UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Botucatu

Instituto de Biociências de Botucatu - Unesp

ANÁLISES *IN SILICO* DE GENES BIOSSINTÉTICOS ORGANIZADOS EM *CLUSTERS* EM GENOMAS DA FAMÍLIA RUBIACEAE

SAMARA MIREZA CORREIA DE LEMOS

BOTUCATU - SP

2023

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Botucatu

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS
(GENÉTICA)

ANÁLISES *IN SILICO* DE GENES BIOSSINTÉTICOS ORGANIZADOS EM *CLUSTERS* EM GENOMAS DA FAMÍLIA RUBIACEAE

NOME DO CANDIDATO: **SAMARA MIREZA CORREIA DE LEMOS**
ORIENTADOR: **PROF. DR. DOUGLAS SILVA DOMINGUES**
COORIENTADOR: **PROF. DR. ALEXANDRE ROSSI PASCHOAL**

Tese apresentada ao Instituto de Biociências, câmpus de Botucatu, UNESP, para obtenção do título de Doutora no Programa de Pós graduação em Ciências Biológicas (Genética).

BOTUCATU - SP
2023

L557a

Lemos, Samara Mireza Correia de
Análises in silico de genes biossintéticos organizados em clusters em genomas da família rubiaceae / Samara Mireza Correia de Lemos. -- Botucatu, 2023
104 p. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Instituto de Biociências, Botucatu
Orientador: Douglas Silva Domingues
Coorientador: Alexandre Rossi Paschoal

1. Genômica. 2. Bioinformática. 3. Rubiaceae. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Biociências, Botucatu. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Agradecimentos

Agradeço a todas as pessoas que estiveram envolvidas em minha vida durante o processo do meu doutoramento. Minha mãe, por todo apoio, formação pessoal e amor. Meus amigos, que foram minha família e contribuíram como ouvintes, terapeutas, professores e portos de segurança.

Agradeço aos meus professores, Douglas, Paschoal, Romain e Marnix. Vocês acreditaram em mim e forneceram muitas oportunidades durante minha formação. Bem como as instituições que me receberam de braços abertos: UNESP e o Instituto de Biociências de Botucatu, o Instituto de Biotecnologia de Botucatu, o departamento de Botânica de Rio Claro, o Instituto de Pesquisa para o Desenvolvimento em Montpellier e a Universidade de Wageningen.

Agradeço à FAPESP como geradora de dados de RNA-seq de *Coffea* (processo 2016/10896-0). Agradeço à CAPES pelo investimento no meu doutorado (processo 88882.461712/2019-01), capacitação internacional (88887.570128/2020-00) e doutorado sanduíche (88887.570702/2020-00). A CAPES mantém viva a ciência no Brasil, por meio de investimentos e acesso à sociedade, por isso sou imensamente grata.

Resumo

As plantas são grandes fontes de compostos químicos. Tais metabólitos são resultantes dos processos metabólicos que garantem o crescimento e o desenvolvimento da planta, regulam a interação planta-ambiente e garantem sua proteção contra patógenos. Os compostos químicos das plantas são amplamente utilizados como medicamentos e produtos industriais. Graças aos avanços tecnológicos dos últimos anos, tecnologias de sequenciamento e estudos de genômica vegetal possibilitaram a descoberta de que genes envolvidos na biossíntese de compostos do metabolismo vegetal podem estar organizados em clusters - os clusters biossintéticos de genes (CMGs). Ainda assim, muitos passos das vias de biossíntese de metabólitos por clusters ainda não são conhecidos, e a literatura carece de estudos que explorem clusters biossintéticos de genes em determinadas famílias vegetais de importância comercial. No presente trabalho identificamos, caracterizamos e comparamos clusters biossintéticos de genes com especial interesse em espécies da família Rubiaceae. A família Rubiaceae é dividida em três subfamílias, Ixoroideae, Rubioideae e Cinchonoideae. Aplicamos abordagens de bioinformática em dados genômicos de oito plantas representantes das três subfamílias da família Rubiaceae e usamos o genoma de tomate (*Solanum lycopersicum*) como grupo externo. Identificamos 2372 possíveis clusters biossintéticos de genes contendo 35715 genes em oito espécies da família Rubiaceae e em *Solanum lycopersicum* (Solanaceae). Utilizando de ferramentas para genômica comparativa, identificamos que os clusters estão distribuídos em 549 famílias de clusters. Investigamos a conservação genômica dos clusters identificados nas subfamílias de Rubiaceae e identificamos uma maior conservação dos clusters na subfamília Ixoroideae. Observamos ainda uma maior conservação entre as subfamílias Ixoroideae e Cinchonoideae do que entre as subfamílias Ixoroideae e Rubioideae. Aplicamos a construção de redes de coexpressão com dados de transcriptoma de seis espécies representantes das três subfamílias de Rubiaceae e identificamos 207 clusters biossintéticos de genes expressando genes chave dos clusters preditos. No total atribuímos status de alta confiança a 204 clusters biossintéticos de genes que tiveram genes chaves coexpressos e estavam dentro de uma família de clusters. Este estudo traz a primeira análise da diversidade genômica de clusters biossintéticos de genes da família Rubiaceae, considerando conservação dentro de suas três principais subfamílias. A predição de clusters destaca o potencial dessas espécies como fonte de novos compostos bioativos de interesse básico e biotecnológico.

Palavras-chave: clusters biossintéticos de genes, genômica comparativa, bioinformática, Rubiaceae

Abstract

Plants produce valuable chemical compounds through metabolic processes that support their growth, help them interact with their environment, and protect them from diseases. These plant-derived chemicals are commonly used in medicines and industrial products. Advancements in technology, such as DNA sequencing and plant genomics, have revealed that many genes responsible for producing specific bioactive compounds are physically organized in close proximity within genomes into groups known as metabolic gene clusters (MGCs). However, there is relatively limited research on these gene clusters in certain plant families that are economically important. The plant family Rubiaceae has great chemical diversity and it is divided into three subfamilies: Ixoroideae, Rubioideae and Cinchonoideae. The objective of this study was to identify, characterize, and compare metabolic gene clusters in species belonging to the Rubiaceae family. We analyzed the genomic data of eight plants representing the three subfamilies from the Rubiaceae family and compared it with the tomato genome (*Solanum lycopersicum*) as a reference. A total of 2,372 potential metabolic gene clusters, which contained a total of 35,715 genes were identified. We categorized these clusters into 549 cluster families through comparative genomics. We also examined how these clusters were conserved within subfamilies of Rubiaceae. Notably, we found that the Ixoroideae subfamily showed a higher degree of conservation in these clusters compared to the Rubioideae subfamily. Furthermore, the conservation was greater between the Ixoroideae and Cinchonoideae subfamilies than between the Ixoroideae and Rubioideae subfamilies. Additionally, we constructed coexpression networks using transcriptome data from six species representing the three subfamilies of Rubiaceae. This allowed us to identify 207 metabolic gene clusters that expressed key genes from the predicted clusters. We identified 204 metabolic gene clusters that had coexpressed key genes and belonged to the same cluster family - these clusters were considered high confidence clusters. In summary, this study represents the first comprehensive analysis of the genetic diversity of metabolic gene clusters within the Rubiaceae family, considering their preservation in the three main subfamilies. Our findings highlight the potential of these plant species as sources of new bioactive compounds with both fundamental and biotechnological applications.

Keywords: metabolic gene clusters, comparative genomics, bioinformatics, Rubiaceae

Lista de figuras

Lista de tabelas

---

2 Capítulo 2

Sumário

CAPÍTULO 1

1.1 Introdução

1.1.1 Organização genômica vegetal

Os genomas das plantas são diversos e podem ser complexos. Diferente de outros eucariotos, as células vegetais possuem seu material genético distribuído não somente no núcleo e nas mitocôndrias, mas também nos cloroplastos (Grotewold et al., 2015). Quando levamos em consideração apenas o material genético do núcleo, há uma certa constância no número de genes codificantes de proteínas, mas outros fatores podem apresentar uma grande variabilidade - como o tamanho do genoma, sua densidade, o número de elementos repetitivos, o número de cromossomos em que o genoma está organizado e a ploidia de cada organismo (Grotewold et al., 2015). Assim, entender a organização do genoma vegetal envolve entender a disposição dos genes, elementos regulatórios e sequências não codificantes ao longo dos cromossomos. Ao longo dos últimos 20 anos, as tecnologias de sequenciamento foram essenciais para a descoberta e montagem dos genomas vegetais. A integridade, exatidão e contiguidade dos genomas dependem da finalidade para a qual foram produzidos, das tecnologias utilizadas e dos recursos dedicados à tarefa. A planta modelo *Arabidopsis thaliana* foi a primeira planta a ter uma montagem de genoma, no ano 2000 pela iniciativa do genoma de Arabidopsis (The Arabidopsis Genome Initiative, 2000). Hoje, com dados atualizados do repositório TAIR - *The Arabidopsis Information Resource*, sabemos que a planta contém 27 mil genes e aproximadamente 135 milhões de pares de bases. É considerada um organismo modelo devido ao tamanho do genoma reduzido e organizado em apenas 5 cromossomos, com 85% do genoma em regiões codificantes. Apesar do desenvolvimento de novas tecnologias de sequenciamento terem possibilitado o aumento na disponibilidade de montagens genômicas, determinadas espécies com genomas poliplóides e ricos em sequências repetitivas ainda são um desafio à montagem de genomas. Um exemplo de genoma vegetal complexo já montado é o do trigo (*Triticum aestivum*), que é hexaplóide e tem um tamanho de aproximadamente 16 bilhões de pares de bases (International Wheat Genome Sequencing Consortium, 2018). A montagem de genomas complexos evoluiu graças ao uso combinado de grandes bibliotecas de inserção,

tecnologias de sequenciamento que geram leituras mais longas (Pacific Biosciences, Oxford Nanopore), técnicas de captura de cromatina e novas abordagens de bioinformática (Kersey, 2019). Além da evolução nas tecnologias de montagem de genomas, houve também a evolução de tecnologias no estudo dos elementos genômicos. As variações na organização genômica refletem a necessidade das plantas de responder rapidamente a mudanças ambientais, regulando a expressão gênica de forma flexível e adaptativa. Devido a essas respostas, as plantas produzem metabólitos especializados com uma vasta gama de diversidade funcional e estrutural. O estudo da genômica funcional baseado em genômica, transcriptômica e metabolômica é uma ferramenta poderosa para decodificar o papel dos genes codificadores de proteínas e sua contribuição na codificação de enzimas de vias bioquímicas - estas fundamentais para o entendimento do metabolismo vegetal. Dos cerca de 1 milhão de metabólitos estimados a serem sintetizados pelas plantas (Afendi et al., 2012), conhecemos as vias biossintéticas de apenas cerca de 0,1% (Schläpfer et al., 2017). Nesse sentido, estratégias computacionais que atribuam funções metabólicas a genes em plantas são de grande interesse. Atualmente existem bancos de dados resultantes de genômica funcional contendo informações sobre números de enzimas associadas ao metabolismo. No caso das plantas citadas acima como exemplos, *A. thaliana* possui ~8700 (AraCyc v. 17.2.0; Mueller et al., 2003) e o trigo possui 30600 (BreadwheatCyc v. 3.0.2; Caspi et al., 2018).

.

1.1.2 Metabolismo Vegetal

O metabolismo vegetal refere-se ao conjunto de processos bioquímicos que ocorrem nas plantas para manter a vida, crescimento e reprodução. Envolve um grande número de reações químicas que ocorrem nas células vegetais, permitindo a síntese de moléculas, o armazenamento e utilização de energia, a regulação do crescimento e desenvolvimento e a interação com o ambiente. Atividades metabólicas responsáveis por processos vitais na célula, como a fotossíntese, respiração, fixação de nitrogênio e a biossíntese de carboidratos e lipídeos compõem o metabolismo primário vegetal (Hartmann, 2007). No entanto, diversos processos mais específicos, envolvendo a relação de determinada espécie com o ambiente e interação com outros organismos, são mediados por outros componentes do metabolismo - o metabolismo especializado (Hartmann, 2007). Ele recebe este nome por poder conferir caráter único a uma dada espécie vegetal. Normalmente os produtos do metabolismo especializado são classificados de acordo com a sua via metabólica e divididos em três

classes de moléculas principais: os compostos fenólicos, terpenos e os compostos contendo nitrogênio (Erb & Kliebenstein, 2020). Há ainda o grupo dos hormônios vegetais, classificados como pequenos compostos que regulam processos como a produção de outros metabólitos, interagindo com as proteínas receptoras. (Erb & Kliebenstein, 2020). Existem vários exemplos de metabólitos especializados que possuem grande valor comercial já que podem ter propriedades medicinais, aromáticas, corantes, repelentes ou tóxicas, e muitos têm sido utilizados pelo ser humano para diversos fins, como na medicina, indústria alimentícia e produção de perfumes. Por exemplo, os flavonóides, alcalóides, fenóis e terpenos extraídos de plantas do gênero *Passiflora* são explorados na indústria farmacêutica por seus efeitos calmantes e analgésicos (Farag et al., 2016). O mesmo ocorre com a artemisinina, um terpeno produzido pela planta *Artemisia annua* com propriedades antimaláricas (Paddon et al., 2013). Outro exemplo é o paclitaxel, um diterpeno produzido pela planta *Taxus brevifolia* com ações anticancerígenas (Howat et al., 2014). Na indústria alimentícia, alguns dos responsáveis pelos sabores marcantes do chá (*Camellia sinensis*) são polifenóis, cafeína, teaninas, óleos voláteis e outros metabólitos (Xia et al., 2017). Quanto ao chocolate, vários componentes químicos dos grãos crus de cacau (*Theobroma cacao*) participam na formação de sabores específicos de acordo com o processo de produção (Aprotosoaie et al., 2016) como alcalóides, metilxantinas, polifenóis, proteínas e carboidratos.

1.1.3 Organização Genômica do Metabolismo Vegetal

Diversos estudos são realizados para compreender como funcionam os processos que geram os metabólitos nas plantas (Figura 1). Um exemplo clássico é a elucidação da produção de benzoxazinóides como metabólitos especializados das gramíneas, que foram identificados no início dos anos 1960 como metabólitos que funcionam como pesticidas naturais e exibem propriedades alelopáticas (Jonczyk et al., 2008). Em milho, a biossíntese dos benzoxazinóides ramifica-se do metabolismo primário na conversão de indol-3-glicerol fosfato em indol pela a enzima codificada pelo gene BX1. A introdução de quatro átomos de oxigênio na porção indol que produz o benzoxazinóide DIBOA é catalisada por quatro enzimas monooxigenases dependentes do citocromo P450, produzidas pelos genes BX2 a BX5. O DIBOA resultante é glicosilado pelas enzimas UDP-glicosiltransferases, produtos dos genes BX8 e BX9. A hidroxilação na posição C-7 do glicosídeo é catalisada pelo 2ODD, gene BX6. A etapa final da via é a O-metilação de TRIBOA-glc por BX7 para produzir DIMBOA-glc (Jonczyk et al., 2008). Outro estudo investigou a evolução dos

genes envolvidos nos processos de metabolismo primário e especializado em 16 organismos modelo, de algas a angiospermas (Chae et al. 2014). Foi possível observar que genes que codificam funções metabólicas especializadas duplicam-se em um grau muito maior e por diferentes mecanismos, quando comparados aos genes do metabolismo primário (Chae et al. 2014). Há também evidências de agrupamento parcial de genes de outras vias do metabolismo vegetal e de duplicação de pares de genes relacionados funcionalmente (Nützmann et al., 2018). Além disso, genes do metabolismo especializado exibem padrões específicos de expressão e de organização física no genoma (Chae et al. 2014). Em um estudo do genoma de mirtilo (*Vaccinium corymbosum*), foi identificado que os genes envolvidos na biossíntese de antioxidantes apresentaram um padrão de expressão distinto e específico por fase de desenvolvimento (Colle et al., 2019). A maioria dos genes associados à biossíntese de antioxidantes como flavonóides e antocianinas possui pelo menos uma duplicação em tandem, com tamanhos variando de 2 a 10 cópias de genes (Colle et al., 2019). Apesar de ser um desafio a descoberta de genes que codificam uma mesma via metabólica, muitos dos genes envolvidos no metabolismo especializado vegetal estão organizados em diversas formas (Figura 2), sendo as mais comuns as repetições em tandem - padrões de repetição sequencial de bases nitrogenadas, ou *clusters* - aglomerados de genes (Nützmann et al., 2018; Smit & Lichman, 2022). Pares ou matrizes de genes parálogos - genes homólogos que derivam de um ancestral comum através de um evento de duplicação - e podem estar localizados proximamente no genoma em formato de repetições em tandem. O mesmo pode ocorrer com genes adjacentes de origens evolutivas distintas envolvidos em vias metabólicas (Figura 2; Smit & Lichman, 2022). Os clusters biossintéticos de genes (CMGs), podem ser definidos como três ou mais genes não-homólogos, co-localizados no genoma e que codificam enzimas da mesma via biossintética. CMGs podem incluir genes *core* ou biossintetizantes que codificam enzimas que sintetizam a estrutura central do composto metabólico, e genes acessórios, que codificam enzimas de adaptação (que modificam a estrutura central do composto), fatores de transcrição e transportadores (Smit & Lichman, 2022). *clusters biossintéticos* não surgem pela transferência horizontal de genes, como usualmente ocorre em bactérias, mas pelo recrutamento de genes de outras regiões do genoma por meio de duplicação e neofuncionalização, em mecanismos ainda pouco conhecidos e explorados (Nützmann et al., 2018). Além disso, os genes dentro dos CMGs são geralmente regulados de forma coordenada e isso se torna uma ferramenta valiosa na caracterização funcional da via metabólica associada a esses genes (Zhan et al., 2022). Um exemplo de CMG contendo genes que agem sequencialmente numa via metabólica é o Dhurrin identificado em *Sorgum bicolor.* Este CMG possui três genes adjacentes de diferentes origens evolutivas que juntos são suficientes para formar o glicosídeo cianogênico

dhurrin a partir da tirosina (Darbani et al., 2016). O primeiro CMG de plantas foi identificado em 1997 (Figura 1) e desde então mais de 40 CMGs foram reportados no repositório MIBiG - Minimum Information about a Biosynthetic Gene cluster (Zhan et al., 2022; Terlouw et al., 2023).

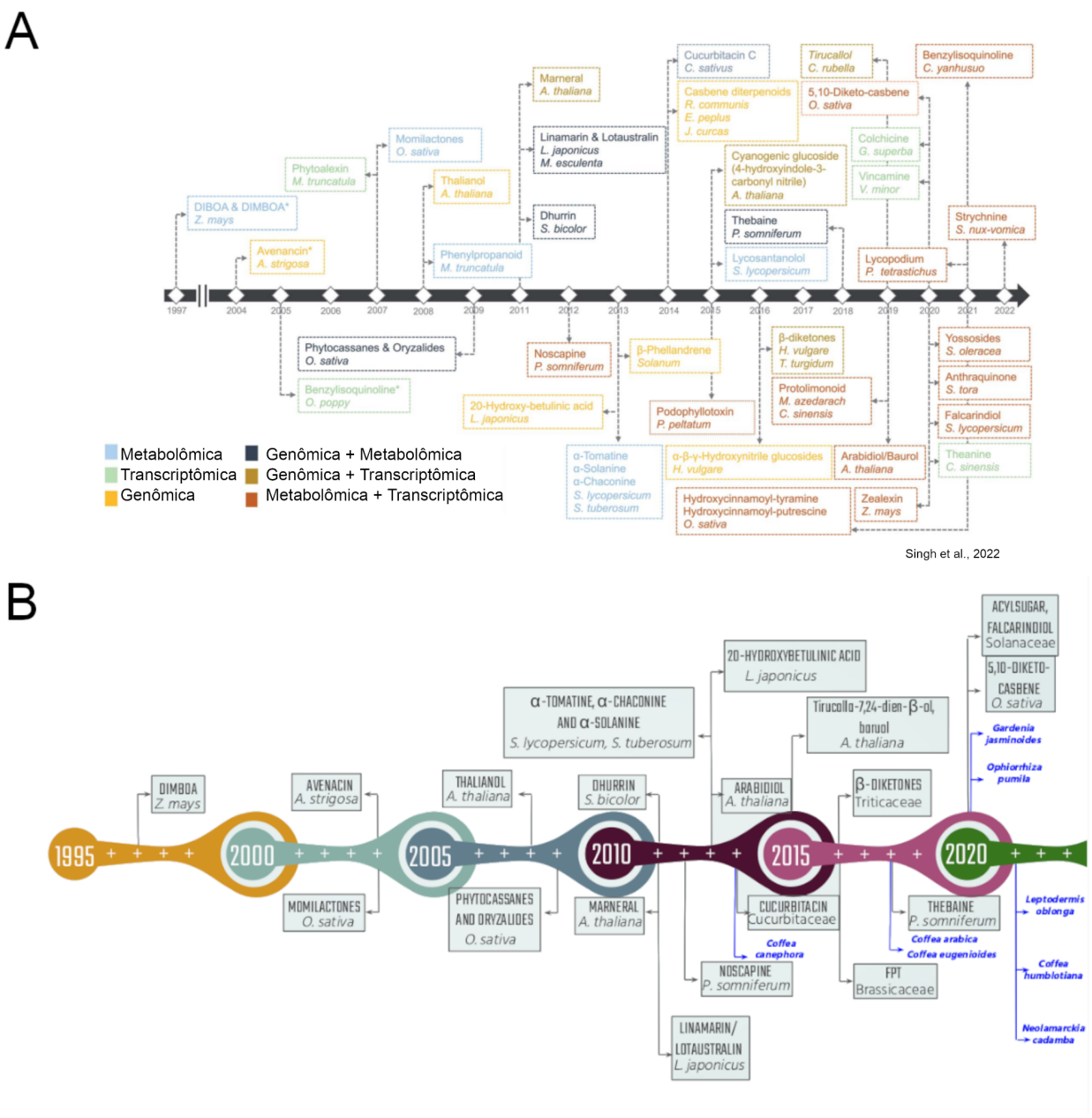

Figura 1 - Identificação de vias metabólicas e clusters biossintéticos de genes em plantas. A: Linha do tempo contendo vias metabólicas identificadas em plantas com o uso de diferentes técnicas ômicas (Singh et al., 2022). B: Linha do tempo contendo clusters biossintéticos de genes em plantas. A disponibilidade de sequências genômicas de sete plantas da família Rubiaceae está destacada em azul.

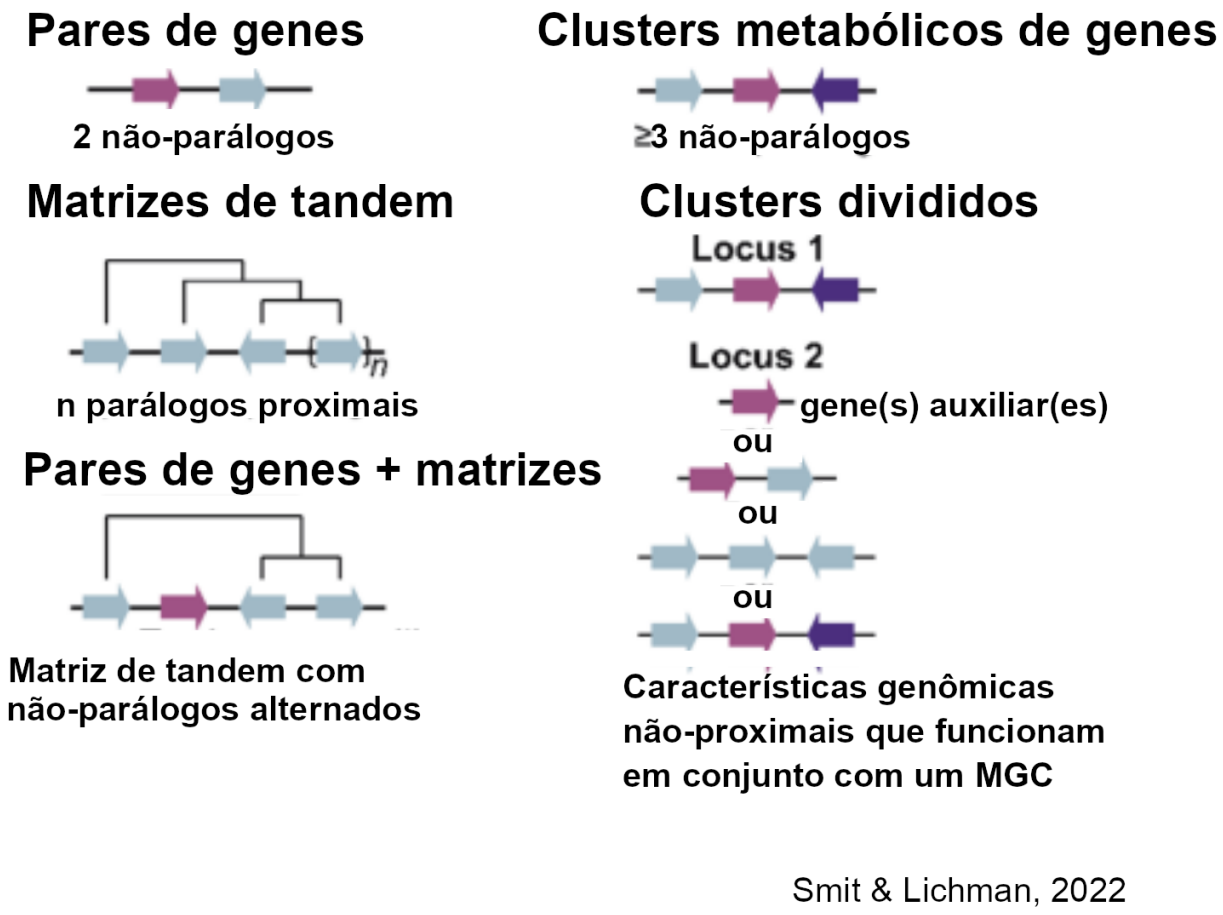

Figura 2 - Características genômicas do metabolismo especializado em plantas. Genes não parálogos são indicados por setas coloridas diferentes com linhas de conexão indicativas de uma região genômica compartilhada. Linhas em forma de árvore ilustram relacionamentos parálogos. Essa figura é uma versão modificada da figura 1 de Smit & Lichman, 2022.

1.1.4 Ferramentas de Bioinformática na Identificação de CMGs

Nos últimos anos foram desenvolvidos pipelines para a identificação de clusters biossintéticos de genes (em inglês, biosynthetic gene clusters) como o Plant Cluster Finder (Schläpfer et al., 2017), plantiSMASH (Kautsar et al., 2017) e phytoClust (Töpfer et al., 2017). O PlantClusterFinder detecta CMGs em genomas sequenciados, utilizando um arquivo de localização de genes disponibilizado pelo usuário e um banco de dados de vias metabólicas criado previamente, para identificar genes codificantes de enzimas localizados proximamente em um cromossomo. O PlantiSMASH emprega uma coleção de 62 modelos ocultos de Markov representantes das famílias

de enzimas relacionadas ao metabolismo de plantas, para minerar sequências genômicas fisicamente co-localizadas levando em consideração a distância intergênica do genoma minerado (Kautsar et al., 2018). O PhytoClust emprega uma estratégia similar, com uma coleção de 26 modelos ocultos de Markov representantes das famílias de enzimas relacionadas ao metabolismo especializado de plantas, para minerar determinadas sequências genômicas fisicamente co-localizadas. Usando o pipeline plantiSMASH, um estudo explorou a relação entre CMGs e eventos evolutivos, como a duplicação do genoma e a poliploidização em *Arabidopsis thaliana* e *Cleome violacea* (Ma et al., 2019). Os autores identificaram CMGs em regiões de sintenia entre os dois genomas, compartilhando um cluster metabólico (Ma et al., 2019). Em outro estudo em plantas de *Nicotiana tabacum*, foi desenvolvido um pipeline integrando dados genômicos, transcriptômicos e proteômicos para identificar clusters de genes metabólicos envolvidos na biossíntese de fitoalexinas e capsidiol, que são metabólitos importantes na defesa da planta contra patógenos (Chen et al., 2019).

Além de identificarem dois CMGs envolvidos na síntese desse terpenoide, verificaram sua conservação em outras espécies de Nicotiana (Chen et al., 2019). Vias metabólicas funcionais normalmente compartilham padrões de expressão gênica nos mesmos tecidos e podem ser ativadas ou reprimidas pelos mesmos estímulos. Genes agrupados em MGCs podem ser regulados mais facilmente do que genes que não estão co-localizados, ou ainda, o agrupamento pode acelerar o recrutamento de genes para o mesmo regulon (unidade genética constituída por um grupo não contíguo de genes sob o controle de um gene regulador) durante a evolução metabólica Smit & Lichman (2022). De fato, os clusters mais bem caracterizados em plantas apresentam coexpressão de seus genes (Polturak et al., 2022; Smit & Lichman, 2022).

Em análises de coexpressão gênica, uma série de abordagens computacionais e estatísticas são empregadas para medir grupos de genes que estão sendo expressos numa mesma condição, órgão e momento. Abordagens de coexpressão foram aplicadas na descoberta de diversos genes relacionados ao metabolismo especializado de plantas, complementando a descoberta de genes que regulavam CMGs ou estavam dentro de um CMG, por exemplo em *Arabidopsis thaliana, Zea mays, Brachypodium distachyon, Artemisia annua* e *Nicotiana tabacum* (Delli-Ponti et al., 2021; Li et al., 2023). No pipeline Clust, por exemplo, a metodologia de coexpressão é aplicada na busca por clusters diretamente a dados transcriptômicos. Por considerar que nem sempre genes que são coexpressos em condições específicas serão coexpressos em outras situações, Clust possui melhor desempenho comparado a outras ferramentas (Abu-Jamous & Kelly, 2018). Outro recurso disponível para a descoberta de CMGs é o repositório MIBiG - Minimum Information about a

Biosynthetic Gene cluster (Terlouw et al., 2023). O repositório contém um formato de dados padronizado que descreve as informações mínimas necessárias para caracterizar exclusivamente um CMG e atualmente possui 41 registros de CMGs em plantas.

1.1.5 Plantas da família Rubiaceae

A família Rubiaceae é a quarta maior família de angiospermas e possui mais de 600 gêneros e 13000 espécies (Bremer & Eriksson, 2009; Martins & Nunes, 2015). Em um estudo de filogenia realizado com dados de mais de 500 espécies e mais de 300 gêneros da família, foi identificado a subdivisão em três subfamílias: Cinchonoideae, Ixoroideae, Rubioideae (Bremer & Eriksson, 2009). Algumas plantas da subfamília Ixoroideae produzem alcalóides de grande valor comercial, especialmente do gênero *Coffea* (Denoeud et al., 2014; Perrois et al., 2015). Todas as espécies de *Coffea* são diplóides com número cromossômico x = 11, com exceção de *Coffea arabica*, que é um alotetraplóide resultante de cruzamento recente entre *Coffea eugenioides* e *Coffea canephora* (Scalabrin et al., 2020). Recentemente sequências genômicas de várias espécies dessa família foram disponibilizadas publicamente: *Coffea arabica, Coffea canephora, Coffea eugenioides, Coffea humblotiana, Gardenia jasminoides, Gynochthodes officinalis, Leptodermis oblonga, Mitragyna speciosa, Ophiorrhiza pumila, Neolamarckia cadamba, Sherardia arvensis, Oldenlandia corymbosa* (https://www.ncbi.nlm.nih.gov/genome/?term=rubiaceae) e há outros projetos de sequenciamento em andamento (Kochko et al., 2015; Xu et al., 2020; Zhao et al., 2021; Rai et al., 2021; Guo et al., 2021; Raharimalala et al., 2021; Brose et al., 2021; Julca et al., 2023). Estudos sobre vias de biossíntese de metabólitos em Rubiaceae atualmente são bastante pontuais, como o da crocina em *G. jasminoides* (Xu et al., 2020), cafeína em *C. canephora* (Denoeud et al., 2014; Perrois et al., 2015), diterpenos em *C. arabica* (Ivamoto et al., 2017) e monoterpenos indol alcalóides em *O. pumila* e *N. cadamba* (Zhao et al., 2021; Rai et al., 2021). Recentemente um MGC envolvido na biossíntese de ciclopeptídeos em *C. arabica* foi identificado (Lima et al., 2023), mas ainda não existem abordagens de larga escala que levem em consideração localização física de genes e que comparem as espécies da família, um estudo completo que identifique, caracterizando e analisando os clusters biossintéticos de genes sob um viés evolutivo. Como o Brasil é o maior produtor e exportador de café do mundo (International Coffee Organization, 2021) e é também referência de pesquisa nessa cultura, conhecimentos sobre os componentes genômicos e metabólicos do gênero *Coffea* são de grande valia para estudos em melhoramento vegetal e biotecnologia. Apresentamos

neste estudo o uso de um conjunto de abordagens computacionais para investigação de clusters biossintéticos de genes nos genomas de oito espécies de plantas da família Rubiaceae (Figura 3). A caracterização de genes envolvidos no metabolismo especializado de plantas é uma oportunidade de agregar conhecimento básico em um tema insuficientemente abordado em genômica, bem como a aplicação desse estudo em espécies de interesse agrícola trazem o potencial da descoberta de mecanismos moleculares de relevância para estratégias biotecnológicas de melhoramento e aumento da produção de compostos metabólicos que podem auxiliar em mudanças de perfil de sabor e aroma da bebida ou resistência a pragas e doenças.

1.2 Objetivos

1.2.1 Geral:

Identificar clusters biossintéticos de genes nos genomas de oito plantas da família Rubiaceae e *Solanum lycopersicum*.

1.2.2 Específicos:

- Caracterizar os clusters biossintéticos de genes identificados.

- Comparar os clusters biossintéticos de genes identificados considerando aspectos evolutivos.

CAPÍTULO 2

2.1 Organização da tese

Materiais, métodos, resultados e discussão estão organizados em um capítulo apresentado no formato de artigo científico. O artigo intitulado “Genome Mining of Metabolic Gene Clusters in the Rubiaceae family” descreve os resultados da pesquisa sobre clusters biossintéticos de genes em oito espécies da família Rubiaceae. No artigo também é apresentada uma análise genômica comparativa onde o genoma de tomate foi usado como grupo externo. Esse artigo científico foi publicado na revista *Computational and Structural Biotechnology Journal*. Os materiais suplementares estão disponíveis no endereço online: https://doi.org/10.5281/zenodo.8221450. Para acessá-los, basta copiar o endereço do link e colar na barra de endereços do navegador. No anexo foram incluídos três artigos científicos e um capítulo de livro nos quais fui co-autora. No artigo científico intitulado “Transcriptomic alterations in roots of two contrasting *Coffea arabica* cultivars after hexanoic acid priming” (anexo 1) aplicamos o uso de RNA-seq para analisar o transcriptoma de raízes de duas cultivares de *Coffea arabica* sob ação do elicitor ácido hexanóico, identificando genes diferencialmente expressos que modulam as respostas da planta ao estresse. No artigo científico intitulado “Maize resistance to witchweed through changes in strigolactone biosynthesis” (anexo 2) identificamos as vias de biossíntese e um cluster metabólico de genes produzindo o hormônio strigolactona em milho. Aplicando experimentos de bioinformática e biologia molecular, identificamos que a modulação de alguns genes da via pode reduzir a produção desse hormônio. No artigo científico intitulado “The genome and population genomics of allopolyploid *Coffea arabica* reveal the diversification history of modern coffee cultivars” (anexo 3) apresentamos novas montagens do genoma de *Coffea arabica*, e análises de genômica comparativa com *Coffea canephora* e *Coffea eugenioides*. No capítulo “A Bioinformatics Tool for Efficient Retrieval of High-Confidence Terpene Synthases (TPS) and Application to the Identification of TPS in *Coffea* and *Quillaja*”, do livro “Plant Secondary Metabolism Engineering” (anexo 4) construímos uma ferramenta para identificar terpeno sintases em sequências proteômicas, testando em *Coffea* e *Quillaja*.

As referências citadas ao longo do capítulo bem como na seção de Introdução estão apresentadas na seção Referências, ao final desta tese.

2.2 Artigo - Genome Mining of Metabolic Gene Clusters in the Rubiaceae family

Samara Mireza Correia de Lemos[1], Alexandre Rossi Paschoal[2], Romain Guyot[3], Marnix Medema[4], Douglas Silva Domingues[1,5*]

[1]Graduate Program in Biological Sciences (Genetics), Institute of Biosciences, São Paulo State University, UNESP, Botucatu, Brazil
[2]Department of Computer Science, Federal University of Technology-Parana, Cornelio Procopio 86300-000, Brazil
[3]UMR DIADE, Institut de Recherche pour le Développement (IRD), Université Montpellier, CIRAD, Montpellier, France
[4]Bioinformatics Group, Wageningen University, 6708 PB Wageningen, The Netherlands
[5]Department of Genetics, “Luiz de Queiroz” College of Agriculture, University of São Paulo, ESALQ/USP, Piracicaba, Brazil
*Corresponding author: Douglas Silva Domingues, address: Depto. de Genética, ESALQ/USP, Av Pádua Dias, 11 Piracicaba, SP, Brazil 13418-900, phone number: +55(19)34294125 #30, e-mail: dougsd@usp.br

Abstract

The Rubiaceae plant family, comprising 3 subfamilies and over 13,000 species, is known for producing significant bioactive compounds such as caffeine and monoterpene indole alkaloids. Despite an increase in available genomes from the Rubiaceae family over the past decade, a systematic analysis of the metabolic gene clusters (MGCs) encoded by these genomes has been lacking. In this study, we aim to identify and analyze metabolic gene clusters within complete Rubiaceae genomes through a comparative analysis of eight species. Applying two bioinformatics pipelines, we identified 2372 candidate MGCs, organized into 549 gene cluster families (GCFs). To enhance the reliability of these findings, we developed coexpression networks and conducted orthology analyses. Using genomic data from *Solanum lycopersicum* (Solanaceae) for comparative purposes, we provided a detailed view of predicted metabolic enzymes, pathways, and coexpression networks. We bring some examples of MGCs and GCFs involved in biological pathways of terpenes, saccharides and alkaloids. Such insights lay the groundwork for discovering new compounds and associated MGCs within the Rubiaceae family, with potential implications in developing more robust crop species and expanding the understanding of plant metabolism. This large-scale exploration also provides a new perspective on the evolution and structure-function relationship of these clusters, offering opportunities for the highly efficient utilization of these unique metabolites. The outcome of this study contributes to a broader comprehension of the biosynthetic pathways, elucidating multiple aspects of specialized metabolism and offering innovative avenues for biotechnological applications.

Keywords: Metabolic Gene Cluster, Comparative Genomics, Rubiaceae

1. Introduction

Plant natural compounds are the main source of bioactives for medicinal, pharmaceutical, agricultural and industrial applications (Twaij & Hasan, 2022). The antimalarial artemisinin, anti-cancer paclitaxel, the codeine analgesic and anti-diabetic metformin are some of the many examples of plant-derived pharmaceuticals (Srivastav et al., 2020; Twaij & Hasan, 2022). In ecosystems, plant bioactive compounds have many ecological functions, such as adaptation to the abiotic and biotic environment, defense against pests and pathogens, competition for nutrients and signaling for seed dispersal pollinators (Pichersky & Lewinsohn, 2011; Maeda & Fernie, 2021; Rieseberg et al., 2022). It is estimated that more than 200,000 known metabolites are products of plant metabolism (Kessler & Kalske, 2018). In biosynthetic pathways, which are a series of biochemical steps that will result in a metabolite, genes involved can either be dispersed across multiple chromosomes or be organised in a physically proximate manner. Although there are many compounds, only a few have well-established metabolic pathways supported by genomic information. Over the past decade, the discovery of biosynthetic compounds has been aided by the development of genome mining and omics approaches, as reviewed by Singh et al. (2022) and Zhao & Rhee (2022). Genes that compose a biosynthetic pathway can be organized as pairs, tandem arrays and biosynthetic or metabolic gene clusters (Smit & Lichman, 2022). A metabolic gene cluster is formed when a set of at least three genes that are of distinct evolutionary origin and are co-localized in the genome contribute to a specific metabolic pathway, ideally acting sequentially (Medema & Osbourn, 2016; Smit & Lichman, 2022). The structure of an MGC often encodes enzymes responsible for creating the core metabolite and tailoring enzymes that modify this structure along with regulatory transcription factors and transporters that carry metabolites and necessary precursors. There are several examples of MGCs discovered using omics approaches in plants, as reviewed in Wang et al., 2022 and Zhan et al., 2022. For example, there is a gene cluster involved in the production of terpenoids in Tomato. The MGC consists of one alcohol oxidase, 5 terpene synthases, 2 cis-prenyl transferases and one functional cytochrome P450 that work together to produce mono and diterpenes in the petiole part of the leaf (Matsuba et al., 2013; Matsuba et al., 2015). Tohge and Fernie (2020) well illustrate in their review cases where the genomic clustering of specialized metabolite genes result in the synthesis of a given compound. More recently, with the large-scale analysis of MGCs across several genomes it was possible to group putative homologous MGCs into gene cluster families (GCFs) (Kautsar et al., 2021). Gene cluster families are groups of MGCs that are functionally closely related and encode the production of the same or very similar molecules (Kautsar et al., 2021). This concept has been employed in genome mining for bacteria (Mohite et al., 2022) and fungi (Robey et al., 2021). Most of the studies that identify plant MGCs focus on one species, as the case of MGCs identification in tobacco (Chen et al., 2019) or selected species from distinct families (Schläpfer et al., 2017), and more recently, MGCs discovery is starting to be incorporated more frequently in studies reporting the assembly and annotation of new plant genomes (Li et al., 2022; Yang et al., 2022; Wu et al., 2022; Liu et al., 2023).

The Rubiaceae is a plant family in the Magnoliopsida class, containing 3 subfamilies (Bremer & Eriksson, 2009), more than 600 genera and 13,000 species (Martins & Nunes, 2015). The first genome of a Rubiaceae family plant, *Coffea canephora*, was published in 2014 (Denoeud et al., 2014). Since then, several other genomes have been documented in the literature (Figure 1; Burge, 2020; Lau et al., 2020; Xu et al., 2020; Brose et al., 2021; Zhao et al., 2021; Guo et al., 2021; Rai et al., 2021; Raharimalala et al. 2021; Wang et al., 2021; Canales et al., 2022; Julca et al., 2023). Plants from this family produce not only well-known alkaloids like caffeine, but other metabolites with great pharmacological potential, including several terpenoids (Lau et al., 2020; Adewole et al., 2021). For example, camptothecin is a monoterpene indole alkaloid produced by *Ophiorrhiza pumila* that possesses antitumor activities (Shi et al., 2020).

There is a small number of studies analysing the genomic basis of plant bioactive compounds synthesis in the Rubiaceae family. Some examples are the study of the crocin metabolic pathway in *Gardenia jasminoides* (Xu et al., 2020), caffeine in *Coffea canephora* (Denoeud et al., 2014; Perrois et al., 2015), monoterpene indole alkaloids (MIAs) in *Ophorrhiza pumila* (Rai et al., 2021), cadambine in *Neolamarckia cadamba* (Zhao et al., 2021), and ursolic acid in *Oldenlandia corymbosa* (Julca et al., 2023).

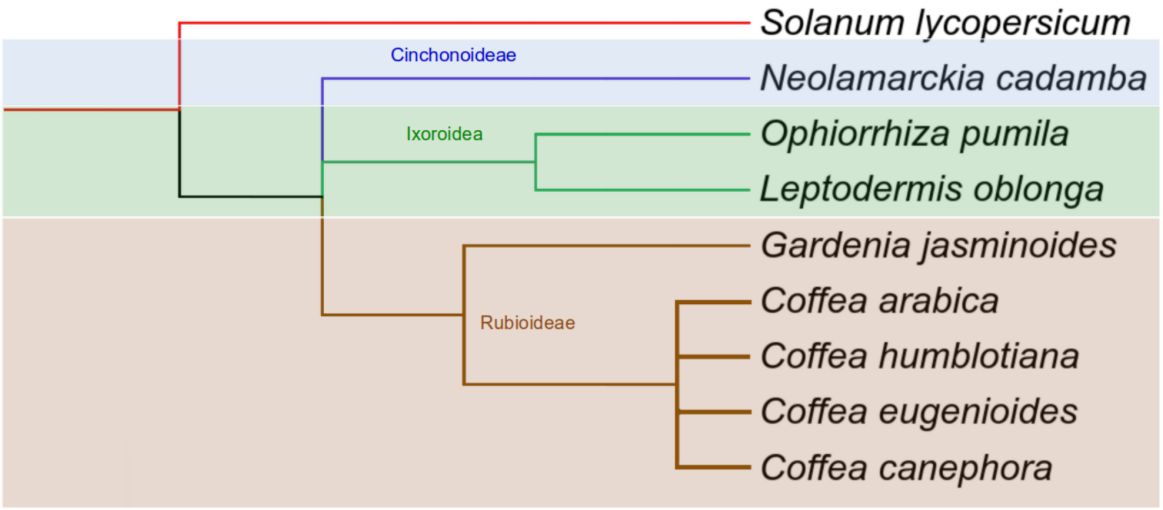

Figure 1 - Phylogenetic representation of the eight plant species of Rubiaceae with three subfamilies and tomato (*Solanum lycopersicum*).

For some Rubiaceae species, i.e. *Ophiorrhiza pumilla* and *Neolamarckia cadamba*, genome analysis included the prediction of MGCs. In *O. pumilla*, it was found specific clusters with highly coexpressed genes, indicating their possible role in MIA biosynthesis (Rai et al., 2021); however, a comparative family-wide analysis is still lacking.

The study advances the field of computational plant genomics by conducting a pioneering detailed comparative analysis of metabolic gene clusters (MGCs) and cluster families in a plant family, using the Rubiaceae family as a case study. By incorporating methods such as orthology assessment, gene family expansions, coexpression analysis, and a comparative analysis of metabolic gene clusters (MGCs), this investigation will facilitate the priorization of previously unknown pathways to understand the synthesis of bioactive compounds in plants. This pioneering comparative genomics research within the

Rubiaceae family seeks to lay a foundational framework for the identification of new compounds and their corresponding MGCs. The potential benefits of uncovering these MGCs are manifold, including the improvement of crop resilience and the exploration of novel bioactive substances, which could have profound implications for applications in both agricultural and medicinal contexts.

2. Materials and methods

2.1 Genomic and annotation data

For our analysis, we considered Rubiaceae genomes with high-quality assemblies (chromosome-level sequencing and a BUSCO score of over 97%) with publicly available deduced proteomes and GFF-formatted genome coordinate files (Table 1; Supplementary File 1). By December 1st, 2022, eight species from three subfamilies met these criteria: *Coffea arabica, Coffea canephora, Coffea eugenioides, Coffea humblotiana, Gardenia jasminoides, Leptodermis oblonga, Ophiorrhiza pumila* and *Neolamarckia cadamba*. Table 1 summarizes these sources. We adopted *Solanum lycopersicum*, a Solanaceae, as an outgroup for our comparative genomics analysis, as it is phylogenetically closer to Rubiaceae than *Arabidopsis thaliana* (Brassicaceae), providing a more relevant comparison, with previous predictions of MGCs (Matsuba et al., 2015; Fan et al., 2020; Zhou & Pichersky, 2020). The genome sequence and annotation files from *Solanum lycopersicum* release SL4.0 (Hosmani et al., 2019) were downloaded from https://solgenomics.net/organism/Solanum_lycopersicum/genome/.

Table 1 Species used in the present study

Species	Subfamily	Assembly	Authors	Source
Coffea arabica	Rubioideae	Cara_1.0	Johns Hopkins University	https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_003713225.1
Coffea canephora	Rubioideae	AUK_PRJEB4211_v1	Denoeud et al., 2014	https://www.ncbi.nlm.nih.gov/data-hub/genome/GCA_900059795.1/
Coffea eugenioides	Rubioideae	Ceug_1.0	Johns Hopkins University	https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_003713205.1/
Coffea humblotiana	Rubioideae	release 1.0	Raharimalala et al. 2021	https://solgenomics.net/organism/Coffea_humblotiana/genome
Gardenia jasminoides	Rubioideae	release 1.0	Xu et al., 2020	https://genomevolution.org/coge/api/v1/genomes/62692/sequence
Leptodermis oblonga	Ixoroidea	release 1.0	Guo et al., 2021	https://www.ncbi.nlm.nih.gov/data-hub/genome/GCA_016801395.1/
Ophiorrhiza pumila	Ixoroidea	release 1.0	Rai et al., 2021	https://pumila.kazusa.or.jp/
Neolamarckia cadamba	Cinchonoideae	release 1.0	Zhao et al., 2021	https://figshare.com/s/ed20e0e82a4e7474396b

2.2 Identification of Metabolic Gene Clusters (MGCs) through Genome Mining

Our goal was to identify metabolic gene clusters using two genome mining tool approaches (PlantClusterFinder and PlantiSMASH), compare the results, and apply criteria to select high confidence MGCs (Figure 2). To acquire a set of high-confidence MGCs, we consider results only from genomes with defined chromosomes. We used E2P2 v4.0 (Hawkins et al. 2021) and annotated protein sequences to identify enzymes associated with plant metabolic pathways and then used Pathway Tools v. 26 (Karp et al. 2022) and the PathoLogic software with default settings to generate metabolic pathway databases. These predicted pathways were manually filtered to only include those present in plants. To predict metabolic gene clusters, we modified a method based on previous studies by Schläpfer et al. (2017) and Chen et al. (2019) (see Figure 2). The output file from E2P2 was used with Pathway Tools to create species-specific metabolic pathway databases. These databases were then exported and inputted into PlantClusterFinder (PCF) version 1.3 (https://github.com/carnegie/PlantClusterFinder, Schläpfer et al. 2017), which identifies groups of metabolic genes located contiguously on the same scaffold using sliding window searching. Default parameters were used for PlantClusterFinder.
Finally, we used PfamScan v. 1.6 (https://github.com/gpertea/gsrc/blob/master/scripts/pfam_scan.pl; El-Gebali et al. 2019) to determine protein domains for all genes identified by PlantClusterFinder.
We also used PlantiSMASH v. 1.0, a computational pipeline that predicts plant MGCs using specific HMM profiles (Kautsar et al. 2017), to identify MGCs. We input the genome sequences and annotation files in GFF3 format and applied the dynamic cutoff parameter for analysis.

2.3 Clustering and evolutionary analysis of metabolic gene clusters in Rubiaceae

After identifying MGCs in Rubiaceae plants, we aimed to determine if they showed evolutionary conservation. To achieve this, first we compared the genomes of the study using Orthofinder v. 2.3.8 (Emms & Kelly, 2019) with default parameters to infer orthology and Orthovenn3 (Sun et al., 2023) with default parameters to infer gene families expansions and contractions and synteny analysis. We then grouped the MGCs into gene cluster families, identified protein domains of predicted metabolic enzymes, and checked for orthology relationships. To classify the high-confidence MGCs into families, we utilized the Biosynthetic Gene Similarity Clustering and Prospecting Engine (BiG-SCAPE) v. 1.1.0 (Navarro-Muñoz et al., 2020). We applied the "mix" and "no-classify" parameters and set a cutoff value of 1.0 as a raw distance. Chae et al. (2014) proposed a system to classify MGCs into 13 primary functional classes. Later, Schläpfer et al. (2017) applied this system to classify MGCs detected with the PlantClusterFinder tool. In our study, we used the same strategy to classify the MGCs that we identified using the PCF tool (refer to Figure 4A for more details).

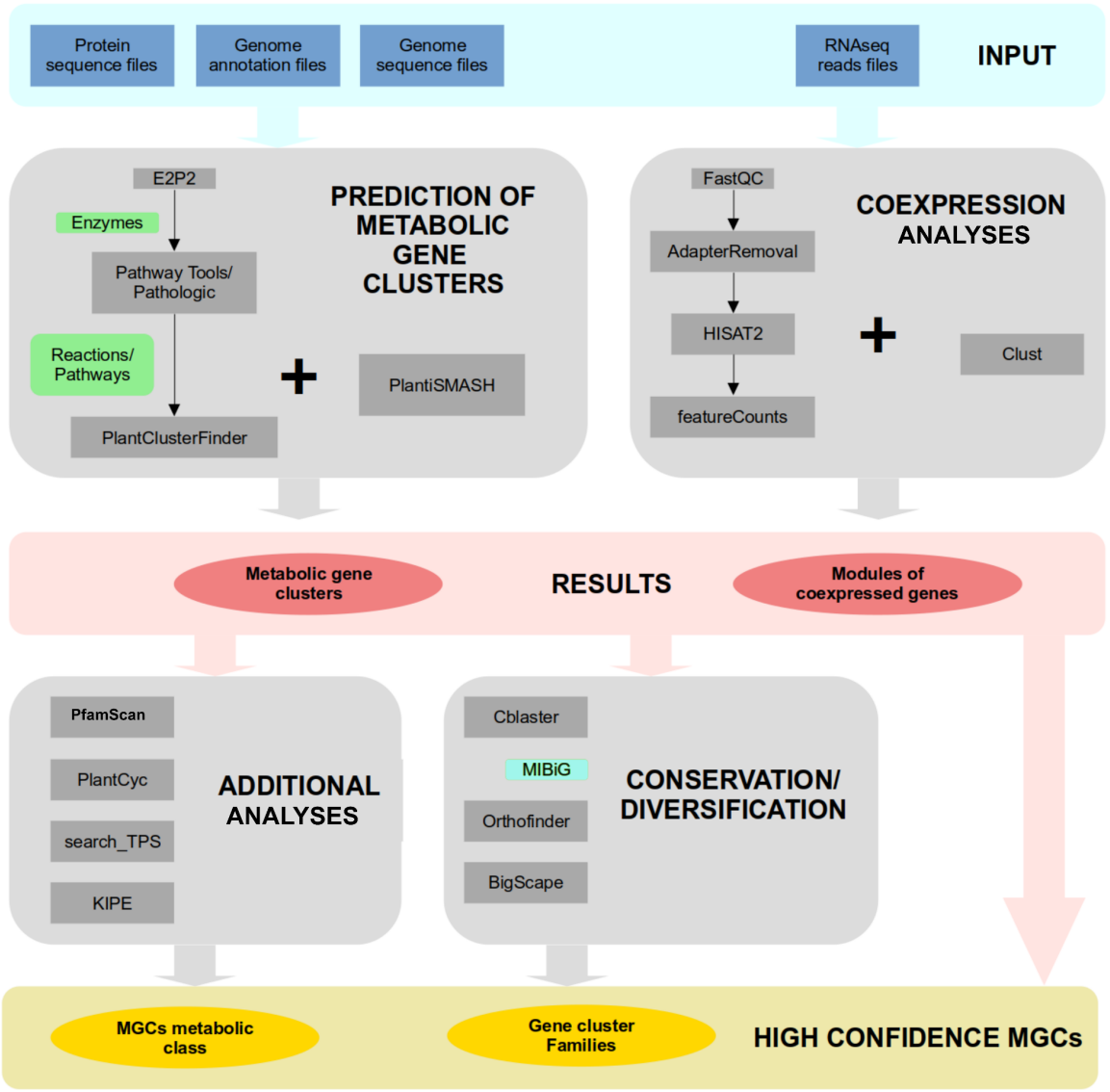

Figure 2 - Overview of the pipeline to predict metabolic gene clusters.

2.4 RNA-Seq data and Coexpression analysis

To determine if the predicted MGC genes were coexpressed, we constructed coexpression networks using transcriptome data for six Rubiaceae species: *Coffea arabica, Coffea canephora, Coffea eugenioides, Gardenia jasminoides, Neolamarckia cadamba* and *Ophiorrhiza pumila*. For this, we used 65 libraries from 4 RNA-Seq experiments available in the European Nucleotide Archives (ENA) and the National Genomics Data Center (NGDC). In Table 2, we detail conditions of these experiments:

Table 2 Information of the RNA-Seq experiments used in this study

Project ID	Source	Experiment description	Author
PRJEB32533	ENA	Transcriptome of seeds in three developmental stages from *Coffea arabica, Coffea canephora* and *Coffea eugenioides*	Stavrinides et al., 2020
PRJNA352919	ENA	Transcriptome of *Ophiorrhiza pumila* hairy roots	Udomsom et al., 2016
PRJCA003540	NGDC	Transcriptome of *Neolamarckia cadamba* roots under aluminum stress	Dai et al., 2020
PRJNA688705	ENA	Transcriptome of *Gardenia jasminoides* fruits in two developmental stages.	Pan et al., 2021

For *Coffea arabica*, *Coffea canephora*, and *Coffea eugenioides*, we performed coexpression analysis using RNA-Seq of seeds at three developmental stages (Stavrinides et al., 2020). The study generated 27 RNA-Seq libraries (3 species, 3 replicates, 3 seed stages). Seed stages corresponded to the following phenological phases: ST5 (seeds from green fruits, peak of reserve deposition and start of endosperm hardening), ST6 (seeds during fruit veraison), and ST7 (seeds from mature cherry fruits with red pericarp). The RNA-Seq experiments in the *Gardenia jasminoides* dataset (Pan et al., 2021) used peel and sarcocarp samples from both green and red fruits, collected in triplicates. The study resulted in 12 RNA-Seq libraries, which were grouped into four tissues for coexpression analysis: GFS (Sarcocarps of green fruits), GFP (Peels of green fruits), FS (Sarcocarps of red fruits) and FP (Peels of red fruits). In the *Ophiorrhiza pumila* dataset (Udomsom et al., 2016), the expression of two ERF transcription factors was suppressed in hairy roots through RNA interference. The study generated 6 RNA-Seq libraries (3 conditions, 2 replicates each), which were divided into the following codes for coexpression analysis: Gusi (Hairy roots transformed with GUS); ERF1i (Hairy roots with suppressed OpERF1) ; ERF2i (Hairy roots with suppressed OpERF2). In the *Neolamarckia cadambia* RNA-Seq experiment (Dai et al., 2020), roots were treated with 400 μM +Al for 1, 3, and 7 days, while controls were grown without Al3+. The experiment made in 20 RNA-Seq libraries, each defined by one of four time sets and two conditions: untreated roots at 0, 1, 3, and 7 days old (AL0, AL1, AL3, AL7) and treated roots at 1, 3, and 7 days old (AL1t, AL3t, AL7t). We analyzed the RNA-Seq raw data using FastQC v0.11.8 tool (Wingett & Andrews, 2018) and removed low quality reads and adapters with the AdapterRemoval v2.3.0 software (Schubert et al. 2016). Then, we mapped the data against the respective genome using HISAT2 v2.2.0 (Kim et al. 2019) with default parameters. Finally, we used the featureCounts v2.0.0 tool (Liao et al. 2014) to count and normalize the transcripts.

We carried out a coexpression analysis with the Clust v1.12.0 tool (Abu-Jamous & Kelly, 2018) using the raw count data from the RNA-Seq experiments (as listed in Table 2). The

parameters used were k-means clustering method, tightness weight of 1.0, and Q3s outliers threshold of 2.0. For a cluster to be considered among those with coexpressed genes, at least three biosynthetic genes should be in the same coexpression module.

3. Results and Discussion

3.1 Genome-wide prediction of metabolic gene clusters in the Rubiaceae family

The surge in large-scale transcriptomic and genomic datasets has opened new dimensions in plant comparative genomics. This work demonstrates the application of omics techniques and bioinformatics tools for discovering metabolic gene clusters. Our analysis of MGCs across eight Rubiaceae species plus *Solanum lycopersicum* allowed us to predict a total of 2,372 metabolic gene clusters using two pipelines. In figure 3, we show the distribution of these genes among clusters and species.

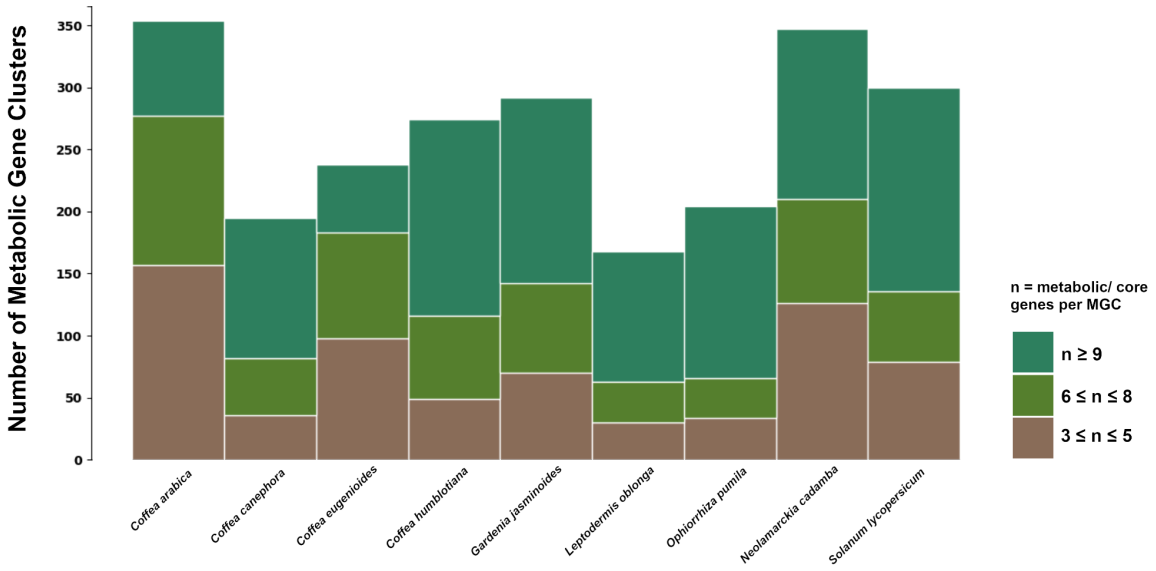

Figure 3 - Number of all predicted metabolic gene clusters of different sizes (number of clustered metabolic genes) across 8 plants from the Rubiaceae family and one Solanaceae species.

Using the PlantClusterFinder pipeline, we identified a total of 1931 metabolic gene clusters containing 31,392 genes, with detailed results in Table 3 and supplementary Table S1. We identified an average of 214 MGCs per species, with the lowest number occurring in *L. oblonga* (118 MGCs) and the highest number occurring in *N. cadamba* (295 MGCs). The predicted MGCs ranged from 5 to 2551 kb with an average size of 178 kb. The average number of genes per MGC was 17. A total of 22,556 genes had an attributed E.C. number and 22,713 had at least one attributed reaction number. The authors from the genome assembly of *O. pumila* study (Rai et al., 2021) also predicted MGCs using PlantClusterFinder and 1069 genes identified in MGCs by their study were identified in our analysis.

Table 3 Overview of results from the PlantClusterFinder pipeline

Species	MGCs	Average nº of genes
Solanum lycopersicum	255	17
Neolamarckia cadamba	295	14
Ophiorrhiza pumila	162	23
Leptodermis oblonga	118	26
Gardenia jasminoides	238	17
Coffea humblotiana	223	19
Coffea canephora	149	21
Coffea eugenioides	200	8
Coffea arabica	291	8

Using the PlantiSMASH pipeline we predicted 441 MGCs, which contained 5,776 genes (Table 4; supplementary Table S2). On average, 49 MGCs were predicted per species, with the lowest number occurring in *C. eugenioides* (38 MGCs) and the highest occurring in *C. arabica* (63 MGCs). The predicted MGCs ranged from 18 to 960 kb, with an average size of 175 kb. The average number of genes per MGC was 13. The authors from the genome assembly of *N. cadamba* study (Zhao et al., 2021) also predicted MGCs using PlantiSMASH and 622 genes identified in MGCs by their study were identified in our analysis.

Table 4 Overview of results from the PlantiSMASH pipeline

Species	Number of Predicted MGC's	Average nº of genes
S. lycopersicum	45	11
N. cadamba	52	10
O. pumila	42	16
L. oblonga	50	15
G. jasminoides	54	12
C. humblotiana	51	13
C. canephora	46	11
C. arabica	63	12
C. eugenioides	38	12

These pipelines are based on different methodologies and algorithms, leading to substantial discrepancies in the MGCs they predicted. PlantClusterFinder identified 41.7%

of the MGCs that were predicted by PlantiSMASH, but only 7.7% of MGCs detected by PlantClusterFinder were detected again by PlantiSMASH. These differences were reported before (Schläpfer et al., 2017; Chen et al., 2019) and underscore the importance of considering multiple methods in MGC discovery and the inherent complexities in these types of analyses.

We determined whether clusters found in Rubiaceae species were homologous to MGCs from the curated database within the "Minimum Information about Biosynthetic Gene clusters" (MiBIG) repository (Terlouws et al., 2023). Using this approach, we successfully recovered all MGCs from *S. lycopersicum,* indicating the effectiveness of our methodological approach. Although this repository contains 43 verified MGCs for the Viridiplantae group, none belong to the Rubiaceae family. To identify any plant MGCs in the MiBIG repository that could share similarities with MGCs in Rubiaceae species, we employed the cblaster tool (version 1.3.16; Gilchrist et al., 2021). Out of the 43 plant MGCs in MiBIG, 23 had partial matches to MGCs from Rubiaceae species (Supplementary File 1), with low similarity (identity below 40%). Among the partially identified cases, 13 involve conserved TPS-CYP gene pairs, which are a common structure in clusters related to terpenoid metabolism (Boutanaev et al., 2014; Bharadwaj et al., 2021; Smit & Lichman, 2022). The fact that no MGCs from other plant species were conserved in the Rubiaceae underscores the unique biosynthetic diversity identified in the genomes of this family.

3.2 Classification of Rubiaceae MGCs

In our study, we used a strategy based on a system proposed by Chae et al., (2014) to classify the MGCs that we identified using the PCF tool (refer to Figure 4A for more details). Each predicted enzyme is then designated a 'signature' or 'tailoring' classification. Out of the total MGCs we initially identified, we found that 175 of them (9%) included both 'signature' and 'tailoring' enzymes.

The PlantiSMASH pipeline attributes a biochemical class to each predicted MGC in saccharides, terpenes, alkaloids, lignans, polyketides, putative or mixed. This classification follows a criteria based on the number of core and accessory genes identified with specific pHMMs within a MGC (Kautsar et al., 2017). In other MGC predictions with PlantiSMASH (*N. cadamba* - Zhao et al., 2021; Tobacco - Rabara et al., 2023), the most identified classes were also saccharides and terpenes. In our analysis, the most frequently occurring biochemical class was saccharides (Figure 4), with a total of 152 clusters identified, averaging 16.8 per species. The class with the lowest number of clusters was polyketides, with a total of 20 and an average of 2.2 per species. We identified 39 clusters as hybrids and the metabolic class was undetermined for 103 clusters (Figure 4B).

Given that the selected methods employ different techniques for predicting Metabolic Gene Clusters (MGCs), we utilized an integrative analysis to consolidate the results. The PlantiSMASH tool predicts MGCs using profile Hidden Markov Models (pHMMs), so to harmonize this with the PlantClusterFinder predictions, we performed a search for protein domains in each gene within an MGC. This was executed using PfamScan (please refer to supplementary Table S3 for more details).

The protein domain family for cytochrome P450 (PF00067) was the most frequently detected, followed by the UDP-glucuronosyl and UDP-glucosyl transferase (PF00201), as well as the 2OG-Fe(II) oxygenase superfamily (PF03171).

In addition to the protein domain search, we performed a search for Metacyc plant pathways for each gene within a Metabolic Gene Cluster (MGC) that was predicted by PlantiSMASH. We chose to do this because the PlantClusterFinder (PCF) pipeline employs this methodology (see supplementary Table S3 for additional information). The most frequently identified pathway was the Secologanin and Strictosidine biosynthesis pathway (PWY-5290). Following closely were the Sesaminol Glucoside/lignan biosynthesis pathway (PWY-7139), the Quercetin Glucoside/flavonoid biosynthesis (PWY-7129), and the flavonoid biosynthesis pathway (PWY1F-FLAVSYN).

To predict and identify terpene synthases and enzymes involved in flavonoid biosynthesis within Metabolic Gene Clusters (MGCs), we employed two specialized tools: search_TPS (version 1.0; Domingues et al., 2022) and KIPEs (version 0.35; Pucker et al., 2020). After conducting a thorough analysis, we identified several MGCs with distinct types of synthases: 50 MGCs contained monoterpene synthases, 30 displayed diterpene synthases, and 75 had sesquiterpene synthases. Furthermore, we found 199 MGCs that included genes related to flavonoid metabolism. Comprehensive details of these findings are provided in Supplementary Table S3.

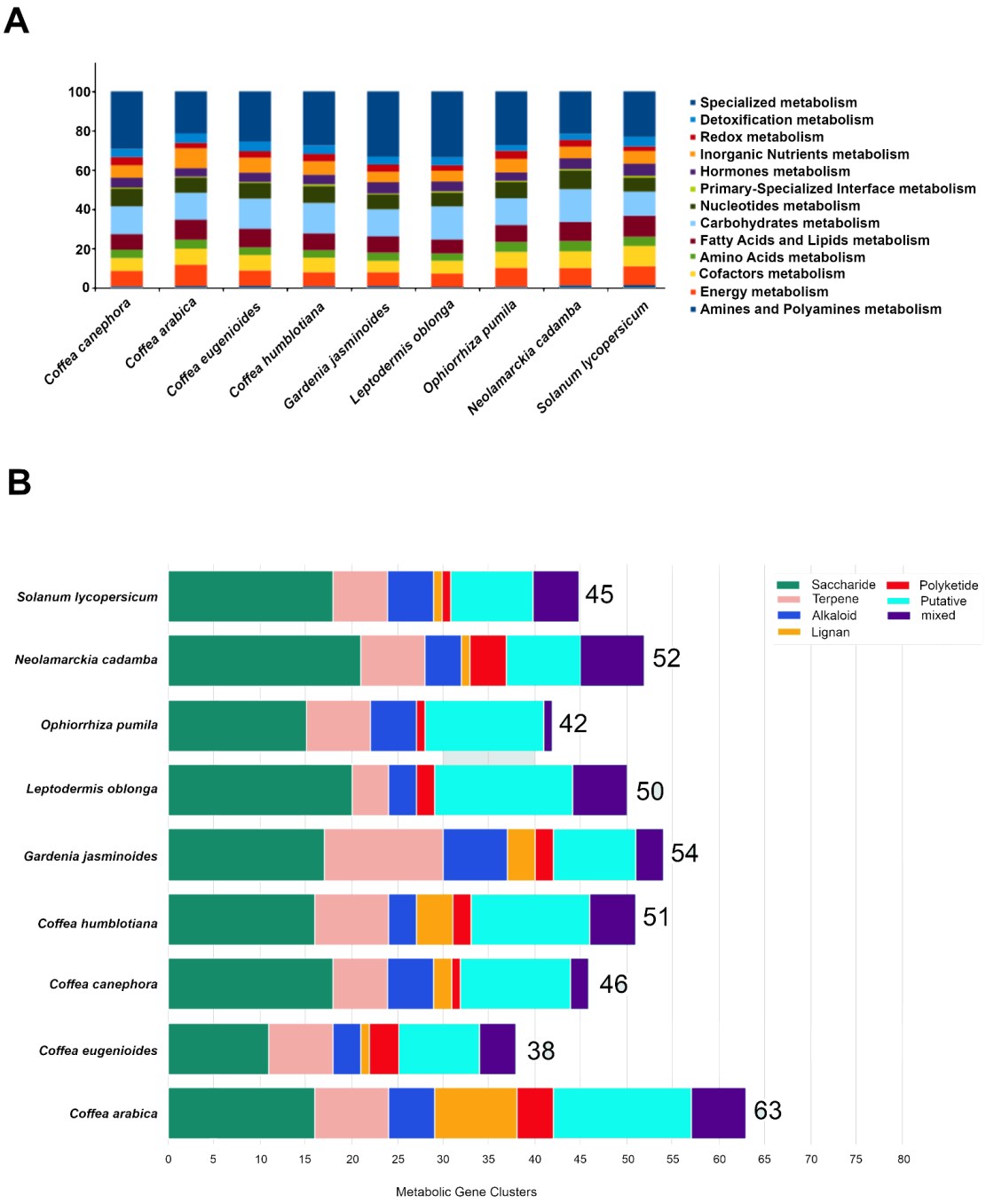

Figure 4 - Percentual distribution of metabolic domains in the MGCs predicted by PCF (A). Overview of MGCs predicted with the PlantiSMASH pipeline and classified into biochemical classes (B).

The MGC predictions unveiled diverse and often complex structures. In figure 5, we show examples of MGCs. In terms of their functional classification, both pipelines detected a high number of saccharide and terpene MGCs, with saccharides being the most prevalent biochemical class identified.

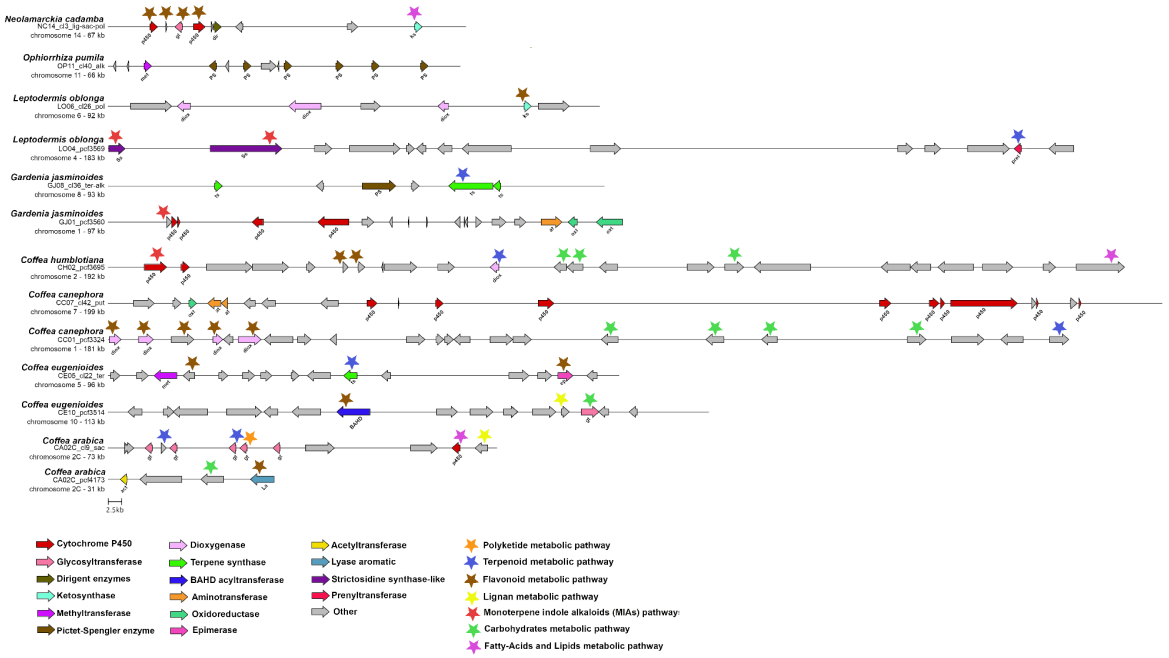

Figure 5 - Example candidate MGCs identified in this study. Two example candidate MGCs per species are shown, MGCs identified with PlantiSMASH and PlantClusterFinder. The examples cover a diverse range of enzymatic classes and predicted metabolic pathways.

3.3 Conservation and diversification of metabolic gene clusters in Rubiaceae

We performed a comparative genomic analysis with all plants of the study to assess conservation and diversification of metabolic gene clusters in Rubiaceae. With an orthology analysis we detected a total of 30,170 orthogroups (supplementary Table S12). A total of 10,925 orthogroups containing all species and 8,152 species-specific orthogroups were identified (Figure 6A). All nine species had species-specific orthogroups.

To investigate gene content changes, we examined the rates and direction of changes in orthogroup size among each of the species. Across the Rubiaceae phylogeny, most species have higher numbers of orthogroup contractions than expansions, except for *N. cadamba, C. arabica* and *C. eugenioides* (Figure 6B). Orthogroups in the *C. arabica* genome exhibit the highest number of expansions and contractions followed by *N. cadamba*.

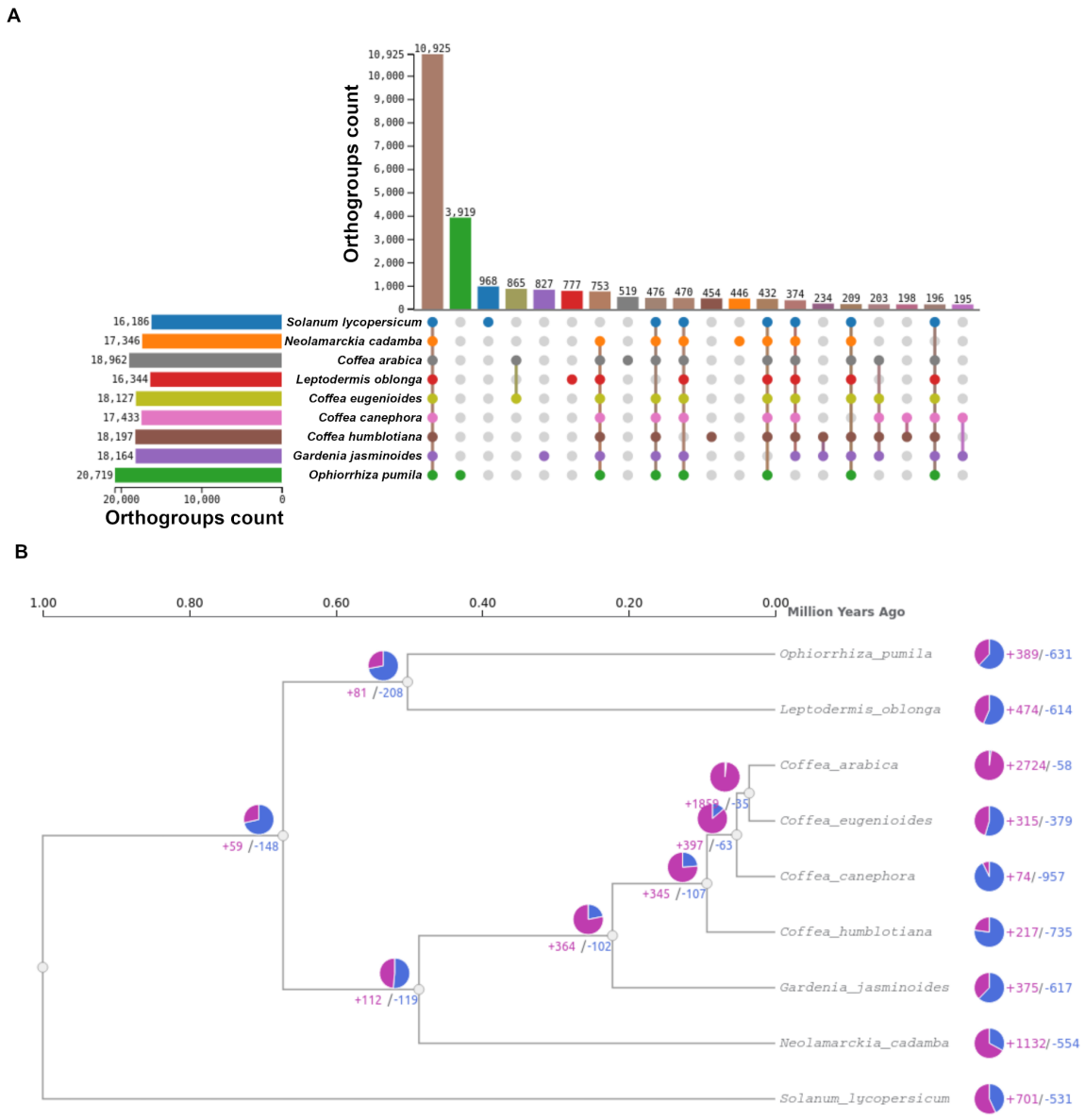

Figure 6 - Gene family contraction and expansion analysis of eight species from Rubiaceae family plus Solanum lycopersicum. (A) The UpSet table displays the count of orthogroups for each species, along with the count of unique orthogroups and the count of shared orthogroups among different species. (B) A pie chart was utilized to visualize gene families with altered gene numbers, representing the expanded gene families (depicted in purple) and contracted gene families (depicted in blue).

Synteny analysis among the nine species identified the biggest collinearity between *C. arabica* and *C. eugenioides* with 50932 (72.08%) collinear gene pairs. The smallest collinearity was identified between *C. canephora* and *N. cadamba* with 3075 (3.23%) collinear gene pairs.

All genes predicted in MGCs were distributed in 3121 orthogroups (supplementary Table S4).

In order to track the conservation within Rubiaceae MGCs, we constructed a similarity network of MGCs and identified a total of 549 gene cluster families (GCFs) (Figure 7; supplementary Table S5). The average number of MGCs per family was 4 and the

maximum number of MGCs in a family was 16. Figure 7 summarizes families that were found in at least four species, with the most conserved MGCs.

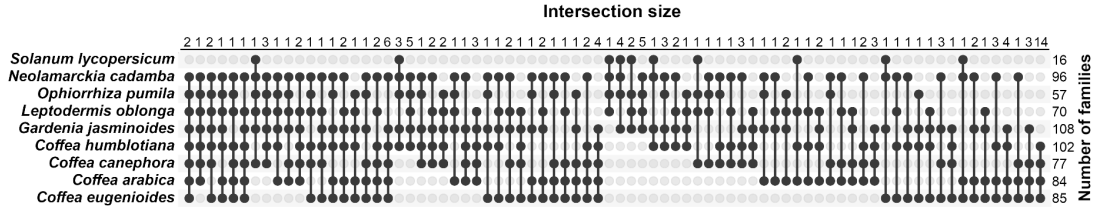

Figure 7 - Gene cluster families presence in four or more species across Rubiaceae and tomato.

The results of the orthology analysis were used to validate the gene cluster families prediction - since it would be expected that genes in a given GCF would be in the same orthogroups. Of the total 549 predicted GCFs, 179 were formed by a single MGC per species (or 2 in the case of the tetraploid Arabica coffee).

It is well known that MGCs of the same family can evolve to produce different molecules, through neofunctionalization of genes or by gene recruitment or loss (Polturak et al., 2022). Thus in most of the cases of Rubiaceae MGCs, the homologous genes are distributed in different genomic positions or forming different MGCs across the plants. We observed an example of this by comparing our results with the genome assembly of *Gardenia jasminoides* study (Xu et al., 2020). Xu et al. describe syntenic regions between *G. jasminoides* and *Coffea canephora* containing crocin biosynthetic genes, in which specific duplications of the genome can explain the synthesis of this compound in Gardenia and the absence in Coffea. One of these syntenic regions contains a tandem array of genes encoding the UDP-glycosyltransferase family (UGT) (PF00201.21) in *Gardenia*, on chromosome 9, which has an orthologous relationship with UGTs identified on chromosome 2 of *C. canephora* (Xu et al., 2020). We identified a MGC in the chromosome 2 of *C. canephora*, CMG CC02_cl15_sac, predicted to be involved in saccharide biosynthesis, which corresponds to this region analyzed by Xu et al. (2020). *C. canephora* MGC was identified in our analysis because, in addition to UGTs, this region of the genome has a gene from the cytochrome P450 family (PFAM domain PF00067.25) that is not present in the homologous region of chromosome 9 in *G. jasminoides*. This data suggests a genomic diversification that led to a metabolic diversification between these two species in this homologous region of the genome.

We highlight two cases of MGCs that conserve core biosynthetic genes across more than five species, with shared orthogroups (Figure 8, Figure 9). As we used *Solanum lycopersicum* as an outgroup for comparative genomics analysis, we observed an example of a conserved-like MGC involved in the metabolism of saccharides. In the gene cluster family FAM-1539 we observed that both the core and accessory genes of MGCs from six Rubiaceae plants demonstrate a degree of conservation with *S. lycopersicum*, indicating their potential importance in the production process of a compound predicted as saccharide. The example of the family FAM-1539, which has retained MGCs in seven species: *N. cadamba, O. pumila, L. oblonga, C. canephora, C. arabica, C. eugenioides*, and *S. lycopersicum*. Each species carries one MGC from this GCF, with the exception of *C. arabica* which has two. These MGCs are linked to saccharide metabolism.

MGCs from FAM-1539 comprises three core biosynthetic genes (a Glycosyltransferase, a Squalene epoxidase, and an Aminotransferase) and seven accessory genes, as displayed

in Figure 8. Notably, we observed that both the core and accessory genes demonstrate a degree of conservation, indicating their potential importance in the compound production process for this specific set of MGCs. In the case of tomato (*S. lycopersicum*), the glycosyltransferase gene is a cellulose synthase (Solyc04g077470, domain PF13632.9), an enzyme usually involved in the synthesis of matrix polysaccharides such as xyloglucan. The tomato aminotransferase is an ACC synthase paralog (Solyc04g077410, domain PF00155.24), a key enzyme implicated in the synthesis of ethylene.

We analyzed the expression of orthologs of two core genes of the tomato MGC in this gene cluster family (cellulose synthase and ACC synthase) and observed that they are not the most expressed of their respective gene families. We also observed that such orthologs are not coexpressed, however, this conservation suggests that the synthetic processes of polysaccharides and hormones are physically linked in the genome of several species of Rubiaceae. Future functional studies interrupting or overexpressing these genes would help in the final understanding of the function of this MGC.

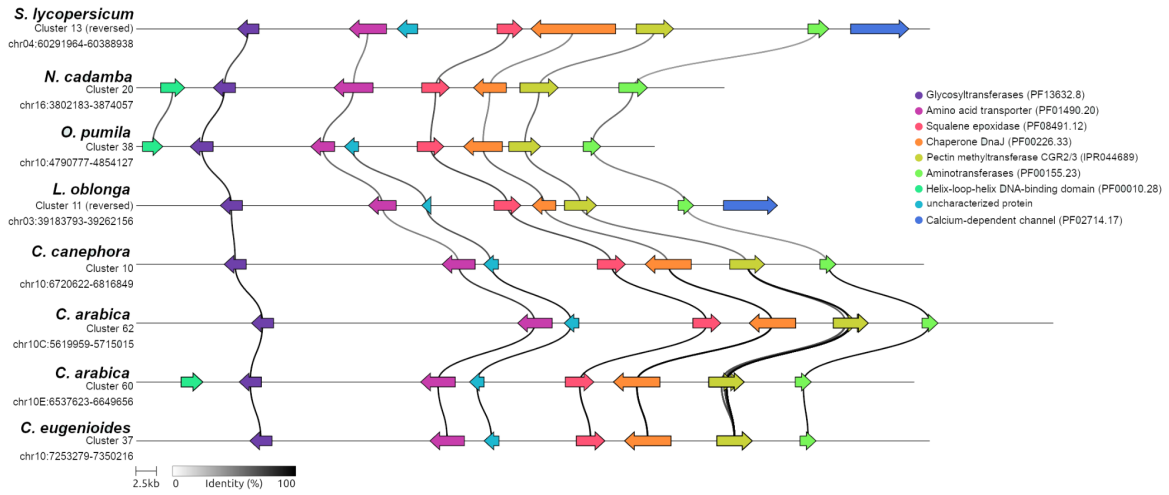

Figure 8 - Gene cluster family FAM-1539 with conserved MGCs in *S. lycopersicum, N. cadamba, O. pumila, L. oblonga, C. canephora, C. arabica* and *C. eugenioides*. Each arrow represents a gene. The width of the links connecting genes represents the percentity of identity.

The second highlighted example involves the partial preservation of a tomato terpenoid MGC in Rubiaceae plants (FAM-1569). This cluster of genes in tomato (*Solanum lycopersicum*) has been found to be involved in the synthesis of mono, sesqui and diterpenes. The tomato MGC contain five complete terpene synthase genes (TPS18, TPS19, TPS20, TPS21, and TPS41), two complete cis-prenyl transferases (CPTs), a cytochrome P450s, an aldehyde oxidase, and three alcohol acyl transferase genes (Matsuba et al., 2013). This cluster evolved via gene duplication, divergence, alterations in substrate specificity, and acquisition of cis-prenyl transferase genes in wild tomato species, such as *Solanum habrochaites*, *S. pennellii*, and *S. pimpinellifolium*. FAM-1569 (Figure 9) includes, besides tomato, MGCs from *N. cadamba, C. humblotiana, C. canephora, C. arabica, and C. eugenioides*.

The clustered tomato diterpene synthase genes TPS18 and TPS21, and the monoterpene synthases TPS19 and TPS20, all classified as e/f (Zhou & Pichersky, 2020) forms an orthogroup with *N. cadamba, O. pumila, G. jasminoides, C. humblotiana, C. canephora, C. arabica*, and *C. eugenioides* TPSs present in this cluster. The class c diterpene synthase gene TPS41 also shares orthogroups. Although the tomato cytochrome P450 have

orthologs only in tomato, the MGCs from family FAM-1569 do contain cytochrome P450 genes in orthogroups that were exclusive to plants from the Rubiaceae family here analyzed. For FAM-1569, we identified both coexpression modules containing TPS and P450 genes from the same MGC in *C. canephora, C. arabica* and *N. cadamba*.

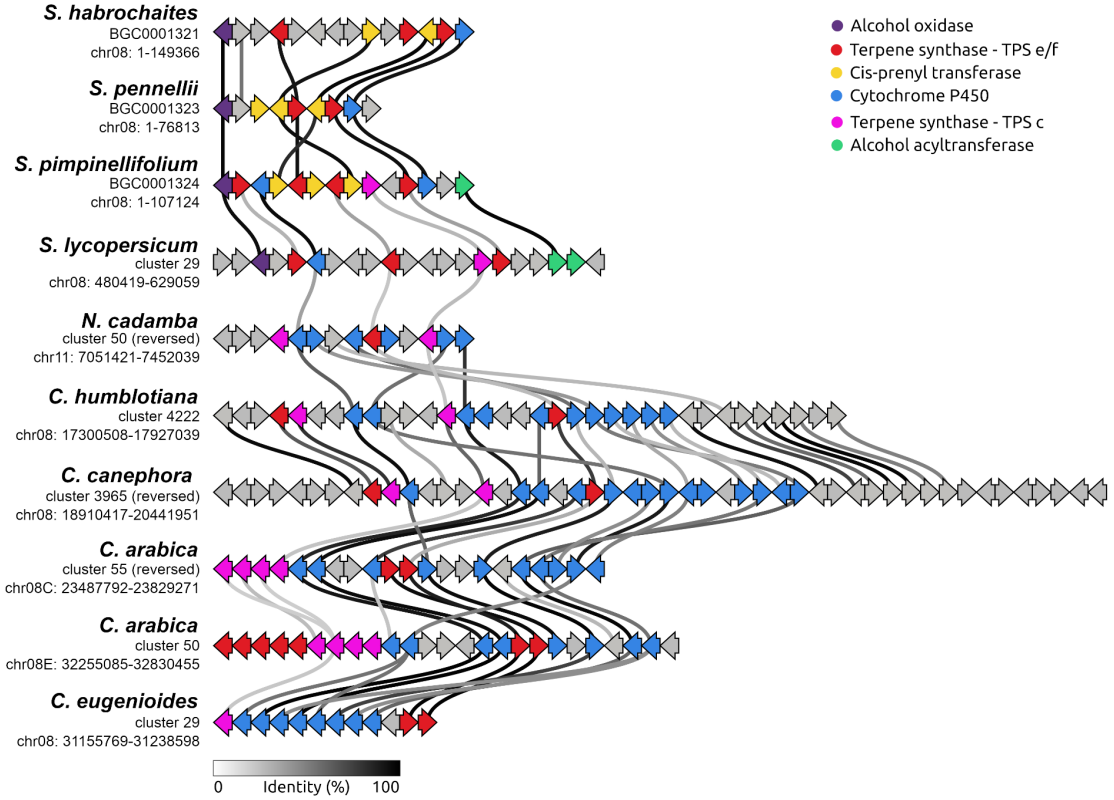

Figure 9 - Gene cluster family FAM-1569 with the tomato lycosantalonol MGC conserved in wild species of tomato. This MGC is partially conserved in five Rubiaceae species. This representation was plotted disregarding the scale factor due to the large difference between the sizes of the represented MGCs.

In tomato, the TPS18 gene synthesizes an unknown diterpene, TPS19 and TPS20 genes synthesize monoterpenes and the TPS21 gene synthesizes lycosantonolol. Those class e/f TPSs genes forms an orthogroup with *N. cadamba, O. pumila, G. jasminoides, C. humblotiana, C. canephora, C. arabica,* and *C. eugenioides* TPSs. The tomato class c diterpene synthase TPS41 gene also shares orthogroups with *N. cadamba, O. pumila, C. humblotiana, C. canephora, C. arabica,* and *C. eugenioides* TPSs. The tomato cytochrome P450 gene CYP71BN1 does not share orthogroups with plants of our study. Nevertheless, we observed that the cytochrome P450 genes from FAM-1569 MGCs share orthogroups with all plants in our study. Additionally, the CYP450 genes were identified as members of CYP71 and CYP76 clans, known to be involved in specialized diterpene metabolism (Bathe & Tissier, 2019). The tomato cis-prenyl transferase gene (CPT1) has orthologs in *N. cadamba* and *O. pumila,* but such orthologs are distributed in other regions of their respective genomes. Future functional studies would help in the final understanding of the function of these genes.

The Rubiaceae genomes sampled in this study comprises three subfamilies strongly supported by previous phylogenies: Rubioideae (*L. oblonga* and *O. pumila*), Ixoroideae (*Coffea* spp. and *G. jasminoides*) and Cinchonoideae (*N. cadamba*) (Bremer & Eriksson, 2009). Taking account that members of the same subfamily should have more shared GCFs, plants of the Ixoroideae subfamily (*Coffea* spp. and *G. jasminoides*) had the major number of shared GCFs, which represent the most conserved metabolic gene clusters. Our results also suggest a major conservation among MGCs from Ixoroideae and Cinchonoideae subfamilies (here represented by *N. cadamba*), than Ixoroideae and Rubioideae subfamilies (here represented by *L. oblonga* and *O. pumila*). A total of 80 GCFs had representatives in all three subfamilies. When comparing with *S. lycopersicum*, we observed a higher number of conserved GCFs between *S. lycopersicum* and the Rubioideae subfamily, followed by *S. lycopersicum* and the Ixoroideae subfamily and finally, *S. lycopersicum* and the Cinchonoideae subfamily.

3.4 Cross-Species analysis of metabolic gene clusters unveils coexpression modules that contribute to set high confidence MGCs

A total of 1453 genes were found using both MGC discovery approaches. They were distributed in 217 clusters identified with PlantiSMASH and 211 clusters identified with PCF (supplementary Table S4). Coexpression analysis has been used to identify candidate genes associated with metabolic pathways. Genes that participate in the same metabolic pathway often display coordinated expression patterns when the environment changes (Singh et al., 2022; Zhao & Rhee, 2022). Thus, we conducted coexpression analysis using publicly available RNA-Seq experiments for *Coffea arabica, Coffea canephora, Coffea eugenioides, Gardenia jasminoides, Ophiorrhiza pumila,* and *Neolamarckia cadamba*.

Our analysis resulted in 19 coexpression modules across the five species (Figure s1-s6, Supplementary File 1). We examined whether genes within MGCs shared coexpression modules. Consequently, we considered MGCs with core genes in the same coexpression module as high-confidence MGCs.

In total, we identified 207 MGCs where at least three core metabolic genes were located in the same coexpression module. Of this total, 204 MGCs were also part of a gene cluster family, indicating conservation among other species in the study.

The coexpression analysis for the *C. arabica* dataset yielded two coexpression modules, with a total of 11 MGCs showing coexpression and conservation in other species. These MGCs were considered high-confidence (supplementary Table S6). For the *C. canephora* dataset, the coexpression analysis resulted in four coexpression modules, with 74 MGCs considered high-confidence (supplementary Table S7). The C. eugenioides dataset revealed four coexpression modules, with 9 MGCs considered high-confidence (supplementary Table S8). In the *G. jasminoides* dataset, the coexpression analysis identified three coexpression modules, and 11 MGCs were deemed high-confidence (supplementary Table S9). The *O. pumila* dataset yielded three modules in the coexpression analysis, with 17 MGCs considered high-confidence (supplementary Table S10). Lastly, the coexpression analysis for the *N. cadamba* dataset resulted in three

coexpression modules, and a total of 82 MGCs were considered high-confidence (supplementary Table S11).

In addition to this, our cross-species analysis demonstrated the power of using coexpression modules for MGC identification. Genes within a metabolic pathway are often coexpressed, and finding these coexpression modules can provide strong evidence for the functional relevance of the predicted MGCs. In our study, we identified 207 high-confidence MGCs where at least three core metabolic genes were located in the same coexpression module. This approach not only enhances the confidence in MGC predictions but also provides a functional context to understand how these genes may work together in metabolic processes.

In conclusion, our analysis has successfully elucidated the complex landscape of MGCs across multiple plant species, paving the way for more targeted and in-depth studies in the future. The identification of coexpression modules also highlights the relevance of such cross-species comparative methods in unraveling potential functional associations and underlying genetic influences in metabolic pathways. Our findings underscore the potential in harnessing this knowledge to enhance plant breeding programs and develop strategies for improved plant metabolic engineering.

CRediT authorship contribution statement

Samara M. Correia de Lemos: Conceptualization, Methodology, Formal analysis, Data curation, Investigation, Writing – original draft, Writing – review & editing. **Alexandre R. Paschoal**: Conceptualization, Methodology, Investigation, Resources, Writing – review & editing, Supervision. **Romain Guyot**: Methodology, Formal analysis, Investigation, Writing – review & editing, Supervision. **Marnix Medema**: Methodology, Formal analysis, Investigation, Writing – review & editing, Supervision. **Douglas S. Domingues**: Conceptualization, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by resources supplied by the Center for Scientific Computing (NCC/GridUNESP) of the São Paulo State University (UNESP). The authors acknowledge the Wageningen University Bioinformatics Group at Wageningen University for providing computing resources that have contributed to the research results reported within this paper. The authors acknowledge the IRD itrop HPC (South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper.

Funding

This study was financed by CAPES-PrInt Program (process 88887.570702/2020-00). SMCL was financed in part by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES)—Finance Code 001. NAPI Bioinformática from Fundação Araucária (# 66.2021). DSD research in coffee genomics and transcriptomics is supported

by São Paulo State Research Foundation (FAPESP, grants #2016/10896-0 and #2018/08042-8) and by the Program for Support of New Faculty 2022/2023, at University of São Paulo. These funding agencies had no role in study design, the collection, analysis, and interpretation of data, or manuscript writing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.5281/zenodo.8221450

3 Conclusão

Nesta tese utilizamos ferramentas de bioinformática para investigar sobre a organização genômica e a conservação genética de oito plantas da família Rubiaceae e utilizando *Solanum lycopersicum* como grupo externo.. Realizamos o primeiro estudo de genômica comparativa desta família sob o viés de elementos específicos do genoma: clusters biossintéticos de genes.

Utilizamos as duas principais ferramentas de bioinformática para identificar clusters biossintéticos de genes e percebemos que as duas metodologias adotam conceitos diferentes sobre clusters biossintéticos de genes, portanto o número de CMGs preditos variou de acordo com a ferramenta utilizada.

Identificamos CMGs conservados dentro da família Rubiaceae, mas entendemos que a maioria dos CMGs preditos são pouco conservados, o que indica que podem estar envolvidos no metabolismo especializado de cada espécie.

O uso de redes de coexpressão é amplamente utilizado na predição de CMGs. Identificamos um baixo número de experimentos de RNA-seq das espécies selecionadas, e isso foi um desafio à construção de redes de coexpressão para agregar status de alta confiança aos CMGs preditos.

Os CMGs de alta confiança identificados na presente tese são importante ponto de partida para identificação e caracterização funcional de vias bioquímicas de compostos de interesse, podendo ser ponto de partida para análises de genômica funcional e metabolômica, permitindo assim a descoberta da base molecular de síntese de metabólitos vegetais.

Referências

Abu-Jamous B, Kelly S. Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. Genome Biology (2018) 19:172. DOI: https://doi.org/10.1186/s13059-018-1536-8.

Adewole KE, Attah AF, Adebayo JO, 2021. Morinda lucida Benth (Rubiaceae): A review of its ethnomedicine, phytochemistry and pharmacology. J Ethnopharmacol. 276: 114055. https://doi.org/10.1016/j.jep.2021.114055.

Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, Ikeda S, Takahashi H, Altaf-Ul-Amin M, Darusman LK, Saito K, Kanaya S. KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. Plant Cell Physiol. 2012 Feb;53(2):e1. doi: 10.1093/pcp/pcr165. Epub 2011 Nov 28. PMID: 22123792.

Aprotosoaie AC, Luca SV, Miron A. Flavor Chemistry of Cocoa and Cocoa Products—An Overview. Comprehensive Reviews in Food Science and Food Safety. (2016) Vol. 15. DOI: 10.1111/1541-4337.12180.

Arabidopsis Genome Initiative (2000) Analysis of the Genome Sequence of the Flowering Plant Arabidopsis thaliana. Nature, 408, 796-815.

Bathe U, Tissier A, 2019. Cytochrome P450 enzymes: A driving force of plant diterpene diversity. Phytochemistry. 161: 149-162. https://doi.org/10.1016/j.phytochem.2018.12.003.

Bharadwaj R, Kumar SR, Sharma A, Sathishkumar R, 2021. Plant Metabolic Gene Clusters: Evolution, Organization, and Their Applications in Synthetic Biology. Front Plant Sci. 12: 697318. https://doi.org/10.3389/fpls.2021.697318.

Boutanaev AM, Moses T, Zi J, Nelson DR, Mugford ST, Peters RJ, Osbourn A, 2015. Investigation of terpene diversification across multiple sequenced plant genomes. Proc Natl Acad Sci U S A. 112(1): E81-8. https://doi.org/10.1073/pnas.1419547112.

Bremer B, Eriksson T, 2009. Time tree of Rubiaceae: phylogeny and dating the family, subfamilies and tribes. Int J Plant Sci. 170, 766–793. http://dx.doi.org/10.1086/599077.

Brose J, Lau KH, Dang TTT, Hamilton JP, Martins LDV, Hamberger B, Hamberger B, Jiang J, O'Connor SE, Buell CR. The Mitragyna speciosa (Kratom) Genome: a resource for data-mining potent pharmaceuticals that impact human health. G3 (Bethesda). 2021 Apr 15;11(4):jkab058. doi: 10.1093/g3journal/jkab058. PMID: 33677570; PMCID: PMC8759815.

Burge D, 2020. Conservation genomics and pollination biology of an endangered, edaphic-endemic, octoploid herb: El Dorado bedstraw (Galium californicum subsp. sierrae; Rubiaceae). PeerJ. 8: e10042. https://doi.org/10.7717/peerj.10042.

Canales NA, Pérez-Escobar OA, Powell RF, Töpel M, Kidner C, Nesbitt M, Maldonado C, Barnes CJ, Rønsted N, Przelomska NAS, Leitch IJ, Antonelli A, 2022. A highly contiguous, scaffold-level nuclear genome assembly for the fever tree (Cinchona pubescens Vahl) as a novel resource for Rubiaceae research. GigaByte. 2022: gigabyte71. https://doi.org/10.46471/gigabyte.71.

Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Midford PE, Ong Q, Ong WK, Paley S, Subhraveti P, Karp PD. The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res. 2018 Jan 4;46(D1):D633-D639. doi: 10.1093/nar/gkx935. PMID: 29059334; PMCID: PMC5753197.

Chae L, Kim T, Nilo-Poyanco R, Rhee SY. Genomic Signatures of Specialized Metabolism in Plants. Science (2014) Vol. 344, Issue 6183, pp. 510-513. DOI: 10.1126/science.1252076.

Chen X, Liu F, Liu L, Qiu J, Fang D, Wang W, Zhang X, Ye C, Timko MP, Zhu Q, Fan L, Xiao B. Characterization and Evolution of Gene Clusters for Terpenoid Phytoalexin Biosynthesis in Tobacco. Planta (2019). DOI: https://doi.org/10.1007/s00425-019-03255-7.

Colle M, Leisner CP, Wai CM, Ou S, Bird KA, Wang J, Wisecaver JH, Yocca AE, Alger EI, Tang H, Xiong Z, Callow P, Ben-Zvi G, Brodt A, Baruch K, Swale T, Shiue L, Song G, Childs KL, Schilmiller A, Vorsa N, Buell CR, VanBuren R, Jiang N, Edger PP. Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry, GigaScience, Volume 8, Issue 3, March 2019, giz012, https://doi.org/10.1093/gigascience/giz012.

Dai B, Chen C, Liu Y, Liu L, Qaseem MF, Wang J, Li H, Wu AM, 2020. Physiological, Biochemical, and Transcriptomic Responses of Neolamarckia cadamba to Aluminum Stress. Int J Mol Sci. 21(24): 9624. https://doi.org/10.3390/ijms21249624.

Darbani, B., Motawia, M., Olsen, C. et al. The biosynthetic gene cluster for the cyanogenic glucoside dhurrin in Sorghum bicolor contains its co-expressed vacuolar MATE transporter. Sci Rep 6, 37079 (2016). https://doi.org/10.1038/srep37079.

Delli-Ponti R, Devendra S, Marek M. Using Gene Expression to Study Specialized Metabolism - A Practical Guide, Frontiers in Plant Science: 11, 2021. DOI: https://www.frontiersin.org/article/10.3389/fpls.2020.625035.

Denoeud F, Carretero-Paulet L, Dereeper A. et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. SCIENCE. (2014) VOL. 345, ISSUE 6201: 1181-1184. https://doi.org/10.1126/science.1255274.

Domingues DS, Oliveira LS, Lemos SMC, Barros GCC, Ivamoto-Suzuki ST, 2022. A Bioinformatics Tool for Efficient Retrieval of High-Confidence Terpene Synthases (TPS) and Application to the Identification of TPS in Coffea and Quillaja. Methods Mol Biol. 2469: 43-53. https://doi.org/10.1007/978-1-0716-2185-1_4.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD, 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47(D1): D427-D432. https://doi.org/10.1093/nar/gky995.

Emms DM, Kelly S, 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20(1): 238. https://doi.org/10.1186/s13059-019-1832-y.

Erb M, Kliebenstein DJ. Plant Secondary Metabolites as Defenses, Regulators, and Primary Metabolites: The Blurred Functional Trichotomy. Plant Physiol. 2020 Sep;184(1):39-52. doi: 10.1104/pp.20.00433. Epub 2020 Jul 7. PMID: 32636341; PMCID: PMC7479915.

Fan P, Wang P, Lou Y, Leong BJ, Moore BM, Schenck CA, Combs R, Cao P, Brandizzi F, Shiu S, Last RL, 2020. Evolution of a plant gene cluster in Solanaceae and emergence of metabolic diversity. eLife 9:e56717. https://doi.org/10.7554/eLife.56717.

Farag MA, Otify A, Porzel A, Michel CG, Elsayed A, Wessjohann LA. Comparative metabolite profiling and fingerprinting of genus Passiflora leaves using a multiplex approach of UPLC-MS and NMR analyzed by chemometric tools. Anal Bioanal Chem (2016). DOI: 10.1007/s00216-016-9376-4.

Forman V, Luo D, Geu-Flores F, Lemcke R, Nelson DR, Kampranis SC, Staerk D, Møller BL, Pateraki I, 2022. A gene cluster in Ginkgo biloba encodes unique multifunctional cytochrome P450s that initiate ginkgolide biosynthesis. Nat Commun. 13(1): 5143. https://doi.org/10.1038/s41467-022-32879-9.

Gilchrist CLM, Booth TJ, van Wersch B, van Grieken L, Medema MH, Chooi YH, 2021. cblaster: a remote search tool for rapid identification and visualization of homologous gene clusters. Bioinform Adv. 1(1): vbab016. https://doi.org/10.1093/bioadv/vbab016.

Grotewold E, Chappell J, Kellogg EA (2015). Plant Genes, Genomes and Genetics. Wiley Blackwell. https://doi.org/10.1002/9781118539385.

Guo XM, Wang ZF, Zhang Y, Wang RJ. Chromosomal-level assembly of the Leptodermis oblonga (Rubiaceae) genome and its phylogenetic implications. Genomics. 2021 Sep;113(5):3072-3082. DOI: 10.1016/j.ygeno.2021.07.012. PMID: 34246693.

Hartmann, T. From waste products to ecochemicals: fifty years research of plant secondary metabolism. Phytochemistry 68 (2007) 2831–2846. DOI: 10.1016/j.phytochem.2007.09.017.
Howat S, Park B, Oh S, Jin Y, Lee E, Loake GJ. Paclitaxel: biosynthesis, production and future prospects. New Biotechnology (2014) Volume 31, Number 3. DOI: http://dx.doi.org/10.1016/j.nbt.2014.02.010.

Hawkins C, Ginzburg D, Zhao K, Dwyer W, Xue B, Xu A, Rice S, Cole B, Paley S, Karp P, Rhee SY, 2021. Plant Metabolic Network 15: A resource of genome-wide metabolism databases for 126 plants and algae. J Integr Plant Biol. 63(11): 1888-1905. https://doi.org/10.1111/jipb.13163.

Hosmani PS, Flores-Gonzalez M, van de Geest H, Maumus F, Bakker LV, Schijlen E, van Haarst J, Cordewener J, Sanchez-Perez G, Peters S, Fei Z, Giovannoni JJ, Mueller LA, Saha S, 2019. An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. bioRxiv 767764. https://doi.org/10.1101/767764.

International Coffee Organization (ICO). http://www.ico.org/Market-Report-22-23-e.asp

International Wheat Genome Sequencing Consortium (IWGSC). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science. 2018 Aug 17;361(6403):eaar7191. doi: 10.1126/science.aar7191.

Ivamoto ST, Reis O, Júnior, Domingues DS, dos Santos TB, de Oliveira FF, Pot D, et al. (2017) Transcriptome Analysis of Leaves, Flowers and Fruits Perisperm of Coffea arabica L. Reveals the Differential Expression of Genes Involved in Raffinose Biosynthesis. PLoS ONE 12(1):e0169595. doi:10.1371/journal.pone.0169595.

Jonczyk R, Schmidt H, Osterrieder A, Fiesselmann A, Schullehner K, Haslbeck M, Sicker D, Hofmann D, Yalpani N, Simmons C, Frey M, Gierl A. Elucidation of the final reactions of DIMBOA-glucoside biosynthesis in maize: characterization of Bx6 and Bx7. Plant Physiol. 2008 Mar;146(3):1053-63. doi: 10.1104/pp.107.111237. Epub 2008 Jan 11. PMID: 18192444; PMCID: PMC2259038.

Julca I, Mutwil-Anderwald D, Manoj V, Khan Z, Lai SK, Yang LK, Beh IT, Dziekan J, Lim YP, Lim SK, Low YW, Lam YI, Tjia S, Mu Y, Tan QW, Nuc P, Choo LM, Khew G, Shining L, Kam A, Tam JP, Bozdech Z, Schmidt M, Usadel B, Kanagasundaram Y, Alseekh S, Fernie A, Li HY, Mutwil M. Genomic, transcriptomic, and metabolomic analysis of Oldenlandia corymbosa reveals the biosynthesis and mode of action of anti-cancer metabolites. J Integr Plant Biol. 2023 Jun;65(6):1442-1466. doi: 10.1111/jipb.13469. Epub 2023 Apr 4. PMID: 36807520.

Karp PD, Paley S, Krummenacker M, Kothari A, Wannemuehler MJ, Phillips GJ, 2022. Pathway Tools Management of Pathway/Genome Data for Microbial Communities. Front Bioinform. 2: 869150. https://doi.org/10.3389/fbinf.2022.869150.

Kautsar SA, Duran HGS, Blin K, Osbourn A, Medema MH. PlantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Research, 2017, Vol. 45, Web Server issue W55–W63. DOI: 10.1093/nar/gkx305.

Kautsar SA, Suarez Duran HG, Medema MH. Genomic Identification and Analysis of Specialized Metabolite Biosynthetic Gene Clusters in Plants Using PlantiSMASH. Methods Mol Biol. 2018; 1795:173-188. doi: 10.1007/978-1-4939-7874-8_15. PMID: 29846928.

Kautsar SA, Blin K, Shaw S, Weber T, Medema MH, 2021. BiG-FAM: the biosynthetic gene cluster families database. Nucleic Acids Research, 49(D1): D490–D497, https://doi.org/10.1093/nar/gkaa812.

Kersey PJ, Plant genome sequences: past, present, future, Current Opinion in Plant Biology, Volume 48, 2019, Pages 1-8, ISSN 1369-5266, https://doi.org/10.1016/j.pbi.2018.11.001.

Kessler A, Kalske A, 2018. Plant Secondary Metabolite Diversity and Species Interactions. Annual Review of Ecology, Evolution, and Systematics, 49: 115-138. https://doi.org/10.1146/annurev-ecolsys-110617-062406.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL, 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 37(8): 907-915. https://doi.org/10.1038/s41587-019-0201-4.

Kochko A. de; Crouzillat D. Arabica Coffee Genome Consortium. Aims and goals of the Arabica Coffee Genome Consortium (ACGC), (2015).

Lau KH, Bhat WW, Hamilton JP, Wood JC, Vaillancourt B, Wiegert-Rininger K, Newton L, Hamberger B, Holmes D, Hamberger B, Buell CR, 2020. Genome assembly of Chiococca alba uncovers key enzymes involved in the biosynthesis of unusual terpenoids. DNA Res. 27(3): dsaa013. https://doi.org/10.1093/dnares/dsaa013.

Li CY, Yang L, Liu Y, Xu ZG, Gao J, Huang YB, Xu JJ, Fan H, Kong Y, Wei YK, Hu WL, Wang LJ, Zhao Q, Hu YH, Zhang YJ, Martin C, Chen XY, 2022. The sage genome provides insight into

the evolutionary dynamics of diterpene biosynthesis gene cluster in plants. Cell Rep. 40(7):111236. https://doi.org/10.1016/j.celrep.2022.111236.

Li C, Dong L, Durairaj J, Guan JC, Yoshimura M, Quinodoz P, Horber R, Gaus K, Li J, Setotaw YB, Qi J, De Groote H, Wang Y, Thiombiano B, Floková K, Walmsley A, Charnikhova TV, Chojnacka A, Correia de Lemos S, Ding Y, Skibbe D, Hermann K, Screpanti C, De Mesmaeker A, Schmelz EA, Menkir A, Medema M, Van Dijk ADJ, Wu J, Koch KE, Bouwmeester HJ. Maize resistance to witchweed through changes in strigolactone biosynthesis. Science. 2023 Jan 6;379(6627):94-99. doi: 10.1126/science.abq4775. Epub 2023 Jan 5. PMID: 36603079.

Liao Y, Smyth GK, Shi W, 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 30(7): 923-30. https://doi.org/10.1093/bioinformatics/btt656.

Lima ST, Ampolini BG, Underwood EB, Graf TN, Earp CE, Khedi IC, Pasquale MA, Chekan JR. A Widely Distributed Biosynthetic Cassette Is Responsible for Diverse Plant Side Chain Cross-Linked Cyclopeptides. Angew Chem Int Ed Engl. 2023 Feb 6;62(7):e202218082. doi: 10.1002/anie.202218082. Epub 2023 Jan 12. PMID: 36529706; PMCID: PMC10107690.

Liu C, Smit SJ, Dang J, Zhou P, Godden GT, Jiang Z, Liu W, Liu L, Lin W, Duan J, Wu Q, Lichman BR, 2023. A chromosome-level genome assembly reveals that a bipartite gene cluster formed via an inverted duplication controls monoterpenoid biosynthesis in Schizonepeta tenuifolia. Mol Plant. 16(3): 533-548. https://doi.org/10.1016/j.molp.2023.01.004.

Ma Y, Schranz ME, Suárez-Duran H. Comparative analysis of biosynthetic gene clusters (BGCs) in *Arabidopsis thaliana* and *Cleome violacea*. Tese (Doutorado em Bioinformática) – Wageningen University, Netherlands. 2019. Disponível em: https://library.wur.nl/WebQuery/theses/2251263.

Maeda HA, Fernie AR, 2021. Evolutionary History of Plant Metabolism. Annu Rev Plant Biol. 72:185-216. https://doi.org/10.1146/annurev-arplant-080620-031054.

Martins D, Nunez CV, 2015. Secondary metabolites from Rubiaceae species. Molecules. 20(7): 13422-95. https://doi.org/10.3390/molecules200713422.

Matsuba Y, Nguyen TT, Wiegert K, Falara V, Gonzales-Vigil E, Leong B, Schäfer P, Kudrna D, Wing RA, Bolger AM, Usadel B, Tissier A, Fernie AR, Barry CS, Pichersky E, 2013. Evolution of a complex locus for terpene biosynthesis in solanum. Plant Cell. 25(6): 2022-36. https://doi.org/10.1105/tpc.113.111013.

Matsuba Y, Zi J, Jones AD, Peters RJ, Pichersky E, 2015. Biosynthesis of the diterpenoid lycosantalonol via nerylneryl diphosphate in Solanum lycopersicum. PLoS One. 10(3): e0119302. https://doi.org/10.1371/journal.pone.0119302.

Medema MH, Osbourn A, 2016. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. Nat Prod Rep. 33(8): 951-62. https://doi.org/10.1039/c6np00035e.

Mohite OS, Lloyd CJ, Monk JM, Weber T, Palsson BO, 2022. Pangenome analysis of Enterobacteria reveals richness of secondary metabolite gene clusters and their associated gene sets. Synth Syst Biotechnol. 7(3): 900-910. https://doi.org/10.1016/j.synbio.2022.04.011.

Mueller LA, Zhang P, Rhee SY. AraCyc: a biochemical pathway database for Arabidopsis. Plant Physiol. 2003 Jun;132(2):453-60. doi: 10.1104/pp.102.017236.

Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W, Fernandez-Guerra A, Cappelini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gomez F, Medema MH, 2020. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol. 16(1): 60-68. https://doi.org/10.1038/s41589-019-0400-9.

Nützmann H, Scazzocchio C, Osbourn A. Metabolic Gene Clusters in Eukaryotes. Annu. Rev. Genet. 2018. 52:7.1–7.25. DOI: https://doi.org/10.1146/annurev-genet-120417-031237.

Paddon CJ, Westfall PJ, Pitera DJ, Benjamin K. et al. High-level semi-synthetic production of the potent antimalarial artemisinin. Nature vol. 496, pages 528–532(2013). DOI:10.1038/nature12051.

Pan Y, Zhao X, Wang Y, Tan J, Chen DX, 2021. Metabolomics integrated with transcriptomics reveals the distribution of iridoid and crocin metabolic flux in Gardenia jasminoides Ellis. PLoS One. 16(9): e0256802. https://doi.org/10.1371/journal.pone.0256802.

Perrois C, Strickler SR, Mathieu G, Lepelley M, Bedon L, Michaux S, Husson J, Mueller L, Privat I. Differential regulation of caffeine metabolism in Coffea arabica (Arabica) and Coffea canephora (Robusta). Planta. 2015 Jan;241(1):179-91. doi: 10.1007/s00425-014-2170-7. Epub 2014 Sep 24. PMID: 25249475; PMCID: PMC4282694.

Pichersky E, Lewinsohn E, 2011. Convergent evolution in plant specialized metabolism. Annu Rev Plant Biol. 62: 549-66. https://doi.org/10.1146/annurev-arplant-042110-103814.

Polturak G, Liu Z, Osbourn A. New and emerging concepts in the evolution and function of plant biosynthetic gene clusters. Current Opinion in Green and Sustainable Chemistry, 2022. 33:100568, ISSN 2452-2236, DOI: https://doi.org/10.1016/j.cogsc.2021.100568.

Pucker B, Reiher F, Schilbert HM, 2020. Automatic Identification of Players in the Flavonoid Biosynthesis with Application on the Biomedicinal Plant *Croton tiglium*. *Plants*. 9(9): 1103. https://doi.org/10.3390/plants9091103.

Rabara RC, Kudithipudi C, Timko MP, 2023. Identification of Terpene-Related Biosynthetic Gene Clusters in Tobacco through Computational-Based Genomic, Transcriptomic, and Metabolic Analyses. *Agronomy*. 13(6): 1632. https://doi.org/10.3390/agronomy13061632.

Raharimalala N, Rombauts S, McCarthy A. et al. The absence of the caffeine synthase gene is involved in the naturally decaffeinated status of Coffea humblotiana, a wild species from Comoro archipelago. Sci Rep 11, 8119 (2021). DOI: https://doi.org/10.1038/s41598-021-87419-0.

Rai A, Hirakawa H, Nakabayashi R. et al. Chromosome-level genome assembly of Ophiorrhiza pumila reveals the evolution of camptothecin biosynthesis. Nat Commun 12, 405 (2021). https://doi.org/10.1038/s41467-020-20508-2.

Rieseberg TP, Dadras A, Fürst-Jansen JMR, et al., 2023. Crossroads in the evolution of plant specialized metabolism, Seminars in Cell & Developmental Biology, 134: 37-58, ISSN 1084-9521. https://doi.org/10.1016/j.semcdb.2022.03.004.

Robey MT, Caesar LK, Drott MT, Keller NP, Kelleher NL, 2021. An interpreted atlas of biosynthetic gene clusters from 1,000 fungal genomes. Proc Natl Acad Sci U S A. 118(19): e2020230118. https://doi.org/10.1073/pnas.2020230118.

Scalabrin S, Toniutti L, Di Gaspero G, Scaglione D, Magris G, Vidotto M, Pinosio S, Cattonaro F, Magni F, Jurman I, Cerutti M, Suggi Liverani F, Navarini L, Del Terra L, Pellegrino G, Ruosi MR, Vitulo N, Valle G, Pallavicini A, Graziosi G, Klein PE, Bentley N, Murray S, Solano W, Al Hakimi A, Schilling T, Montagnon C, Morgante M, Bertrand B. A single polyploidization event at the origin of the tetraploid genome of Coffea arabica is responsible for the extremely low genetic variation in wild and cultivated germplasm. Sci Rep. 2020 Mar 13;10(1):4642. doi: 10.1038/s41598-020-61216-7. PMID: 32170172; PMCID: PMC7069947.

Schläpfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, Dreher K, Chavali AK, Nilo-Poyanco R, Bernard T, Kahn D, Rhee SY. Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants, Plant Physiology, Volume 173, Issue 4, April 2017, Pages 2041–2059, DOI https://doi.org/10.1104/pp.16.01942.

Schubert M, Lindgreen S, Orlando L, 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 9: 88. https://doi.org/10.1186/s13104-016-1900-2.

Singh KS, van der Hooft JJJ, van Wees SCM, Medema MH, 2022. Integrative omics approaches for biosynthetic pathway discovery in plants. Nat Prod Rep. 39(9): 1876-1896. https://doi.org/10.1039/d2np00032f.

Smit SJ, Lichman BR. Plant biosynthetic gene clusters in the context of metabolic evolution. Nat Prod Rep. 2022 Apr 20. doi: 10.1039/d2np00005a. Epub ahead of print. PMID: 35441651.

Srivastav VK, Egbuna C, Tiwari M, 2020. Chapter 1 - Plant secondary metabolites as lead compounds for the production of potent drugs. Phytochemicals as Lead Compounds for New Drug Discovery, Elsevier, ISBN 9780128178904, 3-14. https://doi.org/10.1016/B978-0-12-817890-4.00001-9.

Stavrinides AK, Dussert S, Combes MC, Fock-Bastide I, Severac D, Minier J, Bastos-Siqueira A, Demolombe V, Hem S, Lashermes P, Joët T, 2020. Seed comparative genomics in three coffee species identify desiccation tolerance mechanisms in intermediate seeds. J Exp Bot. 71(4): 1418-1433. https://doi.org/10.1093/jxb/erz508.

Sun J, Lu F, Luo Y, Bie L, Xu L, Wang Y. OrthoVenn3: an integrated platform for exploring and visualizing orthologous data across genomes. Nucleic Acids Res. 2023 Jul 5;51(W1):W397-W403. doi: 10.1093/nar/gkad313. PMID: 37114999; PMCID: PMC10320085.

Terlouw BR, Blin K, Navarro-Muñoz JC, Avalon NE, Chevrette MG, Egbert S, Lee S, Meijer D, Recchia MJJ, Reitz ZL, van Santen JA, Selem-Mojica N, Tørring T, Zaroubi L, Alanjary M, Aleti G, Aguilar C, Al-Salihi SAA, Augustijn HE, Avelar-Rivas JA, Avitia-Domínguez LA, Barona-Gómez F, Bernaldo-Agüero J, Bielinski VA, Biermann F, Booth TJ, Carrion Bravo VJ, Castelo-Branco R, Chagas FO, Cruz-Morales P, Du C, Duncan KR, Gavriilidou A, Gayrard D, Gutiérrez-García K, Haslinger K, Helfrich EJN, van der Hooft JJJ, Jati AP, Kalkreuter E, Kalyvas N, Kang KB, Kautsar S, Kim W, Kunjapur AM, Li YX, Lin GM, Loureiro C, Louwen JJR, Louwen NLL, Lund G, Parra J, Philmus B, Pourmohsenin B, Pronk LJU, Rego A, Rex DAB, Robinson S, Rosas-Becerra LR, Roxborough ET, Schorn MA, Scobie DJ, Singh KS, Sokolova N, Tang X, Udwary D, Vigneshwari A, Vind K, Vromans SPJM, Waschulin V, Williams SE, Winter JM, Witte TE, Xie H, Yang D, Yu J, Zdouc M, Zhong Z, Collemare J, Linington RG, Weber T, Medema MH, 2023. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. Nucleic Acids Res. 51(D1): D603-D610. https://doi.org/10.1093/nar/gkac1049.

Töpfer N, Fuchs LM, Aharoni A. The PhytoClust tool for metabolic gene clusters discovery in plant genomes. Nucleic Acids Res. 2017 Jul 7;45(12):7049-7063. doi: 10.1093/nar/gkx404. PMID: 28486689; PMCID: PMC5499548.

Twaij BM, Hasan MN, 2022. Bioactive Secondary Metabolites from Plant Sources: Types, Synthesis, and Their Therapeutic Uses. International Journal of Plant Biology. 2022; 13(1):4-14. https://doi.org/10.3390/ijpb13010003.

Udomsom N, Rai A, Suzuki H, Okuyama J, Imai R, Mori T, Nakabayashi R, Saito K, Yamazaki M, 2016. Function of AP2/ERF Transcription Factors Involved in the Regulation of Specialized

Metabolism in Ophiorrhiza pumila Revealed by Transcriptomics and Metabolomics. Front Plant Sci. 7: 1861. https://doi.org/10.3389/fpls.2016.01861.

Wang J, Xu S, Mei Y, Cai S, Gu Y, Sun M, Liang Z, Xiao Y, Zhang M, Yang S, 2021. A high-quality genome assembly of Morinda officinalis, a famous native southern herb in the Lingnan region of southern China. Hortic Res. 8, 135. https://doi.org/10.1038/s41438-021-00551-w.

Wang P, Schumacher AM, Shiu SH, 2022. Computational prediction of plant metabolic pathways. Curr Opin Plant Biol. 66: 102171. https://doi.org/10.1016/j.pbi.2021.102171.

Wingett SW and Andrews S, 2018. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Research 7:1338. https://doi.org/10.12688/f1000research.15931.2.

Wu S, Malaco Morotti AL, Wang S, Wang Y, Xu X, Chen J, Wang G, Tatsis EC, 2022. Convergent gene clusters underpin hyperforin biosynthesis in St John's wort. New Phytol. 235(2): 646-661. https://doi.org/10.1111/nph.18138.

Xia EH, Zhang HB, Sheng J, Li K, Zhang QJ, Kim C, Zhang Y, Liu Y, Zhu T, Li W, Huang H, Tong Y, Nan H, Shi C, Shi C, Jiang JJ, Mao SY, Jiao JY, Zhang D, Zhao Y, Zhao YJ, Zhang LP, Liu YL, Liu BY, Yu Y, Shao SF, Ni DJ, Eichler EE, Gao LZ. The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis. Mol Plant. 2017 Jun 5;10(6):866-877. doi: 10.1016/j.molp.2017.04.002. Epub 2017 May 2. PMID: 28473262.

Xu Z, Pu X, Gao R, et al. Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants. BMC Biol. 2020;18(1):63. Published 2020 Jun 18. doi:10.1186/s12915-020-00795-3.

Yang X, Zhang L, Guo X, Xu J, Zhang K, Yang Y, Yang Y, Jian Y, Dong D, Huang S, Cheng F, Li G, 2023. The gap-free potato genome assembly reveals large tandem gene clusters of agronomical importance in highly repeated genomic regions. Mol Plant. 16(2): 314-317. https://doi.org/10.1016/j.molp.2022.12.010.

Zhan C, Shen S, Yang C, Liu Z, Fernie AR, Graham IA, Luo J. Plant metabolic gene clusters in the multi-omics era. Trends Plant Sci. 2022 Oct;27(10):981-1001. doi: 10.1016/j.tplants.2022.03.002. Epub 2022 Mar 30. PMID: 35365433.

Zhao K, Rhee SY, 2022. Omics-guided metabolic pathway discovery in plants: Resources, approaches, and opportunities. Curr Opin Plant Biol. 67: 102222. https://doi.org/10.1016/j.pbi.2022.102222.

Zhao X, Hu X, OuYang K, Yang J, Que Q, Long J, Zhang J, Zhang T, Wang X, Gao J, Hu X, Yang S, Zhang L, Li S, Gao W, Li B, Jiang W, Nielsen E, Chen X, Peng C. Chromosome-level assembly of the Neolamarckia cadamba genome provides insights into the evolution of cadambine biosynthesis. Plant J. 2022 Feb;109(4):891-908. doi: 10.1111/tpj.15600.

Zhou F, Pichersky E, 2020. The complete functional characterisation of the terpene synthase family in tomato. New Phytol. 226(5): 1341-1360. https://doi.org/10.1111/nph.16431.

Anexos

frontiers | Frontiers in Genetics

TYPE Data Report
PUBLISHED 28 October 2022
DOI 10.3389/fgene.2022.925811

OPEN ACCESS

EDITED BY
Andrew H Paterson,
University of Georgia, United States

REVIEWED BY
Isabel Marques,
University of Lisbon, Portugal
Marcio Alves-Ferreira,
Federal University of Rio de Janeiro, Brazil

*CORRESPONDENCE
Douglas S. Domingues,
dougsd@usp.br

SPECIALTY SECTION
This article was submitted to Plant Genomics,
a section of the journal
Frontiers in Genetics

RECEIVED 21 April 2022
ACCEPTED 13 October 2022
PUBLISHED 28 October 2022

CITATION
Budzinski IGF, Camargo PO, Lemos SMC, Guyot R, Calzado NF, Ivamoto-Suzuki ST and Domingues DS (2022), Transcriptomic alterations in roots of two contrasting Coffea arabica cultivars after hexanoic acid priming. *Front. Genet.* 13:925811. doi: 10.3389/fgene.2022.925811

Transcriptomic alterations in roots of two contrasting *Coffea arabica* cultivars after hexanoic acid priming

Ilara G. F. Budzinski[1], Paula O. Camargo[1], Samara M. C. Lemos[1,2], Romain Guyot[3], Natália F. Calzado[1], Suzana T. Ivamoto-Suzuki[1] and Douglas S. Domingues[1,4]*

[1]Group of Genomics and Transcriptomes in Plants, Department of Biodiversity, Institute of Biosciences, São Paulo State University, UNESP, Rio Claro, Brazil, [2]Graduate Program in Biological Sciences (Genetics), Institute of Biosciences, São Paulo State University, UNESP, Botucatu, Brazil, [3]Institut de Recherche pour le Développement (IRD), Université Montpellier, Montpellier, France, [4]Department of Genetics, "Luiz de Queiroz" College of Agriculture, University of São Paulo, ESALQ/USP, Piracicaba, Brazil

KEYWORDS

RNA-seq, coffee, hexanoic acid, priming agent, elicitation, root

Introduction

Plants have the capacity to enter a state of alert that enables them to respond rapidly and robustly after exposure to stress (Aranega-Bou et al., 2014). This phenomenon is known as priming and can be described as an induced state whereby plants are pre-exposed to an inducing agent (elicitor), thus improving their perception and/or amplification of defense response-inducing signals (Aranega-Bou et al., 2014; Tugizimana et al., 2018). Hexanoic acid (Hx), a monocarboxylic acid, is a natural priming agent with proven efficiency in a wide range of host plants and pathogens (Llorens et al., 2016), including coffee pathogens. Coffee (*Coffea* spp.) is one of the most important agricultural commodities in the world. Brazil is the largest producer and exporter of *Coffea arabica* L. (Brazilian Coffee Exporters Council, 2021). The genus *Coffea* comprises 124 species (Davis et al., 2011). The most planted one is *C. arabica*, the only allotetraploid species in the genus. As many other plants, *Coffea* spp. are sensitive to a diverse range of biotic and abiotic stress. It is known that priming leads to changes at the transcriptional, physiological, metabolic and epigenetic levels (Baccelli et al., 2020). A transcriptional reprogramming may occur after priming stimulation, affecting a huge number of genes (Cervantes-Gámez et al., 2016; Baccelli et al., 2020). Within this context, our aim was to investigate the effect *per se* of Hx application. We hypothesize if Hx application could modulate genes related to defense response, in *C. arabica*, being a potential eliciting agent to this crop. To test this, Hx was applied in the roots of two Brazilian *C. arabica* cultivars: Catuaí Vermelho and Obatã. Cultivars were chosen based on their distinct breeding histories and contrasting resistance to rust, the major disease in Arabica coffee worldwide (Talhinhas et al., 2017). Catuaí Vermelho is susceptible to rust, and is one of the most planted cultivars in Brazil, while Obatã is described as a moderately resistant cultivar (Del Grossi et al., 2013). In the present work, transcriptomic analysis of roots were performed, revealing different molecular responses. Based on FPKM ratio and

statistical analyses, 1,545 differentially expressed genes (DEGs) were found. Functional annotation of DEGs through Blast2GO showed that primary, organic substance and cellular metabolic processes were mainly affected by priming, in both cultivars. Here, we present an RNA-seq dataset containing raw files and an initial exploration of differentially expressed genes in two *C. arabica* cultivars. Besides, these data could contribute to the identification of key genes differentially expressed in response to Hx.

Material and methods

Plant material

Plant material and experimental setup used in this work was the same described in a previous publication from our group (Budzinski et al., 2021).

Two commercial cultivars of *C. arabica* (five-month-old plants) were used, Catuaí Vermelho IAC 144 and Obatã IAC 1669-20. Both cultivars are inbred lines of *C. arabica* (Maluf et al., 2005); however, Catuaí is derived from a cross between Catuaí Amarelo 476 × Mundo Novo 374-19, while Obatã is derived from interspecific crosses between (Villa Sarchi × Hybrid of Timor) × Catuaí Vermelho; clarifying that Villa Sarchi is a *C. arabica* cultivar and Hybrid of Timor is a natural *C. arabica* x *C. canephora* hybrid (Lashermes et al., 2000; Maluf et al., 2005). These cultivars were chosen due to their contrasting response to rust, with Obatã being the resistant one (Maluf et al., 2005; Krohling et al., 2018). Plants were selected based on size uniformity and were transferred to pots containing 3 L of aerated nutrient solution (ANS), adapted from Clark, 1975) by de Carvalho et al. (2013). The experiment was carried out as described in Silva et al. (2020), under controlled temperature (23 ± 2°C) and light/dark cycle (12h/12h, photosynthetically active photon flux density of ~400 μmol $m^{-2}.s^{-1}$). The following treatments were applied: (a) ANS (control); (b) ANS + hexanoic acid (Merck, final concentration 0.55 mM) for 48 h. Three plants per pot were grown into six plastic pots in which three pots received each treatment. The experiments were repeated 3 times to obtain biological replicates. The potted plants were grouped in "pools" (made of 9–18 plants), which were considered a biological replicate. Three biological replicates were used. We collected plant secondary roots within the 3rd hour of the light period and stored at -80°C for further analyses.

Total RNA extraction and quality control

All steps from total RNA extraction until gene expression analysis were the same as described in Budzinski et al. (2021).

Total RNA from root pools were isolated using the RNeasy Plant kit (Qiagen, Hilden, North Rhine-Westphalia, Germany). Total RNA samples were purified using the RNeasy Minielute Cleanup kit (Qiagen, Hilden, North Rhine-Westphalia, Germany). The purity of RNA was determined using a NanoDrop ND-100 spectrophotometer (Thermo Scientific, San Jose, CA, United States). RNA concentrations were measured by a Qubit fluorometer (Thermo Fisher Scientific, Wilmington, DE, United States).

Library preparation, and RNA-seq

Poly(A) RNA sequencing library was prepared following Illumina's TruSeq-stranded-mRNA sample preparation protocol (Illumina Technologies, SanDiego, CA). Paired-end sequencing (2 X 150 bp) was performed on Illumina's NovaSeq 6000 sequencing system at LC Sciences (Houston, TX, United States). Data was deposited into the European Nucleotide Archive (ENA), submission PRJEB52366.

RNAseq analysis and gene expression analysis

All steps mentioned here are the same as described in Budzinski et al. (2021). Adaptor contamination, low quality bases and undetermined bases were removed by using Cutadapt (Martin, 2011) and in house PERL scripts. Sequence quality was verified using FastQC (Andrews, 2010). HISAT2 (Kim et al., 2015) was used to map reads to the *Coffea arabica* genome (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/713/225/GCF_003713225.1_Cara_1.0/).

StringTie (Pertea et al., 2015) was used to assemble the mapped reads and to detect the expression level for mRNAs by calculating FPKM. The differentially expressed genes (DEGs) were selected with log2 (fold change) >1 or log2 (fold change) <-1 and with statistical significance (p value <0.05) by R package edgeR (Robinson et al., 2010). A second analysis was done on the differentially expressed mRNAs and only the ones with FPKM (ratio) ≥ 2 or FPKM (ratio) ≤ -2; coefficient of variation ≤30% and average FPKM ≥5 were selected for further analyses. Genes found specifically in one condition (control or plants exposed to Hx) were also described as DEGs.

Sequence annotation and gene ontology (GO) enrichment analysis of DEGs were performed using *Blast2GO* (Conesa et al., 2005), at the *BioBam* (Götz et al., 2008) platform. Sequences were annotated by blasting nucleotide sequences against the NCBI NR database (BLASTX, evalue ≤1.10^{-5}). The hypergeometric distribution was used to test whether the GO function set was significantly enriched ($p < 0.05$). Pathway mapping was done using MapMan software (Thimm et al., 2004) with the *Arabidopsis thaliana* mapping file (http://mapman.gabipd.org/). TAIR IDs were retrieved from NCBI (https://www.ncbi.nlm.nih.gov).

10.3389/fgene.2022.925811

TABLE 1 Summary of sequencing data quality

Sample	Raw data	Valid data	Valid data gb (G)	Valid ratio (reads)	Q30%	GC content%
CC_root1	53,455,648	38,232,894	5.73	71.52	99.03	46
CC_root2	52,404,378	38,309,030	5.75	73.1	99.02	45.5
CC_root3	51,672,842	36,455,566	5.47	70.55	98.99	45
OC_root1	42,011,570	37,025,650	5.55	88.13	97.6	45
OC_root2	42,381,098	37,588,930	5.64	88.69	97.36	45
OC_root3	41,533,684	36,290,882	5.44	87.38	97.36	45.5
CHX_root1	44,422,470	32,902,840	4.94	74.07	97.41	51
CHX_root2	51,270,376	50,002,294	7.5	97.53	97.84	51
CHX_root3	46,166,432	40,680,326	6.1	88.12	98.59	52
OHX_root1	33,581,294	32,733,990	4.91	97.48	98.12	51
OHX_root2	41,411,214	35,239,206	5.29	85.1	98.57	51
OHX_root3	32,963,354	31,991,596	4.8	97.05	98.18	52

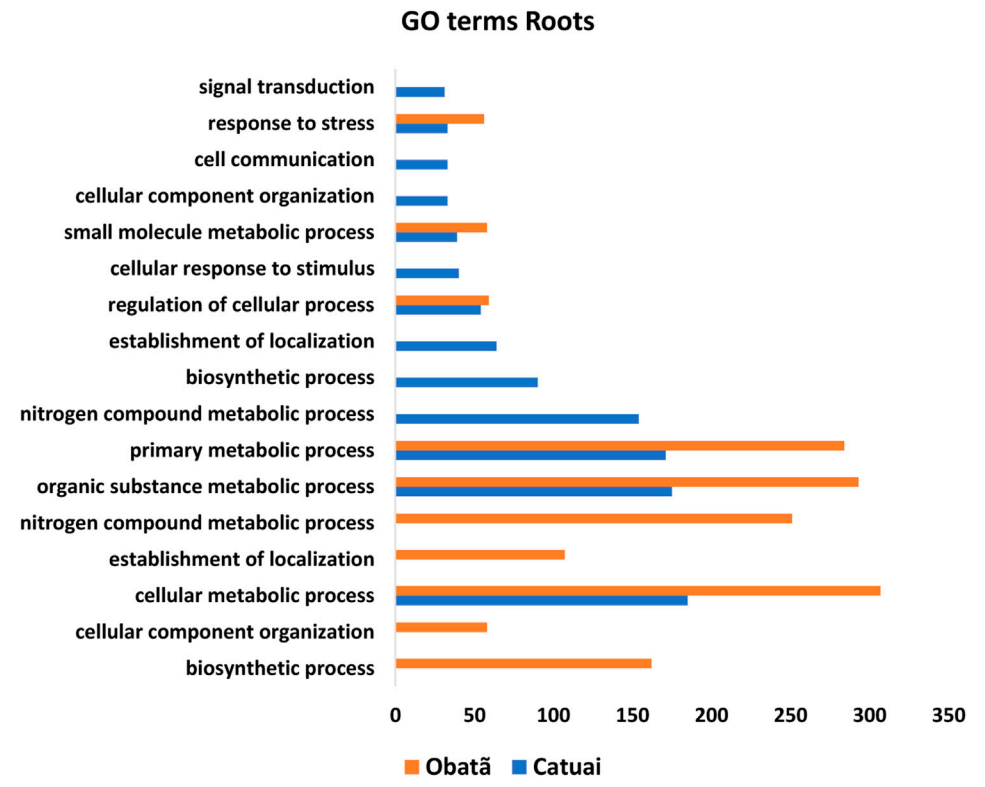

FIGURE 1
Gene ontology (GO) functional enrichment analysis of DEGs from C. *arabica* Catuaí and Obatã cultivars.

Overall data annotation, differentially expressed genes and gene ontology analysis

Quality control and mapping information are available in Table 1. About 67.12 Gb total clean bases were obtained by RNA-seq after quality check, with an average of 5.6 Gb for each sample. The lowest value of Q30 (percentage of bases with sequencing error rate lower than 1∘) was 97.36%. The GC content ranged from 45 to 52%.

As a preliminary analysis to identify genes and functional categories potentially modulated by Hx application, the first step of our work was to identify the DEGs based on FPKM and statistical analysis. Based on FPKM ratio and statistical analyses, 1,545 DEGs were found in total, 557 and 988 in Catuaí and Obatã, respectively (Supplementary Table S1). From these, 157 DEGs were found in both cultivars, while 400 and 831 DEGs were specifically found in Catuaí and Obatã cultivars, respectively (Supplementary Tables S2, S3). We hypothesize that the discrepancy between the number of specific DGEs, found in each cultivar, is related to differences in rust resistance, reinforcing that molecular mechanisms of defense are differentially recruited depending on cultivar tolerance. Most of the DEGs have a role in plant defense, indicating the modulation of this mechanism in roots by priming. Blast2GO analysis showed that primary, organic substance and cellular metabolic processes were mainly affected by priming, followed by response to stress, small molecule metabolic process and regulation of cellular process (Figure 1, Supplementary Table S5). Pathway analysis of DEGs using MapMan showed differences in the activity of cellular metabolisms due to Hx (Supplementary Table S3). The dataset presented here indicates that hexanoic acid modulates plant defense mechanisms in *C. arabica*. Moreover, we are providing useful data for further investigations on *C. arabica* root responses to Hx.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ebi.ac.uk/ena, PRJEB52366. All supplementary files are available on https://doi.org/10.5281/zenodo.6467813.

Author contributions

Conceptualization, Project Administration, Funding Acquisition, Supervision: DD. Data Curation, Investigation: PC, SL, RG, NC, STI-S. Formal Analysis, Validation, Visualization: IB. Writing—Original Draft Preparation, Writing—Review and Editing: IB, DD.

Funding

This research was funded by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), grant number #2016/10896-0 and CAPES-PrInt Program 2346/2018 (process 88881.310767/2018-01). IB, NC and SL were financed in part by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES)—Finance Code 001.

Acknowledgments

IB acknowledges the scholarship granted from the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES), in the scope of the Program CAPES-PrInt, process number 88887.310463/2018-00, International Cooperation Project number 88887.512173/2020-00. SL also acknowledges a CAPES fellowship, process number 88887.570128/2020-00. STI-S. acknowledges FAPESP for providing a post-doctoral fellowship, process number #2017/01455-2. DD also acknowledges CNPq for a research productivity fellowship (process number #312823/2019-3).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.925811/full#supplementary-material

References

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Aranega-Bou, P., Leyva, M., Finiti, I., Garcia-Agustin, P., and Gonzalez-Bosch, C. (2014). Priming of plant resistance by natural compounds. Hexanoic acid as a model. *Front. Plant Sci.* 5, 488. doi:10.3389/fpls.2014.00488

Baccelli, I., Benny, J., Caruso, T., and Martinelli, F. (2020). The priming fingerprint on the plant transcriptome investigated through meta-analysis of RNA-Seq data. *Eur. J. Plant Pathol.* 156, 779–797. doi:10.1007/s10658-019-01928-3

Brazilian Coffee Exporters Council (Cecafe) (2021). Production. Available at: https://www.cecafe.com.br/en/about-coffee/production.

Budzinski, I. G. F., Camargo, P. O., Rosa, R. S., Calzado, N. F., Ivamoto-Suzuki, S. T., and Domingues, D. S. (2021). Transcriptome analyses of leaves reveal that hexanoic acid priming differentially regulate gene expression in contrasting coffea arabica cultivars. *Front. Sustain. Food Syst.* 5. doi:10.3389/fsufs.2021.735893

Cervantes-Gámez, R. G., Bueno-Ibarra, M. A., Cruz-Mendívil, A., Calderón-Vázquez, C. L., Ramírez-Douriet, C. M., Maldonado-Mendoza, I. E., et al. (2016). Arbuscular mycorrhizal symbiosis-induced expression changes in Solanum lycopersicum leaves revealed by RNA-seq analysis. *Plant Mol. Biol. Rep.* 34, 89–102. doi:10.1007/s11105-015-0903-9

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., Robles, M., et al. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi:10.1093/bioinformatics/bti610

Clark, R. B. (1975). Characterization of phosphatase of intact maize roots. *J. Agric. Food Chem.* 23, 458–460.

Davis, A. P., Tosh, J., Ruch, N., and Fay, M. F. (2011). Growing coffee: Psilanthus (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of Coffea. *Bot. J. Linn. Soc.* 167 (4), 357–377. doi:10.1111/j.1095-8339.2011.01177.x

de Carvalho, K., Bespalhok Filho, J. C., dos Santos, T. B., de Souza, S. G. H., Vieira, L. G. E., Pereira, L. F. P., et al. (2013). Nitrogen starvation, salt and heat stress in coffee (*Coffea arabica* L.): Identification and validation of new genes for qPCR normalization. *Mol. Biotechnol.* 53, 315–325.

Del Grossi, L., Sera, T., Sera, G. H., Fonseca, I., Ito, D. S., Shigueoka, L. H., et al. (2013). Rust resistance in Arabic coffee cultivars in northern paraná. *Braz. Arch. Biol. Technol.* 56, 27–33. doi:10.1590/s1516-89132013000100004

Götz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36, 3420–3435. doi:10.1093/nar/gkn176

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360.

Krohling, C. A., Matiello, J. B., de Almeida, S. R., Eutrópio, F. J., and de Siqueira Carvalho, C. H. (2018). Adaptation of progenies/cultivars of arabica coffee (coffea arabica l.) in mountainous edafoclimatic conditions. *Coffee Sci.* 13, 198–209. doi:10.25186/cs.v13i2.1417

Lashermes, P., Andrzejewski, S., Bertrand, B., Combes, M. C., Dussert, S., Graziosi, G., et al. (2000). Molecular analysis of introgressive breeding in coffee (Coffea arabica L.). *Theor. Appl. Genet.* 100, 139–146. doi:10.1007/s001220050019

Llorens, E., Camañes, G., Lapeña, L., and García-Agustín, P. (2016). Priming by hexanoic acid induce activation of mevalonic and linolenic pathways and promotes the emission of plant volatiles. *Front. Plant Sci.* 7, 495. doi:10.3389/fpls.2016.00495

Maluf, M. P., Silvestrini, M., Ruggiero, L. M. C., Filho, O. G., and Colombo, C. A. (2005). Genetic diversity of cultivated Coffea arabica inbred lines assessed by RAPD, AFLP and SSR marker systems. *Sci. Agric.* 62, 366–373. doi:10.1590/s0103-90162005000400010

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* 17, 10–12.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295.

Robinson, M., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616

Silva, N., Ivamoto-Suzuki, S. T., Camargo, P. O., Rosa, R. S., Pereira, L. F. P., and Domingues, D. S. (2020). Low-copy genes in terpenoid metabolism: The evolution and expression of MVK and DXR genes in angiosperms. *Plants* 9, 525. doi:10.3390/plants9040525

Talhinhas, P., Batista, D., Diniz, I., Vieira, A., Silva, D. N., Loureiro, A., et al. (2017). The coffee leaf rust pathogen hemileia vastatrix: One and a half centuries around the tropics. *Mol. Plant Pathol.* 18 (8), 1039–1051. doi:10.1111/mpp.12512

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., et al. (2004). MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37, 914–939.

Tugizimana, F., Mhlongo, M. I., Piater, L. A., and Dubery, I. A. (2018). Metabolomics in plant priming research: The way forward? *Int. J. Mol. Sci.* 19 (6), 1759. doi:10.3390/ijms19061759

PLANT SCIENCE

Maize resistance to witchweed through changes in strigolactone biosynthesis

C. Li[1], L. Dong[1]*, J. Durairaj[2]†, J.-C. Guan[3], M. Yoshimura[4,5,6], P. Quinodoz[5], R. Horber[5], K. Gaus[5], J. Li[7], Y. B. Setotaw[7], J. Qi[7], H. De Groote[8], Y. Wang[1], B. Thiombiano[1], K. Floková[1,9], A. Walmsley[1], T. V. Charnikhova[1], A. Chojnacka[1], S. Correia de Lemos[2,10], Y. Ding[11], D. Skibbe[12], K. Hermann[5], C. Screpanti[5], A. De Mesmaeker[5], E. A. Schmelz[11], A. Menkir[13], M. Medema[2], A. D. J. Van Dijk[2], J. Wu[7], K. E. Koch[3], H. J. Bouwmeester[1]*

Maize (*Zea mays*) is a major staple crop in Africa, where its yield and the livelihood of millions are compromised by the parasitic witchweed *Striga*. Germination of *Striga* is induced by strigolactones exuded from maize roots into the rhizosphere. In a maize germplasm collection, we identified two strigolactones, zealactol and zealactonoic acid, which stimulate less *Striga* germination than the major maize strigolactone, zealactone. We then showed that a single cytochrome P450, ZmCYP706C37, catalyzes a series of oxidative steps in the maize-strigolactone biosynthetic pathway. Reduction in activity of this enzyme and two others involved in the pathway, ZmMAX1b and ZmCLAMT1, can change strigolactone composition and reduce *Striga* germination and infection. These results offer prospects for breeding *Striga*-resistant maize.

Food security is a growing challenge in the face of climate change and increasing food needs (*1*). Maize (*Zea mays*) is one of the most important staple crops in the world, especially in Africa. There, its yield is compromised by the parasitic witchweeds *Striga hermonthica* and *Striga asiatica*. Damage from these *Striga* species threatens the livelihood of millions of people, particularly in sub-Saharan regions (fig. S1) (*2*, *3*). *Striga* seeds lay dormant in soil until their germination is triggered by strigolactones (SLs), signaling compounds exuded by the roots of plants, including maize. The first known SL, strigol, was discovered in the 1960s in the root exudates of cotton (*4*). In addition to having been co-opted as a cue for root-parasitic plants, SLs serve as host signals for beneficial arbuscular mycorrhizal fungi (AMF) and are plant hormones with developmental roles (*5*–*9*).

Thus far, more than 35 different SLs have been discovered, all containing the conserved D-ring (Fig. 1A) (*10*–*12*). The canonical SLs include two groups, the "strigol-type" and "orobanchol-type," whereas noncanonical SLs lack the A-, B-, and/or C-rings (*10*–*12*). Plants usually exude a blend of different SLs, and the composition of the root exudate can vary greatly between and sometimes also within plant species. Many of the SLs display substantial differences in their biological activity, such as the induction of AMF hyphal branching and parasitic plant germination (*9*, *13*–*15*). The biological importance of SL blends is far from understood, but in sorghum (*Sorghum bicolor*), a change in SLs from 5-deoxystrigol to orobanchol decreased *Striga* germination and increased field resistance (*16*).

The mechanisms of SL biosynthesis have only been partially elucidated. Three enzymes—DWARF 27 (D27) and two carotenoid cleavage dioxygenases l(CCDs), CCD7 and CCD8—catalyze the conversion of β-carotene to carlactone (CL) (Fig. 1A) (*17*, *18*). In *Arabidopsis*, CL is oxidized to form carlactonoic acid (CLA) by a cytochrome P450 (CYP) monooxygenase, CYP711A1, encoded by More Axillary Growth 1 (MAX1) homolog AtMAX1 (*19*). *Arabidopsis* has a single copy of this MAX1, whereas maize has three homologs, and rice has five (*18*, *20*). Although both the *Arabidopsis* AtMAX1 and the maize ZmMAX1b form CLA from CL, the rice MAX1 homologs, Os900 and Os1400, instead convert CL to 4-deoxyorobanchol (4DO) and orobanchol, respectively (*18*, *21*). Dicots also form orobanchol, but from CLA rather than CL, and with a different cytochrome P450, CYP722C. A homolog of this CYP722C can also produce 5-deoxystrigol from CLA (*22*, *23*).

Maize roots exude at least six SLs, tw[illegible] which have been structurally identifie[illegible] zealactone and zeapyranolactone (Fig. 1A) (*24*–*26*). However, the identities of the other four SLs remained elusive, as well as the biosynthetic differences between the six and their individual roles in *Striga* germination. In this study, we reveal natural variation in the maize SL blend, identify three new maize SLs, elucidate the entire maize SL biosynthetic pathway, and show that changes in the composition of the SL blend correspond to differences in *Striga* germination and infection. These findings create a pathway for reducing the notorious agricultural problem of *Striga* infection through breeding maize for favorable SL composition.

Natural variation in strigolactone production by maize

To assess the extent of variation in the production of SLs by maize, we grew a collection of maize genotypes, sampled their root exudate, and analyzed SLs with multiple reaction monitoring (MRM) liquid chromatography–tandem mass spectroscopy (LC/MS/MS) (Fig. 1B and figs. S2 and S3) (*24*, *25*). Quantities of exuded SLs varied among these lines (Fig. 1B and fig. S3). Moreover, one of the genotypes, NP2222, displayed a distinctive SL profile, lacking detectable levels of all but two SLs, an unknown SL and designated compound 5 (Fig. 1B and fig. S3). Compound 5 was previously noted in maize root exudate (*24*), but its low abundance and chemical instability hampered structural characterization. Therefore, on the basis of nuclear magnetic resonance (NMR) spectra and retrosynthetic analysis (*24*, *27*–*29*), we postulated structures and subsequently synthesized compound 5 as well as the other unknown SL (figs. S4 to S12). The synthetic products were identical to the natural ones in maize root exudate and were designated zealactol (compound 5) and zealactonoic acid (ZA) (the other unknown SL) (figs. S9 and S12). Bioassay of *Striga* germination showed that both zealactol and ZA were less inductive than zealactone (Fig. 1C), an outcome that highlights how strongly minute differences in SL structure can alter their biological activity. These findings are further supported by work on sorghum (*16*). To unravel the mechanistic basis for these differences in SL blends, we revealed the biosynthetic pathway of maize SLs.

Three maize genes encode the carlactone biosynthetic pathway

Through homology, we identified the maize orthologs *D27*, *CCD7*, and *CCD8*, which catalyze the formation of CL from β-carotene in other plant species (tables S1 and S2). To confirm *ZmCCD8* function, we analyzed root exudate of two independent *zmccd8* mutants (in W22 and Mo17 backgrounds) (*30*). Zealactone was not detected, although it was the major SL in

[1]Plant Hormone Biology Group, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, Netherlands. [2]Bioinformatics Group, Wageningen University & Research, 6708 PB Wageningen, Netherlands. [3]Horticultural Sciences Department, University of Florida, Gainesville, FL 32611, USA. [4]Laboratorium für Organische Chemie, Department of Chemistry and Applied Biosciences, ETH Zürich, 8093 Zürich, Switzerland. [5]Syngenta Crop Protection AG, Schaffhauserstrasse 101, CH-4332 Stein, Switzerland. [6]Kyoto University, iCeMS, Yoshida Ushinomiya-cho, Sakyo-ku, Kyoto 606-8501, Japan. [7]Department of Economic Plants and Biotechnology, Yunnan Key Laboratory for Wild Plant Resources, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China. [8]International Maize and Wheat Improvement Center (CIMMYT), PO Box 1041-00621, Nairobi, Kenya. [9]Laboratory of Growth Regulators, Institute of Experimental Botany, The Czech Academy of Sciences and Faculty of Science, Palacký University, Šlechtitelů 27, 783 71 Olomouc, Czech Republic. [10]Plant genomics and transcriptomics group, Institute of Biosciences, Sao Paulo State University, 13506-900 Rio Claro, Brazil. [11]Section of Cell and Developmental Biology, University of California at San Diego; La Jolla, CA 92093, USA. [12]Seeds Research, Syngenta Crop Protection, LLC, Research Triangle Park, NC 27709, USA. [13]International Institute of Tropical Agriculture, PMB 5320 Oyo Road, Ibadan, Nigeria.
***Corresponding author. Email: h.j.bouwmeester@uva.nl (H.J.B.), l.dong2@uva.nl (L.D.)**
†Present address: Biozentrum, University of Basel, Spitalstrasse 41, 4056 Basel, Switzerland.

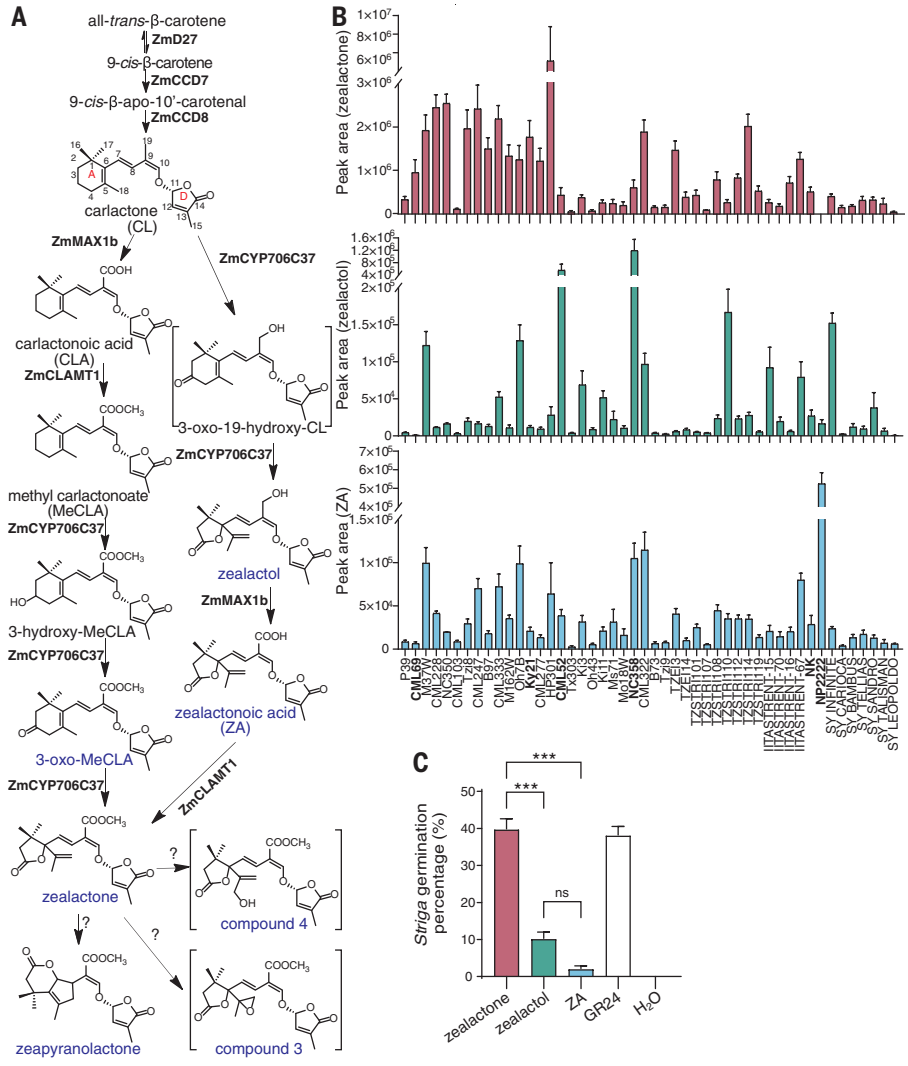

Fig. 1. Discovery of two strigolactones with low *Striga* germination–inducing activity from maize line screening. (**A**) Strigolactone (SL) biosynthetic pathway of maize. The enzymes identified in this study are shown in bold. SLs detected in maize root exudate are indicated in blue. Structures in square brackets are putative. (**B**) Detection of three maize SLs (zealactone, mass/charge ratio (*m/z*) 377 > 97; zealactol, *m/z* 331 > 97; ZA, *m/z* 363 > 249) in root exudate of a collection of maize lines. Names of lines selected for further analysis are indicated in bold. Data for the other four maize SLs are shown in fig. S3. (**C**) Induction of germination of *Striga* by zealactone, zealactol, and ZA (0.347 μM). GR24 (0.335 μM) and water were used as positive and negative control, respectively. Bars indicate means ± SEM. Ns, not significant ($P > 0.05$), ***$P < 0.001$, one-way ANOVA test followed by Tukey's multiple comparisons test comparing the mean of each column with the mean of every other column.

wild-type exudate (fig. S13A), showing that *ZmCCD8* is a key enzyme in maize SL biosynthesis (*17*, *31*, *32*). The transient expression of *ZmD27* (*GRMZM2G158175*), *ZmCCD7* (*GRMZM2G158657*), and *ZmCCD8* (*GRMZM2G446858*) together in *Nicotiana benthamiana* led to accumulation of CL (Figs. 1A and 2A, fig. S14A, and table S3), which is consistent with results from rice and tomato orthologs (*21*, *33*).

Identification of gene candidates for carlactone conversion

On the basis of the structures of the maize SLs identified thus far (Fig. 1A and fig. S2) (*24*–*26*), we postulated the involvement of a methyl transferase and several CYPs in the pathway downstream of CL. Several bioinformatic approaches were combined to select candidate genes for further functional characterization.

Mutual Rank (MR)–based global gene coexpression analysis (*34*, *35*) showed that of the three maize *MAX1* homologs, only *ZmMAX1b* tightly coexpressed with *ZmCCD8* (fig. S15), making it the strongest candidate for the next biosynthetic step. Analysis of root exudate from a *zmmax1a zmmax1c* double mutant (supplementary materials) showed wild-type levels of zealactone, thus excluding both homologs from being the biosynthetic genes we sought (fig. S13B). Earlier research also demonstrated that ZmMAX1b (GRMZM2G023952) converts CL to CLA more efficiently than does ZmMAX1a (GRMZM2G018612) or ZmMAX1c (GRMZM2G070508) (*18*). The amounts of CL in leaf extracts decreased after coinfiltration of *ZmMAX1b* with *ZmD27*, *ZmCCD7*, and *ZmCCD8* in *N. benthamiana*, (Fig. 2A), confirming that ZmMAX1b uses CL as a substrate (*18*). However, only traces of the expected product, CLA, were detected in this expression system (Fig. 2B and fig. S14B). To resolve this enigma, *N. benthamiana* extracts were analyzed with LC–quadrupole time-of-flight (QTOF)–MS. Prominent peaks of CLA-hexose and CLA-dihexose conjugates were detected in samples expressing the maize CL pathway genes together with *ZmMAX1b*. These conjugates were lacking in control samples and other gene combinations (Fig. 2C and table S4). Similar conjugation has been demonstrated for the transient production of other acidic compounds with *N. benthamiana* (*36*, *37*).

For selection of remaining candidate genes, we combined three approaches: (i) MR–based coexpression with *ZmCCD8* and *ZmMAX1b* as baits (fig. S15), (ii) coexpression modules in MaizeGGM2016 (*38*), and (iii) differential gene expression in a *zmccd8* mutant (Fig. 2D). For the latter, we assumed that SL pathway genes downstream of *CCD8* would be transcriptionally regulated in the *zmccd8* mutant (*33*). The *ZmCCD7*, *ZmCCD8*, and *ZmMAX1b* genes clustered together in MaizeGGM2016 module 091, suggesting that the 32 other genes in this module were candidates for the missing pathway genes (table S5). In the roots of *zmccd8* seedlings, 1301 genes were differentially expressed (DEGs) (less than or equal to twofold change, false discovery rate (FDR) < 0.05) compared with the B73 wild type (tables S5 and S6). These three approaches shared a seven-gene overlap (Fig. 2D and table S2) in which three [*GRMZM2G033126*, *GRMZM2G158342*, and *GRMZM2G023952* (*ZmMAX1b*)] formed a putative gene cluster on chromosome 3 (Fig. 2, D and E, and fig. S15) (*39*). Genes homologous to these also cluster in other *Poaceae* species (fig. S16), but the functional importance is unknown. So too is the identity of SLs produced by some of these species, such as switchgrass.

ZmCLAMT1 is a carlactonoic acid methyltransferase

Because SLs zealactone and zeapyranolactone are methylesters, their proposed precursor has been methyl carlactonoate (MeCLA) (*24*). Thus, we sought a methyltransferase gene that causes the formation of MeCLA from CLA. We bioinformatically identified a top candidate (*GRMZM2G033126*) (Fig. 2, D and E), which

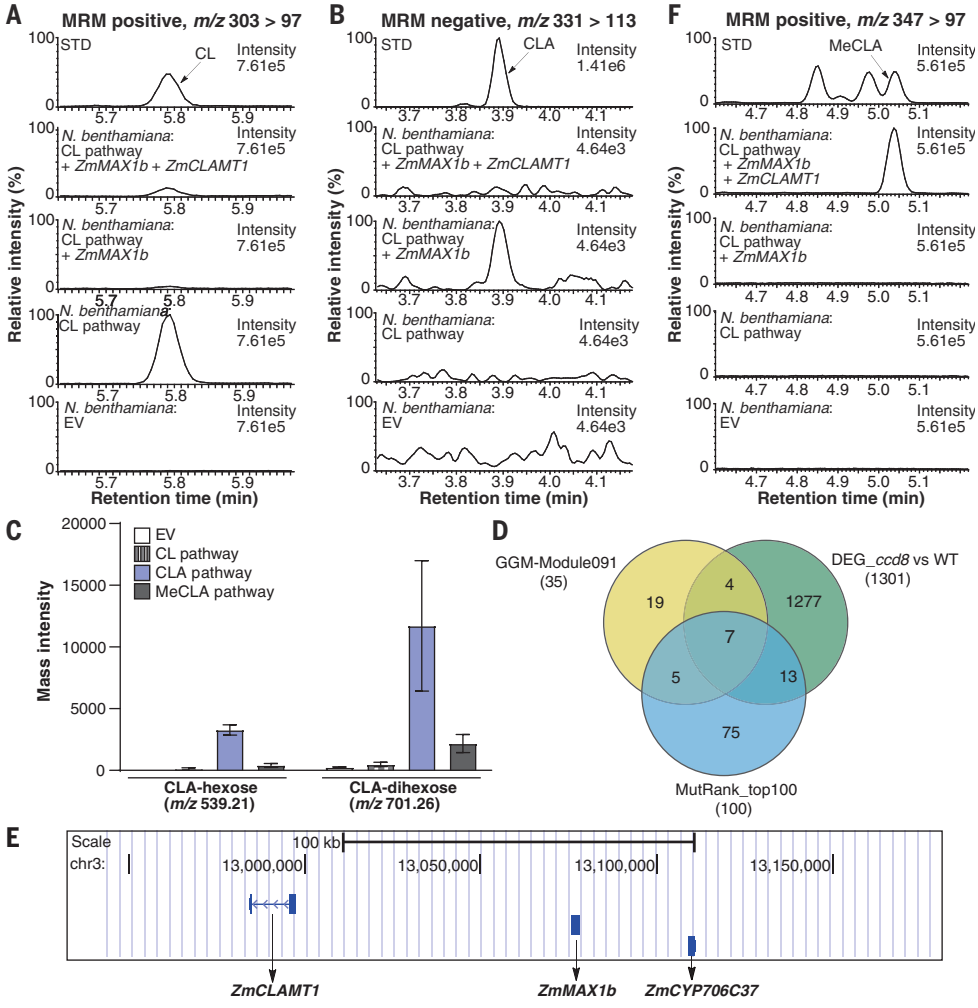

Fig. 2. Identification of gene candidates for maize strigolactone biosynthesis. (**A** and **B**) Representative MRM-LC/MS/MS chromatograms of carlactone (CL), $[M+H]^+$ *m/z* 303 > 97 (A), and carlactonoic acid (CLA), $[M-H]^-$ *m/z* 331 > 113 (B), in *N. benthamiana* leaf samples transiently expressing maize strigolactone (SL) precursor pathway genes. (**C**) Untargeted metabolomics to identify CLA conjugates in *N. benthamiana* leaf samples. *m/z* 539.21: CLA + hexose + formic acid − H_2O; *m/z* 701.26: CLA + 2 hexose + formic acid − H_2O (**D**) Venn diagram of candidate gene numbers from several analyses: module091 from maizeGGM, genes differentially expressed in *zmccd8* roots (compared with wild type), and the top 100 genes coexpressed with *ZmCCD8* and *ZmMAX1b* (*34*, *35*). (**E**) Putative SL biosynthetic gene cluster on chromosome 3 consisting of *ZmCLAMT1*, *ZmMAX1b*, and *ZmCYP706C37*, adapted from screenshot from UCSC Genome Browser on *Z. mays* (B73 RefGen_v3) Assembly (zm3) (http://genome.ucsc.edu) (*39*). (**F**) Representative chromatograms of methylcarlactonoic acid (MeCLA), $[M+H]^+$ *m/z* 347 > 97, in *N. benthamiana* leaf samples. STD, standard; EV, empty vector infiltrated control sample. CL pathway, maize carlactone biosynthetic pathway genes, *ZmD27*, *ZmCCD7*, and *ZmCCD8*. CLA pathway, CL pathway genes + *ZmMAX1b*. MeCLA pathway, CLA pathway genes + *ZmCLAMT1*. Bars indicate mean ± SEM.

successfully produced MeCLA in *N. benthamiana* when transiently expressed together with genes for the maize CLA pathway (Fig. 2F). We therefore identified *GRMZM2G033126* as a carlactonoic acid methyltransferase gene and named the enzyme ZmCLAMT1 (Fig. 1A). The maize gene is an ortholog of *At4g36470*, which was recently found to encode a carlactonoic acid methyltransferase CLAMT in *Arabidopsis* (*40*, *41*).

ZmCYP706C37 catalyzes formation of several maize strigolactones

The other candidate genes were coinfiltrated by different combinations of precursor-pathway genes. Coinfiltration of *ZmCYP706C37* (*GRMZM2G158342*) (*42*) by those encoding the MeCLA pathway decreased levels of MeCLA, indicating that this CYP can use MeCLA as a substrate (fig. S17A) and produce zealactone (Fig. 3A and fig. S2). To check for other possible biosynthetic pathways, we also coexpressed *ZmCYP706C37* with genes encoding the CL pathway enzymes. This combination resulted in production of zealactol (Fig. 4A and fig. S17B). Formation of both zealactone and zealactol involves complex rearrangement of the SL A ring and, for zealactol, a hydroxylation at C19 as well. To exclude the possibility of endogenous enzymes from *N. benthamiana* contributing to these complex conversions, we expressed *ZmCYP706C37* in yeast, isolated its microsomes, and analyzed product formation with different substrates (Figs. 3B and 4B). This approach confirmed that ZmCYP706C37 can convert MeCLA to zealactone and CL to zealactol (Fig. 1A).

To form zealactone from MeCLA, ZmCYP706C37 must catalyze several consecutive oxidative reactions with 3-hydroxy-MeCLA and 3-oxo-MeCLA as putative intermediates (Figs. 1A and 3C). The latter two compounds were previously synthesized as intermediates in the total synthesis of heliolactone (*43*). We used them here as substrates in our *ZmCYP706C37*-expressing yeast-microsome assay, and both were successfully converted to zealactone (Fig. 3D). We developed an MRM method for detection of these compounds (fig. S2) and identified them as intermediate products in the conversion of MeCLA to zealactone (fig. S18). Moreover, analysis of maize root exudate revealed that 3-oxo-MeCLA is also a natural maize SL previously referred to as compound 6 (fig. S19 and Fig. 1A) (*24*). These results demonstrate that a single enzyme, ZmCYP706C37, can catalyze the many oxidative steps necessary for the conversion of MeCLA to zealactone that were previously hypothesized to require several enzymes (Figs. 1A and 3C) (*24*).

For additional insight into the parallel biosynthetic pathway of CL to zealactol, we further analyzed samples from *N. benthamiana* and yeast microsome assays with untargeted metabolomics and MRM-LC-MS/MS. This process revealed another putative intermediate, 3-oxo-19-hydroxy-CL (compound 7) (Fig. 1A and figs. S2 and S20 and table S7). LC-QTOF-MS analysis showed that the accurate mass of compound 7 is consistent with its putative structure (fig. S20). On the basis of these data, we included compound 7 as an intermediate in the postulated steps required to convert CL to zealactol (Fig. 4C and fig. S21).

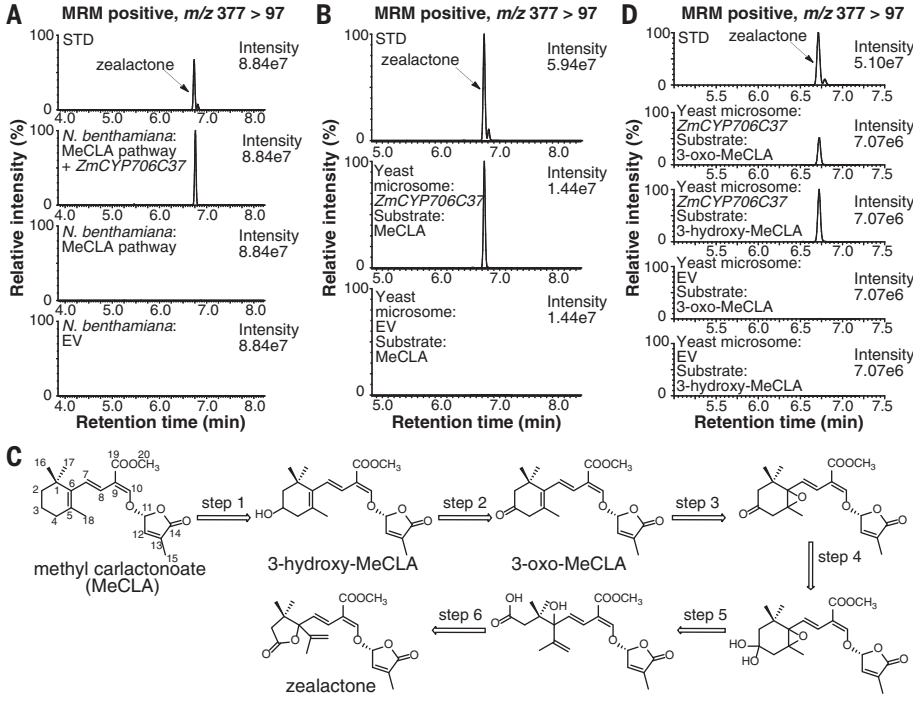

Fig. 3. Zealactone biosynthesis. (**A**) Representative MRM–LC/MS/MS chromatograms of zealactone, $[M+H]^+m/z$ 377 > 97, in *N. benthamiana* leaf samples. (**B** and **D**) Representative MRM–LC/MS/MS chromatograms of zealactone from in vitro assays with yeast microsomes expressing *ZmCYP706C37* or empty vector (EV) with methyl carlactonoate (MeCLA), 3-hydroxy-MeCLA, or 3-oxo-MeCLA as substrate. (**C**) Proposed enzymatic conversion of methyl carlactonoate (MeCLA) to zealactone.

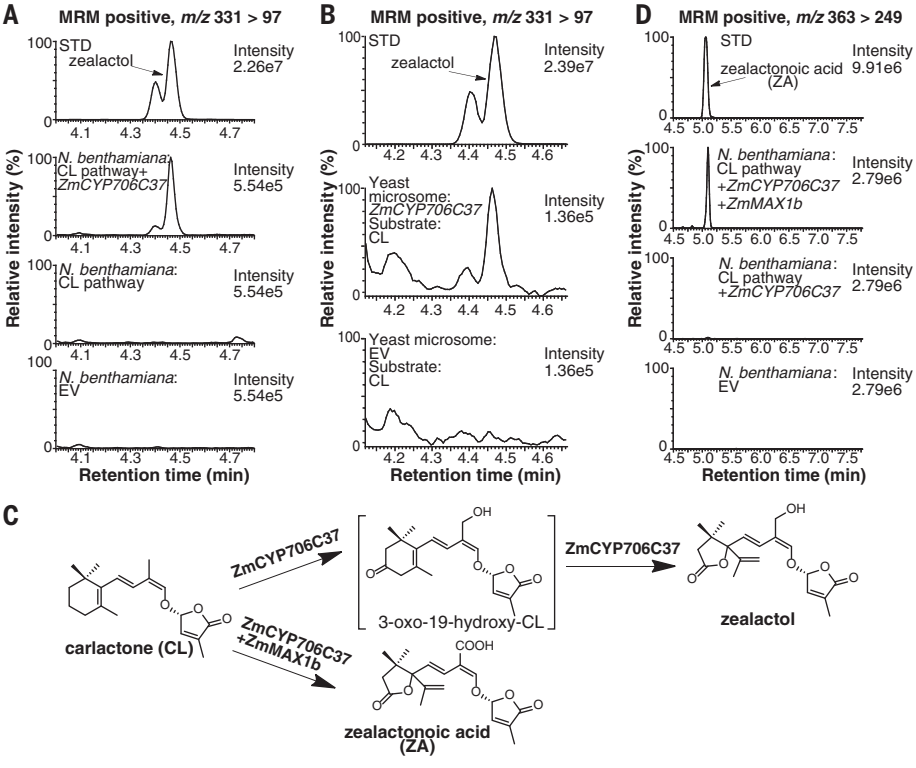

Fig. 4. Zealactol and zealactonoic acid biosynthesis. (**A**) Representative MRM–LC/MS/MS chromatograms of zealactol, $[M+H\text{-}H_2O]^+m/z$ 331 > 97, in *N. benthamiana* leaf samples. (**B**) Representative MRM–LC/MS/MS chromatograms of zealactol from in vitro assays with yeast microsomes expressing *ZmCYP706C37* or empty vector (EV) with carlactone (CL) as substrate. (**C**) Reactions from CL to zealactol and ZA catalyzed by ZmCYP706C37 and ZmMAX1b. Structure in square brackets is putative. (**D**) Representative MRM–LC/MS/MS chromatograms of ZA, $[M+H]^+m/z$ 363 > 249, in *N. benthamiana* leaf samples. STD, standard; EV, empty vector control. CL pathway, maize carlactone biosynthetic pathway genes, *ZmD27*, *ZmCCD7*, and *ZmCCD8*.

Moreover, agroinfiltration of the CL pathway genes with *ZmCYP706C37* and *ZmMAX1b* resulted in production of ZA, a result also confirmed with LC-QTOF-MS (Fig. 4, C and D, and fig. S22).

Last, analysis of root exudate from a *zmcyp706c37* mutant [EMS4-045ad8, stop-codon gained (fig. S23A)] showed no detectable levels of zealactol, ZA, zealactone, or three other SLs derived from the latter (fig. S23B) (*44*). Although 3-oxo-MeCLA was detectable in the mutant exudate, it was present at a much lower level than in that of the wild type. Instead, CLA and MeCLA accumulated in the mutant exudate, whereas they are absent in the wild type exudate (fig. S23, C and D). Together, these data support our functional characterization of ZmCYP706C37.

Biosynthetic control of the maize strigolactone blend

To determine how the different maize SLs are biosynthetically related, we applied 3-hydroxy-MeCLA, 3-oxo-MeCLA, and zealactol to seedlings

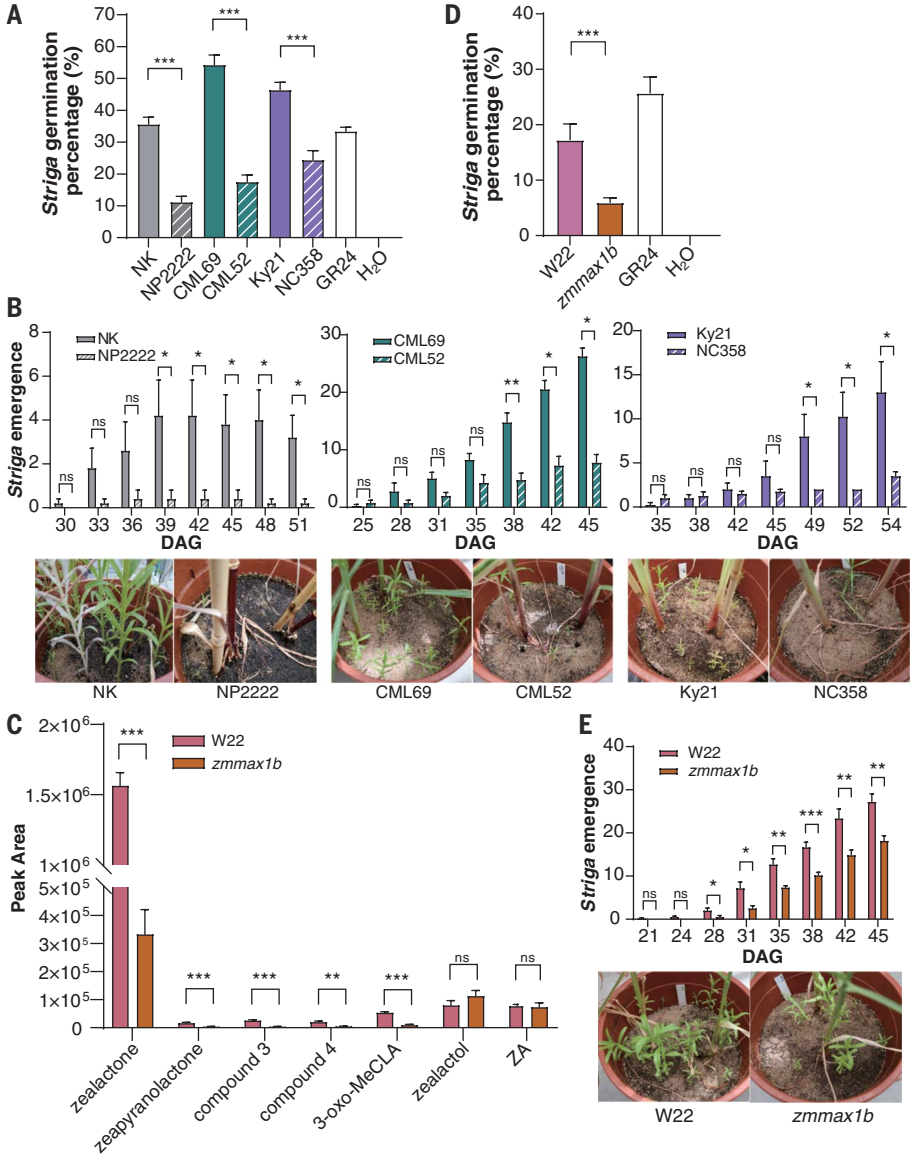

Fig. 5. Changes in the maize strigolactone blend result in changes in *Striga* resistance. (**A** and **D**) Induction of *Striga* germination by root exudates of selected maize lines. GR24 (0.335 μM) and water were used as positive and negative control, respectively. (**B** and **E**) *Striga* infection of selected maize lines. Emerged *Striga* numbers were recorded; representative photos highlight the differences. DAG, days after germination of maize. (**C**) SL levels in the root exudate of *zmmax1b* and its wild type, W22. Bars indicate means ± SEM, ns = not significant ($P > 0.05$), *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-tailed, unpaired *t* test.

of another commercial line, NK Falkone, which were treated with fluridone, an inhibitor of SL biosynthesis (*45*). Each of these three compounds complemented zealactone production (fig. S24A), confirming that they can serve as biosynthetic precursors for zealactone. Combined transient expression of *ZmMAX1b* and *ZmCLAMT1* in *N. benthamiana* leaves and subsequent infiltration of zealactol also showed that the latter can be converted to zealactone by ZmMAX1b together with ZmCLAMT1 (Fig. 1A and fig. S25). Application of zealactone to fluridone-treated plants led to the formation of zeapyranolactone and two other maize SLs, designated compounds 3 and 4, suggesting that zealactone is their precursor (Fig. 1A and fig. S24, B to D) (*24*).

Next, we sought mechanisms underlying the distinctive maize SL profile of NP2222 (fig. S26). This line produces zealactone in fluridone-treated seedlings, as does NK Falkone, but only from MeCLA and 3-oxo-MeCLA, not from zealactol (figs. S24A and S26A), suggesting inactivity of MAX1b and/or CLAMT1. As previously noted, ZA accumulated in the root exudate of NP2222 (Fig. 1B and fig. S26D), indicating dysfunction of CLAMT1. Zealactol added to either NK Falkone or NP2222 was converted to ZA, showing that ZmMAX1b is active in NP2222 (fig. S26, B and C). Inspection of the *CLAMT1* sequence in a proprietary NP2222 genome database revealed a large insertion in the second exon of this gene, and reverse transcriptase polymerase chain reaction (RT-PCR) showed that regions flanking the insertion were not transcribed (fig. S26E). These collective data indicate disfunction of CLAMT1 in NP2222.

To analyze biological consequences of the different SL profiles, several maize lines were selected for *Striga* germination and infection assays. The NP2222 root exudate induced much lower germination than that of NK Falkone. Results were consistent with their respective SL profiles and differences in germination-inducing activity of the individual SLs (Figs. 1C and 5A and fig. S26D). CML52 and NC358, both with high proportions of zealactol and ZA, induced significantly less *Striga* germination than did CML69 and Ky21, which produced mostly zealactone despite similar total SL peak areas (Figs. 1C and 5A, and fig. S27, A and B). These differences were also reflected in a *Striga* infection assay with a containerized system, in which *Striga* emergence was less for low-zealactone genotypes (Fig. 5B). In addition to their SL blend, these lines may have other genetic differences that could affect these results. However, we also analyzed a gene-suppression mutant of *ZmMAX1b* (transposon insertion in a W22 background) (fig S28, A and B). This mutant exuded significantly less izealactone and zealactone-derived SLs, whereas the level of zealactol was higher than in the W22 control (Fig. 5C). The *zmmax1b* mutant also induced less *Striga* germination and emergence (Fig. 5E). Results confirm that a change in activity of specific SL biosynthetic enzymes in maize can change the SL composition and confer *Striga* resistance. Although the underlying mechanisms are completely different, these findings resemble those of *lgs* sorghum (*16*) and present a promising prospect for *Striga* resistance breeding in maize. The *zmmax1b* mutant did not exhibit a branching phenotype, in contrast to *zmccd8* (fig. S28C). Also, *zmcyp706c37*, which is located parallel to or downstream of ZmMAX1b, did not display an obvious branching phenotype either. This all suggests that the downstream SLs are not nor precursors of the branching inhibiting hormone and are therefore safe breeding targets that will not result in unwanted pleiotropic effects.

Conclusions

We have shown that two parallel SL biosynthetic pathways operate in maize and that both pathways produce the major maize SL, zealactone. Changes in flux through these pathways can alter the maize SL profile by shifting the balance between zealactone and zealactol

plus ZA. Zealactol and ZA induce much less *Striga* germination, thus imparting a strong reduction in *Striga* infection to genotypes that exude more zealactol and ZA than zealactone. Future research should investigate whether these changes in the SL blend affect colonization by AM fungi, which was not observed for *lgs* sorghum (*16*). Our results offer a perspective for breeding *Striga* resistance through modification of the SL blend in maize and thus potentially reducing the devastating effects of this parasitic weed in Africa.

REFERENCES AND NOTES

1. T. Wheeler, J. von Braun, *Science* **341**, 508–513 (2013).
2. B. Badu-Apraku, F. M.A.B., *Advances in Genetic Enhancement of Early and Extra-Early Maize for Sub-Saharan Africa* (Springer Cham, 2017).
3. I. Dörr, *Ann. Bot. (Lond.)* **79**, 463–472 (1997).
4. C. E. Cook, L. P. Whichard, B. Turner, M. E. Wall, G. H. Egley, *Science* **154**, 1189–1190 (1966).
5. A. Besserer *et al*., *PLOS Biol.* **4**, e226 (2006).
6. V. Gomez-Roldan *et al*., *Nature* **455**, 189–194 (2008).
7. S. Al-Babili, H. J. Bouwmeester, *Annu. Rev. Plant Biol.* **66**, 161–186 (2015).
8. M. Umehara *et al*., *Nature* **455**, 195–200 (2008).
9. K. Akiyama, K. Matsuzaki, H. Hayashi, *Nature* **435**, 824–827 (2005).
10. H. Bouwmeester, C. Li, B. Thiombiano, M. Rahimi, L. Dong, *Plant Physiol.* **185**, 1292–1308 (2021).
11. K. Yoneyama *et al*., *J. Exp. Bot.* **69**, 2231–2239 (2018).
12. K. Mashiguchi, Y. Seto, S. Yamaguchi, *Plant J.* **105**, 335–350 (2021).
13. K. Akiyama, S. Ogasawara, S. Ito, H. Hayashi, *Plant Cell Physiol.* **51**, 1104–1117 (2010).
14. N. Mori, K. Nishiuma, T. Sugiyama, H. Hayashi, K. Akiyama, *Phytochemistry* **130**, 90–98 (2016).
15. H. I. Kim *et al*., *J. Pestic. Sci.* **35**, 344–347 (2010).
16. D. Gobena *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4471–4476 (2017).
17. A. Alder *et al*., *Science* **335**, 1348–1351 (2012).
18. K. Yoneyama *et al*., *New Phytol.* **218**, 1522–1533 (2018).
19. S. Abe *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **111**, 18084–18089 (2014).
20. C. Cardoso *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2379–2384 (2014).
21. Y. Zhang *et al*., *Nat. Chem. Biol.* **10**, 1028–1033 (2014).
22. T. Wakabayashi *et al*., *Sci. Adv.* **5**, eaax9067 (2019).
23. T. Wakabayashi *et al*., *Planta* **251**, 97 (2020).
24. T. V. Charnikhova *et al*., *Phytochemistry* **137**, 123–131 (2017).
25. T. V. Charnikhova *et al*., *Phytochem. Lett.* **24**, 172–178 (2018).
26. X. Xie *et al*., *J. Pestic. Sci.* **42**, 58–61 (2017).
27. M. Yoshimura *et al*., *Helv. Chim. Acta* **103**, e2000017 (2020).
28. M. C. Dieckmann, P.-Y. Dakas, A. De Mesmaeker, *J. Org. Chem.* **83**, 125–135 (2018).
29. T. Kumagai *et al*., *Heterocycles* **36**, 1729–1734 (1993).
30. J. C. Guan *et al*., *Plant Physiol.* **160**, 1303–1317 (2012).
31. K. C. Snowden *et al*., *Plant Cell* **17**, 746–759 (2005).
32. W. Kohlen *et al*., *New Phytol.* **196**, 535–547 (2012).
33. Y. Zhang *et al*., *New Phytol.* **219**, 297–309 (2018).
34. E. Poretsky, A. Huffaker, *PeerJ* **8**, e10264 (2020).
35. S. Stelpflug *et al*, *The Plant Genome* **9**, plantgenome2015.04.0025 (2016).
36. L. Dong *et al*., *Metab. Eng.* **20**, 198–211 (2013).
37. X. Xu *et al*., *J. Exp. Bot.* **72**, 5462–5477 (2021).
38. S. Ma, Z. Ding, P. Li, *BMC Plant Biol.* **17**, 131 (2017).
39. W. J. Kent *et al*., *Genome Res.* **12**, 996–1006 (2002).
40. T. Wakabayashi *et al*., *Planta* **254**, 88 (2021).
41. K. Mashiguchi *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2111565119 (2022).
42. Y. Li, K. Wei, *BMC Plant Biol.* **20**, 93 (2020).
43. M. Yoshimura *et al*., *Helv. Chim. Acta* **102**, e1900211 (2019).
44. X. Lu *et al*., *Mol. Plant* **11**, 496–504 (2018).
45. J. A. López-Ráez *et al*., *New Phytol.* **178**, 863–874 (2008).

ACKNOWLEDGMENTS

We acknowledge S. Al Babili from King Abdullah University of Science and Technology and D. Werck-Reichhart from the University of Strasbourg for helpful discussions, as well as L. Hagmann from Syngenta for his support in NMR analyses and interpretation. **Funding:** This work was funded by the China Scholarship Council (CSC) PhD scholarship 201706300041 (C.L.), the European Research Council (ERC) Advanced grant CHEMCOMRHIZO 670211 (H.J.B.), the Dutch Research Council (NWO/OCW) Gravitation program Harnessing the second genome of plants (MiCRop) 024.004.014 (H.J.B.), the Marie Curie fellowship NEMHATCH 793795 (L.D.). K.E.K and J.G. acknowledge funding from the US National Science Foundation (NSF) Plant Genome Research Program (PGRP) (1421100 and 1748105). **Author contributions:** C.L., L.D., and H.J.B. conceived and designed the project. C.L. discovered and characterized the candidate genes, grew the plants, collected and analyzed the root exudate, cloned the genes, performed agroinfiltration, yeast microsome assay and plant compound treatment assays, and coordinated the project; K.F., T.V.C. and A.C. developed LC-MS methods and helped with SL analysis; T.V.C., J.D., and A.D.J.V.D. helped to establish the biosynthesis mechanisms; J.G. and K.E.K. developed and provided maize seeds (NAM, *zmccd8*, *zmmax1azmmax1c*, and *zmmax1b*) and analyzed RNA-seq and related data. B.T. and L.D. supported the metabolomics analysis; M.Y., K.G., A.D.M. synthesized zealactol and provided zealactone, 3-hydroxy-MeCLA, and 3-oxo-MeCLA; P.Q., R.H., and A.D.M. synthesized zealactonoic acid; J.L., Y.B.S., J.Q., and J.W. grew the *zmcyp706c37* EMS mutants, performed genotyping, selfing, and root exudate collection; H.D.G. collected and prepared the maps of maize and *Striga* occurrence. Y.W. helped with the agroinfiltration and yeast microsome assays; C.L., A.W., and B.T. performed the *Striga* germination and infection bioassays; S.M.C.d.L. and M.H.M. carried out the gene cluster analysis; Y.D. and E.A.S. provided support on coexpression analysis; D.K., K.H. and C.S. provided all commercial maize seeds from Syngenta and coordinated the collaboration with Syngenta. A.M. provided African inbred maize lines. C.L., L.D., and H.J.B. wrote the manuscript, with contributions from other authors. **Competing interests:** M.H.M. is a consultant to Corteva Agriscience, but that company was not involved in this work. All the other authors declare that they have no competing interests. **Data and materials availability:** The maize mutants *zmccd8* and *zmmax1azmmax1c* were obtained via a material transfer agreement (MTA) with the University of Florida Board of Trustees. The RNA-seq data of *zmccd8* and B73 root tissues are available in the NCBI database (BioProject PRJNA757767) under accession numbers of SRR15613590, SRR15613591, SRR15613599, SRR15613593, SRR15613594, and SRR15613595. All the other data are presented in the main text and in the Supplementary Materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abq4775
Materials and Methods
Figs. S1 to S28
Tables S1 to S8
References (*46*–*70*)

View/request a protocol for this paper from *Bio-protocol*.

Submitted 12 April 2022; accepted 30 November 2022
10.1126/science.abq4775

Maize resistance to witchweed through changes in strigolactone biosynthesis

C. Li, L. Dong, J. Durairaj, J.-C. Guan, M. Yoshimura, P. Quinodoz, R. Horber, K. Gaus, J. Li, Y. B. Setotaw, J. Qi, H. De Groote, Y. Wang, B. Thiombiano, K. Floková, A. Walmsley, T. V. Charnikhova, A. Chojnacka, S. Correia de Lemos, Y. Ding, D. Skibbe, K. Hermann, C. Screpanti, A. De Mesmaeker, E. A. Schmelz, A. Menkir, M. Medema, A. D. J. Van Dijk, J. Wu, K. E. Koch, and H. J. Bouwmeester

Science **379** (6627), . DOI: 10.1126/science.abq4775

Diversity reveals infection resistance

Parasitic witchweed (*Striga*) reduces the yield of maize grown in infected fields. Strigolactones from maize roots encourage *Striga* germination. Li *et al*. analyzed the natural variation in types of strigolactones exuded from maize roots. Maize genotypes that produced mainly zealactol suffered less *Striga* infection than those that produced mainly zealactone. A single cytochrome P450 catalyzes several of the oxidative steps in strigolactone biosynthesis, including conversion of precursors to either zealactol or zealactone. —PJH

The genome and population genomics of allopolyploid *Coffea arabica* reveal the diversification history of modern coffee cultivars

Jarkko Salojärvi[1,2,3,*], Aditi Rambani[4,†], Zhe Yu[5,†], Romain Guyot[6,7,†], Susan Strickler[4,†], Maud Lepelley[8], Cui Wang[2], Sitaram Rajaraman[2], Pasi Rastas[9], Chunfang Zheng[5], Daniella Santos Muñoz[5], João Meidanis[10], Alexandre Rossi Paschoal[11], Yves Bawin[12], Trevor Krabbenhoft[13], Zhen Qin Wang[13], Steven Fleck[13], Rudy Aussel[8,14], Laurence Bellanger[8], Aline Charpagne[15], Coralie Fournier[15], Mohamed Kassam[15], Gregory Lefebvre[15], Sylviane Métairon[15], Déborah Moine[15], Michel Rigoreau[8], Jens Stolte[15], Perla Hamon[6], Emmanuel Couturon[6], Christine Tranchant-Dubreuil[6], Minakshi Mukherjee[13], Tianying Lan[13], Jan Engelhardt[16], Peter Stadler[17], Samara Mireza Correia De Lemos[18], Suzana Ivamoto Suzuki[19], Ucu Sumirat[20], Wai Ching Man[21], Nicolas Dauchot[22], Simon Orozco-Arias[7], Andrea Garavito[23], Catherine Kiwuka[24], Pascal Musoli[24], Anne Nalukenge[24], Erwan Guichoux[25], Havinga Reinout[26], Martin Smit[26], Lorenzo Carretero-Paulet[27], Oliveiro Guerreiro Filho[28], Masako Toma Braghini[28], Lilian Padilha[29], Gustavo Hiroshi Sera[30], Tom Ruttink[12,33], Robert Henry[31], Pierre Marraccini[32], Yves Van de Peer[33,34,35,40], Alan Andrade[36], Douglas Domingues[18], Giovanni Giuliano[37], Lukas Mueller[4], Luiz Filipe Pereira[38], Stephane Plaisance[39], Valerie Poncet[6], Stephane Rombauts[33,40], David Sankoff[5], Victor A. Albert[13,*], Dominique Crouzillat[8,*], Alexandre de Kochko[6,*], Patrick Descombes[15,*]

[1] School of Biological Sciences, Nanyang Technological University, Singapore 637551, Singapore
[2] Organismal and Evolutionary Biology Research Programme, University of Helsinki, 00014 Helsinki, Finland
[3] Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore 637551, Singapore
[4] Boyce Thompson Institute, University of Cornell, Ithaca NY 14853, US
[5] Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada K1N 6N5
[6] Institut de Recherche pour le Développement (IRD), Université de Montpellier, 34394 Montpellier, France
[7] Department of Electronics and Automation, Universidad Autónoma de Manizales, Manizales 170002, Colombia
[8] Société des Produits Nestlé SA, Nestlé Research, 37097 Tours CEDEX 2, France
[9] Institute of Biotechnology, University of Helsinki, 00014 Helsinki, Finland
[10] Institute of Computing, University of Campinas, 13083-852 Campinas, Sao Paulo, Brazil
[11] Department of Computer Science, The Federal University of Technology – Paraná (UTFPR), 86300-000, Cornélio Procópio, Brazil
[12] Plant Sciences Unit, Flanders research Institute for Agriculture, Fisheries and Food (ILVO), 9090 Melle, Belgium
[13] Department of Biological Sciences, University at Buffalo, New York, USA
[14] Centre d'Immunologie de Marseille-Luminy, Aix Marseille Université, France
[15] Société des Produits Nestlé SA, Nestlé Research, 1015 Lausanne, Switzerland
[16] Department of Computer Science, University of Leipzig, 04107 Leipzig, Germany
[17] Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, 04107 Leipzig, Germany
[18] Group of Genomics and Transcriptomes in Plants, São Paulo State University, UNESP, Rio Claro, SP, Brazil, 13506-900
[19] Centro de Ciências Agrárias, Universidade Estadual de Londrina, 86057-970 Londrina, Brazil
[20] Indonesian Coffee and Cocoa Research Institute (ICCRI), Jember 68118 Indonesia
[21] Texas A&M University, Urbana, Illinois 61801, USA
[22] Research Unit in Plant Cellular and Molecular Biology, University of Namur, Namur 5000, Belgium
[23] Departamento de Ciencias biológicas, Facultad de Ciencias Exactas y Naturales, Universidad de Caldas, Manizales, Colombia

[24] National Agricultural Research Organization (NARO), Uganda
[25] Biodiversité Gènes & Communautés, INRA, 33610 CESTAS, France
[26] Hortus Botanicus Amsterdam, 1018 DD Amsterdam, Netherlands
[27] Departamento de Biología y Geología, Universidad de Almería, Almería, Spain
[28] Instituto Agronômico (IAC) Centro de Café 'Alcides Carvalho', Fazenda Santa Elisa, Caixa Postal 28, Campinas (SP), Brasil 13012 – 970
[29] Embrapa Café / Instituto Agronômico (IAC) Centro de Café 'Alcides Carvalho', Fazenda Santa Elisa, Caixa Postal 28, Campinas (SP), Brasil 13012 - 970
[30] Instituto de Desenvolvimento Rural do Paraná- IAPAR, 86047-902 Londrina, Brasil
[31] Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Brisbane 4072, Australia
[32] CIRAD - UMR DIADE (IRD-CIRAD-Université de Montpellier) BP 64501, F-34394 Montpellier Cedex 5, France
[33] Department of Plant Biotechnology and Bioinformatics, Ghent University
[34] Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa
[35] College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing, 210095, China
[36] Embrapa Café/Inovacafé Laboratory of Molecular Genetics Campus da UFLA-MG, 37200-900 Lavras-MG, Brazil
[37] Italian National Agency for New technologies, Energy and Sustainable Economic Development (ENEA), Casaccia Res. Ctr., 00123 Roma, Italy
[38] Embrapa Café / Lab. Biotecnologia, Área de Melhoramento Genético, Londrina – PR, Brasil - 86047-902
[39] VIB Nucleomics Core, B-3000 Leuven, Belgium
[40] Center for Plant Systems Biology, VIB, 9052 Ghent, Belgium

* Correspondence should be addressed to: Patrick Descombes (patrick.descombes@rd.nestle.com), Alexandre de Kochko (alexandre.dekochko@ird.fr), Jarkko Salojärvi (jarkko@ntu.edu.sg), Victor A. Albert (vaalbert@buffalo.edu), or Dominique Crouzillat (dcrouzillat@gmail.com).

[†]These authors contributed equally to this work

ABSTRACT

Coffea arabica, an allotetraploid hybrid of *C. eugenioides* and *C. canephora,* is the source of approximately 60% of coffee products worldwide. Cultivated accessions have undergone several population bottlenecks resulting in low genetic diversity. We present chromosome-level assemblies of a di-haploid *C. arabica* accession and modern representatives of its diploid progenitors, *C. eugenioides* and *C. canephora*. The three species exhibit largely conserved genome structures between diploid parents and descendant subgenomes, which show a mosaic pattern of dominance, similar to other polyploid crop species. Resequencing of 39 wild and cultivated accessions suggests a founding polyploidy event ~610,000 years ago, followed by several subsequent bottlenecks, including a population split ~30.5 kya and a period of migration between Arabica populations until ~8.9 kya. Analysis of lines historically introgressed with *C. canephora* highlights loci that may contribute to their superior pathogen resistance and lay the groundwork for future genomics-based breeding of *C. arabica*.

INTRODUCTION

Polyploidy is a powerful evolutionary force that has shaped genome evolution across all eukaryote lineages[1], possibly because increased gene content and resultant genome plasticity may be advantageous in times of global change[2]. Such whole genome duplications (WGDs) are

particularly characteristic of plants, which depending on the lineage, may have experienced several ancestral whole genome multiplication events over their history[3]. Polyploids are important for agriculture and plant breeding[4], with a great proportion of crop species being polyploid[5-11]. Our understanding of genome evolution following a WGD is still incomplete, but it appears to depend on the compatibility of the parental species; i.e., whether the event induces genomic shock[12,13], dominance by one of the subgenomes (either in the form of biased gene loss[14], expression dominance[15] or homoeologous exchange[9,16]), or if the polyploid gradually adapts to a new ploidy level[17]. Regardless, the most common fate of polyploids appears to be diploidization over the subsequent millions of years[18].

With an estimated production of 10 million metric tons per year, coffee is one of the most traded commodities in the world, and of extraordinary economic importance to countries in tropical regions of South America, South-East Asia and Africa, where it is widely cultivated. The most broadly appreciated coffee is produced from the allotetraploid species *Coffea arabica*, especially from cultivars belonging to the Bourbon or Typica lineages and their hybrids[19]. *C. arabica* (2n = 4x = 44 chromosomes, with a genome composition of CCEE) resulted from a natural hybridization event between the ancestors of present-day *C. canephora* (Robusta coffee, subgenome C) and *C. eugenioides* (subgenome E; each with 2n = 2x = 22). An accurate timing as well as precise location for the founding WGD have been difficult to pinpoint, but the event has been estimated to have taken place between 10,000 to one million years ago[20-23], with the Robusta-derived subgenome of *C. arabica* shown to be closely related to *C. canephora* accessions from northern Uganda[24].

Arabica cultivation was initiated in 15th -16th century Yemen, which held the world monopoly on coffee production at the time (**Fig. S1**). This domination was broken around 1600 AD by an Indian monk, Baba Budan, who smuggled the fabled "seven seeds" out of Yemen[25], thus establishing Indian *C. arabica* cultivar lineages. In the 17th century, the Dutch obtained Arabica plants either from Sri Lanka or from India - these became the founding population of the contemporary Typica group – from which Arabica cultivation in Java and Southeast Asia was established. One plant from the same stock was shipped to Amsterdam in 1706, where one of its descendants was later donated to Louis XIV of France. Seeds from this cultivar were subsequently used to establish Arabica cultivation in the Caribbean, starting from Martinique in 1723. The French also began cultivating Arabica on the island of Bourbon (presently Réunion) from seeds obtained from Mocha, Yemen[26]. Only one plant from this population survived by 1720, a single parent of the contemporary Bourbon group. Most important Arabica cultivars today are thought to be descendants of these Typica or Bourbon lineages, except for a few wild ecotypes whose origin can be traced back to natural forests in Ethiopia. Due to Arabica's recent allotetraploid origin and strong bottlenecks that occurred during its early cultivation and global spread, cultivated *C. arabica* harbors a particularly low genetic diversity with an effective population size (N_e) estimated to range between 10,000-50,000 individuals[20].

Due to the low genetic diversity resulting from these historical single-plant bottlenecks, modern Arabica cultivars are susceptible to many plant pests and diseases. As a result, the classic Bourbon-Typica lineages can only be cultivated successfully in a few regions around the world, where climate is permissive and pathogens less abundant. In 1927, a spontaneous hybrid between *C. canephora* x *C. arabica* was identified on the island of Timor[27], showing strong resistance to coffee leaf rust (*Hemileia vastatrix*). Hence, many contemporary cultivars carry introgressions from *C. canephora* via Timor hybrid-based breeding, aimed at boosting pathogen resistance. However, this introgressive breeding strategy has also produced unwanted side-effects, such as decreased quality of the coffee beverage[28] or loss of pathogen resistance in further backcrosses to Arabica cultivars[29,30]. Modern genomic tools and a detailed understanding of the origin and breeding history of contemporary varieties are therefore vital to enhance the development of new cultivars better adapted to climate change and agricultural practices[31,32] , and to permit implementation of effective bioengineering strategies[33].

Here, we present chromosome-level assemblies of *C. arabica* (CA) and representatives of its progenitor species, *C. canephora* (CC) and *C. eugenioides* (CE). Whole-genome resequencing data of 39 wild and cultivated Arabica accessions facilitated in-depth analysis of *C. arabica* breeding history and dissemination routes, as well as the identification of candidate genomic regions associated with pathogen resistance.

Results and Discussion

Chromosome-level assemblies and annotations of *Coffea arabica* and its diploid progenitors

For producing chromosome-level assemblies, we chose the di-haploid line ET-39[34] for the *C. arabica* genome, a previously sequenced doubled haploid accession[35] for *C. canephora*, and the wild accession Bu-A for *C. eugenioides.* Contig-level assemblies were produced using Pacific Biosciences long reads followed by polishing with Illumina reads, (Online methods and **Supplementary sections 2.1-2.2**). The assemblies spanned 672 Mb (*C. canephora*), 645 Mb (*C. eugenioides*) and 1,088 Mb (*C. arabica*), respectively. The *C. canephora* assembly and this first version for *C. arabica* were scaffolded with Dovetail chromosome conformation capture (Hi-C) technology into 11 and 22 pseudo-chromosomes, respectively, spanning 82.7% and 62.5% respectively of the projected genome sizes (**Table 1**).

One possible reason for the relatively low proportion of scaffolds assigned to CA pseudomolecules in this first CA assembly was the high sequence similarity between the two subgenomes, which resulted in partially collapsed scaffolds for which assigning short contigs to either of the subgenomes was difficult. To improve our *C. arabica* assembly quality, we generated a second assembly (hereinafter called *C. arabica* HiFi), using PacBio HiFi technology followed by Hi-C scaffolding (Online methods and **Supplementary section 2.3**). After gap filling, the final 1,198 Mb assembly consisted of 132 scaffolds with an N50 of 53.7 Mb, of which 1,192 Mb (93.1% of the predicted genome size based on cytological evidence[36]) was anchored to 22 pseudochromosomes (**Table 1**). Gene space completeness, assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO)[37] was >96% for all assemblies. Importantly, 93.2% of the BUSCO genes were duplicated in the HiFi assembly (**Table 1**), indicating that homoeologous regions had been separately assembled and that most of the genes that duplicated during the allopolyploidization event still retain their duplicate state.

The CC and CE genomes contained, respectively, 67.5% and 59.7% Transposable Elements (TEs) (**Supplementary section 3.2**). These proportions were 63.1% and 63.8% for *C. arabica* subgenomes CC (subCC) and EE (subEE), respectively, possibly indicating TE transfer between the two subgenomes, for example via homoeologous exchange. The vast majority of TEs were members of the Long Terminal Repeat (LTR) retrotransposon superfamilies Gypsy and Copia, with the former accounting for most of the difference between CC and CE. CC contained considerably more recent LTR TE insertion elements than CE, suggesting that those elements had remained active after the divergence of the species. Again, the CC and EE subgenomes of *C. arabica* showed greater similarity to each other in recent LTR TE insertion content than the two progenitor genomes. No major evidence was found for LTR TE mobilization following the Arabica allopolyploidy event, in contrast to what has been observed in tobacco[38], with the results instead following the pattern observed in *Brassica* synthetic allotetraploids[39].

High quality gene annotation was done using both short- and long-read RNA-Seq data (**Supplementary Section 3.1**), *ab initio* gene prediction, as well as homolog information from a previously published *C. canephora* assembly[35] (**Supplementary section 3.4**). The resulting annotations consisted of 28,857, 33,505, 56,670, and 69,314 gene models for the CC, CE, CA, and CA HiFi, respectively (**Table 1**). Altogether ~97% of CC, 89% of CA and 99.6% of CA HiFi gene models were placed on the pseudo-chromosomes, of which 33,618 and 35,449, respectively, were specific to the CC and EE subgenomes (**Table 1**). Annotation completeness,

assessed using the BUSCO pipeline, was ≥95% for CE and CC, and reached 97.3% for CA HiFi. Automated prediction was followed by manual curation of gene models along key biosynthetic pathways (**Supplementary section 3.4**).

Strong conservation of the Arabica subgenomes

We next examined subgenome-wise structural evolution following the allotetraploidy event in *C. arabica*. The split into subgenomes was first verified by comparing synonymous mutation (K_s) values of syntenic gene pairs (**Supplementary section 3.6**). Comparison of the Arabica CC and EE subgenomes against their *C. canephora* and *C. eugenioides* counterparts revealed high conservation in terms of chromosome number, centromere position and numbers of genes per chromosome (**Fig. 1, Supplementary section 4**).

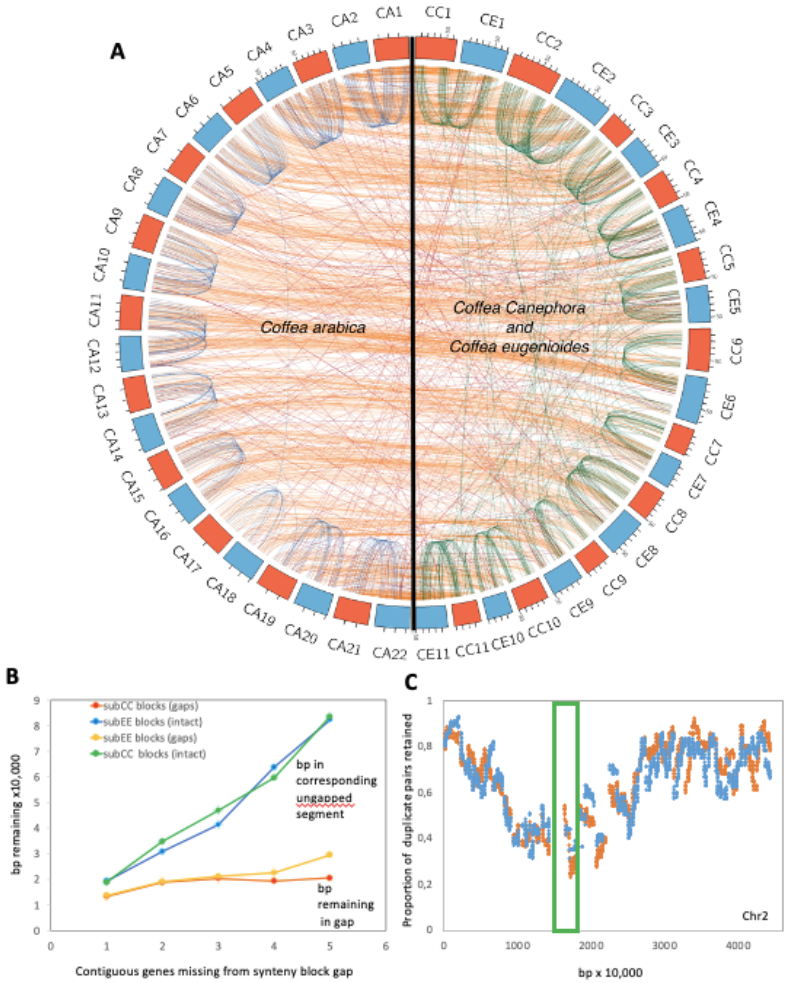

Figure 1. *Patterns of synteny, fractionation and gene loss in* C. arabica *(CA) and its progenitor species* C. canephora *(CC) and* C. eugenioides *(CE).* ***A.*** *Corresponding syntenic blocks between CA subgenomes subCC (orange) and subEE (blue), and with the CC (orange) and CE (blue) genomes.* ***B.*** *bp in intergenic DNA in synteny block gaps caused by fractionation in a subCC-subEE comparison, compared to numbers of bp in homoeologous unfractionated regions, as a function of numbers of consecutive genes deleted.* ***C.*** *Gene retention rates in synteny blocks plotted along subCC chromosome 2; subCC is plotted in orange and subEE in blue. The green box indicates the pericentromeric region.*

Patterns of gene loss following the *gamma* paleohexaploidy event displayed high structural conservation between CC and CE during the 4-6 million years (My) since their initial species

split[22,23] (**Supplementary section 4**). Likewise, the structures of the *C. arabica* subgenomes were highly conserved between each other, with moderate additional fractionation since the Arabica-founding allotetraploidy event (**Fig. 1**). This additional fractionation is reflected also in the BUSCO genes, of which around 5% have reverted to the diploid state (**Table 1**). Syntenic comparisons revealed that rather than pseudogenization of individual genes, genomic excision events removing one or several genes at a time, in similar proportions across the two subgenomes (**Fig. 1B, Supplementary section 4**), have been the main driving force in genome fragmentation both before and after the polyploidy event, in agreement with a previous study[40].

Fractionation occurred mostly in pericentromeric regions, whereas chromosome arms showed more moderate paralogous gene deletion (**Fig 1C, Supplementary section 4**). In support of the dosage-balance hypothesis[41], subgenomic regions with high duplicate retention rates were significantly enriched for genes that originated from the Arabica WGD (Fisher exact test, p<2.2e-16). In contrast, low duplicate retention rate regions significantly overlapped with genes originating from small-scale (tandem) duplications (**Table S9**). Genes with high retention rates were enriched in Gene Ontology categories such as "cellular component organization or biogenesis", "primary metabolic process", "developmental process" and "regulation of cellular process", while low retention rate genes were enriched in categories such as "RNA-dependent DNA biosynthetic process" and "defense response" (in both subgenomes) and "spermidine hydroxycinnamate conjugate biosynthetic process" (involved in plant defense[42]) and "plant-type hypersensitive response" (in the EE subgenome) (**Tables S10-S13**). Similar functional biases have previously been observed in many other plant species[43-46].

The *C. arabica* allopolyploidy event did not seemingly affect the rate of genome fractionation, which remained roughly constant when comparing fractionation in progenitor species versus CA subgenomes after the event (**Supplementary Section 4 and Fig. S29**). Also, overall TE frequencies remain similar in progenitor and descendant species, thereby showing no evidence for a WGD-induced "genomic shock" resulting in rampant activation of TE elements (**Supplementary section 3.2**) or genome rearrangements in the modern allopolyploid. The observed *C. arabica* genome evolution instead follows more closely the "harmonious coexistence" pattern[47] observed in *Arabidopsis* hybrids[17,48].

Gene -specific subgenome expression dominance

Subgenome dominance may also demonstrate itself through dosage balance changes via homoeologous exchange (HE), or expression bias patterns[47]. However, in some cases subgenomes coexist with a more-or-less equal contribution to phenotype[49,50]. To study patterns of expression bias we first identified syntelogous gene pairs between the CA, CC and CE genomes and then removed the pairs with evidence for homoeologous exchange in the CA subgenomes (see under **Origin and domestication of Arabica coffee**, below)[51] (**Supplementary section 5**). The global expression patterns did not show significant subgenome expression dominance patterns in different bean developmental stages, even when accounting for homoeologous exchanges between the two subgenomes (**Tables S15, S16**). However, as in most allopolyploids, individual gene families may still exhibit gene expression partitioning by subgenome[49]. In Arabica, of particular interest are the gene families encoding the enzymes that contribute to its biochemical and aromatic cup qualities, such as *N*-methyltransferases (*NMT*), terpene synthases (*TPS*), and the fatty acid desaturase 2 (*FAD2*) gene family. The availability of a high-quality genome, in which the homoeologs are mapped to the two subgenomes, allowed us to study in detail the subgenome expression dominance in the above gene families (**Fig. 2**).

Caffeine is a purine alkaloid responsible for the psychotropic properties of many plant derived beverages, such as coffee, tea, yerba mate and guaranà[52]. The metabolite appeared in different plant genera through convergent evolution mediated by *NMT* gene duplication[35,53,54]. Caffeine biosynthesis is mediated by three different *NMT* genes (**Fig. 2A**), all of which retain duplicates from CC- and CE-derived subgenomes in the CA genome. Furthermore, the *DXMT* gene,

298 catalyzing the last step in caffeine biosynthesis, is tandemly duplicated in subEE, and both
299 copies are expressed in fully ripe fruits at decreased levels compared to subCC (**Fig. 2B-C**)
300 providing a molecular basis for the lower caffeine content of CA with respect to CC [55-57]. More
301 generally, the *NMT* gene family in CA shows clear, though not extensive, mosaicism in
302 subgenome-wise expression.
303
304 Terpenes derived from geranyl diphosphate (**Fig. 2A**) strongly contribute to coffee aroma[58].
305 Here, we observed expression dominance by subEE for genes encoding *α*-terpineol synthases
306 (**Fig. 2B-C**) and a putative nerolidol synthase. In contrast, a tandem duplication and expression
307 of several putative *α*-farnesene synthases in subCC suggests expression dominance there,
308 whereas the genes encoding for isoprene synthase show also dominance for subCC, but only at
309 the green maturation stage.
310
311 Polyunsaturated fatty acids form an energy reserve in coffee beans and contribute to coffee
312 flavor, aroma, and its shelf life[59]. *FATTY ACID DESATURASE 2* (*FAD2*) encodes the key
313 enzymes that desaturate oleic acid to linoleic acid, the major unsaturated fatty acid in coffee
314 (**Fig. 2A**). *Arabidopsis* encodes a single, constitutively expressed *FAD2* gene, whereas some
315 oil-producing plants, and coffee, have multiple copies of *FAD2*, some of which are mainly
316 expressed in seeds. Based on phylogenetic analyses, these seed-type *FAD2*s evolved from
317 housekeeping FAD2-encoding genes by gene duplication[60,61]. In Arabica, the housekeeping
318 *FAD2* syntelogs were present in both subgenomes (**Fig. 2B-C**), with expression dominated by
319 subEE, whereas seed-type *FAD2* was duplicated in both subgenomes, where overall, the
320 duplicates show expression patterns suggesting dominance by subCC. Interestingly, the genic
321 region encoding the N-terminal signal peptide of the only highly expressed subEE homolog
322 (Cara002g032150) differed from those of the other seed-type paralogs; this could result, for
323 example, from a gene conversion event[62,63].
324

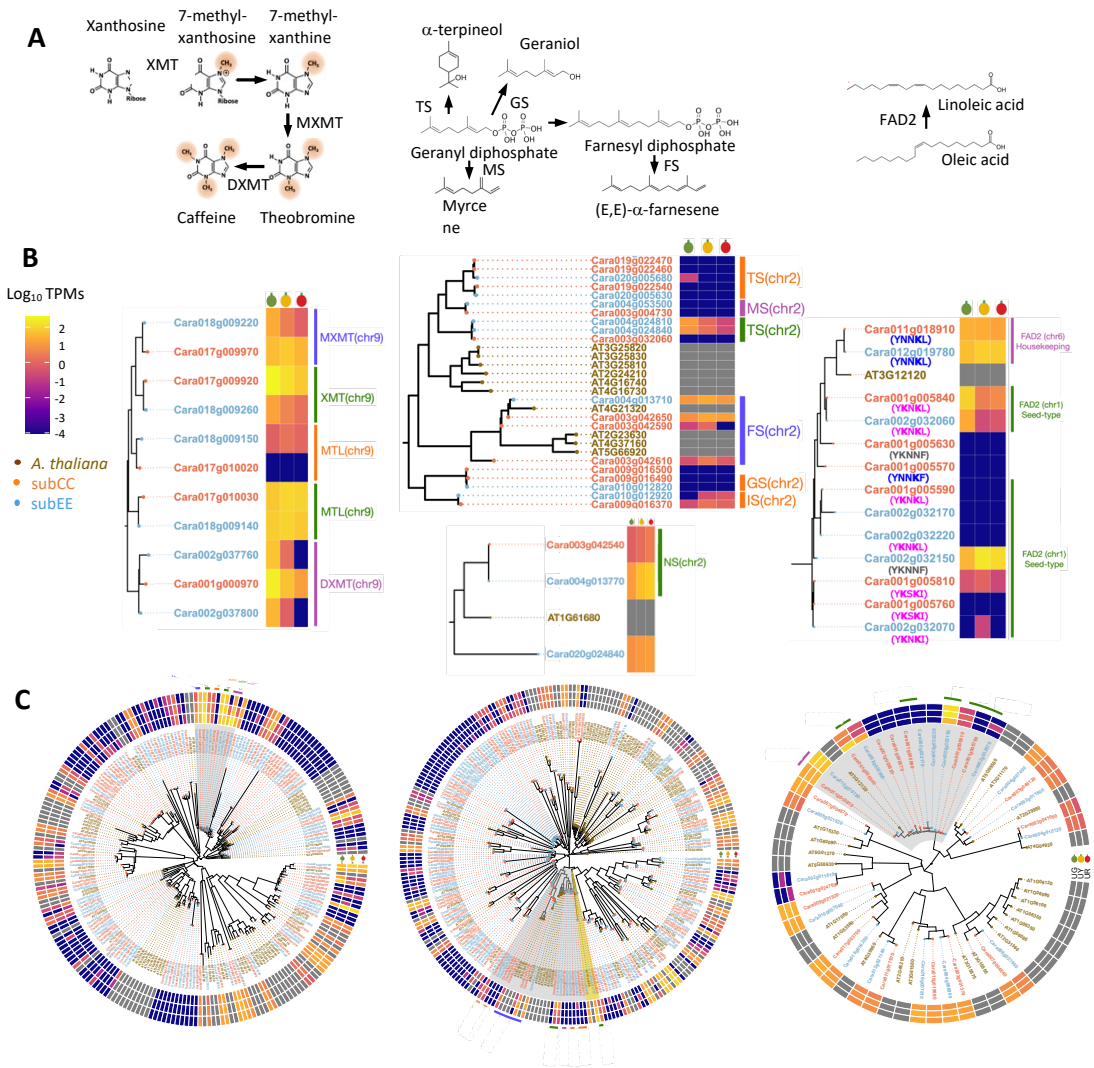

Figure 2. *Composition and expression of* exemplar C. arabica *gene families contributing to bean quality traits.* ***A****. Schematic biosynthesis of caffeine (left), terpenoids (middle), and unsaturated fatty acids (right)* ***B.*** *Phylogenies and expression during fruit development of C. arabica genes for N-methyltransferases (NMTs) mediating caffeine biosynthesis (left), terpene synthases (TPS) (middle), and fatty acid dehydrogenase 2 (FAD2) (right). RNA sequencing was carried out for three biological replicates from three different fruit maturation stages (green, yellow, and red) of the K7 cultivar.* ***C.*** *Genome-wide NMT (left), TPS (middle), and FAD2 (right) phylogenies and expression patterns during fruit development. The genes located in the two subgenomes are indicated by font color; subCC (red) and subEE (blue). Grey areas highlight the parts of phylogenies shown in B. XMT: xanthosine methyltransferase; MXMT: 7-methylxanthine methyltransferase; DXMT: 1,7-dimethylxanthine methyltransferase; MTL: N-methyltransferase-like; FS: (E,E)-a-farnesene synthase; GS: Geraniol synthase; IS: Isoprene synthase; MS: myrcene synthase; TS: (-)-a-terpineol synthase; FAD2: Fatty acid desaturase 2.*

In conclusion, the *C. arabica* genome exhibits little, if any, subgenome expression dominance, while individual genes demonstrate fine-scale expression partitioning by subgenome, as expected. Importantly, dominance was observed within specialized metabolic pathways contributing to coffee aromatic quality, consistent with earlier studies on these and other gene families[64-67]. These studies suggest that subgenome dominance, while averaging out equally in a global sense, follows a mosaic pattern in *C. arabica* wherein different sub-processes are locally dominated by each of the subgenomes. Similar subgenomic coexistence has also been identified in other neopolyploids such as rapeseed[11] and cotton[50]. This phenomenon could be

due to the high conservation of the genome structures between the two progenitor species and the recency of the polyploid event.

Origin and domestication of Arabica coffee

To obtain a genomic perspective on the evolutionary history of wild and cultivated coffee varieties, we sequenced the complete genomes of 39 *C. arabica* accessions (**Supplementary section 6.1**, **Table S18**), as well as the 18th century type specimen, kindly provided by the Linnaean Society of London. However, sequencing of this accession yielded only 1.5x coverage, which limited its use in subsequent analyses.

The set of 15 modern cultivars differed in their breeding histories, including members from both the Typica and Bourbon groups, crosses between the two groups carried out in Brazil (e.g., Mundo Novo[19] and its further crosses), one Indian cultivar most likely derived from the "Seven Seeds" smuggled into India around 1600 (Jackson 1, JK1), and the recent cultivar Geisha, which originates from an Ethiopian forest and was put directly into cultivation in the 1930's. Additionally, we included five *C. canephora* x *C. arabica* crosses from the Timor hybrid lineage that have different levels of back-crosses to CA cultivars. The 17 wild accessions were collected from the Eastern and Western sides of the Great Rift Valley during FAO[68] and IRD[69] missions during 1960's (**Table S18, Fig. 3A**).

C. arabica displays disomic inheritance with bivalent pairing of homologous chromosomes, which largely prevents recombination between subgenomes[70]. However, homoeologous exchange (HE) has been observed in several neopolyploids[9,51,71], and since TE contents in the two subgenomes were found to be intermediate between the CA and CE progenitors, we explored the extent of HE among our CA accessions. Overall, all accessions shared a fixed allele bias toward subEE at one end of chromosome 7 (**Supplementary section 5**). The genes in this region were enriched for chloroplast-associated functions (**Table S17**). Since the plastid genome in CA was inherited from CE, HE in this region was likely selected due to compatibility issues between nuclear and chloroplast genes encoding chloroplast-localized proteins[72]. Surprisingly, all but one accession (BMJM) showed significant (*p*-values <<9.8e-37; Chi-square test) 3:1 allele bias towards subCC. These patterns were present in both wild and cultivated Arabicas, suggesting that the allelic bias is an adaptive trait and not associated with breeding. The pattern similarity further suggests the biases originated in a common ancestor of all sampled CA accessions, possibly from the establishment of a single stable tetraploid individual subsequent to the initial polyploid event. The site frequency spectrum displayed a strong bias towards recent HE as well as to events shared by all individuals (**Fig. S38**), suggesting that HE events are under strong selection. As demonstrated in BMJM, rare HE events were also present in cultivars; in the case of BMJM, the bias towards subEE was due to a single crossover in chromosome 1. Altogether, the results suggest that in a polyploid species with low genetic diversity such as Arabica, HE could be one possible cause for the phenotypic variation observed among extremely closely related accessions[73].

We next studied the patterns of adaptation and drift within each of the subgenomes separately. Since subEE is most affected by HE events, interpretations were first sought from subCC results. In order to assess how well our set of samples represented the overall variation in the existing *C. arabica* populations, we combined our own variant data with more broadly sampled but variant-sparse genotyping-by-sequencing (GBS) data[20]. Principal component analysis (PCA) showed that our wild Ethiopian samples represented well the genetic diversity among the larger set of GBS samples from the same region, whereas our cultivar samples all grouped together with wild Yemeni individuals (**Fig. S39**).

Among the 17 wild samples, the overall genomic diversity was extremely low, concordant with the earlier GBS-based study, with $\pi_{subCC} = 6.99 \times 10^{-4}$ and $\pi_{subEE} = 8.7 \times 10^{-4}$, both indicative of small effective population sizes. For both subgenomes, Tajima's D was negative (Tajima's

D_{subCC} = -0.41 and Tajima's D_{subEE} = -0.33), suggesting an expanding population, possibly following one or more population bottlenecks. The genetic diversity among cultivars did not differ from wild population samples, as demonstrated by low fixation index (F_{ST}) values, 0.094 for subCC and 0.090 for subEE, respectively. Possibly reflecting the known bottlenecks associated with Arabica cultivation history, nucleotide diversity for cultivars was even lower (π_{subCC} = 5.6 x 10^{-4}, π_{subEE} = 6.6 x 10^{-4}) than in wild populations, and subsequent wide-spread cultivation may have led to less negative Tajima's D scores (Tajima's D_{subCC} = -0.31 and D_{subEE} = -0.25).

For a first view of Arabica population structure, SNP tree estimation and ADMIXTURE analysis were carried out separately for each of the subgenomes to infer ancestral populations (**Fig. 3B**). For subCC, a three-population solution yielded the best cross-validation score and showed grouping into Typica-Bourbon cultivars (Population 1, interpreted to represent the cultivar population), wild accessions (Population 2; wild population), and Timor hybrid-based cultivars (Population 3; hybrid individuals with CC introgression). The Typica and Bourbon groups were all assigned to the same population, whereas old cultivars with fewer breeding cycles (BMJM, Erecta, TIP1, and JK1) showed admixed states. The recently established Geisha cultivar demonstrated roughly equal proportions expected from recent admixture. Earlier marker-based studies demonstrated that Indian varieties encompass Typica and Bourbon variation[20], and our results support this finding. The wild individuals were assigned to a different population than the cultivars, whereas the Linnaean sample grouped together with the cultivars, supporting its hypothesized origin from the Dutch East Indies[25]. Among the wild samples, roughly half showed admixture with the cultivar population. A complementary analysis using PCA showed results concordant with the ADMIXTURE analysis (**Fig. S42**). The subEE subgenome displayed largely similar admixture patterns as subCC, but here the Timor hybrid lines grouped together with the other cultivars, suggesting that in this hybrid the *C. canephora* introgression has occurred only into subgenome C. This admixture was also well reflected in the maximum likelihood tree inferred from independent SNPs (**Figs. 3B, S43**). Notably, the Linnaean sample grouped together with the old Typica cultivars Erecta and TIP1.

To identify the demographic events that shaped overall Arabica coffee evolution, we modeled the population history of *C. arabica* accessions using the SMC++[74] and pairwise sequentially Markovian coalescent[75] models (**Fig. 3D**; **Figs. S44-S48**). To avoid possible confounding effects of admixture, the analyses initially focused on non-admixed wild individuals. Both subgenomes concordantly showed two bottlenecks during their paleohistory (**Fig. 3D**). Using a mutation rate of 7.77 x 10^{-9} / (bp*generation)[76] and a generation time of 21 years[77], the most recent bottleneck initiated around 5,000 years ago (5 kya), and an additional, longer period of lower population size was modeled between 20-100 kya (**Fig. 3D**). Because a large part of modern human evolution occurred in East Africa, the past geoclimatic history of the region has been well studied. As such, evidence has been found for an extended drought and cooler climate in this region at 40-70 kya, coinciding with human migration out from Africa[78] (**Fig. 3E**). Later, during the African humid period (AHP), around 6-15 kya[79], growth conditions were likely more beneficial for *C. arabica*. This period largely coincides with the population increases seen in SMC++ for both wild and cultivar populations (**Fig. 3D**). While demographic modeling provides accurate estimates of changes in historical population sizes, timings of events should be treated with caution since there is considerable uncertainty in mutation rate, factors contributing to it[80], and generation time estimates in plants[81]. In the case of *C. arabica*, the generation time was estimated from empirical data[77], but the precise mutation rate is not known.

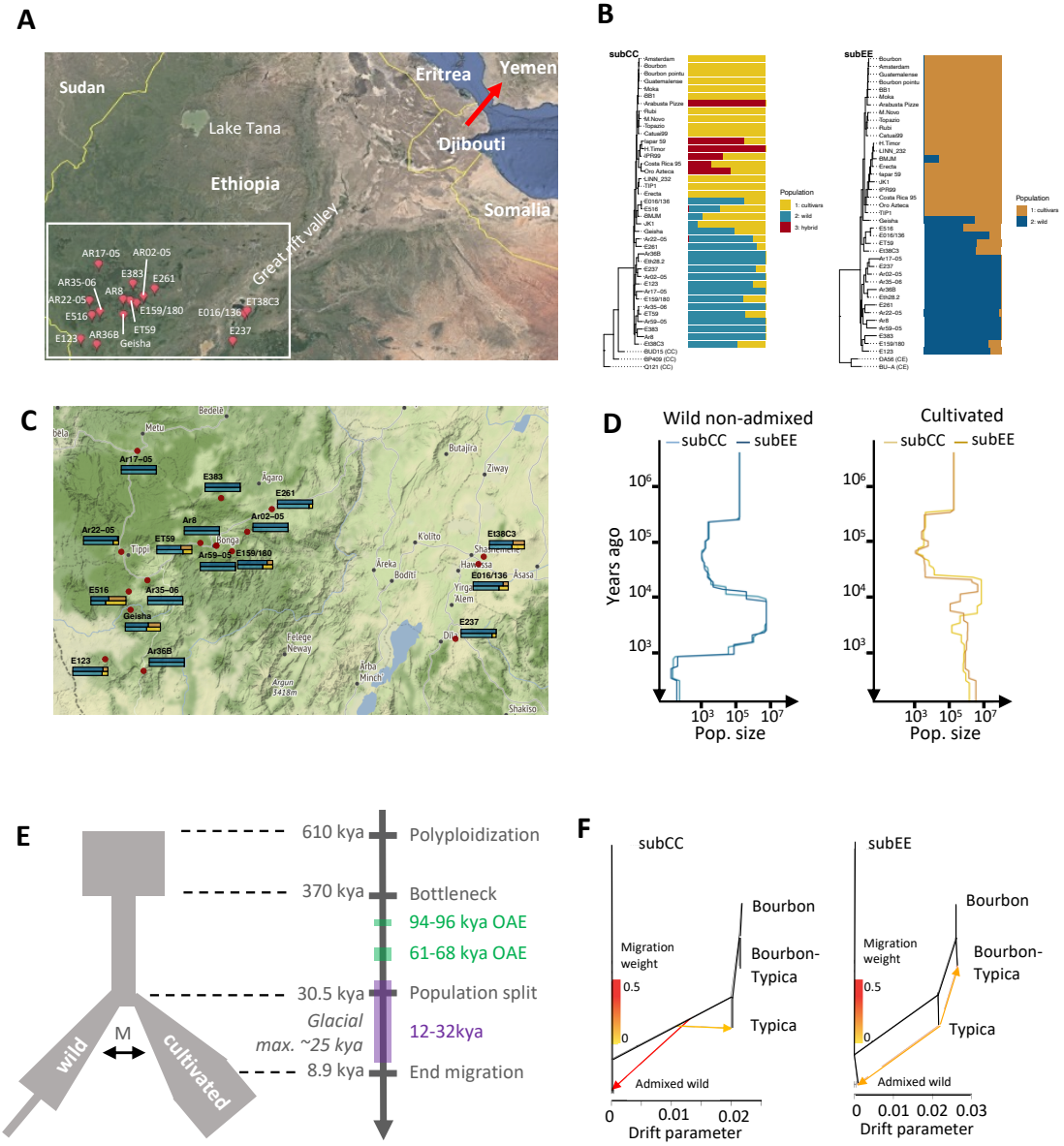

Figure 3. *Population history of Coffea arabica.* ***A.*** *Geographic origin of resequenced wild C. arabica accessions. The red arrow indicates the probable route of the migration to Yemen in historical times.* ***B.*** *Ancestral population assignments of C. arabica accessions for the CC subgenome (left) and EE subgenome (right). Relationships among individuals are described with the phylogenetic tree obtained from independent SNPs.* ***C.*** *Magnification of panel A, showing the admixture values for each of the accessions in subgenome C (top) and E (bottom); the colors correspond to the analysis in panel B.* ***D.*** *Population sizes of wild and cultivated accessions, inferred using SMC++, suggest genetic bottlenecks at ~370 and 1kya (limited to non-admixed wild individuals).* ***E.*** *FastSimcoal2 output, suggesting a population split ~30.5 kya, followed by a period of migration between the populations until ~8.9 kya. This timing corresponds with increased population diversity in cultivars at a similar time, calculated using SMC++. Green rectangles along the timeline show "windows of opportunity", times when Yemen was connected with the African continent wherein human migrations to the Arabian Peninsula may have occurred. The purple rectangle shows the last ice age.* ***F.*** *Directional gene flow analysis using Orientagraph suggests two hypotheses: gene flow from the shared ancestral population of all cultivars to the Ethiopian wild individuals (subCC), or gene flow from the Typica lineage to Ethiopia (subEE).*

To gain more insight into population splits, we modeled Arabica population history with approximate Bayesian computation using FastSimcoal2[82], keeping the wild population and cultivars as two separate lineages. In the best-fitting model (**Fig. 3E**), the wild population, including the admixed wild individuals, was predicted to split from the cultivar founding population at 1,450 generations ago (~30 kya), while the two populations maintained some gene flow (in terms of migration) until ~8-9 kya. Such gene flow between the two diverging populations could explain the modeled increase in effective population size between 8-20 kya (**Fig. 3D-E)**.

When compared against major climatic events, the wild vs. cultivated population split was predicted to occur before the latest glacial maximum (20-27 kya), with migration maintained until the end of the AHP. Since our sampling of wild individuals did not include pure wild representatives of the modern cultivated population, the precise place of origin for the latter remains unknown. However, an extended period of migration between the two populations is most parsimonious if they were separated only by a relatively small geographic distance, such as along the two sides of the African Great Rift Valley (**Fig. 4A**). It is possible that the second ancestral population could have extended as far as Yemen, ~1,000 km away, and in that case, the end of migration between the two populations may have coincided with the end of the AHP and widening of the Bab al-Mandab strait (separating Yemen and Africa) to tens of kilometers due to rising sea levels[78]. A wild native *C. arabica* population has been identified in Yemen[83], which could support this hypothesis. Moreover, the Linnaean sample, together with the Typica and Bourbon cultivars, originate from the native Yemen population, as suggested by the ADMIXTURE, PCA, and SNP analyses (**Fig. 3B**, **S42, S43**). More recently, when inter-lineage migration ended, both wild and cultivar populations underwent strong independent bottlenecks at ~1 kya; this was observed when analyzing the SMC++ curves for subgroups of wild, wild admixed, Typica and Bourbon individuals (**Figs. S46-S48**). All trajectories showed diverging behaviors between the subgenomes at ~8 kya, roughly the time at which migration was modeled to end. These differences may be due to a limited number of migrating individuals with different histories; in contrast, the wild, non-admixed individuals showed a strong bottleneck and overall low effective population sizes for both subgenomes.

To look for evidence of the original Arabica speciation event we also modeled older population bottlenecks. The wild population displayed strong coalescence at three different time points, with the first two at 11,200 and 17,700 generations ago. The oldest bottleneck, shown independently in both subgenomes, suggested coalescence to an extremely small population around 29,000 generations ago (610 kya). This is close to some previous estimates for the allopolyploidy event in recent literature[22,23], and also corresponds with the "flatlining" of all the SMC++ results (indicating full coalescence), the crown age of the *C. arabica* accessions in the subCC SNP tree (**Fig. S43**), as well as divergence estimates based on gene fractionation and the distribution of non-synonymous mutations (**Fig. S29**). Interestingly, the SMC++ analysis for introgressed cultivars did not show similar coalescence but rather a steady decline in ancestral population size at this time horizon. Widespread inbreeding is known to accelerate coalescence[84], and we have recently shown this to affect demographic modeling as well[85]. Admixture and introgression, in contrast, introduce intra- and interspecific polymorphisms into genomes. Reflecting this, the population histories of admixed Arabica individuals demonstrate shifts towards more ancient coalescence times (**Fig. S45**), and in the case of Timor hybrids, introgression may have resulted in ancestral polymorphisms introduced back into CA through the hybridization event, resulting in alleles with deep coalescence.

In summary, roughly concordant with some previous studies, our analyses suggest that the Arabica allopolyploidy event occurred 610 kya or earlier, when considering that the inbreeding present in *Coffea* populations would accelerate coalescence[84,85]. Earlier work suggesting more recent timings, such as 20 kya[20], could be an underestimate due to multiple population bottlenecks in cultivated and wild populations.

Kinship estimation identifies the probable origins of cultivars

The low genetic diversity and known data on Arabica cultivar history provided us with an opportunity to explore the pedigree of coffee cultivars from SNP data using Kinship-based INference for Gwas (KING)[86]. The analysis identified all known breeding relationships (**Fig. 4, S49; Tables S18-S20**); for example, Bourbon and Typica group cultivars were all 1st degree related and showed parental relationship to Bourbon-Typica crosses. Strikingly, the Typica and Bourbon individuals were also 1st degree related, suggesting direct parent-offspring relationships, and similarly, the Indian variant JK1 was 1st degree related to more modern cultivars. Besides confirming the likely Yemeni origin of all Arabica cultivars, this finding also underscores the Yemeni-based germplasm's very limited genetic diversity. In general, subEE showed one degree lower (2nd) order relationships, possibly due to a higher initial level of genetic variation in the subgenome or HEs there.

In our Timor hybrid samples, *C. canephora* introgression initially occurred in the Typica background, and the subsequent crosses were to the Bourbon lineage. In subEE the differing degrees of relationship to Typica and Bourbon were visible (**Fig. S49; Table S20**), supporting the overall validity of our results and demonstrating that subEE had not received substantial introgression. On the other hand, in subCC the introgression from CC clearly broke the haplotype blocks such that no relationships to Typica or Bourbon cultivars could be detected (**Fig. 5**). However, further analysis on the subCC chromosomes with low levels of introgression recovered these relationships.

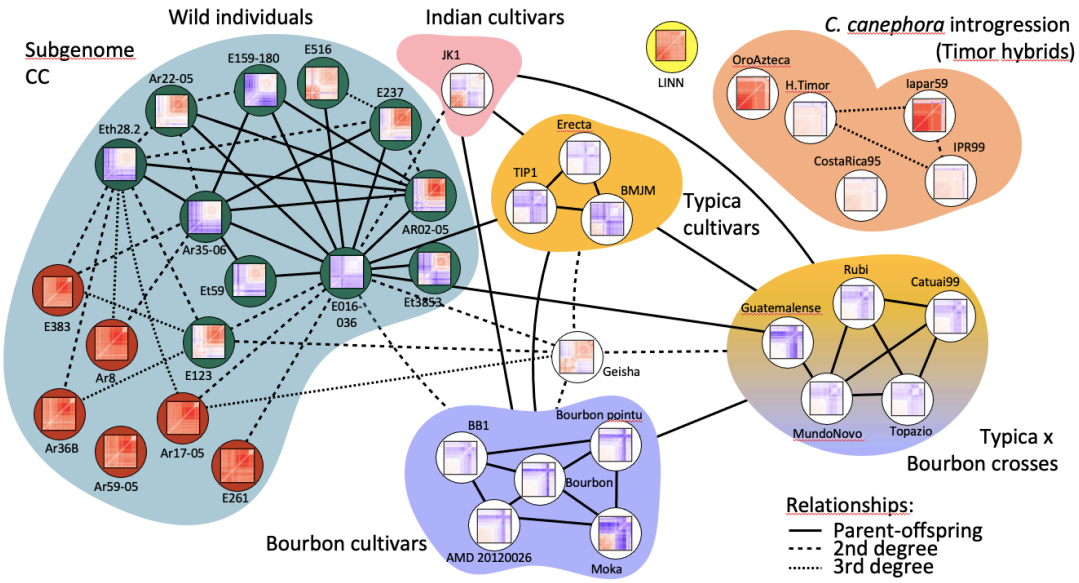

Figure 4. *Kinship estimation of* C. arabica *accessions, inferred from SNPs in the subCC. The degree of relatedness was estimated using Kinship-based INference for GWAS (KING) and describes the number of generations between the related accessions. Thumbnail images show false discovery rate corrected F3 tests of introgression Z-statistics for each of the target individuals. Each cell in the matrix illustrates an F3 test result for the target accession containing introgression from two different sources (x- and y- axis); -blue color illustrates significant gene flow (or allele sharing via identity by descent[87]; IBD) from the two source accessions to the target, while red color illustrates lack of gene flow. For detailed images and the order of the accessions, see* ***Fig. S50****; see also S49, S51 for corresponding analyses in subEE. In the wild accessions, the dark green background highlights the admixedindividuals (****Figure 3B****), while the non-admixed individuals are highlighted with red background. Relationships follow standard nomenclature (eg 2nd degree refers to an individual's*

grandparents, grandchildren etc., whereas 3rd degree refers to great-grandparents, great-grandchildren, etc.

The old cultivar lines JK1 (Indian), Erecta (Indonesian Typica), BMJM (Blue Mountain Jamaica, a Caribbean Typica-like cultivar), TIP1 (Brazilian Typica), and BB1 (Brazilian Bourbon) showed 2nd-4th order relationships with a cluster of closely related wild individuals, centered on E016/136. The set of related wild samples consisted of individuals that were found to be admixed in earlier analyses (**Fig. 3B**). When split into admixed-related vs. non-admixed wild individuals, we observed a clear difference in nucleotide diversities between the two groups, illustrating the effect of admixture on the wild population (**Fig. S41**). A formal F3 test of admixture[88] demonstrated highly negative F3 scores for all cultivars, resulting either from admixture or from identity by descent (IBD) through sampling within a closely interrelated population[87] (**Figs. S50-S51**). The recently established Geisha cultivar was related to the cluster of wild admixed individuals as well as to the Bourbon and Typica groups, suggesting similar population origins. Interestingly, accession E016/136 showed admixture/IBD with both wild and cultivated populations, and also when having either Bourbon or Typica individuals as possible sources. Other members in the cluster of highly interrelated wild individuals showed varying levels of wild-cultivar allele sharing, similar to Geisha.

In a comparison of geographic origins, the wild individuals showed that all samples on the Eastern side of the Great Rift Valley had some levels of admixture and were closely interrelated (**Fig. 3C**). On the Western side of the Great Rift Valley the admixed, related individuals were mostly concentrated around the Gesha region and the road connecting Bonga with Tippi. The wild individual with the closest relationship to cultivars, E016/136, demonstrated a first-degree relationship with several individuals, but only Ar35-06 and Eth28.2 were pure representatives of the wild population in ADMIXTURE analysis (**Fig. 3B**). Therefore, either one of these two individuals may be genetically closest (in our sample) to the true parent of cultivated Arabica. Only the origin of Ar35-06 is known; it was collected from Mizan-Teferi near the Gesha mountain. Among the other closely related samples, the geographic origin of E516 reflects its collection from the field where it was cultivated, but in fact, it was recovered from the Sheka forest near Tippi. Hence, the wild individuals with good quality Arabica coffee may have already been spread by human intervention. Intriguingly, the Geisha cultivar was found in the Gori Gesha forest near the mountain Gesha, close to the location of Ar35-06. Altogether this suggests the Gesha region to have been a hotspot of admixture for CA.

Admixture observed among the wild samples may have occurred either by a recent hybridization event, whereby cultivars from Yemen migrated back to Ethiopia, or instead, the accessions may have hybridized after their collection from the wild. A third alternative is that perhaps the Yemeni population (and hence the cultivars) originate from an ancestral population from the Eastern side of the Great Rift Valley or the Gesha region. Analysis of admixture patterns with Orientagraph[89] (**Fig. 3F**) suggested hybridization with the common ancestor of the Bourbon and Typica lineages in subCC, and hybridization from Typica to the wild individuals in subEE. To identify the more plausible scenario, we assumed that for the case of recent admixture, the introduced haplotypes would form long contiguous regions since recombination would not yet have broken them up. Hence, we identified the genomic regions of Typica origin in wild admixed population using the distance fraction (d_f) statistic[90]. All individuals showed only short blocks, comprising 0.9-1.6% of the genome (**Fig. S52**). As a positive control, we calculated the block lengths in the Timor hybrids, with their known introgression ca. 100 years ago, and obtained considerably longer block lengths (**Fig. S53**). Therefore, it appears that admixture events among wild accessions were not very recent, and that modern coffee cultivars likely originated from a second ancestral population, as supported by the Orientagraph result from subCC (**Fig. 4F**). Subsequent admixture may have contributed towards the development of the Typica lineage and could explain its genotypic differences from Bourbon. Admixture may have occurred in the wild, possibly to the Eastern side of the Great Rift Valley, perhaps leading to the development of the Geisha cultivar.

Domestication of crop plants is generally split into four stages; (i) pre-domestication management of wild populations, (ii) selection of lineages with desirable alleles and their cultivation (iii), geographical radiation and adaptation to different environments, and (iv) targeted breeding[91]. As a globally grown crop plant, *C. arabica* would be placed at the fourth stage of domestication, but its long generation time means that the cultivars have had relatively few breeding cycles; altogether this makes *C. arabica* an interesting target for studying the early stages of cultivation.

Cultivation (stages 1 and 2) involves a population bottleneck, which results in highly reduced genetic diversity in cultivars[92]. Accordingly, in Arabica coffee the mean nucleotide diversity in neutral sites (π_s) was lower among cultivars, while Geisha and older cultivars with fewer breeding cycles demonstrated higher diversity (**Fig. S41**). Among the major cultivar groups, Typica had higher diversity than Bourbon, whereas modern cultivar crosses of the two lines showed intermediate values. The differences in π_s could result from the known single-individual bottleneck in Bourbon, admixture in Typica, and more concentrated breeding efforts on the Bourbon lineage.

Common to many crops, admixture and subsequent heterosis may have played a role when selecting the early cultivars[85], such as BMJM and JK1. In Arabica, this is supported by the fact that admixed wild individuals are preferred for coffee production in Ethiopia, and also by the success of the more recently identified Geisha cultivar, which also derives from the admixed population. The admixture results and nucleotide diversity estimates suggest, however, that the heterosis effect among Bourbon and Typica lines appears to have been lost during coffee breeding (**Fig. S41**). However, the maintenance of current heterozygosity levels may be the driving force behind modern coffee breeding, whereby two individuals are preferentially crossed, instead of the widespread selfing that occurs among wild individuals. Additionally, the more modern cultivars represent crosses between the Bourbon and Typica lineages. Perhaps because of this strategy of maintaining heterozygosity, the inbreeding coefficients in the cultivated accessions were similar to those of wild accessions (**Fig. S40**), differing from the general expectations for a domesticated species[92].

During cultivation, increased genetic drift due to small population size and genomic hitchhiking linked with loci under selection elevate the frequency of deleterious alleles. Furthermore, continued inbreeding leads to eventual purging of deleterious alleles[92]. Concordantly, the cultivated Arabicas showed decreased diversity both in terms of non-coding sites as well as the ratio of non-synonymous diversity to synonymous diversity (**Fig. S41**). To look for pathways affected by selection in cultivars, we calculated the F_{ST} values between cultivars and wild accessions across the genome for all genes and their 2 kb flanking regions and identified the genes with high F_{ST} (95 % quantile) in subCC and subEE, respectively (**Tables S21-S22**). Assuming that deleterious variants (and the ones under selection) will have increased frequencies, we focused our analysis on genes with a large number of shared (>40% individuals having the mutation), derived amino acid changing mutations among cultivars. This screening generated a list of 556 genes that were significantly enriched for only one GO category, "Defense response" (**Table S23**). Sixteen out of 22 genes in this category were NB-ARC domain-containing resistance (R) genes, while two out of the remaining six genes were members of the leucine-rich repeat (LRR) gene family. High diversity in immune related responses is one possible mechanism for preventing extensive spread of pathogens in plant communities[93], and therefore reduced diversity in resistance genes may have compromised modern Arabica cultivar immunity.

To look for other possible phenotypically causative mutations, e.g., in promoter regions, we extended the screening to all derived alleles. This yielded a larger set of 1,908 genes (**Tables S24-S25**) that were enriched for the GO categories "cellular response to nitrogen starvation",

675 "regulation of innate immune response" and "regulation of defense response" (**Table S26**), and
676 contained homologs of ammonium transporters *AMT1* and *AMT2*, important for nitrogen
677 uptake in *Coffea*[94], a homolog of the salicylic acid receptor *NONEXPRESSER OF PR GENES*
678 *1* (*NPR1*), which is required in SA signaling and systemic acquired resistance[95], as well as a
679 homolog of *Arabidopsis LSU2* gene, which was earlier identified as one of the hub genes
680 convergently targeted by effectors of pathogens from different kingdoms[96]. Nitrogen is a
681 principal nutrient required by all plants, and purifying selection on ammonium transporters may
682 have resulted following uniform growth conditions of cultivars in the field. Altogether, the low
683 diversity in R genes as well as key regulatory genes for innate immunity, such as *NPR1* and
684 *LSU2*, may contribute to the higher disease susceptibility observed in non-CC-introgressed
685 cultivated Arabica[29].
686
687
688 **Introgression from *Coffea canephora* had a profound impact on modern *C. arabica***
689 **cultivars**
690 The high level of structural conservation between the CA subgenomes and their diploid
691 progenitors has facilitated spontaneous interspecific hybridization events, as seen in the Timor
692 hybrid lineage. The most striking aspect of the Timor lineage is its resistance to coffee leaf rust,
693 to which the Bourbon and Typica groups are highly susceptible[29]. Modern breeding programs
694 have successfully backcrossed the Timor hybrid lineage to pure CA cultivars, establishing the
695 so-called Catimor and Sarchimor lineages that are very commonly grown by coffee producers
696 in South America. Our sample set included five descendants of the original Timor hybrid,
697 making it possible to analyze the genome effects of recent introgression.
698
699 Overall, the hybridization affected subgenome CC more profoundly, with higher levels of
700 nucleotide divergence (F_{ST}=0.185) than in subEE (F_{ST}=0.0897), when comparing cultivars and
701 hybrids. The divergence from wild populations was even greater, with F_{ST}=0.254 for subCC
702 and F_{ST}=0.138 for subEE. In contrast, F_{ST} values on subEE displayed similar ranges as the wild-
703 cultivar comparison, illustrating that introgression occurred seemingly solely within subCC.
704
705 We searched for introgressed loci in the Timor hybrid by calculating the d_f statistic[90] genome-
706 wide, identifying largely the same regions as the F_{ST} scans (**Fig. 5A**). Plots along the genome
707 demonstrated large introgressed blocks, reflecting recent hybridization (**Fig. 5A, S53**) and
708 covering 7-11% of the genome. The introgressed regions showed significant overlap with
709 regions of higher subgenome fractionation (p=0.001873; **Table S27**). This could be due to
710 heterologous recombination occurring between subCC and *C. canephora* in these regions,
711 possibly resulting in increased levels of non-homologous end-joining due to local differences
712 in genome structure between the two species. An introgressed region shared by all Timor hybrid
713 lines is evident on chromosome 4 where the descendants of two sister lineages (HT832/1 and
714 HT832/2) displayed highly contrasting patterns due to differences in recombination, most likely
715 tracing back to differences in the original sister lineages, followed by varying number of back-
716 cross generations.
717
718 To look for functional mutations affecting rust tolerance in Timor hybrids, we identified all the
719 genes inside the introgressed regions shared by all five individuals. The set of 233 genes (**Table**
720 **S28**) contained members of a tandemly duplicated block of ten resistance-related genes in
721 subCC chromosome 4. In subsequent F_{ST} scans these genes also showed a high level of
722 heterozygosity difference between cultivars and introgressed lines. Functional characterization
723 based on *Arabidopsis* suggests the genes to be homologs of *Arabidopsis RPP8*, an NLR
724 resistance locus that has been found to confer pleiotropic resistance to several pathogens, such
725 as an oomycete and two distinct viruses[97,98]. *RPP8* shows a great amount of variation in
726 *Arabidopsis* alone, and intrachromosomal gene conversion combined with balancing selection
727 has been found to contribute to its exceptional diversity in that species[99]. The same subCC
728 region containing the *RPP8* homologs also included a homolog of the F-Box protein

CONSTITUTIVE EXPRESSER OF PR GENES 1 (CPR1), a negative regulator of defense response that targets resistance proteins[100,101].

We next assessed the possible involvement of the introgressed regions on leaf rust resistance by reanalyzing the data from Florez et al.[102], where RNA sequencing for single replicates of Timor and Caturra genotypes inoculated with *H. vastatrix* was carried out at five time points. To look for possible associations with introgressed regions, the post-inoculation timepoints were treated as biological replicates and differential expression for Caturra vs. Timor was analyzed using DESeq2[103]. To allow for possible variation across time points, we applied a relaxed adjusted *p*-value threshold of 0.2 and focused on genes with fold change >2. The selected set of 723 genes was enriched for 50 GO categories, most of which were associated with defense responses (**Tables S29, S30**). The *RPP8* homolog we identified in subCC chromosome 4 was included within this set, as a member of several significantly enriched GO categories.

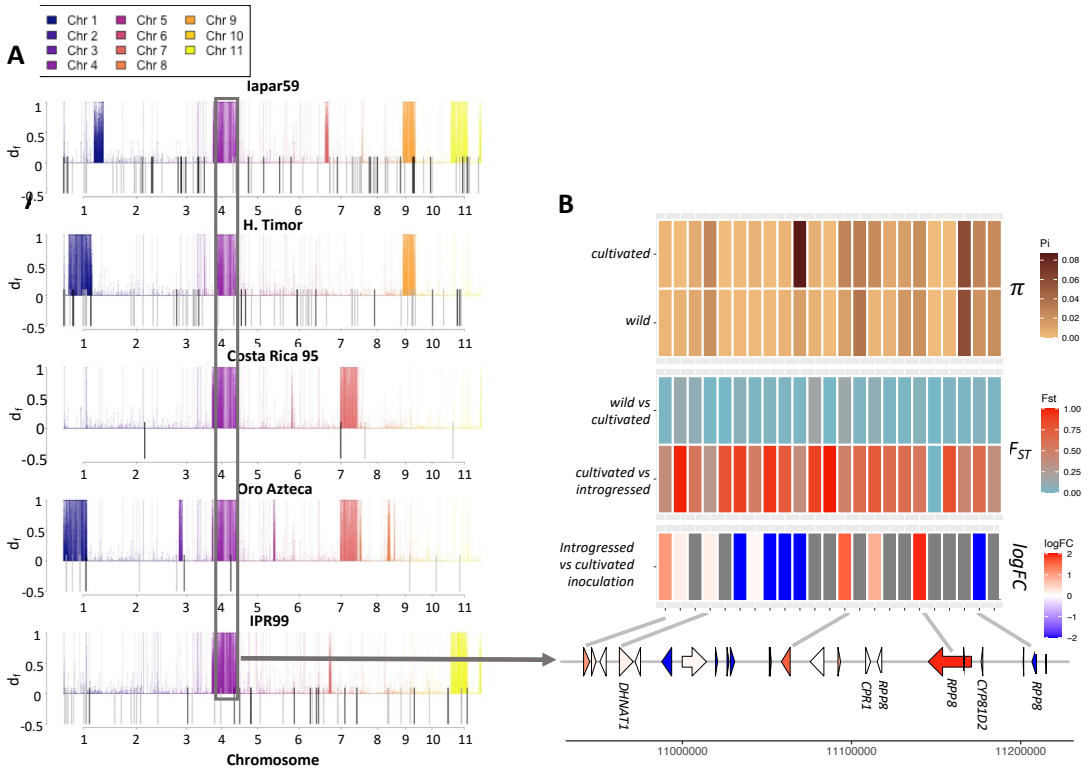

Figure 5. *Introgression of C. canephora into H. vastatrix-resistant C. arabica lineages. A. Introgression d_f statistic estimated for different Timor hybrid derivatives. Colored lines above the axis mark regions of significant introgression in the line under inspection, and are colored by chromosome. The shared introgressed region on Chr 4 is colored in purple and boxed. Transposon Insertion Polymorphisms are represented as lines below the X axis and exhibit overlap with introgressed regions. B. The shared introgressed genomic region on subCC chromosome 4 contains a cluster of R genes and a homolog of a negative regulator of R genes (bottom). The heatmap shows, from the bottom up, (i) log fold change of gene expression after H. vastatrix inoculation, when comparing resistant Timor hybrid lineage against a susceptible cultivar; red color means elevated expression in the hybrid; (ii) fixation index (F_{ST}) values for the introgessed lines vs cultivars and between cultivars and wild*

accessions; (iii) nucleotide diversity for the wild and cultivated accessions for each gene coding region, plus the flanking 2kb upstream and downstream of the region.

Retroelement insertions may contribute to Arabica diversity

Retroelement insertions have been shown to play a role in gene regulation and thereby plant diversity, and for example in tomato, they have been found to be associated with variation of agronomic traits[104]. The five Timor hybrid individuals permitted quantification of the distribution of Transposon Insertion Polymorphisms (TIPs) in individuals with recent introgression. Overall, we identified 157 Gypsy and 63 Copia insertions unique to the Timor hybrids; only 13 Gypsy and 4 Copia elements were shared by at least two accessions, testifying to their recent origins and dynamic nature. The TIPs showed a significant overlap with regions that were targets of introgression in different hybrids (Gypsy p=0.0002107; Copia p=0.03527; **Fig. 5B**). Similar to all insertions, the Timor hybrid-specific TIPs significantly overlapped tandemly duplicated gene regions (Gypsy p<2.2e-16; Copia p=3.99e-06). Many plant secondary metabolites are synthesized by genes arranged in biosynthetic gene clusters, a set of co-located genes contributing to the same pathway[105,106]. Within the 39 accessions, the TIPs significantly also overlapped genes residing in biosynthetic gene clusters (**Table S31**, p<2.2e-16; Fisher exact test). We similarly found TIPs in the Timor hybrid-derived material near genes of the *FAD2* and *TPS* families; it remains to be ascertained, however, whether these TIPs affect sensory characteristics in these cultivars.

Conclusions

Besides providing genomic resources for molecular breeding of one of the most important agricultural commodities, the *C. arabica* genome provides a unique window into the genome evolution of a recently formed allopolyploid stemming from two closely related species. Analysis of repetitive elements did not suggest a genomic shock, but in contrast, a higher LTR turnover rate in CA; this mechanism could possibly originate from CE, since CC demonstrates elevated numbers of LTRs when compared to other sequenced *Coffea* species. No evidence of genomic shock was observed in genome fractionation analyses either, since fractionation rates remained unaltered before and after the allopolyploidy event. Likewise, gene expression analyses showed no global subgenome dominance, but rather mosaic-type dominance effects, suggesting that Arabica biosynthetic pathways and regulatory networks form a mosaic dominated by either one of the subgenomes, similar to what has been observed in other neopolyploid crops such as rapeseed[11] and cotton[50]. However, we detected genome dominance in terms of biased homoeologous exchange from subCC to subEE; such asymmetry has been observed earlier, for example in octoploid strawberry[9]. Since *C. canephora* has one of the widest geographic ranges in the *Coffea* genus whereas *C. eugenioides* is more limited, the biased HE might be adaptive. This was also supported by the site frequency spectrum of HE loci, showing signs of directional selection (**Fig. S37**). Intriguingly, transposable insertion polymorphisms significantly overlap with tandem gene duplications and biosynthetic gene clusters, hinting at their possible role in gene cluster evolution.

The newly sequenced 39 accessions displayed a geographic split along the Eastern versus Western sides of the Great Rift Valley, with cultivated coffee variants all placed with the Eastern population. We identified admixture both in the wild representatives as well as in older, less developed cultivars. It is possible that this admixture has played a role in the development of cultivated coffee, since the old Typica variants, cultivated Ethiopian landraces, and the recent Geisha cultivar all showed varying levels of admixture. Bourbon and Typica variants were found to be parent and child, but these lineages still show clear phenotypic differences; this could be due to Typica being the result of an admixture event, as suggested by analysis of gene flow (**Fig. 3**). Admixture has played a large role in breeding many fruit-bearing cultivars, lychee perhaps being one of the most extreme cases[85].

Theory suggests that domestication imposes a certain associated cost on the crop species, with an increased genetic load from elevated levels of deleterious variants within the cultivar genomes[107]. Interestingly, breeding had not massively reduced the genetic diversity among cultivars, suggesting that at least in Arabica, a certain level of heterozygosity is needed in cultivated lines, probably due to advantages conferred by heterosis[31]. The cost of domestication in Arabica appears to be the reduced diversity of resistance genes, especially in families where selection favors high allelic diversity. This may have contributed to the susceptibility of *C. arabica* cultivars to coffee leaf rust disease. The analysis of five introgressed rust-resistant cultivars from two sister lineages of Timor hybrids allowed us to pinpoint a novel region harboring members of *RPP8* resistance gene family known for their allelic diversity, as well as a general regulator of resistance genes, *CPR1*. These results suggest a novel target locus for improving pathogen resistance in Arabica.

Data availability

Coffee genome assemblies are available at CoGe (https://genomevolution.org/): *C. canephora*: 50947, *C. eugenioides*: 60235, and *C. arabica*: 66663 (Pacbio HiFi) and 53628 (Pacbio). All genome information, including the VCF files with SNP information are available at ftp.solgenomics.net; the genome data is also available at ORCAE (https://bioinformatics.psb.ugent.be/orcae/overview/Coara and https://bioinformatics.psb.ugent.be/gdb/coffea_arabica/).
The sequencing data have been deposited to NCBI under bioproject ID: PRJNA698600.

TABLES

Table 1. Statistics of the *Coffea* assemblies presented in this paper.

Assembly	*C. eugenioides*	*C. canephora*	*C. arabica*	*C. arabica* HiFi
Projected genome size (Mb)*	682	705	1281	1281
Total assembly length (Mb)	661	672	1,088	1,198
% of projected genome	96.9%	95.3%	84.9%	93.5%
N scaffolds	253	3,033	8,474	132
Scaffold N50	61.3 Mb	50.1 Mb	32.7 Mb	53.7 Mb
N contigs	5,736	3,755	11,863	238**
Contig N50 (Mb)	0.40	0.76	0.23	30.0
Pseudochromosomes (Mb)	n.a.	583	801	1192
% of projected genome	n.a.	82.7%	62.5%	93.1%
N. genes	33,505	28,880	56,670	69,314
Genes in pseudochromosomes	n.a.	27,881	50,410	69,067
% genes in pseudochromosomes	n.a.	97%	89%	99.6%
BUSCO genome				
complete	96.7%	97.4%	97.6%	97.9%
single	88.5%	94.8%	20.1%	4.3 %
duplicated	8.2%	2.6%	77.5%	93.6 %
fragmented	1.1%	0.9%	0.8%	0.8 %
missing	2.2%	1.7%	1.6%	1.3 %
total	2,326	2,326	2,326	2,326
BUSCO annotation				
complete	94.9%	96.2%	92.1%	97.3%
single	82.4%	92.8%	33.3%	4.1%
duplicated	12.5%	3.4%	58.8%	93.2%
fragmented	2.1%	1.5%	2.8%	0.8%
missing	3.0%	2.3%	5.1%	1.9%
total	2,326	2,326	2,326	2,326

*From the Plant DNA C-values database: https://cvalues.science.kew.org/; **After gap filling; ***Denoeud et al, 2014; ****Scalabrin et al, 2020.

ACKNOWLEDGMENTS

The authors acknowledge the Natural History Museum in London for providing a sample of the *C. arabica* lectotype. JS acknowledges funding from Academy of Finland (decisions 318288 and 329441) and Nanyang Technological University start-up grant. RG and S.O-A acknowledge funding from Ecos-Nord N°C21MA01 and STICAMSUD 21-STIC-13. PR acknowledges Academy of Finland (grant 343656). ARP acknowledges NAPI Bioinformática from Fundação Araucária and TELearning Project 2021-22 (21-STIC-13) from STIC AmSud. YB acknowledges the funding from Research Foundation - Flanders (FWO, No G056517N). YVdP acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 833522) and from Ghent University (Methusalem funding, BOF.MET.2021.0005.01). GG acknowledges the Horizon Europe program, PRO-GRACE project (n. 101094738). DD and SMCL acknowledge São Paulo State Research Foundation (FAPESP), grant numbers #2016/10896-0 and #2017/01455-2. DS acknowledges the funding from NSERC and the Canada Research Chairs program. VAA acknowledges United States National Science Foundation grants 1442190 and 2030871. PD acknowledges funding from Nestlé Research. JS wishes to acknowledge the High Performance Computation Centre at NTU Singapore and University of Helsinki Linux administrators, as well as the CSC – IT Center for Science, Finland, for computational resources.

AUTHOR CONTRIBUTIONS

Conceived the study: AdK, DC, PD. Provided genetic resources: AA, AN, CK, EC, GHS, HR, LB, LFP, LP, MS, MTB, OGF, PaM, PM, PH, US. Carried out DNA sequencing: AC, CF, DM, GL, JeS, LB, MK, ND, PD, SM. Sequencing of the Linnaean accession: EG. Carried out genome assembly: SS, CW, JS, SP, LM. Genetic mapping: PR, MR, JS. Genome annotation: AR, SS, LM, JS, SR, VP, ZQW, DD, SIS, MM, RA, SMCL, ML, MP, CT-D, GG. Annotation of non-coding RNA: ARP, JE, PS. Transposable element annotation and analysis: SOA, AG, RG. Telomere identification: VAA, WCM. Analyzed genome evolution: ZY, ZC, DSM, RG, JM, DS, LC-P, TL, TK, VAA, SOA, AG, JS. Gene family analysis: ZQW, VP, DD, GG, SF, VAA, SiR, JS. RNA-seq data analysis: AR, SP, SiR, JS. Provided RNA-seq data: RH. Analyzed population data: JS. Analysed GBS data: YB, RG. Arranged online data access: LM, SR. Wrote the first draft: JS, completed with input from DS, VAA, LFP, GG, RG, SR, AdK, PD, VP, LM, DC, DD, SP, AA, as well as PM, YB, TR, YVdP, and all co-authors.

COMPETING INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nature Reviews Genetics* **18**, 411-424 (2017).
2. Van de Peer, Y., Ashman, T.-L., Soltis, P.S. & Soltis, D.E. Polyploidy: an evolutionary and ecological force in stressful times. *The Plant Cell* **33**, 11-26 (2021).
3. Leebens-Mack, J.H. *et al.* One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679-685 (2019).
4. Sattler, M.C., Carvalho, C.R. & Clarindo, W.R. The polyploidy and its key role in plant breeding. *Planta* **243**, 281-296 (2016).
5. Sun, H. *et al.* Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nature Genetics* **54**, 342-348 (2022).
6. Athiyannan, N. *et al.* Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning. *Nature Genetics* **54**, 227-231 (2022).
7. Wu, S. *et al.* Genome sequences of two diploid wild relatives of cultivated sweetpotato reveal targets for genetic improvement. *Nature Communications* **9**, 4580 (2018).
8. Wang, T. *et al.* A complete gap-free diploid genome in Saccharum complex and the genomic footprints of evolution in the highly polyploid Saccharum genus. *Nature Plants* **9**, 554-571 (2023).
9. Edger, P.P. *et al.* Origin and evolution of the octoploid strawberry genome. *Nature Genetics* **51**, 541-547 (2019).
10. Li, F. *et al.* Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. *Nature Biotechnology* **33**, 524-530 (2015).
11. Chalhoub, B. *et al.* Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* **345**, 950-953 (2014).
12. McClintock, B. The Significance of Responses of the Genome to Challenge. *Science* **226**, 792-801 (1984).
13. Sha, Y. *et al.* Genome shock in a synthetic allotetraploid wheat invokes subgenome-partitioned gene regulation, meiotic instability, and karyotype variation. *Journal of Experimental Botany*, erad247 (2023).
14. Thomas, B.C., Pedersen, B. & Freeling, M. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research* **16**, 934-946 (2006).
15. Schnable, J.C., Springer, N.M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences* **108**, 4069 (2011).

16. Gaeta, R.T., Pires, J.C., Iniguez-Luy, F., Leon, E. & Osborn, T.C. Genomic Changes in Resynthesized Brassica napus and Their Effect on Gene Expression and Phenotype. *The Plant Cell* **19**, 3403-3417 (2007).
17. Burns, R. *et al.* Gradual evolution of allopolyploidy in Arabidopsis suecica. *Nature Ecology & Evolution* **5**, 1367-1381 (2021).
18. Conant, G.C., Birchler, J.A. & Pires, J.C. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Current Opinion in Plant Biology* **19**, 91-98 (2014).
19. Carvalho, A. *et al.* Melhoramento do cafeeiro: IV - Café Mundo Novo. *Bragantia* **12**, 97-130 (1952).
20. Scalabrin, S. *et al.* A single polyploidization event at the origin of the tetraploid genome of *Coffea arabica* is responsible for the extremely low genetic variation in wild and cultivated germplasm. *Sci Rep* **10**, 4642 (2020).
21. Cenci, A., Combes, M.-C. & Lashermes, P. Genome evolution in diploid and tetraploid *Coffea* species as revealed by comparative analysis of orthologous genome segments. *Plant Molecular Biology* **78**, 135-145 (2012).
22. Bawin, Y. *et al.* Phylogenomic analysis clarifies the evolutionary origin of *Coffea arabica*. *Journal of Systematics and Evolution* **59**, 953-963 (2020).
23. Yu, Q. *et al.* Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (*Coffea*). *The Plant Journal* **67**, 305-317 (2011).
24. Merot-L'anthoene, V. *et al.* Development and evaluation of a genome-wide Coffee 8.5K SNP array and its application for high-density genetic mapping and for investigating the origin of *Coffea arabica* L. *Plant Biotechnology Journal* **17**, 1418-1430 (2019).
25. Wellman, F.L. *Coffee: botany, cultivation and utilization*, (L. Hill, London, 1961).
26. Lécolier, A., Besse, P., Charrier, A., Tchakaloff, T.-N. & Noirot, M. Unraveling the origin of *Coffea arabica* 'Bourbon pointu' from La Réunion: a historical and scientific perspective. *Euphytica* **168**, 1-10 (2009).
27. Clarindo, W.R., Carvalho, C.R., Caixeta, E.T. & Koehler, A.D. Following the track of "Híbrido de Timor" origin by cytogenetic and flow cytometry approaches. *Genetic Resources and Crop Evolution* **60**, 2253-2259 (2013).
28. Bertrand, B., Guyot, B., Anthony, F. & Lashermes, P. Impact of the *Coffea canephora* gene introgression on beverage quality of *C. arabica*. *Theoretical and Applied Genetics* **107**, 387-394 (2003).
29. Talhinhas, P. *et al.* The coffee leaf rust pathogen *Hemileia vastatrix*: one and a half centuries around the tropics. *Molecular Plant Pathology* **18**, 1039-1051 (2017).
30. World_Coffee_Research. Coffee leaf rust resistant coffee variety overcome in Honduras. in *WCR NEWS* (2017).
31. Marie, L. *et al.* G × E interactions on yield and quality in *Coffea arabica*: new F1 hybrids outperform American cultivars. *Euphytica* **216**, 78 (2020).
32. Bertrand, B., Villegas Hincapié, A.M., Marie, L. & Breitler, J.-C. Breeding for the main agricultural farming of *Arabica* coffee. *Frontiers in Sustainable Food Systems* **5**(2021).
33. Breitler, J.-C. *et al.* CRISPR/Cas9-mediated efficient targeted mutagenesis has the potential to accelerate the domestication of *Coffea canephora*. *Plant Cell, Tissue and Organ Culture (PCTOC)* **134**, 383-394 (2018).
34. Berthaud, J. Etude cytogénétique d'un haploïde de *Coffea arabica* L. *Café, Cacao, Thé (Francia) v. 20 (2) p. 91-96* (1976).
35. Denoeud, F. *et al.* The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345**, 1181-1184 (2014).
36. Pellicer, J. & Leitch, I.J. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytologist* **226**, 301-305 (2020).
37. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A. & Zdobnov, E.M. BUSCO

Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **38**, 4647-4654 (2021).
38. Petit, M. *et al.* Mobilization of retrotransposons in synthetic allotetraploid tobacco. *New Phytologist* **186**, 135-147 (2010).
39. Sarilar, V. *et al.* Allopolyploidy has a moderate impact on restructuring at three contrasting transposable element insertion sites in resynthesized Brassica napus allotetraploids. *New Phytologist* **198**, 593-604 (2013).
40. Yu, Z., Zheng, C., Albert, V.A. & Sankoff, D. Excision dominates pseudogenization during fractionation after whole genome duplication and in gene loss after speciation in plants. *Frontiers in Genetics* **11**(2020).
41. Birchler, J.A. & Veitia, R.A. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *The New phytologist* **186**, 54-62 (2010).
42. Zeiss, D.R., Piater, L.A. & Dubery, I.A. Hydroxycinnamate Amides: Intriguing Conjugates of Plant Protective Metabolites. *Trends in Plant Science* **26**, 184-195 (2021).
43. Salojärvi, J. *et al.* Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature Genetics* **49**, 904-912 (2017).
44. Myburg, A.A. *et al.* The genome of *Eucalyptus grandis*. *Nature* **510**, 356-362 (2014).
45. Rendón-Anaya, M. *et al.* The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. *Proceedings of the National Academy of Sciences* **116**, 17081 (2019).
46. Lan, T. *et al.* Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. *Proceedings of the National Academy of Sciences* **114**, E4435 (2017).
47. Bird, K.A., VanBuren, R., Puzey, J.R. & Edger, P.P. The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytologist* **220**, 87-93 (2018).
48. Göbel, U. *et al.* Robustness of transposable element regulation but no genomic shock observed in interspecific *Arabidopsis* hybrids. *Genome Biology and Evolution* **10**, 1403-1415 (2018).
49. Pfeifer, M. *et al.* Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science* **345**, 1250091 (2014).
50. Yoo, M.J., Szadkowski, E. & Wendel, J.F. Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* **110**, 171-180 (2013).
51. Bird, K.A. *et al.* Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid Brassica napus. *New Phytologist* **230**, 354-371 (2021).
52. Ashihara, H., Mizuno, K., Yokota, T. & Crozier, A. Xanthine Alkaloids: Occurrence, Biosynthesis, and Function in Plants. in *Progress in the Chemistry of Organic Natural Products 105* (eds. Kinghorn, A.D., Falk, H., Gibbons, S. & Kobayashi, J.i.) 1-88 (Springer International Publishing, Cham, 2017).
53. Xia, E.-H. *et al.* The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. *Molecular Plant* **10**, 866-877 (2017).
54. Xu, Z. *et al.* Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants. *BMC Biology* **18**, 63 (2020).
55. Ashihara, H. & Crozier, A. Biosynthesis and catabolism of caffeine in low-caffeine-containing species of *Coffea*. *Journal of Agricultural and Food Chemistry* **47**, 3425-3431 (1999).
56. Campa, C., Doulbeau, S., Dussert, S., Hamon, S. & Noirot, M. Diversity in bean caffeine content among wild *Coffea* species: evidence of a discontinuous distribution. *Food Chemistry* **91**, 633-637 (2005).
57. Ashihara, H. Metabolism of alkaloids in coffee plants. *Brazilian Journal of Plant Physiology* **18**, 1-8 (2008).
58. Del Terra, L. *et al.* Functional characterization of three *Coffea arabica* L.

monoterpene synthases: Insights into the enzymatic machinery of coffee aroma. *Phytochemistry* **89**, 6-14 (2013).
59. Speer, K. & Kölling-Speer, I. The lipid fraction of the coffee bean. *Brazilian Journal of Plant Physiology* **18**, 201-216 (2006).
60. Hernández, M.L., Mancha, M. & Martínez-Rivas, J.M. Molecular cloning and characterization of genes encoding two microsomal oleate desaturases (FAD2) from olive. *Phytochemistry* **66**, 1417-1426 (2005).
61. Dar, A.A., Choudhury, A.R., Kancharla, P.K. & Arumugam, N. The *FAD2* gene in plants: occurrence, regulation, and role. *Frontiers in Plant Science* **8**(2017).
62. Guo, H. *et al.* Extensive and Biased Intergenomic Nonreciprocal DNA Exchanges Shaped a Nascent Polyploid Genome, Gossypium (Cotton). *Genetics* **197**, 1153-1163 (2014).
63. Deb, S.K., Edger, P.P., Pires, J.C. & McKain, M.R. Patterns, mechanisms, and consequences of homoeologous exchange in allopolyploid angiosperms: a genomic and epigenomic perspective. *New Phytologist* **238**, 2284-2304 (2023).
64. Cheng, B., Furtado, A. & Henry, R.J. The coffee bean transcriptome explains the accumulation of the major bean components through ripening. *Scientific Reports* **8**, 11414 (2018).
65. Perrois, C. *et al.* Differential regulation of caffeine metabolism in *Coffea arabica* (*Arabica*) and *Coffea canephora* (Robusta). *Planta* **241**, 179-191 (2015).
66. Vidal, R.O. *et al.* A high-throughput data mining of single nucleotide polymorphisms in *Coffea* species expressed sequence tags suggests differential homeologous gene expression in the allotetraploid *Coffea arabica*. *Plant Physiology* **154**, 1053-1066 (2010).
67. Marraccini, P. Gene expression in coffee. in *Progress in botany* (ed. Cánovas F.M, L.U.R.M.C.P.H.) 43-111 (Springer, Cham, Suisse, 2020).
68. Meyer, F.G., Fernie, L.M., Narasimhaswami, R.L., Monaco, L.C. & Greathead, D.J. FAO Coffee Mission to Ethiopia, 1964-1965. (1968).
69. Halle, F. Echantillonnage du matériel *Coffea arabica* récolté en Ethiopie. *Bulletin - IFCC*, 13-18 (1978).
70. Krug, C.A.M., A.J.T. Cytological observations in *Coffea* – IV. *J Genet* **39**, 189–203 (1940).
71. Chalhoub, B. *et al.* Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* **345**, 950-953 (2014).
72. Lashermes, P. *et al.* Molecular characterisation and origin of the Coffea arabica L. genome. *Molecular and General Genetics MGG* **261**, 259-266 (1999).
73. Wu, Y. *et al.* Genomic mosaicism due to homoeologous exchange generates extensive phenotypic diversity in nascent allopolyploids. *National Science Review* **8**, nwaa277 (2021).
74. Terhorst, J., Kamm, J.A. & Song, Y.S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics* **49**, 303-309 (2017).
75. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496 (2011).
76. Xie, Z. *et al.* Mutation rate analysis via parent–progeny sequencing of the perennial peach. I. A low rate in woody perennials and a higher mutagenicity in hybrids. *Proceedings of the Royal Society B: Biological Sciences* **283**, 20161016 (2016).
77. Moat, J., Gole, T.W. & Davis, A.P. Least concern to endangered: Applying climate change projections profoundly influences the extinction risk assessment for wild *Arabica* coffee. *Global Change Biology* **25**, 390-403 (2019).
78. Lambeck, K. *et al.* Sea level and shoreline reconstructions for the Red Sea: isostatic and tectonic considerations and implications for hominin migration out of Africa. *Quaternary Science Reviews* **30**, 3542-3574 (2011).
79. Kuper, R. & Kröpelin, S. Climate-controlled holocene occupation in the Sahara: motor of Africa's evolution. *Science* **313**, 803-807 (2006).

80. Bashir, T. *et al.* Hybridization Alters Spontaneous Mutation Rates in a Parent-of-Origin-Dependent Fashion in Arabidopsis *Plant Physiology* **165**, 424-437 (2014).
81. Salojärvi, J. *et al.* Author Correction: Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature Genetics* **51**, 1187-1189 (2019).
82. Excoffier, L. *et al.* fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics* **37**, 4882-4885 (2021).
83. Montagnon, C., Mahyoub, A., Solano, W. & Sheibani, F. Unveiling a unique genetic diversity of cultivated *Coffea arabica* L. in its main domestication center: Yemen. *Genetic Resources and Crop Evolution* **68**, 2411-2422 (2021).
84. Nordborg, M. & Donnelly, P. The coalescent process with selfing. *Genetics* **146**, 1185 (1997).
85. Hu, G. *et al.* Two divergent haplotypes from a highly heterozygous lychee genome suggest independent domestication events for early and late-maturing cultivars. *Nature Genetics* **54**, 73-83 (2022).
86. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010).
87. Lan, T. *et al.* Insights into bear evolution from a Pleistocene polar bear genome. *Proceedings of the National Academy of Sciences* **119**, e2200016119 (2022).
88. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).
89. Molloy, E.K., Durvasula, A. & Sankararaman, S. Advancing admixture graph estimation via maximum likelihood network orientation. *Bioinformatics* **37**, i142-i150 (2021).
90. Pfeifer, B. & Kapan, D.D. Estimates of introgression as a function of pairwise distances. *BMC Bioinformatics* **20**, 207 (2019).
91. Meyer, R.S. & Purugganan, M.D. Evolution of crop species: genetics of domestication and diversification. *Nature Reviews Genetics* **14**, 840-852 (2013).
92. Gaut, B.S., Seymour, D.K., Liu, Q. & Zhou, Y. Demography and its effects on genomic variation in crop domestication. *Nature Plants* **4**, 512-520 (2018).
93. Jousimo, J. *et al.* Ecological and evolutionary effects of fragmentation on infectious disease dynamics. *Science* **344**, 1289-1293 (2014).
94. dos Santos, T.B., Baba, V.Y., Vieira, L.G.E., Pereira, L.F.P. & Domingues, D.S. The urea transporter DUR3 is differentially regulated by abiotic and biotic stresses in coffee plants. *Physiology and Molecular Biology of Plants* **27**, 203-212 (2021).
95. Wang, W. *et al.* Structural basis of salicylic acid perception by Arabidopsis NPR proteins. *Nature* **586**, 311-316 (2020).
96. Mukhtar, M.S. *et al.* Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science* **333**, 596-601 (2011).
97. Cooley, M.B., Pathirana, S., Wu, H.J., Kachroo, P. & Klessig, D.F. Members of the *Arabidopsis* HRT/RPP8 family of resistance genes confer resistance to both viral and oomycete pathogens. *The Plant cell* **12**, 663-676 (2000).
98. Mohr, T.J. *et al.* The *Arabidopsis* downy mildew resistance gene *RPP8* is induced by pathogens and salicylic acid and is regulated by W-box *cis* elements. *Molecular Plant-Microbe Interactions®* **23**, 1303-1315 (2010).
99. MacQueen, A. *et al.* Population genetics of the highly polymorphic *RPP8* gene family. *Genes* **10**(2019).
100. Cheng, Y.T. *et al.* Stability of plant immune-receptor resistance proteins is controlled by SKP1-CULLIN1-F-box (SCF)-mediated protein degradation. *Proceedings of the National Academy of Sciences*, 201105685 (2011).
101. Hedtmann, C. *et al.* The Plant Immunity Regulating F-Box Protein CPR1 Supports Plastid Function in Absence of Pathogens. *Frontiers in Plant Science* **8**(2017).
102. Florez, J.C. *et al.* High throughput transcriptome analysis of coffee reveals prehaustorial resistance in response to *Hemileia vastatrix* infection. *Plant Molecular Biology* **95**, 607-623 (2017).

103. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
104. Domínguez, M. *et al.* The impact of transposable elements on tomato diversity. *Nature Communications* **11**, 4058 (2020).
105. Polturak, G. & Osbourn, A. The emerging role of biosynthetic gene clusters in plant defense and plant interactions. *PLOS Pathogens* **17**, e1009698 (2021).
106. Polturak, G., Liu, Z. & Osbourn, A. New and emerging concepts in the evolution and function of plant biosynthetic gene clusters. *Current Opinion in Green and Sustainable Chemistry* **33**, 100568 (2022).
107. Gaut, B.S., Díez, C.M. & Morrell, P.L. Genomics and the Contrasting Dynamics of Annual and Perennial Domestication. *Trends in Genetics* **31**, 709-719 (2015).

Chapter 4

A Bioinformatics Tool for Efficient Retrieval of High-Confidence Terpene Synthases (TPS) and Application to the Identification of TPS in *Coffea* and *Quillaja*

Douglas S. Domingues, Liliane S. Oliveira, Samara M. C. Lemos, Gian C. C. Barros, and Suzana T. Ivamoto-Suzuki

Abstract

Terpenoids are a class of compounds that are found in all living organisms. In plants, some terpenoids are part of primary metabolism, but most terpenes found in plants are classified as specialized metabolites, encoded by terpene synthases (TPS). It is not obvious how to assign the putative product of a given TPS using bioinformatics tools. Phylogenetic analyses easily assign TPS into families; however members of the same TPS family can synthetize more than one terpenoid—and, in many biotechnological applications, researchers are more interested in the product of a given TPS rather than its phylogenetic profile. Automated protein annotation can be used to classify TPS based on their products, despite the family they belong to. Here, we implement an automated bioinformatics method, search_TPS, to identify TPS proteins that synthesize mono, sesqui and diterpenes in Angiosperms. We verified the applicability of the method by classifying wet lab validated TPS and applying it to find TPS proteins in *Coffea arabica*, *C. canephora*, *C. eugenioides*, and *Quillaja saponaria*. Search_TPS is a computational tool based on PERL scripts that carries out a series of HMMER searches against a curated database of TPS profile hidden Markov models. The tool is freely available at https://github.com/liliane-sntn/TPS.

Key words Terpene synthase, Monoterpene, Diterpene, Sesquiterpene, Profile HMMs

1 Introduction

1.1 Overall Context

Terpenes, also known as terpenoids or isoprenoids, comprise the most diverse family of natural products in plants, comprising more than 80,000 compounds [1], involved in attracting pollinators and seed dispersers, in defense against pathogens and herbivores, and in attracting useful soil microorganisms (revised in [2]).

The structural diversity associated with these products of secondary metabolism highlight them as impressive examples of the

Arthur Germano Fett-Neto (ed.), *Plant Secondary Metabolism Engineering: Methods and Protocols*, Methods in Molecular Biology, vol. 2469, https://doi.org/10.1007/978-1-0716-2185-1_4,

divergent evolution in plant metabolites. The evolutionary success of this compound class is in part based on the simplicity of constructing different size molecules. All terpenoids are derived from two five-carbon "building blocks," the isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP). The prenyl diphosphate intermediates built by condensation of these five-carbon units are used as precursors for the biosynthesis of a large number of terpenoids with specialized roles in the interaction of plants and their environment [3]. Specialized terpenoids have a long history of being used as flavors, fragrances, pharmaceuticals, insecticides, and industrial compounds [4].

The tremendous structural diversity of terpenoids in plants is a consequence of divergent biosynthetic gene evolution. There is a single gene encoding a TPS in the bryophyte *Physcomitrella patens*; however, we can find more than 100 genes in some angiosperms, like *Eucalyptus* species [5–7]. In terms of protein subcellular localization, TPS are distributed in plastids, cytosol and mitochondria [8].

Based on structure and biochemical properties, TPS enzymes can be divided in two types, type I and type II [1]. These types are defined by the presence of one to three conserved helical domains: alpha, beta and gamma [9]. The single TPS gene in *Physcomitrella* is a bifunctional TPS containing the three helical domains; duplication, loss of function in single domains and neofunctionalization probably created type-specific TPSs [7].

An extensive phylogenetic analysis of plant TPS [5] has divided them into subfamilies TPS-a to TPS-h. Type I TPS sequences form clades TPS-a, TPS-b, and TPS-d (gymnosperm-specific); TPS-e/f, TPS-g, and TPS-h (specific to *Selaginella* spp.); type II TPSs form clade TPS-c [7]. Based on the structural and evolutionary relationships among terpene synthases (TPSs) proposed by Katunanithi and Zerbe [9], we present here a bioinformatics tool to identify Type I TPS, of subfamilies c, b, d, and e/f in mono and eudicots (*see* **Note 1**). These TPS generate monoterpenes (10 carbon terpenoids), sesquiterpenes (15 carbon terpenoids) and several diterpenes (20 carbon terpenoids).

Classification of TPS in seven subfamilies (a to g) are well-established and easily retrieved under standard phylogenetic analyses and tools, like Terzyme [10]. However, the product of a TPS is not defined solely by phylogenetic subfamilies; for instance, diterpenes can be synthetized either by TPS from subfamilies a, b, and c [9, 11, 12]. In terms of biotechnological uses, most researchers are more interested in defining products of TPS rather than classify them phylogenetically. The determination of TPS products relies on a gene-by-gene analysis, including the expression in heterologous systems, which in an initial screening is a very laborious task. In this sense, a bioinformatics tool, with openly distributed profiles and data, would help researchers to depict the TPS universe

in plants that have genomic and transcriptomic data, but no extensive "terpenome" analysis.

The tool search_TPS was then developed to fill up this gap, identifying monoterpene synthases (monoTPS), sesquiterpene synthases (sesquiTPS) and diterpene synthases (diTPS) in FASTA sequences, available at https://github.com/liliane-sntn/TPS.

1.2 Model Construction

Our tool is an improvement of the Terzyme [10] classification scheme. We downloaded curated TPS protein sequences of monoTPS, diTPS, and sesquiTPS classes indicated by Terzyme database [10]. In total, 400 proteins were obtained (153 monoTPS, 176 sesquiTPS, and 71 diTPS). After redundancy analyses made by in-house PERL scripts, 4 monoTPS and 1 sesquiTPS were discarded, and the final dataset consisted of 395 proteins (149 monoTPS, 175 sesquiTPS, and 71 diTPS). The final dataset was submitted to multiple sequence alignment (MSA) using MUSCLE [13] with default parameters. The constructed multiple sequence alignment (MSA) was used as an input to build a maximum likelihood phylogenetic tree using iqtree [14, 15] with default parameters (this tree is available at https://doi.org/10.5281/zenodo.4542419). As expected, mono, di, and sesquiTPS proteins did not result in product specific clades. A manual inspection of the phylogenetic tree was performed to generate protein clusters for all classes. In total, 9, 20, and 8 clusters were generated for the monoTPS, sesquiTPS, and diTPS, respectively. These clusters were then analyzed for typical PFAM [16] domains for TPS: PF01397 (Terpene synthase N terminal domain), PF03936 (Terpene synthase family, metal binding domain), PF19086 (Terpene synthase C terminal domain), using the hmmsearch program (HMMER package [17]) with default parameters (*see* **Note 2** for further considerations on that). TPS containing at least 2 of these 3 domains were considered for the construction of our tool [Fig. 1]. The similarity regions of the TPS protein sequences detected by the PFAM models were extracted and stored according to their respective PFAM model. Redundant sequences were removed again, and the remaining sequences were used as input to build class-specific HMM profiles using the hmmbuild program (HMMER package) with default parameters. As a result, 27, 60, and 24 class-specific profile HMMs were constructed for monoTPS, sesquiTPS, and diTPS, respectively. These profiles had their score cutoff for high confidence TPS based on the analysis of 395 Terzyme [10] curated TPS. All steps for the construction of class-specific TPS were summarized in Fig. 1.

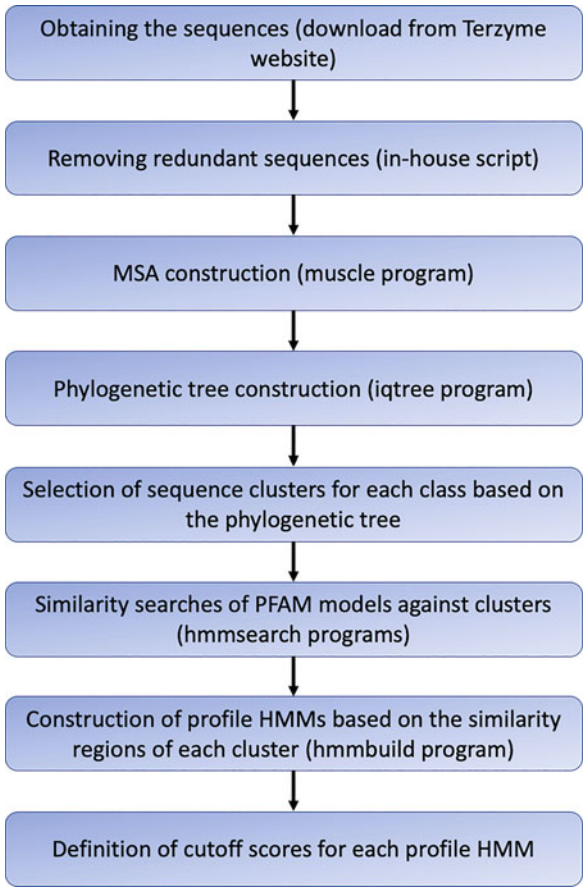

Fig. 1 Steps used to build specific profile HMM for mono-, sesqui-, and diterpene synthases in search_TPS

2 Materials

The tool search_TPS reads models and it works in two steps: search_TPS uses Hidden Markov model (HMM) protein profiles to firstly search proteins that contain typical PFAM TPS domains. If proteins contain at least 2 out of 3 PFAM domains for TPS, then the tool identifies if selected proteins respect the properties of specific models for monoTPS, sesquiTPS, and diTPS (*see* **Note 2**).

2.1 Hardware, System, and Sequence Data

A personal computer or workstation is needed.

The program search_TPS is developed in Perl language. It can be used in a POSIX-compliant operating system (for example: UNIX and Linux distributions) with an installed

Perl interpreter (http://www.perl.org). It analyzes protein sequences provided in one single file. The input file must be a FASTA format (https://en.wikipedia.org/wiki/FASTA_format)

file containing the protein sequence(s) to be analyzed (*see* **Note 3** for further considerations on this).

2.2 Input, Software and Availability

The HMM protein profiles used by search_TPS are available at https://github.com/liliane-sntn/TPS/tree/main/Tutorial. They are described in Introduction.

To run search_TPS, the user needs to install the HMMER3 package (www.hmmer.org) to perform similarity searches of the profile HMMs against the input sequences. The hmmsearch program must be located in a directory listed in the PATH of the operating system. Program search_TPS itself does not need to be installed, the user should only download the search_TPS.pl file available at https://github.com/liliane-sntn/TPS

3 Method

The program receives as input a file containing the sequences of interest in FASTA format, a directory composed of the PFAM models of terpene synthases, a directory composed of class-specific profile HMMs, and a tabular file containing the cut-off score values for each class-specific profile HMM. In Introduction, we specify how we determined HMM scores to identify high-confidence TPS proteins (*see* **Note 4**). In default execution, search_TPS will use three TPS models from PFAM database (PF01397, PF03936, and PF19086) and 27, 60, and 24 profile HMMs specific for monoTPS, sesquiTPS, and diTPS classes, respectively. To identify low confidence TPS sequences, the program uses a standard score value (score = 100), but this value can be set by the user (for more details *see* **Note 5**). In Subheading 3.1, we detail how we determined if parameters were correct [Fig. 2].

The hmmsearch program must be located in a directory listed in the PATH of the operating system.

The basic command for usage is:

```
search_TPS.pl -d <class-scpecific_phmms_dir> -i
<input_fasta_file> -t <table_file> -s <search_option> -p
<pfams_dir> -o <output_dir>
```

Example:

```
search_TPS.pl -d class-specific_hmms -i sequences.fasta -t
score_tables_dir/all.scores -s 1 -p PFAMs_dir -o res
```

Five parameters are mandatory: class-specific HMMs directory, an input FASTA file, a directory for PFAM models, groups of TPS to be searched, and the table file containing cutoff scores for high-confidence TPS:

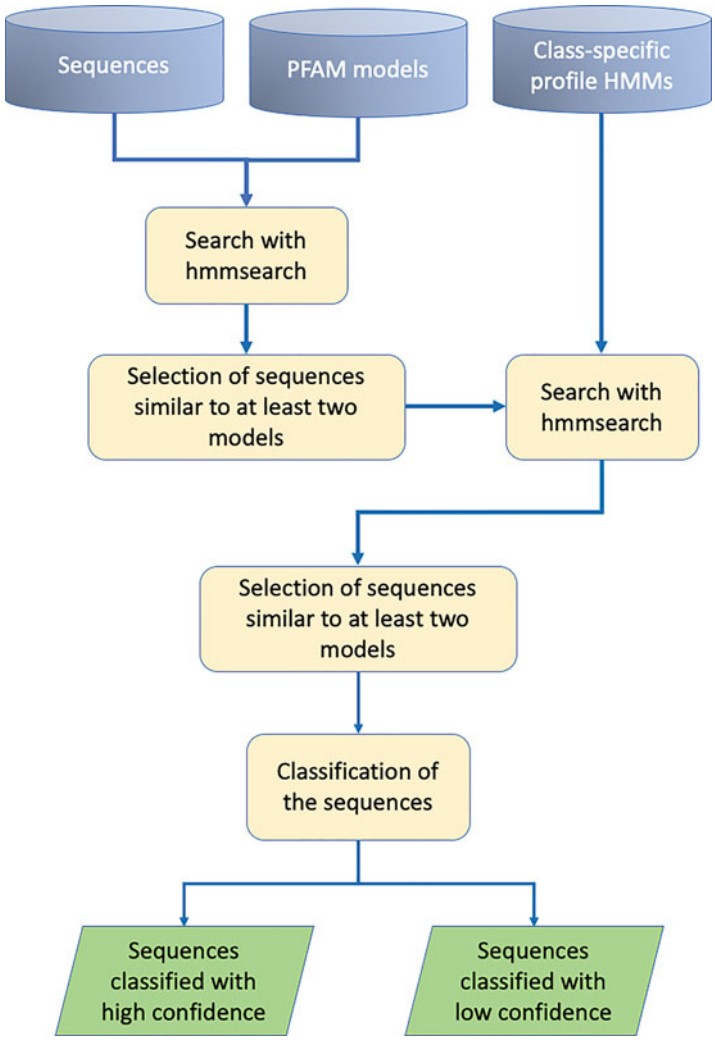

Fig. 2 search_TPS workflow. (**a**) As a first step, search_TPS performs similarity searches using the hmmsearch program of PFAM TPS models against the input strings. Sequences that show similarity with at least two models are selected for the next step. (**b**) The selected sequences are then used as input for similarity searches with hmmsearch using the class-specific profile HMMs. Sequences that present similarity with at least two specific profile HMMs for the same class are selected. (**c**) Finally, the selected sequences are classified based on the highest score value obtained in the similarity searches

- -d <class-scpecific_phmms_dir> : Directory containing specific profile HMMs for monoTP, diTP and sesquiTP.
- -i <input_fasta_file> : Fasta file containing the sequences to be searched.
- -p <pfams_dir> : Directory containing the PFAMs models.
- -s <search_options> : Group of TPS to be searched:

1 - all (monoTP, diTP and sesquiTP)

2 - monoTP

3 - diTP

4 - sesquiTP

– -t <table_file> : File containing the cutoff scores to be used to select the results of each HMM profile.

Optionally, users can select the number of processors to be used, change the low confidence parameter and set output directory. Help and version can be also shown:

– -cpu <num> : Number of threads to be used by hmmsearch.
– -h|help : Show this help message.
– -l <low_confidence_score> : Low confidence score to be considered (default = 100).
– -o <dir> : Output directory (default = output_dir).
– -v|version : Version.

Figure 2 summarizes the main steps of the method.

search_TPS program execution generates several files and subdirectories in an output directory:

– file.log: file that reports the steps of the execution.
– error.log: file containing the error messages. If the execution ends without errors, this file remains empty.
– PFAM_dir: this directory contains files generated by hmmsearch execution with the PFAM models, a tabular file listing the header of the selected sequences and a FASTA file containing the selected sequences.

The subdirectories monoTP_dir, diTP_dir, and sesquiTP_-dir contain the following files.

– hmmsearch.txt: this file contains all results found by the execution of hmmsearch for each profile HMM.
– hmmsearch_high.tab: this file contains the results found by the execution of hmmsearch for each profile HMM selected by the cutoff score of the model (high confidence results).
– hmmsearch_low.tab: this file contains the results found by the execution of hmmsearch program for each profile HMM selected by the low confidence cutoff score (low confidence results).
– high_confidence_results.csv: this file contains all sequences that presented similarity with at least two models of the same class with high confidence. In this case, a sequence could be classified in more than one class with high and low confidence.

- low_confidence_results.csv: this file contains all sequences that presented similarity with at least two models of the same class with low confidence. In this case, a sequence could be classified in more than one class with high and low confidence.
- high_confidence_final_results.csv: this file contains the sequences that presented similarity with at least two models of the same class with high confidence without redundance. In this case, a sequence is classified in only one class, and it is not present in the final file of sequences classified with low confidence.
- low_confidence_final_results.csv: this file contains the sequences that presented similarity with at least two models of the same class with low confidence without redundance. In this case, a sequence is classified in only one class (monoTPS, sesquiTPS, or diTPS), and it is not present in the final file of sequences classified with high confidence.

3.1 Tool Validation

In order to evaluate search_TPS performance, we reannotated TPS from *Arabidopsis thaliana*, *Solanum lycopersicum*, and *Setaria italica*. For all species we used the most recent genome version available at Phytozome (https://phytozome.jgi.doe.gov/) and compared with recent publications that delivered functional annotation of TPS for these species [7, 18].

Using search_TPS we were able to recover all annotated TPS in the three species, including alternative splicing isoforms (Information available at https://doi.org/10.5281/zenodo.4542419). We also annotated extra putative TPS genes not previously identified in these genomes: one in *Arabidopsis* (*ATG1G48820.1*), one in tomatoes (*Solyc07g052135.1*). Additionally, in *S. italica*, we identified 17 new TPS splicing variants and 4 new TPS genes (*Seita.6G088000.1*, *Sect.8G232200.1*, *Sect.9G405300.1*, *Sect.9G448900*).

Therefore, search_TPS is an effective tool to rapidly retrieve TPS genes from a given dataset (*see* **Notes 5**–**7**). In high-confidence TPS, search_TPS is very effective to determine TPS classes (*see* **Note 7**): 14 out of 17 Arabidopsis high-confidence TPS genes were correctly attributed to product class, and 15 out of 21 tomato high-confidence TPS genes were correctly classified. Detailed results are provided in https://doi.org/10.5281/zenodo.4542419 and https://github.com/liliane-sntn/TPS.

3.2 Application of search_TPS

In order to demonstrate the applicability of search_TPS in datasets where there is not any detailed classification, we identified TPS genes in four species that did not have TPS described by any gene annotation process: *Coffea arabica*, *C. canephora*, *C. eugenioides*, and *Quillaja saponaria*. The first three species have genomic data available at NCBI Genomes, and, for *Q. saponaria*, we assembled public RNA-seq data (*see* **Note 6**). All information is also provided

at https://doi.org/10.5281/zenodo.4542419 and https://github.com/liliane-sntn/TPS.

We identified 49 TPS genes in *Coffea canephora*, 64 genes in *C. eugenioides*, and 93 genes in *C. arabica*. In *Quillaja saponaria*, we identified 11 genes. Among those TPS coding sequences, we identified 32, 54, 39, and 5 high-confidence TPS sequences in *C. canephora*, *C. eugenioides*, *C. arabica*, and *Q. saponaria*, respectively. These are an initial group of synthases for further exploration in functional analyses. For practical rules to improve use of search_TPS, *see* **Notes 7** and **8**.

Detailed characterization of protein size, subcellular localization and overall classification is available at https://doi.org/10.5281/zenodo.4542419.

4 Notes

1. search_TPS was developed with the main aim of characterizing genomic, transcriptional, and proteomic resources of angiosperms—especially monocotyledons and eudicotyledons. Thus, we do not know the performance of search_TPS in other taxonomic groups.
2. One advantage of search_TPS is that the first step of the pipeline is composed by selecting proteins containing at least 2 typical TPS domains. This feature prevents search_TPS from annotating pseudogenes and incomplete sequences, which might occur in Terzyme [10].
3. Since search_TPS needs a FASTA file containing proteins, in the case of de novo assembled transcripts users might use Trinotate (http://trinotate.github.io/) or CodAn [19] to obtain translated coding sequences.
4. search_TPS output is based on the best score result of monoTPS, sesquiTPS, and diTPS, when all models run simultaneously. We strongly advise users to test using search_TPS with each TPS model separately.
5. Low-confidence prediction was arbitrarily determined at the score of 100. Users might test other score cutoffs and/or check e-values of hmmsearch in order to better determine the lowest score to classify a TPS.
6. It is our understanding that the main application of the program is to characterize the assembly of new plant transcriptomes. Hence, we reassembled the *Q. saponaria* transcriptome using public data (SRA-NCBI accession number ERX651069) following the "Best Practices for De Novo Transcriptome Assembly with Trinity" (https://informatics.fas.harvard.edu/best-practices-for-de-novo-transcriptome-assembly-with-

trinity.html) for a case study of search_TPS. Detailed results of the assembly are available at https://doi.org/10.5281/zenodo.4542419 .

7. As a rule of thumb, we expect that search_TPS in default parameters would retrieve all TPS coding sequences from a given dataset in Angiosperms. We also expect that high-confidence TPS classification would correctly classify the most probable product from a given TPS. Therefore, we believe high-confidence TPS from search_TPS are the most relevant TPS genes to be selected for a functional characterization, including expression in heterologous systems.
8. Checking protein size, subcellular localization, and phylogenetic analysis are still relevant steps to further characterize TPS genes. For the four species we used here as proof of concept, as well as the three species with well-characterized TPS, we deliver all analyses we made with search_TPS and other tools at https://doi.org/10.5281/zenodo.4542419 and https://github.com/liliane-sntn/TPS.

Acknowledgments

This work was partially financed by the São Paulo State Research Foundation (FAPESP) (grant numbers 2016/10896-0, 2017/01455-2, and 2019/15477-3). SMCL receives a CAPES fellowship (Finance Code - 001). DSD is a CNPq research productivity fellow (#312823/2019-3). Douglas S. Domingues, Liliane S. Oliveira, and Suzana T. Ivamoto-Suzuki contributed equally to this work.

References

1. Christianson DW (2017) Structural and chemical biology of terpenoid cyclases. Chem Rev 118:11795
2. Pichersky E, Raguso RA (2018) Why do plants produce so many terpenoid compounds? New Phytol 220:692–702
3. Tholl D (2015) Biosynthesis and biological functions of terpenoids in plants. Adv Biochem Eng Biotechnol 148:63–106
4. Zerbe P, Bohlmann J (2015) Plant diterpene synthases: exploring modularity and metabolic diversity for bioengineering. Trends Biotechnol 33:419–428
5. Chen F, Tholl D, Bohlmann J et al (2011) The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom: terpene synthase family. Plant J 66:212–229
6. Külheim C, Padovan A, Hefer C et al (2015) The *Eucalyptus* terpene synthase gene family. BMC Genomics 16:450
7. Zhou F, Pichersky E (2020) The complete functional characterisation of the terpene synthase family in tomato. New Phytol 226: 1341–1360
8. Sun P, Schuurink RC, Caissard J-C et al (2016) My way: noncanonical biosynthesis pathways for plant volatiles. Trends Plant Sci 21: 884–894
9. Karunanithi PS, Zerbe P (2019) Terpene synthases as metabolic gatekeepers in the evolution of plant terpenoid chemical diversity. Front Plant Sci 10:1166
10. Priya P, Yadav A, Chand J et al (2018) Terzyme: a tool for identification and analysis of the plant terpenome. Plant Methods 14:4

11. Wang Q, Jia M, Huh J-H et al (2016) Identification of a Dolabellane type diterpene synthase and other root-expressed diterpene synthases in *Arabidopsis*. Front Plant Sci 7:1761
12. Hansen NL, Heskes AM, Hamberger B et al (2017) The terpene synthase gene family in *Tripterygium wilfordii* harbors a labdane-type diterpene synthase among the monoterpene synthase TPS-b subfamily. Plant J 89:429–441
13. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797
14. Trifinopoulos J, Nguyen L-T, von Haeseler A et al (2016) W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res 44:W232–W235
15. Minh BQ, Schmidt HA, Chernomor O et al (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534
16. Mistry J, Chuguransky S, Williams L et al (2021) Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412–D419
17. Eddy SR (2011) Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195
18. Karunanithi PS, Berrios DI, Wang S et al (2020) The foxtail millet (*Setaria italica*) terpene synthase gene family. Plant J 103: 781–800
19. Nachtigall PG, Kashiwabara AY, Durham AM (2020) CodAn: predictive models for precise identification of coding regions in eukaryotic transcripts. Brief Bioinform 22(3):bbaa045. https://doi.org/10.1093/bib/bbaa045