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**).

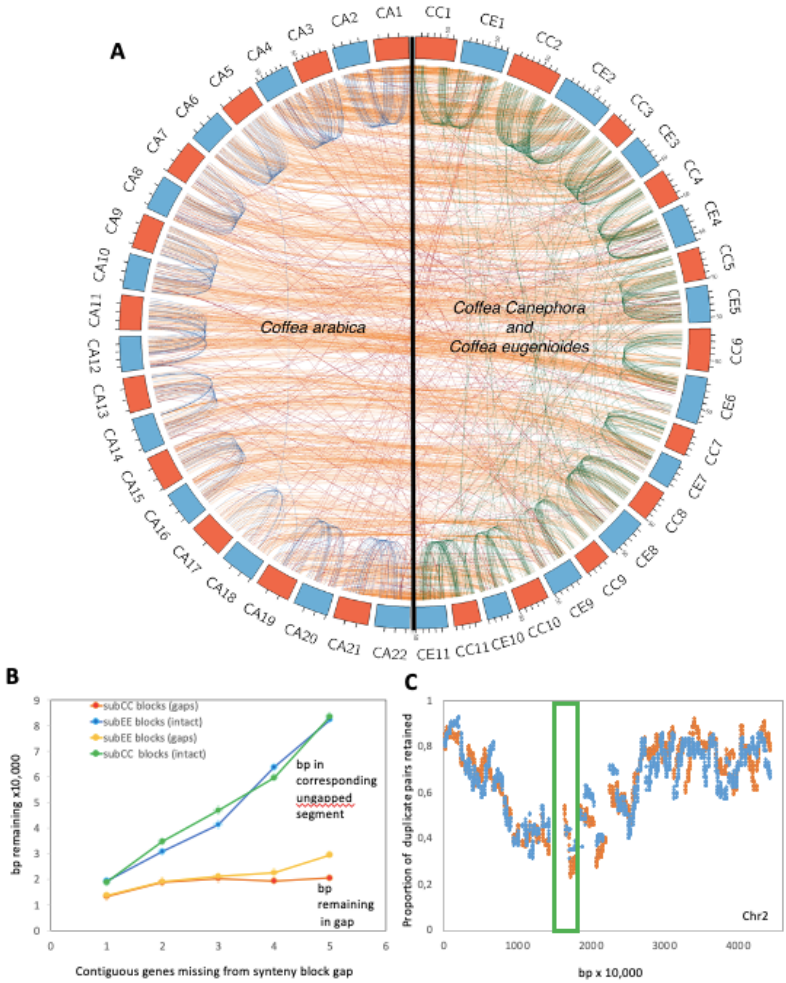

Figure 1. *Patterns of synteny, fractionation and gene loss in* C. arabica *(CA) and its progenitor species* C. canephora *(CC) and* C. eugenioides *(CE).* ***A.*** *Corresponding syntenic blocks between CA subgenomes subCC (orange) and subEE (blue), and with the CC (orange) and CE (blue) genomes.* ***B.***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

catalyzing the last step in caffeine biosynthesis, is tandemly duplicated in subEE, and both copies are expressed in fully ripe fruits at decreased levels compared to subCC (**Fig. 2B-C**) providing a molecular basis for the lower caffeine content of CA with respect to CC [55-57]. More generally, the *NMT* gene family in CA shows clear, though not extensive, mosaicism in subgenome-wise expression.

Terpenes derived from geranyl diphosphate (**Fig. 2A**) strongly contribute to coffee aroma[58]. Here, we observed expression dominance by subEE for genes encoding *α*-terpineol synthases (**Fig. 2B-C**) and a putative nerolidol synthase. In contrast, a tandem duplication and expression of several putative *α*-farnesene synthases in subCC suggests expression dominance there, whereas the genes encoding for isoprene synthase show also dominance for subCC, but only at the green maturation stage.

Polyunsaturated fatty acids form an energy reserve in coffee beans and contribute to coffee flavor, aroma, and its shelf life[59]. *FATTY ACID DESATURASE 2* (*FAD2*) encodes the key enzymes that desaturate oleic acid to linoleic acid, the major unsaturated fatty acid in coffee (**Fig. 2A**). *Arabidopsis* encodes a single, constitutively expressed *FAD2* gene, whereas some oil-producing plants, and coffee, have multiple copies of *FAD2*, some of which are mainly expressed in seeds. Based on phylogenetic analyses, these seed-type *FAD2*s evolved from housekeeping FAD2-encoding genes by gene duplication[60,61]. In Arabica, the housekeeping *FAD2* syntelogs were present in both subgenomes (**Fig. 2B-C**), with expression dominated by subEE, whereas seed-type *FAD2* was duplicated in both subgenomes, where overall, the duplicates show expression patterns suggesting dominance by subCC. Interestingly, the genic region encoding the N-terminal signal peptide of the only highly expressed subEE homolog (Cara002g032150) differed from those of the other seed-type paralogs; this could result, for example, from a gene conversion event[62,63].

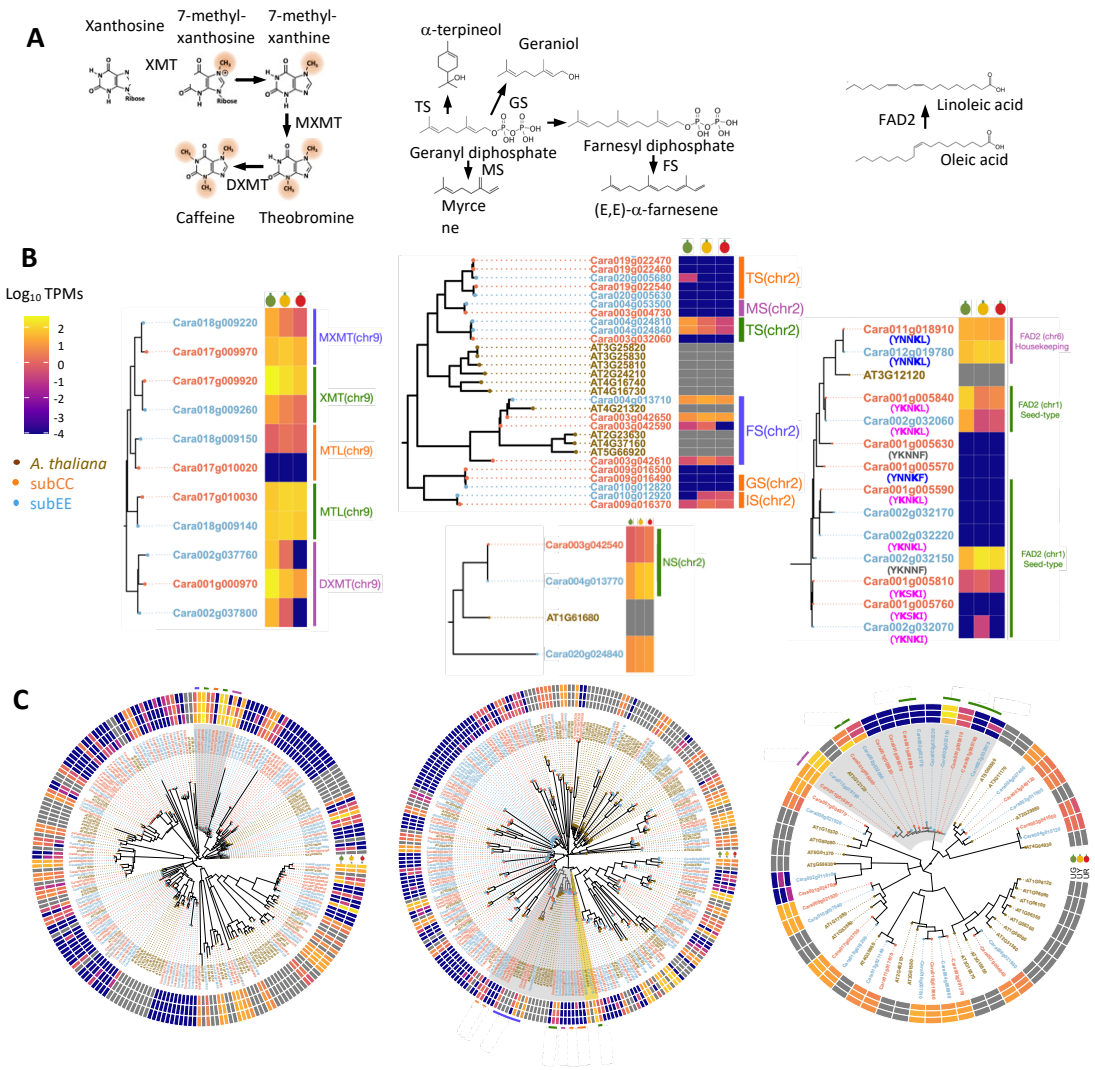

Figure 2. *Composition and expression of* exemplar C. arabica *gene families contributing to bean quality traits.* ***A****. Schematic biosynthesis of caffeine (left), terpenoids (middle), and unsaturated fatty acids (right)* ***B****. Phylogenies and expression during fruit development of C. arabica 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.

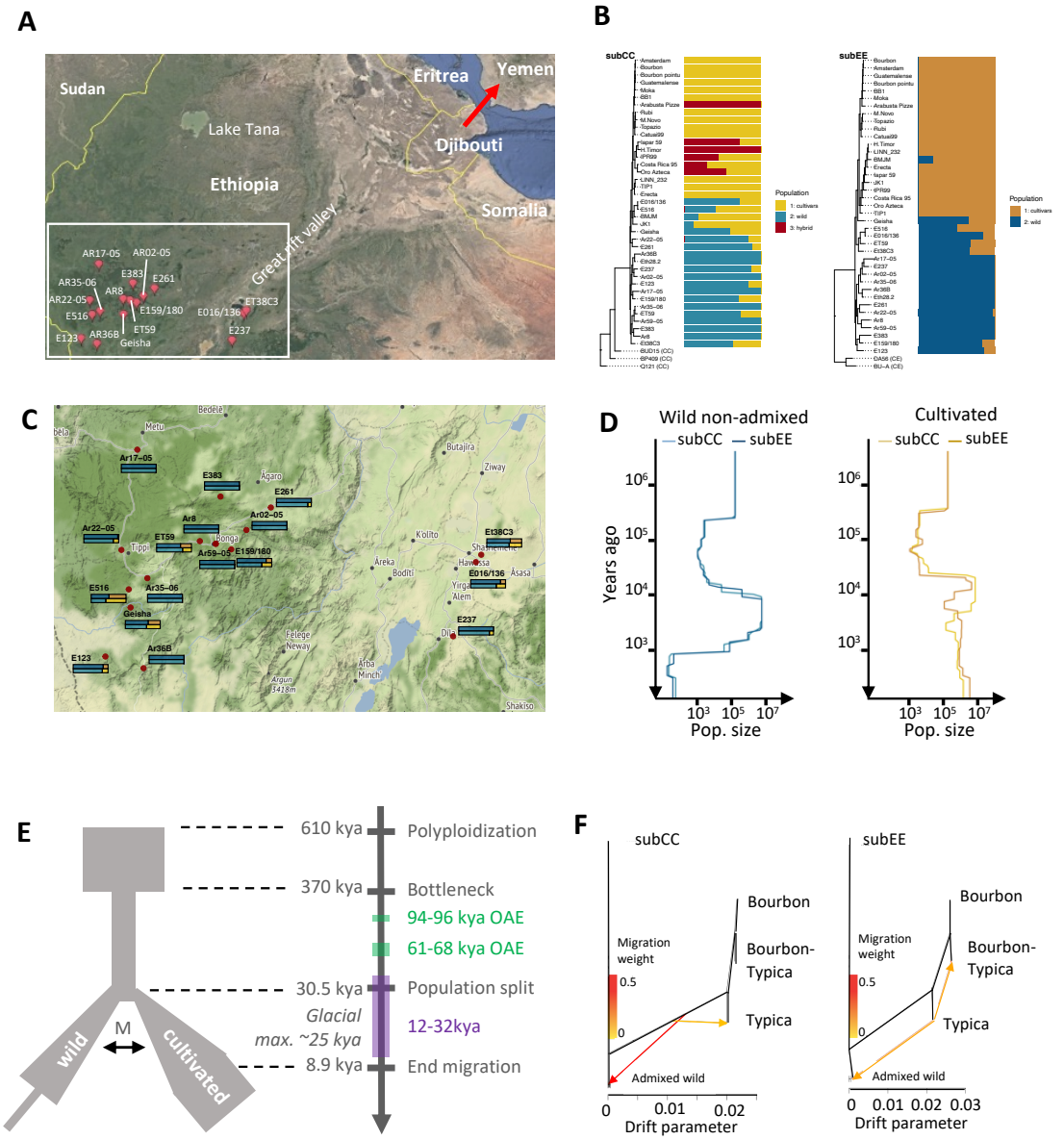

Figure 3. *Population history of Coffea arabica.* ***A.*** *Geographic origin of resequenced wild C. arabica accessions. The red arrow indicates the probable route of the migration to Yemen in historical times.* ***B.*** *Ancestral population assignments of C. arabica accessions for the CC subgenome (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.

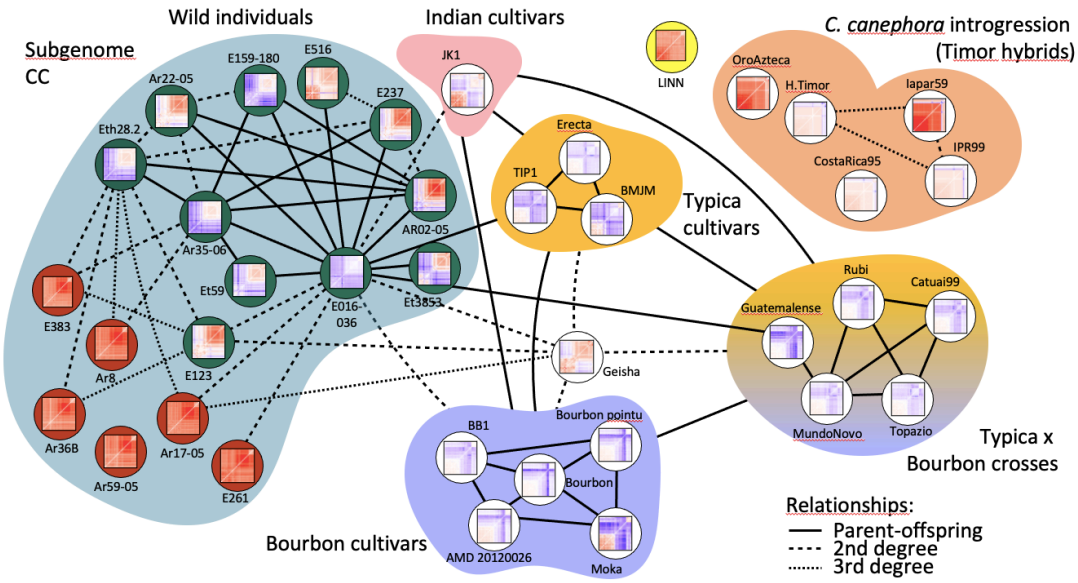

Figure 4. *Kinship estimation of* C. arabica *accessions, inferred from SNPs in the subCC. The degree of relatedness was estimated using Kinship-based INference for GWAS (KING) and describes the number of generations between the related accessions. Thumbnail images show false discovery rate corrected 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.

620
621
622 Domestication of crop plants is generally split into four stages; (i) pre-domestication
623 management of wild populations, (ii) selection of lineages with desirable alleles and their
624 cultivation (iii), geographical radiation and adaptation to different environments, and (iv)
625 targeted breeding[91]. As a globally grown crop plant, *C. arabica* would be placed at the fourth
626 stage of domestication, but its long generation time means that the cultivars have had relatively
627 few breeding cycles; altogether this makes *C. arabica* an interesting target for studying the
628 early stages of cultivation.
629
630 Cultivation (stages 1 and 2) involves a population bottleneck, which results in highly reduced
631 genetic diversity in cultivars[92]. Accordingly, in Arabica coffee the mean nucleotide diversity in
632 neutral sites (π_s) was lower among cultivars, while Geisha and older cultivars with fewer
633 breeding cycles demonstrated higher diversity (**Fig. S41**). Among the major cultivar groups,
634 Typica had higher diversity than Bourbon, whereas modern cultivar crosses of the two lines
635 showed intermediate values. The differences in π_s could result from the known single-
636 individual bottleneck in Bourbon, admixture in Typica, and more concentrated breeding efforts
637 on the Bourbon lineage.
638
639 Common to many crops, admixture and subsequent heterosis may have played a role when
640 selecting the early cultivars[85], such as BMJM and JK1. In Arabica, this is supported by the fact
641 that admixed wild individuals are preferred for coffee production in Ethiopia, and also by the
642 success of the more recently identified Geisha cultivar, which also derives from the admixed
643 population. The admixture results and nucleotide diversity estimates suggest, however, that the
644 heterosis effect among Bourbon and Typica lines appears to have been lost during coffee
645 breeding (**Fig. S41**). However, the maintenance of current heterozygosity levels may be the
646 driving force behind modern coffee breeding, whereby two individuals are preferentially
647 crossed, instead of the widespread selfing that occurs among wild individuals. Additionally, the
648 more modern cultivars represent crosses between the Bourbon and Typica lineages. Perhaps
649 because of this strategy of maintaining heterozygosity, the inbreeding coefficients in the
650 cultivated accessions were similar to those of wild accessions (**Fig. S40**), differing from the
651 general expectations for a domesticated species[92].
652
653 During cultivation, increased genetic drift due to small population size and genomic hitchhiking
654 linked with loci under selection elevate the frequency of deleterious alleles. Furthermore,
655 continued inbreeding leads to eventual purging of deleterious alleles[92]. Concordantly, the
656 cultivated Arabicas showed decreased diversity both in terms of non-coding sites as well as the
657 ratio of non-synonymous diversity to synonymous diversity (**Fig. S41**). To look for pathways
658 affected by selection in cultivars, we calculated the F_{ST} values between cultivars and wild
659 accessions across the genome for all genes and their 2 kb flanking regions and identified the
660 genes with high F_{ST} (95 % quantile) in subCC and subEE, respectively (**Tables S21-S22**).
661 Assuming that deleterious variants (and the ones under selection) will have increased
662 frequencies, we focused our analysis on genes with a large number of shared (>40% individuals
663 having the mutation), derived amino acid changing mutations among cultivars. This screening
664 generated a list of 556 genes that were significantly enriched for only one GO category,
665 "Defense response" (**Table S23**). Sixteen out of 22 genes in this category were NB-ARC
666 domain-containing resistance (R) genes, while two out of the remaining six genes were
667 members of the leucine-rich repeat (LRR) gene family. High diversity in immune related
668 responses is one possible mechanism for preventing extensive spread of pathogens in plant
669 communities[93], and therefore reduced diversity in resistance genes may have compromised
670 modern Arabica cultivar immunity.
671
672 To look for other possible phenotypically causative mutations, e.g., in promoter regions, we
673 extended the screening to all derived alleles. This yielded a larger set of 1,908 genes (**Tables**
674 **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.

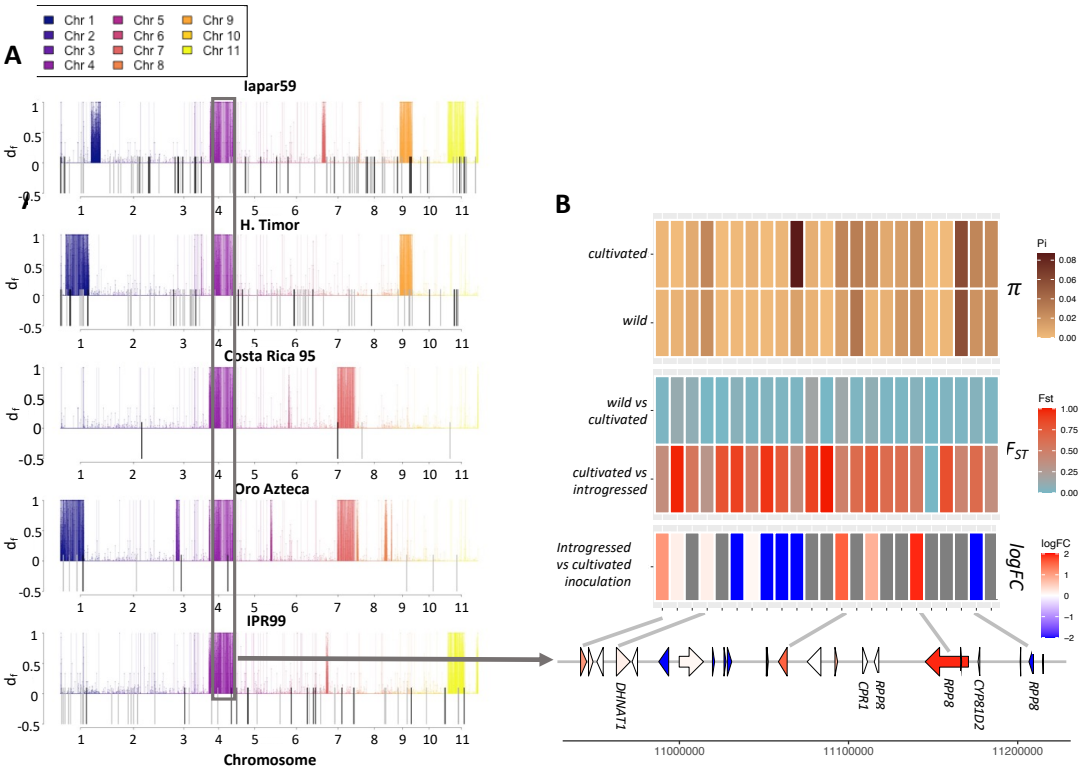

Figure 5. *Introgression of C. canephora into H. vastatrix-resistant C. arabica lineages. A. Introgression d_f statistic estimated for different Timor hybrid derivatives. Colored lines above the axis mark regions of significant introgression in the line under inspection, and are colored by chromosome. 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).