.M., Butchart, S.H.M., Bolam, F.C., Burgess, N.D., Ekstrom, J.M.M., Milner-Gulland, E.J., Hoffmann, M., Ma, K., Macfarlane, N.B.W., Raimondo, D.C., Rodrigues, A.S.L., Shen, X., Strassburg, B.B.N., Beatty, C.R., G´omez-Creutzberg, C., Iribarrem, A., Irmadhiany, M., Lacerda, E., Mattos, B.C., Parakkasi, K., Tognelli, M.F., Bennett, E.L., Bryan, C., Carbone, G., Chaudhary, A., Eiselin, M., da Fonseca, G.A.B., Galt, R., Geschke, A., Glew, L., Goedicke, R., Green, J.M.H., Gregory, R.D., Hill, S.L.L., Hole, D.G., Hughes, J., Hutton, J., Keijzer, M.P.W., Navarro, L.M., Nic Lughadha, E., Plumptre, A.J., Puydarrieux, P., Possingham, H.P., Rankovic, A., Regan, E.C., Rondinini, C., Schneck, J.D., Siikam¨aki, J., Sendashonga, C., Seutin, G., Sinclair, S., Skowno, A.L., Soto-Navarro, C.A., Stuart, S.N., Temple, H.J., Vallier, A.,
Verones, F., Viana, L.R., Watson, J., Bezeng, S., B¨ohm, M., Burfield, I.J., Clausnitzer, V., Clubbe, C., Cox, N.A., Freyhof, J., Gerber, L.R., HiltonTaylor, C., Jenkins, R., Joolia, A., Joppa, L.N., Koh, L.P., Lacher, T.E., Langhammer, P.F., Long, B., Mallon, D., Pacifici, M., Polidoro, B.A., Pollock, C.M., Rivers, M.C., Roach, N.S., Rodr´ıguez, J.P., Smart, J., Young, B.E., Hawkins, F., McGowan, P.J.K.: A metric for spatially explicit contributions to science-based species targets. Nature Ecology & Evolution **5**(6), 836–844 (2021). https://doi.org/10.1038/s41559-021-01432-0

[22] Miraldo, A., Li, S., Borregaard, M.K., Fl´orez-Rodr´ıguez,A., Gopalakrishnan, S., Rizvanovic, M., Wang, Z., Rahbek, C., Marske, K.A., Nogu´es-Bravo, D.: An anthropocene map of genetic diversity.

Science **353**(6307), 1532–1535 (2016) https://www.science.org/doi/pdf/10.1126/science.aaf4381. https://doi.org/10.1126/science.aaf4381

[23] De Kort, H., Prunier, J.G., Ducatez, S., Honnay, O., Baguette, M., Stevens, V.M., Blanchet, S.: Life history, climate and biogeography interactively affect worldwide genetic diversity of plant and animal populations. Nature Communications **12**(1), 516 (2021). https://doi.org/10. 1038/s41467-021-20958-2

[24] Safi, K., Cianciaruso, M.V., Loyola, R.D., Brito, D., Armour-Marshall, K., Diniz-Filho, J.A.F.: Understanding global patterns of mammalian functional and phylogenetic diversity. Philosophical Transactions of the Royal Society B: Biological Sciences **366**(1577), 2536–2544 (2011) https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2011.0024. https://doi.org/10.1098/rstb.2011.0024

[25] Theodoridis, S., Fordham, D.A., Brown, S.C., Li, S., Rahbek, C., NoguesBravo, D.: Evolutionary history and past climate change shape the distribution of genetic diversity in terrestrial mammals. Nature Communications **11**(1), 2557 (2020). https://doi.org/10.1038/s41467-020-16449-5

[26] Owen, N.R., Gumbs, R., Gray, C.L., Faith, D.P.: Global conservation of phylogenetic diversity captures more than just functional diversity. Nature Communications **10**(1), 859 (2019). https://doi.org/10.1038/ s41467-019-08600-8

[27] Nagia, M., Gaid, M., Biedermann, E., Fiesel, T., El-Awaad, I., H¨ansch, R., Wittstock, U., Beerhues, L.: Sequential regiospecific gem-diprenylation of tetrahydroxyxanthone by prenyltransferases from *Hypericum* sp. New Phytologist **222**(1), 318–334 (2019) https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/nph.15611. https://doi.org/10.1111/nph.15611

[28] Ren, Z., Qiu, F., Wang, Y., Yu, W., Liu, C., Sun, Y., Wang, Y., Zhang, X., Xing, S., Tao, S., Huang, Y., Liu, G., Wei, Z., Yu, B., Du, S., Lei, Z., Wei, G.: Network analysis of transcriptome and LC-MS reveals a possible biosynthesis pathway of anthocyanins in *Dendrobium officinale*. BioMed Research International **2020**, 6512895 (2020). https: //doi.org/10.1155/2020/6512895

[29] Xiao, Z., Fang, L., Niu, Y., Yu, H.: Effect of cultivar and variety on phenolic compounds and antioxidant activity of cherry wine. Food Chemistry **186**, 69–73 (2015). https://doi.org/10.1016/j.foodchem.2015.01.050. ISPMF

2015: International Symposium on Phytochemicals in Medicine and Food (Shanghai, China, June 26th –29th, 2015)

[30] Bakrim, S., Benkhaira, N., Bourais, I., Benali, T., Lee, L.-H., El Omari,N., Sheikh, R.A., Goh, K.W., Ming, L.C., Bouyahya, A.: Health benefits and pharmacological properties of stigmasterol. Antioxidants 11(10) (2022). https://doi.org/10.3390/antiox11101912

[31] Griebel, T., Zeier, J.: A role for -sitosterol to stigmasterol conversion in plant–pathogen interactions. The Plant Journal 63(2), 254–268 (2010) https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365313X.2010.04235.x. https://doi.org/10.1111/j.1365-313X.2010.04235.x

[32] Marques, J.V., Kim, K.-W., Lee, C., Costa, M.A., May, G.D., Crow, J.A., Davin, L.B., Lewis, N.G.: Next generation sequencing in predicting gene function in podophyllotoxin biosynthesis. Journal of Biological Chemistry **288**(1), 466–479 (2013). https://doi.org/10.1074/jbc. M112.400689

[33] Lau, W., Sattely, E.S.: Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone. Science **349**(6253), 1224–1228 (2015) https://www.science.org/doi/pdf/10.1126/science.aac7202. https://doi.org/10.1126/science.aac7202

[34] Schnabel, A., Cotinguiba, F., Athmer, B., Vogt, T.: *Piper nigrum* CYP719A37 catalyzes the decisive methylenedioxy bridge formation in piperine biosynthesis. Plants **10**(1) (2021). https://doi.org/10.3390/plants10010128

[35] Ikezawa, N., Tanaka, M., Nagayoshi, M., Shinkyo, R., Sakaki, T., Inouye, K., Sato, F.: Molecular cloning and characterization of CYP719, a methylenedioxy bridge-forming enzyme that belongs to a novel P450 family, from cultured *Coptis japonica* cells. Journal of Biological Chemistry **278**(40), 38557–38565 (2003). https://doi.org/10.1074/ jbc.M302470200

[36] Yadav, S., Sharma, A., Nayik, G.A., Cooper, R., Bhardwaj, G., Sohal, H.S., Mutreja, V., Kaur, R., Areche, F.O., AlOudat, M., Shaikh, A.M., Kovács, B., Mohamed Ahmed, A.E.: Review of shikonin and derivatives: Isolation, chemistry, biosynthesis, pharmacology and toxicology. Frontiers in Pharmacology 13 (2022). https://doi.org/10.3389/fphar.2022.905755

[37] Vadhel, A., Bashir, S., Mir, A.H., Girdhar, M., Kumar, D., Kumar, A., Mohan, A., Malik, T., Mohan, A.: Opium alkaloids, biosynthesis, pharmacology and association with cancer occurrence. Open Biology 13(5), 220355 (2023)

https://royalsocietypublishing.org/doi/pdf/10.1098/rsob.220355. https://doi.org/10.1098/rsob.220355

[38] Theodoridis, S., Rahbek, C., Nogues-Bravo, D.: Exposure of mammal genetic diversity to mid-21st century global change. Ecography **44**(6), 817–831 (2021) https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecog.05588. https: //doi.org/10.1111/ecog.05588

[39] Steffen, W., Richardson, K., Rockström , J., Cornell, S.E., Fetzer, I., Bennett, E.M., Biggs, R., Carpenter, S.R., de Vries, W., de Wit, C.A., Folke, C., Gerten, D., Heinke, J., Mace, G.M., Persson, L.M., Ramanathan, V., Reyers, B., Sörlin , S.: Planetary boundaries: Guiding human development on a changing planet. Science **347**(6223), 1259855 (2015) https://www.science.org/doi/pdf/10.1126/science.1259855. https: //doi.org/10.1126/science.1259855

[40] Theissinger, K., Fernandes, C., Formenti, G., Bista, I., Berg, P.R., Bleidorn, C., Bombarely, A., Crottini, A., Gallo, G.R., Godoy, J.A., Jentoft, S., Malukiewicz, J., Mouton, A., Oomen, R.A., Paez, S., Palsbøll, P.J., Pampoulie, C., Ruiz- López, M.J., Secomandi, S., Svardal, H., Theofanopoulou, C., de Vries, J., Waldvogel, A.-M., Zhang, G., Jarvis, E.D., Bálint , M., Ciofi, C., Waterhouse, R.M., Mazzoni, C.J., Höglund, J., Aghayan, S.A., Alioto, T.S., Almudi, I., Alvarez, N., Alves, P.C., Amorim do Rosario, I.R., Antunes, A., Arribas, P., Baldrian, P., Bertorelle, G., B¨ohne, A., Bonisoli-Alquati, A., Boštjančić , L.L., Boussau,B., Breton, C.M., Buzan, E., Campos, P.F., Carreras, C., Castro, L.F.C., Chueca, L.J., Ciampor, F., Conti, E., Cook-Deegan, R., Croll, D., Cunha,M.V., Delsuc, F., Dennis, A.B., Dimitrov, D., Faria, R., Favre, A., Fedrigo, O.D., Fern´andez, R., Ficetola, G.F., Flot, J.-F., Gabald´on, T., Agius, D.R., Giani, A.M., Gilbert, M.T.P., Grebenc, T., Guschanski, K., Guyot, R., Hausdorf, B., Hawlitschek, O., Heintzman, P.D., Heinze, B., Hiller, M., Husemann, M., Iannucci, A., Irisarri, I., Jakobsen, K.S., Klinga, P., Kloch, A., Kratochwil, C.F., Kusche, H., Layton, K.K.S., Leonard, J.A., Lerat, E., Liti, G., Manousaki, T., Marques-Bonet, T., Matos- Maraví, P., Matschiner, M., Maumus, F., Mc Cartney, A.M., Meiri, S., MeloFerreira, J., Mengual, X., Monaghan, M.T., Montagna, M., Mysl ajek, R.W., Neiber, M.T., Nicolas, V., Novo, M., Ozretić , P., Palero, F., Pârvulescu, L., Pascual, M., Paulo, O.S., Pavlek, M., Pegueroles, C., Pellissier, L., Pesole, G., Primmer, C.R., Riesgo, A., Rüber, L., Rubolini, D., Salvi, D., Seehausen, O., Seidel, M., Studer, B., Theodoridis, S., Thines, M., Urban, L., Vasemägi, A., Vella, A., Vella, N., Vernes, S.C., Vernesi, C., Vieites, D.R., Wheat, C.W., Wörheide , G., Wurm, Y., Zammit, G.: How genomics can help biodiversity conservation. Trends in Genetics **39**(7), 545–559 (2023). https://doi.org/10.1016/j.tig.2023.01.005

[41] Guo, C., He, J., Song, X., Tan, L., Wang, M., Jiang, P., Li, Y., Cao, Z., Peng, C.: Pharmacological properties and derivatives of shikonin—a review in recent years. Pharmacological Research **149**, 104463 (2019). https://doi.org/10.1016/j.phrs.2019.104463

[42] Brigham, L.A., Michaels, P.J., Flores, H.E.: Cell-Specific Production and Antimicrobial Activity of Naphthoquinones in Roots of Lithospermum erythrorhizon1. Plant Physiology

[43] Theodoridis, S., Drakou, E.G., Hickler, T., Thines, M., Nogues-Bravo, D.: Evaluating natural medicinal resources and their exposure to global change. The Lancet Planetary Health **7**(2), 155–163 (2023). https://doi.org/10.1016/S2542-5196(22)00317-5

[44] Howes, M.-J., Quave, C., Collemare, J., Tatsis, E., Twilley, D., Lulekal, E., Farlow, A., al, L., Bell, E.: Molecules from nature: Reconciling biodiversity conservation and global healthcare imperatives for sustainable use of medicinal plants and fungi. Plants People Planet **2**, 463–481 (2020). https://doi.org/10.1002/ppp3.10138

[45] Chen, S.-L., Yu, H., Luo, H.-M., Wu, Q., Li, C.-F., Steinmetz, A.: Conservation and sustainable use of medicinal plants: problems, progress, and prospects. Chinese Medicine **11**(1), 37 (2016). https://doi.org/10.1186/ s13020-016-0108-7

[46] Marienhagen, J., Bott, M.: Metabolic engineering of microorganisms for the synthesis of plant natural products. Journal of Biotechnology **163**(2), 166–178 (2013). https://doi.org/10.1016/j.jbiotec.2012.06.001.

[47] Yang, D., Park, S.Y., Park, Y.S., Eun, H., Lee, S.Y.: Metabolic engineering of *Escherichia coli* for natural product biosynthesis. Trends in Biotechnology **38**(7), 745–765 (2020). https://doi.org/10.1016/ j.tibtech.2019.11.007

[48] Wu, S., Chappell, J.: Metabolic engineering of natural products in plants; tools of the trade and challenges for the future. Current Opinion in Biotechnology **19**(2), 145–152 (2008). https://doi.org/10.1016/j.copbio. 2008.02.007.

[49] Stevenson, P.C., Nicolson, S.W., Wright, G.A.: Plant secondary metabolites in nectar: impacts on pollinators and ecological functions. Functional Ecology 31(1), 65–75 (2017) https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2435.12761. https://doi.org/10.1111/1365-2435.12761

[50] Chomel, M., Guittonny-Larchevêque, M., Fernandez, C., Gallet, C., DesRochers, A., Paré , D., Jackson, B.G., Baldy, V.: Plant secondary metabolites: a key driver of litter decomposition and soil nutrient cycling. Journal of Ecology 104(6), 1527–1541 (2016) https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2745.12644. https://doi.org/10.1111/1365-2745.12644

[51] Steppuhn, A., Gase, K., Krock, B., Halitschke, R., Baldwin, I.T.: Nicotine's defensive function in nature. PLOS Biology 2(8), (2004). https: //doi.org/10.1371/journal.pbio.0020217

[52] Zheng, X.-b., Zhang, Y.-l., Li, Q., Liu, Y. G., Wang, X.-d., Yang, B.-l., Zhu, G.-c., Zhou, C.-f., Gao, Y., Liu, Z.-x.: Effects of 1,8-cineole on neuropathic pain mediated by P2X2 receptor in the spinal cord dorsal horn. Scientific Reports 9(1), 7909 (2019). https://doi.org/10.1038/ s41598-019-44282-4

[53] Lackus, N.D., Lackner, S., Gershenzon, J., Unsicker, S.B., Köllner , T.G.: The occurrence and formation of monoterpenes in herbivore-damaged poplar roots. Scientific Reports 8(1), 17936 (2018). https://doi.org/10.1038/s41598-018-36302-6

[54] Wink, M.: Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. Phytochemistry 64(1), 3–19 (2003). https://doi.org/10.1016/S0031-9422(03)00300-5

[55] Lev-Yadun, S., Gould, K.S.: In: Winefield, C., Davies, K., Gould, K. (eds.) Role of Anthocyanins in Plant Defence, pp. 22–28. Springer, New York, NY (2009). https://doi.org/10.1007/978-0-387-77335-3_2.

[56] Kusumoto, B., Chao, A., Eiserhardt, W.L., Svenning, J.-C., Shiono, T., Kubota, Y.: Occurrence-based diversity estimation reveals macroecological and conservation knowledge gaps for global woody plants. bioRxiv (2023) https://www.biorxiv.org/content/early/2023/04/11/2023.04.09.536180.full.pdf. https://doi.org/10.1101/2023.04.09.536180

[57] Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., Burleigh, J.G., Gitzendanner, M.A., Wafula, E., Der, J.P., dePamphilis, C.W., Roure, B., Philippe, H., Ruhfel, B.R., Miles, N.W., Graham, S.W., Mathews, S., Surek, B., Melkonian, M., Soltis, D.E., Soltis, P.S., Rothfels, C., Pokorny, L., Shaw, J.A., DeGironimo, L., Stevenson, D.W., Villarreal, J.C., Chen, T., Kutchan, T.M., Rolf, M., Baucom, R.S., Deyholos, M.K., Samudrala, R., Tian, Z., Wu, X., Sun, X., Zhang, Y., Wang, J., Leebens-Mack, J., Wong, G.K.-S.: Data access for the 1,000 Plants (1KP)

project. GigaScience **3**(1), 2047–217317 (2014) https://academic.oup.com/gigascience/articlepdf/3/1/2047-217X-3-17/25512207/13742 2014 article 69.pdf. https://doi.org/10.1186/2047-217X-3-17

[58] Frazee, A.C., Jaffe, A.E., Langmead, B., Leek, J.T.: Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics **31**(17), 2778–2784 (2015) https://academic.oup.com/bioinformatics/articlepdf/31/17/2778/49035291/bioinformatics 31 17 2778.pdf. https://doi.org/10.1093/bioinformatics/btv272

[59] Chen, S., Zhou, Y., Chen, Y., Gu, J.: Fastp: an ultrafast all-in-one FASTQ preprocessor. Bioinformatics **34**(17), 884–890 (2018) https://academic.oup.com/bioinformatics/articlepdf/34/17/i884/50582648/bioinformatics 34 17 i884.pdf. https://doi.org/10.1093/bioinformatics/bty560

[60] Ewels, P., Magnusson, M., Lundin, S., Käller, M.: MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics **32**(19), 3047–3048 (2016) https://academic.oup.com/bioinformatics/articlepdf/32/19/3047/49021359/bioinformatics 32 19 3047.pdf. https://doi.org/10.1093/bioinformatics/btw354

[61] Kopylova, E., Noé , L., Touzet, H.: SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics **28**(24), 3211–3217 (2012) https://academic.oup.com/bioinformatics/articlepdf/28/24/3211/48882806/bioinformatics 28 24 3211.pdf. https://doi.org/10.1093/bioinformatics/bts611

[62] Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.-W., Li, Y., Xu, X., Wong, G.K.-S., Wang, J.: SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics **30**(12), 1660–1666 (2014) https://academic.oup.com/bioinformatics/articlepdf/30/12/1660/48927094/bioinformatics 30 12 1660.pdf. https://doi.org/10.1093/bioinformatics/btu077

[63] Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., Notredame, C.: Nextflow enables reproducible computational workflows. Nature Biotechnology **35**(4), 316–319 (2017). https://doi.org/10.1038/ nbt.3820

[64] Morgulis, A., Gertz, E.M., Schäffer, A.A., Agarwala, R.: A fast and symmetric dust implementation to mask low-complexity dna sequences. Journal of Computational Biology **13**(5), 1028–1040 (2006) https://doi.org/10.1089/cmb.2006.13.1028. https://doi.org/10.1089/cmb.2006.13.1028. PMID: 16796549

[65] Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M.: BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**(19), 3210–3212 (2015) https://academic.oup.com/bioinformatics/articlepdf/31/19/3210/49035194/bioinformatics 31 19 3210.pdf. https://doi.org/10.1093/bioinformatics/btv351

[66] Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S.A., Soltis, D.E., Soltis, P.S., Stevenson, D.W., Ullrich, K.K., Wickett, N.J., DeGironimo, L., Edger, P.P., Jordon-Thaden, I.E., Joya, S., Liu, T., Melkonian, B., Miles, N.W., Pokorny, L., Quigley, C., Thomas, P., Villarreal, J.C., Augustin, M.M., Barrett, M.D., Baucom, R.S., Beerling, D.J., Benstein, R.M., Biffin, E., Brockington, S.F., Burge, D.O., Burris, J.N., Burris, K.P., Burtet-Sarramegna, V., Caicedo, A.L., Cannon, S.B., C¸ebi, Z., Chang, Y., Chater, C., Cheeseman, J.M., Chen, T., Clarke, N.D., Clayton, H., Covshoff, S., Crandall-Stotler, B.J., Cross, H., dePamphilis, C.W., Der, J.P., Determann, R., Dickson, R.C., Di Stilio, V.S., Ellis, S., Fast, E., Feja, N., Field, K.J., Filatov, D.A., Finnegan, P.M., Floyd, S.K., Fogliani, B., García, N., Gâteblé, G., Godden, G.T., Goh, F.Q.Y., Greiner, S., Harkess, A., Heaney, J.M., Helliwell, K.E., Heyduk, K., Hibberd, J.M., Hodel, R.G.J., Hollingsworth, P.M., Johnson, M.T.J., Jost, R., Joyce, B., Kapralov, M.V., Kazamia, E., Kellogg, E.A., Koch, M.A., Von Konrat, M., Könyves, K., Kutchan, T.M., Lam, V., Larsson, A., Leitch, A.R., Lentz, R., Li, F.-W., Lowe, A.J., Ludwig, M., Manos, P.S., Mavrodiev, E., McCormick, M.K., McKain, M., McLellan, T., McNeal, J.R., Miller, R.E., Nelson, M.N., Peng, Y., Ralph, P., Real, D., Riggins, C.W., Ruhsam, M., Sage, R.F., Sakai, A.K., Scascitella, M., Schilling, E.E., Schlösser, E.-M., Sederoff, H., Servick, S., Sessa, E.B., Shaw, A.J., Shaw, S.W., Sigel, E.M., Skema, C., Smith, A.G., Smithson, A., Stewart, C.N., Stinchcombe, J.R., Szövényi, P., Tate, J.A., Tiebel, H., Trapnell, D., Villegente, M., Wang, C.N., Weller, S.G., Wenzel, M., Weststrand, S., Westwood, J.H., Whigham, D.F., Wu, S., Wulff, A.S., Yang, Y., Zhu, D., Zhuang, C., Zuidof, J., Chase, M.W., Pires, J.C., Rothfels, C.J., Yu, J., Chen, C., Chen, L., Cheng, S., Li, J., Li, R., Li, X., Lu, H., Ou, Y., Sun, X., Tan, X., Tang, J., Tian, Z., Wang, F., Wang, J., Wei, X., Xu, X., Yan, Z., Yang, F., Zhong, X., Zhou, F., Zhu, Y., Zhang, Y., Ayyampalayam, S., Barkman, T.J., Nguyen, N.-p., Matasci, N., Nelson, D.R., Sayyari, E., Wafula, E.K., Walls, R.L., Warnow, T., An, H., Arrigo, N., Baniaga, A.E., Galuska, S., Jorgensen, S.A., Kidder, T.I., Kong, H.,

Lu-Irving, P., Marx, H.E., Qi, X., Reardon, C.R., Sutherland, B.L., Tiley, G.P., Welles, S.R., Yu, R., Zhan, S., Gramzow, L., Theißen, G., Wong, G.K.-S., Initiative, O.T.P.T.: One thousand plant transcriptomes and the phylogenomics of green plants. Nature **574**(7780), 679–685 (2019). https://doi.org/10.1038/s41586-019-1693-2

[67] Phillips, S.J., Anderson, R.P., Schapire, R.E.: Maximum entropy modeling of species geographic distributions. Ecological Modelling **190**(3), 231–259 (2006). https://doi.org/10.1016/j.ecolmodel.2005.03.026

[68] Jiménez-Valverde, A., Lobo, J.M.: Threshold criteria for conversion of probability of species presence to either–or presence–absence. Acta Oecologica **31**(3), 361–369 (2007). https://doi.org/10.1016/j.actao.2007.02.001