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, and its cultivated accessions have undergone several population bottlenecks. 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, with no obvious global subgenome dominance. We find evidence for a founding polyploidy event 350,000-610,000 years ago, followed by several pre-domestication bottlenecks, resulting in narrow genetic variation. A split between wild accessions and cultivar progenitors occurred ~30.5 kya, followed by a period of migration between the two populations. Analysis of modern varieties, including lines historically introgressed with *C. canephora,* highlights their breeding histories and loci that may contribute to pathogen resistance, laying the groundwork for future genomics-based breeding of *C. arabica*.

Introduction

Polyploidy is a powerful evolutionary force that has shaped genome evolution across many eukaryotic lineages, possibly offering adaptive advantages in times of global change[1,2]. Such whole genome duplications (WGDs) are particularly characteristic of plants[3], and a great proportion of crop species are polyploid[4-11]. Our understanding of genome evolution following WGD is still incomplete, but outcomes can result in genomic shock, in terms of activation of cryptic transposable elements, subgenome-partitioned gene regulation or fractionation, homoeologous exchange, meiotic instability, and even karyotype variation [8, 12-16]. Alternatively, few or none of the above phenomena can materialize, and the two genomes can coexist harmonically, gradually adaptating to new ploidy levels[17]. Regardless, the most common fate of polyploids appears to be fractionation and eventual reversion to the diploid state[18].

With an estimated production of 10 million metric tons per year, coffee is one of the most traded commodities in the world. 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) resulted from a natural hybridization event between the ancestors of present-day *C. canephora* (Robusta coffee, subgenome CC) and *C. eugenioides* (subgenome EE; each with 2n = 2x = 22). The founding WGD has been dated to between 10,000 to one million years ago[20-23], with the Robusta-derived subgenome of *C. arabica* closest related to *C. canephora* accessions from northern Uganda[24]. Arabica cultivation was initiated in 15th -16th century Yemen. (**Ext. data Fig. 1**). Around 1600, "seven seeds" were smuggled out of Yemen[25], establishing Indian *C. arabica* cultivar lineages. A century later, the Dutch began cultivating Arabica in Southeast Asia – thus setting up the founders of the contemporary Typica group. One plant, shipped to Amsterdam in 1706, was used to establish Arabica cultivation in the Caribbean in 1723. Independently, the French cultivated Arabica on the island of Bourbon (presently Réunion)[26], and the descendants of a single plant that survived by 1720 form the contemporary Bourbon group. Contemporary Arabica cultivars descend from these Typica or Bourbon lineages, except for a few wild ecotypes with origins in natural forests in Ethiopia. Due to its recent allotetraploid origin and strong bottlenecks during its history, cultivated *C. arabica* harbors a particularly low genetic diversity[20] and is susceptible to many plant pests and diseases, such as coffee leaf rust (*Hemileia vastatrix*). As a result, the classic Bourbon-Typica lineages can only be cultivated successfully in a few regions around the world. Fortunately, a spontaneous *C. canephora* x *C. arabica* hybrid resistant to *H. vastatrix* was identified on the island of Timor[27] in 1927. Many modern Arabicas contain *C. canephora* introgressions derived from this hybrid, ensuring rust resistance, but having also unwanted side effects, such as decreased beverage quality[28].

Modern genomic tools and a detailed understanding of the origin and breeding history of contemporary varieties are vital to developing new Arabica cultivars, better adapted to climate change and agricultural practices[29-31]. Here, we present chromosome-level assemblies of *C. arabica* and representatives of its progenitor species, *C. canephora* (Robusta) and *C. eugenioides* (hereafter Eugenioides). Whole-genome resequencing data of 41 wild and cultivated accessions facilitated in-depth analysis of Arabica 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, genome fractionation, and subgenome dominance

As reference individuals, we chose the di-haploid Arabica line ET-39[32], a previously sequenced doubled haploid Robusta[33], and the wild Eugenioides accession Bu-A, respectively. Long and short-read-based hybrid assemblies were obtained (Online methods and **Supplementary sections 2.1-2.2**), spanning 672 Mb (Robusta), 645 Mb (Eugenioides) and 1,088 Mb (Arabica), respectively. Upon HiC scaffolding, the Robusta and Arabica assemblies consisted of 11 and 22 pseudochromosomes, and spanned 82.7% and 62.5%, respectively, of the projected genome sizes (**Table 1**). To improve the Arabica assembly, we generated a second assembly using PacBio HiFi technology followed by Hi-C scaffolding (Online methods and **Supplementary section 2.3**). This assembly was 1,198 Mb long, of which 1,192 Mb (93.1% of the predicted genome size based on cytological evidence[34]) was anchored to pseudochromosomes (**Table 1**). Gene space completeness, assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO)[35], was >96% for all assemblies. Importantly, 93.2% of the BUSCO genes were duplicated in the HiFi assembly (**Table 1**), indicating that most of the gene duplicates from the allopolyploidy event were retained.

The Robusta and Eugenioides genomes contained, respectively, 67.5% and 59.7% transposable elements (TEs) (**Supplementary section 3.2**), with Gypsy LTR retrotransposons accounting for most of the difference between the two species. This difference was greatly reduced (63.1% and 63.8%) in the two Arabica subgenomes (subCC and subEE, stemming from Robusta and Eugenioides ancestors, respectively), possibly indicating TE transfer via homoeologous exchange. Robusta contained considerably more recent LTR TE insertion elements than Eugenioides. Again, the two Arabica subgenomes showed greater similarity to each other in recent LTR TE insertions than the two progenitor genomes. No major evidence was found for LTR TE mobilization following Arabica allopolyploidization, in contrast to what has been observed in tobacco[36], but similar to *Brassica* synthetic allotetraploids[37]. Arabica genome evolution observed instead more closely follows the "harmonious coexistence" pattern[38] seen in *Arabidopsis* hybrids[17,39].

High-quality gene annotations, followed by manual curation of specific gene families (**Supplementary Sections 3.1-3.4**), resulted in 28,857, 33,505, 56,670, and 69,314 gene models for the Robusta, Eugenioides, PacBio Arabica, and Arabica HiFi assemblies, respectively (**Table 1**). Altogether ~97% of Robusta and 99.6% of Arabica HiFi gene models were placed on the pseudochromosomes, with 33,618 and 35,449, respectively, to subgenomes subCC and subEE (**Table 1**). Annotation completeness from BUSCO was ≥95% for Eugenioides and Robusta, and reached 97.3% for Arabica HiFi.

Comparison of Arabica subCC and subEE against their Robusta and Eugenioides counterparts revealed high conservation in terms of chromosome number, centromere position and numbers of genes per chromosome (**Fig. 1, Supplementary section 4**). Patterns of gene loss following the *gamma* paleohexaploidy event displayed high structural conservation between Robusta and Eugenioides during the 4-6 million years (My) since their initial species split[22,23] (**Supplementary section 4**). Likewise, the structures of the two Arabica subgenomes were highly conserved between each other, with, since the Arabica-founding allotetraploidy event, only ~5% of BUSCO genes having reverted to the diploid state (**Fig. 1A; Table 1**). Syntenic comparisons revealed that genomic excision events, removing one or several genes at a time in similar proportions across the two subgenomes, have been the main driving force in genome fragmentation both before and after the polyploidy event (**Fig. 1B, Supplementary section 4**). Fractionation occurred mostly in pericentromeric regions, whereas chromosome arms showed more moderate paralogous gene deletion (**Fig 1C, Supplementary section 4**). The Arabica allopolyploidy event seemingly did not affect the rate of genome fractionation, which remained roughly constant when comparing deletions in progenitor species versus Arabica subgenomes after the event. In support of the dosage-balance hypothesis[40], 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 S1**). Genes with high retention rates were enriched in Gene Ontology (GO) 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[41]) and "plant-type hypersensitive response" (in subEE) (**Tables S2-S5**).

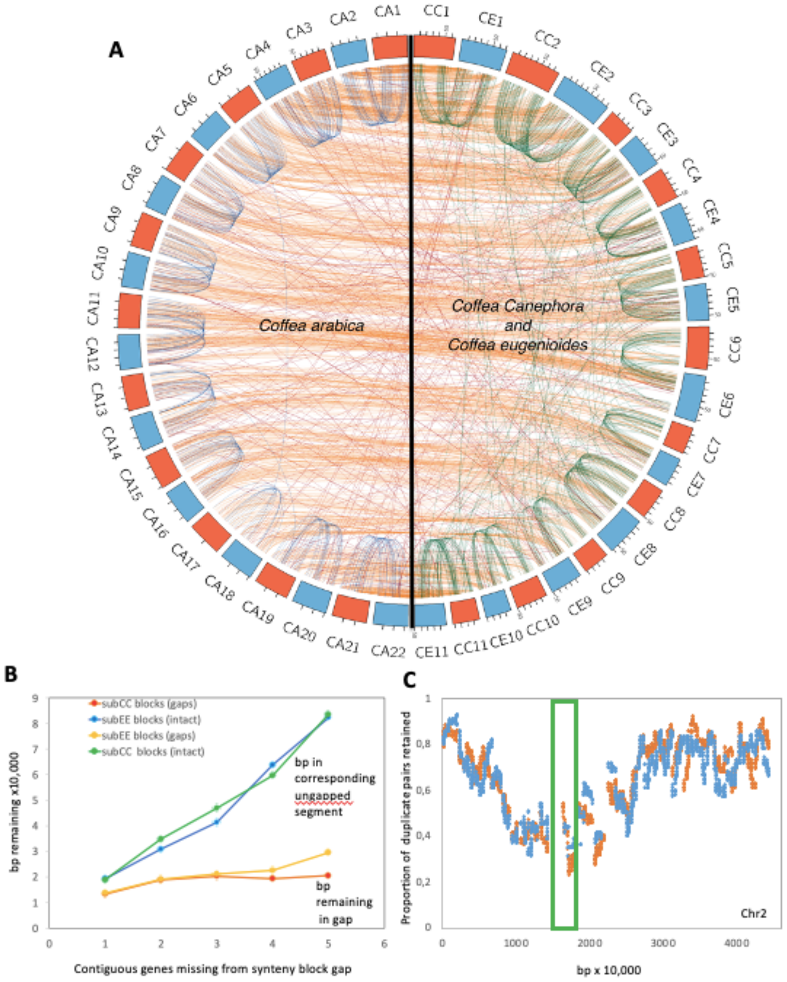

Figure 1. *Patterns of synteny, fractionation and gene loss in* Coffea arabica *(CA) and its progenitor species* C. canephora *(CC) and* C. eugenioides *(CE).* ***A.*** *Corresponding syntenic blocks between CA subgenomes subCC (orange) and subEE (blue), and with the CC (orange) and CE (blue) genomes.* ***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.*

To study possible expression biases between subgenomes, we identified syntelogous gene pairs and removed the pairs showing homoeologous exchanges in the Arabica subgenomes (see under **Origin and domestication of Arabica coffee**, below)[42] (**Supplementary section 5**). Overall, no significant global subgenome expression dominance was observed (**Tables S6-S7**). However, gene families regularly displayed mosaic patterns of expression, including several encoding enzymes that contribute to cup quality, such as *N*-methyltransferase (*NMT*), terpene synthase (*TPS*), and fatty acid desaturase 2 (*FAD2*) families, all having some genes being more expressed in one of the two subgenomes (**Ext. data Fig. 2**), as per a recent study[43]. Similar gene family-wise patterns occur in other evolutionarily recent polyploids such as rapeseed[10] and cotton[44], which are also at their early stages of transitioning back to a diploid state.

Origin and domestication of Arabica coffee

To obtain a genomic perspective on the evolutionary history of Arabica, we sequenced 46 accessions, including three Robusta, two Eugenioides, and 41 Arabica. The latter included an 18th-century type specimen, kindly provided by the Linnaean Society of London, 12 cultivars with different breeding histories, the Timor hybrid and five of its backcrosses to Arabica, and 17 wild and three wild/cultivated accessions collected from the Eastern and Western sides of the Great Rift Valley[45,46] (**Table S8, Fig. 2A**).

Homoeologous exchange (HE) between subgenomes has been observed in several recent polyploids[8,42,47]. Arabica generally displays bivalent pairing of homologous chromosomes and disomic inheritance[48], but since the subgenomes share high similarity, occasional homoeologous pairing and exchange may also occur. We therefore explored the extent of HE among Arabica accessions and its possible contribution to genome evolution. Overall, all accessions shared a fixed allele bias toward subEE at one end of chromosome 7, which contained genes enriched for chloroplast-associated functions (**Ext. data Fig. 3A, Supplementary section 5**, **Table S9**). Since the Arabica plastid genome is derived from Eugenioides[49], HE in this region was likely selected for, due to compatibility issues between nuclear and chloroplast genes encoding chloroplast-localized proteins[50]. Surprisingly, all but one accession (BMJM) showed significant (*p*-value <9.8e-37) 3:1 allelic biases towards subCC. The highly concordant HE patterns, present in both wild and cultivated Arabicas (**Ext. data Fig. 4)**, suggested that i) the allelic bias is an adaptive trait not associated with breeding, and ii) it originated in a common ancestor of all sampled accessions, possibly immediately after the founding allopolyploidy event. Some exchanges, shared by only a few accessions, probably originated more recently (**Ext. data Fig. 3B**). More recent HE events were also found in some cultivars and also showed a bias towards subCC, except for BMJM, which showed bias towards subEE due to a single large crossover in chromosome 1 (**Ext. data Fig. 3A**). An interesting hypothesis for future investigation is that in a low-diversity polyploid species such as Arabica, HE could be a major contributor to phenotypic variation observed among closely related accessions[51].

We next studied population genetic statistics for each of the subgenomes (**Table S10**). The 17 wild samples demonstrated low genomic diversities, indicative of small effective population sizes, while negative Tajima's D suggested an expanding population, possibly following one or more population bottlenecks. The cultivars and wild population samples had similar genetic diversities, as demonstrated by low fixation index (F_{ST}) values. In cultivars, nucleotide diversities were only slightly lower than in wild populations and Tajima's D scores were less negative, suggesting that only minor bottlenecks and subsequent population expansions occurred during domestication.

SNP tree estimation and ADMIXTURE analyses (**Fig. 2B**) identified a three-population solution for subCC: Typica-Bourbon cultivars (Population 1), wild accessions (Population 2), and Timor hybrid-derived cultivars (Population 3). The old BMJM and the recently established Geisha cultivars showed admixed states on both subgenomes, similar to about half of the wild accessions. Indian varieties encompassed both Typica and Bourbon variation, in agreement with previous studies[20]. The Linnaean sample grouped with the cultivars, supporting its hypothesized origin from the Dutch East Indies[25]. A complementary analysis using PCA (**Ext. data Fig. 5)** was in agreement with ADMIXTURE analysis.

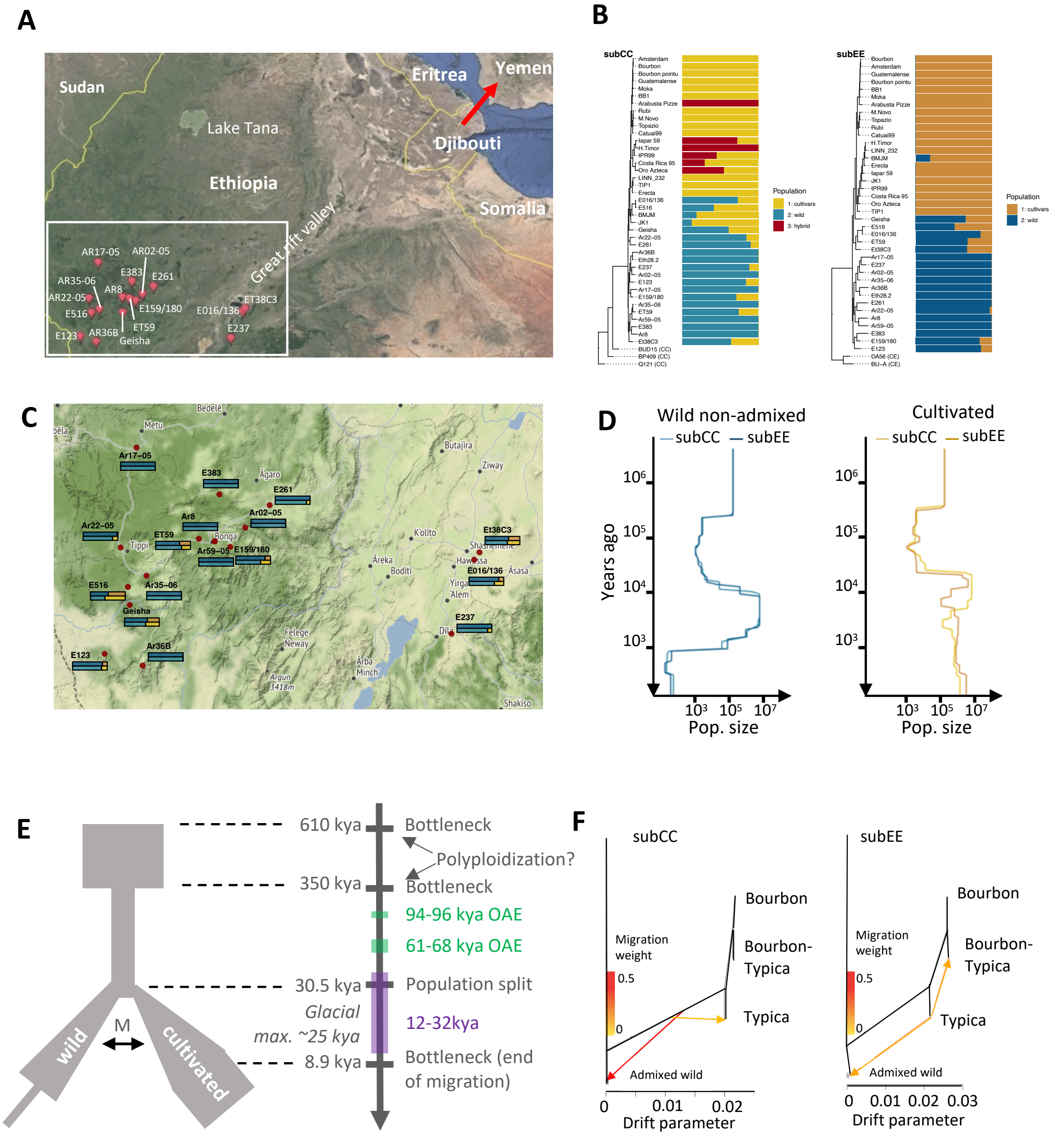

Figure 2. *Population history of* Coffea arabica*.* ***A.*** *Geographic origin of resequenced wild* C. arabica *accessions. The red arrow indicates the probable route of migration to Yemen in historical times.* ***B.*** *Ancestral population assignments of* C. arabica *accessions for subCC (left) and subEE (right). Relationships among individuals are illustrated with phylogenetic trees obtained from independent SNPs. For magnified views of the trees, see* ***Fig. S45****.* ***C.*** *Magnification of panel A, showing the admixture values for each of the accessions in subCC (top) and subEE (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 ~350 and 1 kya (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).*

In wild accessions, both subgenomes concordantly showed two population bottlenecks (**Fig. 2D**) in the SMC++[52] modeling. Assuming a 21-year generation time[53], the oldest bottleneck initiated abruptly around 350 thousand years ago (kya) and ended around 15 kya, at the start of the African humid period (AHP)[54], when climatic conditions were more favourable for Arabica growth. The more recent bottleneck initiated more gradually around 5 kya and lasts to this day. Cultivated accessions, however, exhibited the oldest, but not the more recent bottleneck. In part due to these differences, we also modeled Arabica population history using FastSimcoal2[55], modeling the wild population and cultivars as two separate lineages. In the best-fitting model (**Fig. 2E**) the wild population was predicted to split from the cultivar founding population 1,450 generations ago (~30 kya), i.e., before the last glacial maximum. The original founding event was analyzed using the non-admixed wild individuals, revealing an ancestral population bottleneck at 350 kya (**Ext. data Fig. 6A**). Divergence estimates based on gene fractionation, the distribution of nonsynonymous mutations (**Ext. data Fig. 6B**), and calibrated SNP trees (**Fig. 2B**) suggested the allopolyploid founding event occurred at 610 kya, which is close to previous estimates[22,23]. The 350 kya bottleneck, on the other hand, corresponds to that found in the SMC++ analyses (**Fig. 2D**). We therefore consider 610-350 kya a likely time range for the polyploidization event (**Fig. 2E**). The wild and pre-cultivar lineages maintained some gene flow (in terms of migration) until ~8-9 kya, which may have contributed to the modeled increase in effective population size (**Fig. 2D-E**).

While these data were not able to identify the precise place of origin of the modern cultivated population (see also the following section), the extended period of migration between wild and cultivated accessions suggests that they were separated only by a relatively small geographic distance, such as along the two sides of the African Great Rift Valley (**Fig. 2A-C**). It is also possible that the cultivated lineage could have extended as far as Yemen, and that the end of migration between the two populations could have been caused by the widening of the Bab al-Mandab strait (separating Yemen and Africa) due to rising sea levels[56] at the end of the AHP. A native Arabica population exists in Yemen[57], which could support this hypothesis. The Linnaean sample, together with the Typica and Bourbon cultivars, originate from this second population that was also used to establish cultivation in Yemen, as suggested by the SNP, ADMIXTURE, and PCA analyses (**Fig. 2B**, **Ext. data Fig. 5**).

In conclusion, our analyses suggest that the Arabica allopolyploidy event occurred between 610 and 350 kya, when considering that inbreeding present in *Coffea* populations would accelerate coalescence estimation[58,59]. Earlier work proposing more recent timings, such as 20 kya[20], could be underestimates stemming from confounding effects of population bottlenecks in cultivated and wild lineages.

Origin of modern cultivars

The known breeding history of several of our Arabica cultivars provided us with a gold standard set for deducing the Arabica pedigree using Kinship-based INference for Gwas (KING)[60] (**Fig. 3**). The method correctly identified the relationships between Bourbon and Typica group cultivars and the Bourbon-Typica crosses in subCC. In contrast, the subEE pedigree showed lower (2nd) order relationships, possibly due to HE in that subgenome (**Ext. data Fig. 7**). Timor hybrid-derived accessions did not show significant relationships to mainline cultivars in subCC (likely due to Robusta introgressions in this subgenome that broke the haplotype blocks, see below), while subEE showed 2nd degree relationships to both the Typica and Bourbon groups (**Fig. 3; Ext. data Fig. 7**), confirming that subEE has not received substantial introgression.

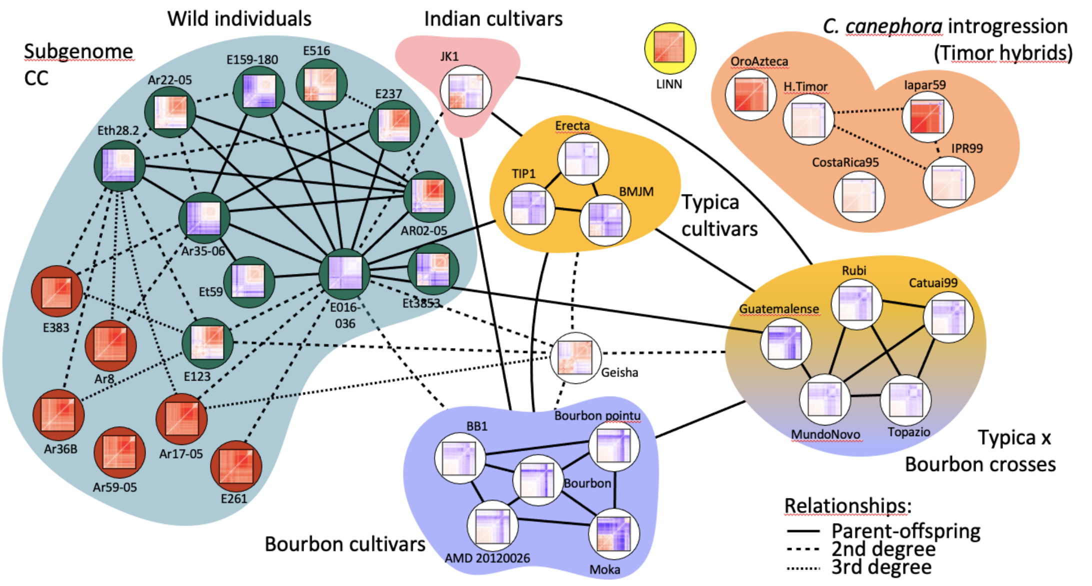

Figure 3. *Kinship estimation of* C. arabica *accessions, inferred from SNPs in subCC. The degree of relatedness was estimated using Kinship-based INference for GWAS (KING) and describes the number of generations between the related accessions. Thumbnail images show false discovery rate corrected F3 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[61]; IBD) from the two source accessions to the target, while red color illustrates lack of gene flow. See* ***Ext. data Fig.*** *7 for corresponding analyses in subEE. In the wild accessions, the dark green background highlights the admixed individuals (****Fig. 3B****), while the non-admixed individuals are highlighted with red background. Relationships follow standard nomenclature (e.g., 2nd degree refers to an individual's grandparents, grandchildren etc., whereas 3rd degree refers to great-grandparents, great-grandchildren, etc.)*

Interestingly, Typica, Bourbon and JK1 individuals were also 1st degree related, suggesting direct parent-offspring relationships. Besides confirming their shared Yemeni origins, this finding also underscores the Yemeni germplasm's limited genetic diversity. Further, the old cultivar lines JK1 (Indian), Erecta (Indonesian Typica), BMJM (Caribbean Typica), TIP1 (Brazilian Typica), and BB1 (Brazilian Bourbon) showed 2nd or higher degree relationships with a cluster of closely related wild admixed accessions, centered on E016/136 (**Fig. 2B**). The recently established Geisha cultivar showed similar relationships to the wild admixed individuals and the Bourbon and Typica groups, suggesting common origins. Interestingly, admixed wild accession E016/136 was closely related to both wild and cultivated populations.

In a comparison of geographic origins, wild individuals from the Eastern side of the Great Rift Valley had some levels of admixture and were closely interrelated, while on the Western side, the admixed, related individuals were mostly concentrated around the Gesha region (**Fig. 2C, Fig. 3**). The E016/136 admixed accession, closest to cultivars, demonstrated a first-degree relationship with several wild accessions, of which only Ar35-06 and Eth28.2 were pure representatives of the wild population (**Fig. 2B**). Therefore, these two accessions are genetically closest, in our sample, to the hypothetical true wild parent of cultivated Arabica, with E016/136 representing an intermediate form. Ar35-06 was collected near Gesha mountain, close to the origin of the modern Geisha cultivar. Altogether these data point to the Gesha region as a hotspot of wild accessions amenable to domestication.

Admixed wild samples may have originated from a recent hybridization event that occurred before or after their collection from the wild. A third alternative is that the Yemeni population (and hence the cultivars) originated from an admixed population from the Eastern side of the Great Rift Valley or the Gesha region. Analysis of admixture patterns with Orientagraph[62] (**Fig. 2F**) suggested hybridization with the common ancestor of the Bourbon and Typica lineages in subCC, and of Typica in subEE. In the case of recent hybridization, introduced haplotypes would exist as long contiguous blocks (as in the Timor hybridization, which occurred 100 years ago), while for older events the blocks would be more fragmented due to crossing over. Analysis using the distance fraction (d_f) statistic[63] showed the latter to be the case (**Ext. data Fig. 8**, indicating that admixture events among wild accessions were not very recent, supporting our third hypothesis.

Domestication and cultivation usually involve strong population bottlenecks based on high wild diversity, resulting in reduced genetic diversity in cultivars[64]. However, Arabica nucleotide diversity was already very low in the wild, probably as a result of earlier bottlenecks (**Figs. 2D-E**), but only marginally reduced in the pre-cultivated lineage (**Ext. data Fig. 9A**). Bourbon had lower diversity than Typica, probably resulting from the known single-individual bottleneck in this group. Also, the inbreeding coefficients in the wild and cultivated accessions were similar (**Ext. data Fig. 9B**), differing from general expectations for a domesticated species[64].

To look for pathways under purifying selection in cultivars, we identified genes with high F_{ST} (95 % quantile) between cultivars and wild accessions. This resulted in a set of 1,908 genes that were enriched for the GO categories "cellular response to nitrogen starvation", "regulation of innate immune response" and "regulation of defense response" (**Table S11**), and contained homologs of ammonium transporters *AMT1* and *AMT2*, important for nitrogen uptake in *Coffea*[65], a homolog of the salicylic acid receptor *NONEXPRESSER OF PR GENES 1* (*NPR1*), required in SA signaling and systemic acquired resistance[66], as well as a homolog of the *Arabidopsis LSU2* gene, previously identified as a hub convergently targeted by effectors of pathogens from different kingdoms[67]. A second screen, focused on genes with a large number of high-impact nonsynonymous mutations shared among cultivars (>40% individuals having the mutation), generated a list of 556 genes that were significantly enriched for only one GO category, "defense response" (**Table S12**). From the 22 genes in this category, 16 were NB-ARC domain-containing resistance (R) genes, and two were members of the leucine-rich repeat (LRR) defense gene family. High diversity in immune related responses is one possible pathogen resistance mechanism in plant communities[68], and therefore reduced diversity may have compromised modern Arabica cultivar immunity.

The high level of conservation between the Arabica subgenomes and their diploid progenitors may have facilitated spontaneous interspecific hybridization events. This was the case for the Timor hybrid, a spontaneous Robusta x Arabica hybrid resistant to *Hemileia vastatrix*[27]. Our sample set included five descendants of the original Timor hybrid, obtained by backcrossing to Arabica. As expected, the hybridization affected subCC more profoundly, with much higher levels of nucleotide divergence apparent (F_{ST}=0.185) than in subEE (F_{ST}=0.0897), when comparing cultivars and hybrids. The divergence from wild populations was even greater, with F_{ST}=0.254 for subCC and F_{ST}=0.138 for subEE, illustrating that introgression occurred almost exclusively within subCC.

In the Timor hybrids, the regions found with d_f statistics[63] largely overlapped the introgressed loci identified using F_{ST} scans (**Fig. 4A**) and were found in large blocks, reflecting recent hybridization, and covering 7-11% of the genome (**Fig. 4A, Ext. data Fig. 8**). Transposon Insertion Polymorphisms (TIPs) also overlapped with introgressed regions (Gypsy p=0.0002; Copia p=0.035; Fisher exact test), confirming their recent origin from Robusta (**Fig. 4B).** The introgressed regions overlapped with regions of higher subgenome fractionation (p=0.001873; **Table S13**), possibly due to heterologous recombination between subCC and Robusta, resulting in unequal crossing-over.

An introgressed region shared by all Timor hybrid lines was evident on chromosome 4 (**Fig. 4A**). We identified a set of 233 genes shared by all hybrids (**Table S14**). The set contained members of three colocalized tandemly duplicated blocks of resistance-related genes on chromosome 4, subCC, and showed high F_{ST} values between cultivars and introgressed lines. A tandem array of five genes were homologs of *Arabidopsis RPP8*, an NLR resistance locus conferring pleiotropic resistance to several pathogens[69,70]. *RPP8* shows a great amount of variation in *Arabidopsis* alone, where intrachromosomal gene conversion combined with balancing selection contributes to its exceptional diversity[71]. The same subCC region also included a tandem array of ten homologs of *CONSTITUTIVE EXPRESSER OF PR GENES 1* (*CPR1*), a negative regulator of defense response that targets resistance proteins[72,73]. Finally, we identified three duplicates encoding leaf rust 10 disease-resistance locus receptor-like protein kinases (LRK10L). The LRK10L are a gene family that is widespread across plants. First identified as a protein kinase in a locus contributing leaf rust resistance in wheat[74], they found to be upregulated during various biotic and abiotic stresses[75] and confirmed as positive regulators of wheat hypersensitive resistance response to stripe rust fungus[75] and powdery mildew[76].

The high F_{ST} values between cultivated and introgressed, but not wild individuals (**Fig. 4B**), indicate that the wild population cannot be the source for allelic asymmetries. Nucleotide diversities further illustrate this point; some genes demonstrate lower nucleotide diversity in wild individuals, suggesting these genes to have experienced selective sweeps. To further narrow down candidate genes involved in leaf rust resistance, we reanalyzed comparative gene expression data from susceptible and resistant accessions after *H. vastatrix* inoculation[77]. This analysis identified 723 differentially expressed genes, most of which were associated with defense responses (**Fig. 4B, Tables S14-S14b**). The combination of high F_{ST} values, nucleotide diversities, and differential expression data highlight several strong candidate genes (one *RPP8*, six *CPR1* and one *LRK10L*) at this locus.

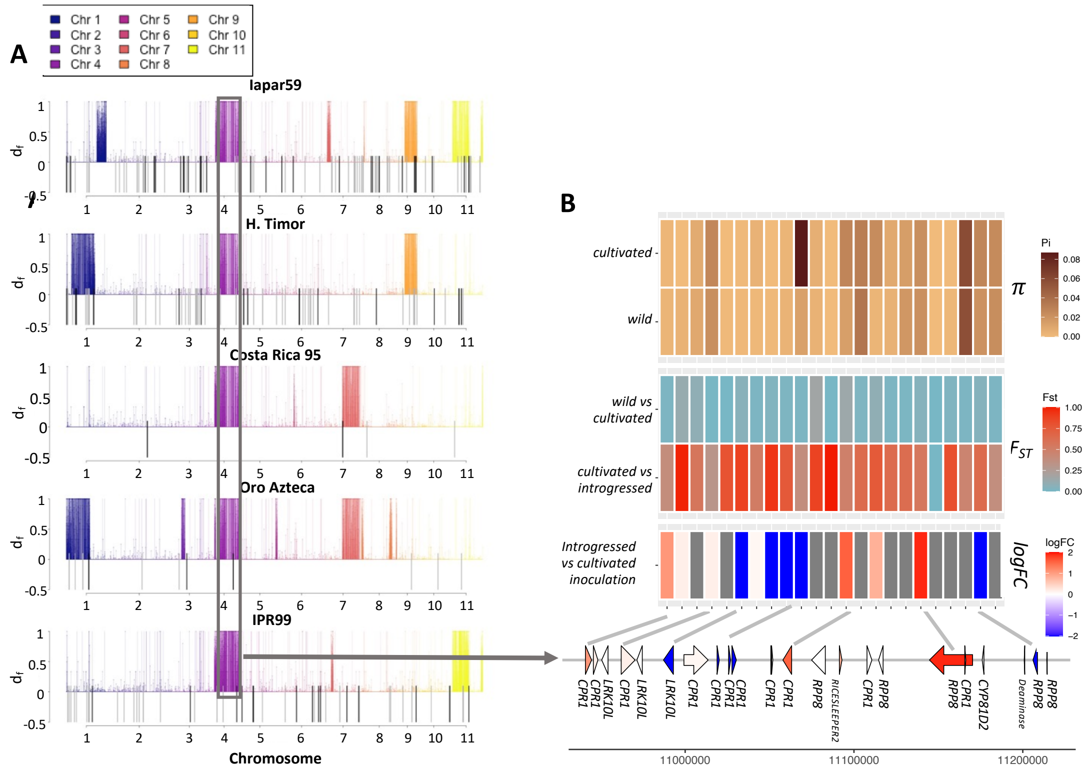

***Figure 4.** Introgression of* Coffea canephora *into* H. vastatrix-*resistant* C. arabica *lineages. **A.** Introgression d_f statistic estimated for different Timor hybrid derivatives. Colored lines above the axis mark regions of significant introgression in the line under inspection, and are colored 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 (RPP8), a cluster of homologs of a negative regulator of R genes (CPR1), and a cluster of homologs of leaf rust resistance 10 kinases (LRK10L) (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, and blue decreased expression. (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.*

Conclusions and outlook

Besides providing genomic resources for molecular breeding of one of the most important agricultural commodities, our Arabica, Robusta and Eugenioides genomes provide a unique window into the genome evolution of a recently formed allopolyploid stemming from two closely related species. Our Arabica data did not suggest a genomic shock induced by allopolyploidy, but instead, only higher LTR transposon turnover rate. Genome fractionation rates remained basically unaltered before and after the allopolyploidy event. Likewise, no global subgenome dominance in gene expression was observed, but rather a mosaic-type pattern as in other recent polyploids[44,47], affecting the expression of individual gene family members. However, similar to octoploid strawberry[8], we detected genome dominance in terms of biased homoeologous exchanges favoring subCC. Since Robusta has one of the widest geographic ranges in the *Coffea* genus, whereas Eugenioides is more range-limited, this biased HE might be adaptive. This hypothesis was supported by the site frequency spectrum of HE loci, showing signs of directional selection (**Ext. data Fig. 3**). Intriguingly, transposable insertion polymorphisms significantly overlapped with tandem gene duplications and biosynthetic gene clusters, hinting at their possible roles in cluster evolution.

Domestication of perennial species like Arabica coffee differs markedly from that of annual crops, consisting instead of three phases: selection of outstanding genotypes from wild forests, clonal propagation and cultivation, then breeding and diversification[78]. In addition to being a perennial crop, Arabica is also a predominantly autogamous allopolyploid, which puts it in a class of its own. We show here that genetic diversity was already very low among wild accessions, due to multiple pre-domestication bottlenecks, and that the genotypes selected for cultivation by humans (both the ancient cultivated Ethiopian landraces, and the recent Geisha cultivar) already were somewhat admixed between divergent lineages. The resequenced 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. Such admixture has played a large role in breeding many fruit-bearing crops, the non-polyploid allogamous perennial lychee being one of the most extreme cases[59].

The prevalent autogamy of Arabica, combined with the multiple genetic bottlenecks it underwent in the wild, may have selectively purged deleterious alleles, explaining the capacity of the species to survive single-plant bottlenecks that occurred during its cultivation. An additional element buffering deleterious alleles was probably Arabica's allopolyploidy itself, which provided some level of heterosis[79]. However, the narrow genetic basis of both cultivated and wild modern Arabica constitutes a major drawback, as well as an obstacle for its breeding using wild genepool diversity. On the other hand, the extensive collinearity of its CC and EE subgenomes with those of its Robusta and Eugenioides progenitors is likely to facilitate introgression of interesting traits from these species, as already happened historically in the Timor spontaneous hybrid. The high-quality genome sequences of the three species provided in this work, together with the identification of the genomic region conferring resistance to coffee leaf rust, constitute a cornerstone for the breeding of novel Arabica varieties with superior adaptability and pathogen resistance.

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 a 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). AA acknowledges funding from INCT-Café-CNPq/Fapemig. DD and SMCL acknowledge São Paulo State Research Foundation (FAPESP), grant numbers #2016/10896-0 and #2017/01455-2. DS acknowledges funding from the NSERC and the Canada Research Chairs programs. 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 GG, DS, VAA, LFP, 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. Sun, H. *et al.* Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nature Genetics* **54**, 342-348 (2022).
5. Athiyannan, N. *et al.* Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning. *Nature Genetics* **54**, 227-231 (2022).
6. Wu, S. *et al.* Genome sequences of two diploid wild relatives of cultivated sweetpotato reveal targets for genetic improvement. *Nature Communications* **9**, 4580 (2018).
7. 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).
8. Edger, P.P. *et al.* Origin and evolution of the octoploid strawberry genome. *Nature Genetics* **51**, 541-547 (2019).
9. 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).

10. Chalhoub, B. *et al.* Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* **345**, 950-953 (2014).
11. Sattler, M.C., Carvalho, C.R. & Clarindo, W.R. The polyploidy and its key role in plant breeding. *Planta* **243**, 281-296 (2016).
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. Marie, L. *et al.* G × E interactions on yield and quality in *Coffea arabica*: new F1 hybrids outperform American cultivars. *Euphytica* **216**, 78 (2020).
30. 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).

31. 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).

32. Berthaud, J. Etude cytogénétique d'un haploïde de *Coffea arabica* L. *Café, Cacao, Thé (Francia) v. 20 (2) p. 91-96* (1976).

33. Denoeud, F. *et al.* The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345**, 1181-1184 (2014).

34. 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).

35. 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).

36. Petit, M. *et al.* Mobilization of retrotransposons in synthetic allotetraploid tobacco. *New Phytologist* **186**, 135-147 (2010).

37. 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).

38. 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).

39. 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).

40. 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).

41. 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).

42. Bird, K.A. *et al.* Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid Brassica napus. *New Phytologist* **230**, 354-371 (2021).

43. Combes, M.-C., Joët, T., Stavrinides, A.K. & Lashermes, P. New cup out of old coffee: contribution of parental gene expression legacy to phenotypic novelty in coffee beans of the allopolyploid Coffea arabica L. *Annals of Botany* **131**, 157-170 (2023).

44. Yoo, M.J., Szadkowski, E. & Wendel, J.F. Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* **110**, 171-180 (2013).

45. Meyer, F.G., Fernie, L.M., Narasimhaswami, R.L., Monaco, L.C. & Greathead, D.J. FAO Coffee Mission to Ethiopia, 1964-1965. (1968).

46. Halle, F. Echantillonnage du matériel *Coffea arabica* récolté en Ethiopie. *Bulletin - IFCC*, 13-18 (1978).

47. Chalhoub, B. *et al.* Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* **345**, 950-953 (2014).

48. Krug, C.A.M., A.J.T. Cytological observations in *Coffea* – IV. *J Genet* **39**, 189–203 (1940).

49. Cros, J. *et al.* Phylogenetic Analysis of Chloroplast DNA Variation in *Coffea* L. *Molecular Phylogenetics and Evolution* **9**, 109-117 (1998).

50. Lashermes, P. *et al.* Molecular characterisation and origin of the Coffea arabica L. genome. *Molecular and General Genetics MGG* **261**, 259-266 (1999).

51. Wu, Y. *et al.* Genomic mosaicism due to homoeologous exchange generates extensive phenotypic diversity in nascent allopolyploids. *National Science Review* **8**, nwaa277 (2021).

52. 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).

53. 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).

54. Kuper, R. & Kröpelin, S. Climate-controlled holocene occupation in the Sahara: motor of Africa's evolution. *Science* **313**, 803-807 (2006).

55. Excoffier, L. *et al.* fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics* **37**, 4882-4885 (2021).

56. 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).

57. 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).

58. Nordborg, M. & Donnelly, P. The coalescent process with selfing. *Genetics* **146**, 1185 (1997).

59. 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).

60. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010).

61. Lan, T. *et al.* Insights into bear evolution from a Pleistocene polar bear genome. *Proceedings of the National Academy of Sciences* **119**, e2200016119 (2022).

62. Molloy, E.K., Durvasula, A. & Sankararaman, S. Advancing admixture graph estimation via maximum likelihood network orientation. *Bioinformatics* **37**, i142-i150 (2021).

63. Pfeifer, B. & Kapan, D.D. Estimates of introgression as a function of pairwise distances. *BMC Bioinformatics* **20**, 207 (2019).

64. 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).

65. 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).

66. Wang, W. *et al.* Structural basis of salicylic acid perception by Arabidopsis NPR proteins. *Nature* **586**, 311-316 (2020).

67. Mukhtar, M.S. *et al.* Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science* **333**, 596-601 (2011).

68. Jousimo, J. *et al.* Ecological and evolutionary effects of fragmentation on infectious disease dynamics. *Science* **344**, 1289-1293 (2014).

69. 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).

70. 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).

71. MacQueen, A. *et al.* Population genetics of the highly polymorphic *RPP8* gene family. *Genes* **10**(2019).

72. 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).

73. 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).

74. Feuillet, C., Schachermayr, G. & Keller, B. Molecular cloning of a new receptor-like

kinase gene encoded at the Lr10 disease resistance locus of wheat. *The Plant Journal* **11**, 45-52 (1997).

75. Zhou, H. *et al.* Molecular analysis of three new receptor-like kinase genes from hexaploid wheat and evidence for their participation in the wheat hypersensitive response to stripe rust fungus infection. *The Plant Journal* **52**, 420-434 (2007).

76. Xia, T. *et al.* Efficient expression and function of a receptor-like kinase in wheat powdery mildew defence require an intron-located MYB binding site. *Plant Biotechnology Journal* **19**, 897-909 (2021).

77. 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).

78. 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).

79. Chen, Z.J. Molecular mechanisms of polyploidy and hybrid vigor. *Trends in Plant Science* **15**, 57-71 (2010).