Complex history of admixture during citrus domestication revealed by genome analysis

Authors: G. Albert Wu[1†], Simon Prochnik[1†], Jerry Jenkins[2], Jerome Salse[3], Uffe Hellsten[1], Florent Murat[3], Xavier Perrier[4], Manuel Ruiz[4], Simone Scalabrin[5], Javier Terol[8], Marco Aurélio Takita[6], Karine Labadie[7], Julie Poulain[7], Arnaud Couloux[7], Kamel Jabbari[7], Federica Cattonaro[5], Cristian Del Fabbro[5], Sara Pinosio[5], Andrea Zuccolo[5,25], Jarrod Chapman[1], Jane Grimwood[2], Francisco R. Tadeo[8], Leandro H. Estornell[8], Juan V. Muñoz-Sanz[8], Victoria Ibanez[8], Amparo Herrero-Ortega[8], Pablo Aleza[9], Julián Pérez Pérez[10], Daniel Ramón[10], Dominique Brunel[7,11], François Luro[12], Chunxian Chen[13],William G. Farmerie[14], Brian Desany[15], Chinnappa Kodira[15], Mohammed Mohiuddin[15], Tim Harkins[15‡,] Karin Fredrikson[15], Paul Burns[16], Alexandre Lomsadze[16], Mark Borodovsky[16,17], Giuseppe Reforgiato[18], Juliana Freitas-Astúa[6,19], Francis Quetier[7,20], Luis Navarro[9], Mikeal Roose[21], Patrick Wincker[7,20,22], Jeremy Schmutz[2], Michele Morgante[5,23], Marcos Antonio Machado[6], Manuel Talon[8], Olivier Jaillon[7,20,22], Patrick Ollitrault[4], Frederick Gmitter[13]*, Daniel Rokhsar[1,24]*

Affiliations:

[1]US-Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598.

[2]HudsonAlpha Biotechnology Institute, 601 Genome Way Northwest, Huntsville AL 35806.

[3]INRA/UBP UMR 1095 GDEC, 5 chemin de Beaulieu, 63100 Clermont Ferrand, France.

[4]CIRAD, UMR AGAP, F-34398 Montpellier, France.

[5]Istituto di Genomica Applicata, Via J. Linussio 51, Udine 33100, Italy.

[6]Centro de Citricultura Sylvio Moreira, IAC, Cordeirópolis, SP, Brazil.

[7]Commissariat à l'Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, 2 rue

Gaston Crémieux, BP5706, 91057 Evry, France.

[8]Centro de Genomica, Instituto Valenciano de Investigaciones Agrarias (IVIA), 46113 Moncada, Valencia, Spain.

[9]Centro de Protección Vegetal y Biotecnología-IVIA, Ctra. Moncada-Náquera Km 4.5, 46113 Moncada, Valencia, Spain.

[10]Lifesequencing SL, C/ Catedrático Agustín Escardino 9, edificio B2,Parc Cientific Universitat de Valencia, 46980-Paterna; Valencia, Spain.

Complex history of admixture during citrus domestication revealed by genome analysis

[11]INRA, US EPGV_1279, 2 rue Gaston Cremieux, 91057, Evry, France.

[12]INRA GEQA, 20230, San Giuliano, France.

[13]Citrus Research and Education Center (CREC), Institute of Food and Agricultural Sciences (IFAS), University of Florida, Lake Alfred, FL 33850, USA.

[14]Interdisciplinary Center for Biotechnology Research, University of Florida, PO Box 103622, Gainesville, FL 32610, USA.

[15]454 Life Sciences, A Roche Company, 15 Commercial Street, Branford CT 06405.

[16]Wallace H. Coulter Department of Biomedical Engineering & School of Computational Science & Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

[17]Department of Biological and Medical Physics, Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia.

[18]Consiglio per la Ricerca e la Sperimentazione in Agricoltura (CRA-ACM), Acireale, Italy. [19]Embrapa Cassava and Fruits Embrapa Cassava and Fruits, Cruz das Almas, BA, Brazil.

[20]Département de Biologie, Université d'Evry, UMR 8030, CP5706, Evry, France.

[21]Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA.

[22]Centre National de Recherche Scientifique (CNRS), UMR 8030, CP5706, Evry, France.

[23]Department of Agriculture and Environmental Sciences, University of Udine, Via delle Scienze, Udine 33100, Italy.

[24]Division of Genetics, Genomics, and Development, University of California, Berkeley, CA 94720.

[25]Institute of Life Sciences, Scuola Superiore Sant'Anna, 56127 Pisa, Italy.

*Correspondence to: D.S.R. (dsrokhsar@gmail.com); F.G.G. (fgmitter@ufl.edu).

‡ Current address: Life Technologies Corp., Grand Island, NY 14072, USA.

† These authors contributed equally

* To whom correspondence may be addressed. Daniel Rokhsar, US-Department of Energy Joint Genome Institute/LBNL, 2800 Mitchell Drive, Walnut Creek, CA, 94598, USA. dsrokhsar@gmail.com

July 1, 2014

Complex history of admixture during citrus domestication revealed by genome analysis

ACKNOWLEDGMENTS:

Work by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Complex history of admixture during citrus domestication revealed by genome analysis.

G. Albert Wu[1]†, Simon Prochnik[1]†, Jerry Jenkins[2], Jerome Salse[3], Uffe Hellsten[1], Florent Murat[3], Xavier Perrier[4], Manuel Ruiz[4], Simone Scalabrin[5], Javier Terol[8], Marco Aurélio Takita[6], Karine Labadie[7], Julie Poulain[7], Arnaud Couloux[7], Kamel Jabbari[7], Federica Cattonaro[5], Cristian Del Fabbro[5], Sara Pinosio[5], Andrea Zuccolo[5,25], Jarrod Chapman[1], Jane Grimwood[2], Francisco R. Tadeo[8], Leandro H. Estornell[8], Juan V. Muñoz-Sanz[8], Victoria Ibanez[8], Amparo Herrero-Ortega[8], Pablo Aleza[9], Julián Pérez Pérez[10], Daniel Ramón[10], Dominique Brunel[7,11], François Luro[12], Chunxian Chen[13],William G. Farmerie[14], Brian Desany[15], Chinnappa Kodira[15], Mohammed Mohiuddin[15], Tim Harkins[15‡], Karin Fredrikson[15], Paul Burns[16], Alexandre Lomsadze[16], Mark Borodovsky[16,17], Giuseppe Reforgiato[18], Juliana Freitas-Astúa[6,19], Francis Quetier[7,20], Luis Navarro[9], Mikeal Roose[21], Patrick Wincker[7,20,22], Jeremy Schmutz[2], Michele Morgante[5,23], Marcos Antonio Machado[6], Manuel Talon[8], Olivier Jaillon[7,20,22], Patrick Ollitrault[4], Frederick Gmitter[13]*, Daniel Rokhsar[1,24]*

Affiliations

[1]US-Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598.

[2]HudsonAlpha Biotechnology Institute, 601 Genome Way Northwest, Huntsville AL 35806.

[3]INRA/UBP UMR 1095 GDEC, 5 chemin de Beaulieu, 63100 Clermont Ferrand, France.

[4]CIRAD, UMR AGAP, F-34398 Montpellier, France.

[5]Istituto di Genomica Applicata, Via J. Linussio 51, Udine 33100, Italy.

[6]Centro de Citricultura Sylvio Moreira, IAC, Cordeirópolis, SP, Brazil.

[7]Commissariat à l'Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, 2 rue Gaston Crémieux, BP5706, 91057 Evry, France.

[8]Centro de Genomica, Instituto Valenciano de Investigaciones Agrarias (IVIA), 46113 Moncada, Valencia, Spain.

[9]Centro de Protección Vegetal y Biotecnología-IVIA, Ctra. Moncada-Náquera Km 4.5, 46113 Moncada, Valencia, Spain.

[10]Lifesequencing SL, C/ Catedrático Agustín Escardino 9, edificio B2,Parc Cientific Universitat de Valencia, 46980-Paterna; Valencia, Spain.

[11]INRA, US EPGV_1279, 2 rue Gaston Cremieux, 91057, Evry, France.

[12]INRA GEQA, 20230, San Giuliano, France.

[13]Citrus Research and Education Center (CREC), Institute of Food and Agricultural Sciences (IFAS), University of Florida, Lake Alfred, FL 33850, USA.

[14]Interdisciplinary Center for Biotechnology Research, University of Florida, PO Box 103622, Gainesville, FL 32610, USA.

[15]454 Life Sciences, A Roche Company, 15 Commercial Street, Branford CT 06405.

[16]Wallace H. Coulter Department of Biomedical Engineering & School of Computational Science & Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

[17]Department of Biological and Medical Physics, Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia.

[18]Consiglio per la Ricerca e la Sperimentazione in Agricoltura (CRA-ACM), Acireale, Italy.

[19]Embrapa Cassava and Fruits Embrapa Cassava and Fruits, Cruz das Almas, BA, Brazil.

[20]Département de Biologie, Université d'Evry, UMR 8030, CP5706, Evry, France.

[21]Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA.

[22]Centre National de Recherche Scientifique (CNRS), UMR 8030, CP5706, Evry, France.

[23]Department of Agriculture and Environmental Sciences, University of Udine, Via delle Scienze, Udine 33100, Italy.

[24]Division of Genetics, Genomics, and Development, University of California, Berkeley, CA 94720.

[25]Institute of Life Sciences, Scuola Superiore Sant'Anna, 56127 Pisa, Italy.

*Correspondence to: D.S.R. (dsrokhsar@gmail.com); F.G.G. (fgmitter@ufl.edu).

‡ Current address: Life Technologies Corp., Grand Island, NY 14072, USA.

† These authors contributed equally.

Abstract

Although Citrus is the most globally significant tree fruit, its domestication history is poorly understood. Cultivated citrus types are believed to comprise selections from and/or hybrids of several wild progenitor species, but the identities of these progenitors, and their contribution to modern cultivars, remain controversial. Here we report the genomes of a collection of mandarins, pummelos, and oranges, including a high quality reference sequence from a haploid Clementine mandarin. By comparative genome analysis we show that these cultivated types can

be derived from two progenitor species. Cultivated pummelos represent selections from a single progenitor species *C. maxima*. Unexpectedly, however, we find that cultivated mandarins are introgressions of *C. maxima* into a distinct second population that we identify with the ancestral wild mandarin species *C. reticulata*. Sweet and sour oranges are found to be interspecific hybrids. Sweet orange, the most widely cultivated citrus, arose as the offspring of previously admixed individuals. In contrast, sour (or Seville) orange is an F1 hybrid of pure *C. maxima* and *C. reticulata* parents, implying that wild mandarins were part of the early breeding germplasm. Surprisingly, we also find that a wild Chinese "mandarin" from Mangshan, China shows substantial sequence divergence from *C. reticulata* and appears to represent a distinct taxon. Understanding the relationships and phylogeny of cultivated citrus through genome analysis will clarify taxonomic relationships and enable previously inconceivable opportunities for sequence-directed genetic improvement.

Citrus are widely consumed worldwide as juice or fresh fruit, providing important sources of vitamin C and other health-promoting compounds. Global production in 2012 exceeded 86 million metric tons, with an estimated value of US$9 billion (http://www.fas.usda.gov/psdonline/circulars/citrus.pdf). The very narrow genetic diversity of cultivated citrus makes it highly vulnerable to disease outbreaks, including citrus greening disease (also known as Huanglongbing) that is rapidly spreading throughout the world's major citrus producing regions[1]. Understanding the population genomics and domestication of citrus will enable strategies for improvements to citrus including resistance to greening and other diseases.

The domestication and distribution of edible citrus types began several thousand years ago in Southeast Asia and spread globally following ancient land and sea routes. The lineages that gave rise to most modern cultivated varieties, however, are lost in undocumented antiquity, and their identities remain controversial[2, 3]. Several features of *Citrus* biology and cultivation make deciphering these origins difficult. Cultivated varieties are typically propagated clonally by grafting and through asexual seed production (apomixis *via* nucellar polyembryony) to maintain desirable combinations of traits (Fig. 1). Thus many important cultivar groups have characteristic basic genotypes that presumably arose through interspecific hybridization and/or successive introgressive hybridizations of wild ancestral species. These domestication events predated the global expansion of citrus cultivation by hundreds or perhaps thousands of years, with no record of the domestication process. Diversity within such groups arises through accumulated somatic mutations, generally without sexual recombination, either as limb sports on trees or variants among apomictic seedling progeny.

Two wild species are believed to have contributed to domesticated pummelos, mandarins and oranges. Based on morphology and genetic markers, "pummelos" have generally been identified with the wild species *C. maxima* (Burm.) Merrill that is indigenous to Southeast Asia. Although "mandarins" are similarly widely identified with the species *C. reticulata* Blanco [4-6], wild populations of *C. reticulata* have not been definitively described. Various authors have taken different approaches to classifying mandarins, and several naming conventions have been developed[7, 8]. Here we emphasize that the term "mandarin" is a commercial or popular designation referring to citrus with small, easy-peeling, sweet fruit, and not necessarily a

taxonomic one. We use the qualifier "traditional" to refer to mandarins without previously suspected admixture from other ancestral species, to distinguish them from mandarin types that are known or believed to be recent hybrids. For clarity we use "×" in the systematic name of such known hybrids (see *e.g.*, Ref. 9). Recognizing that genome sequencing and diversity analysis has provided insights into the domestication history of several other fruit crops[10, 11], cereals[12, 13] and other crops (reviewed in Ref. 14), we sequenced and analyzed the genomes of a diverse collection of cultivated pummelos, mandarins, and oranges to test the pummelo-mandarin species hypothesis and to uncover the origins of several important citrus cultivars.

Results

To provide a genomic platform for analyzing *Citrus*, we generated a high quality reference genome from ~7× Sanger dideoxy whole genome shotgun coverage of a haploid derivative of Clementine "mandarin" (*C.* × *clementina* cv. Clemenules) (Supplementary Note 2). The use of haploid material (derived from a single ovule after induced gynogenesis[15]) removes complications that arise when assembling outbred diploid genomes. The resulting 301.4 Mbp reference sequence is nearly complete, with superior assembly contiguity (contig L50 = 119 kbp) and scaffolding (scaffold L50 before pseudochromosome construction = 6.8 Mbp) compared to a recently published sweet orange draft sequence[16] (Supplementary Table 2.4). The long scaffolds allowed us to construct pseudochromosomes by assigning 96% of the assembly to a location on the nine citrus chromosomes using the latest *Citrus* genetic map[17], compared with only 79% in the sweet orange draft[16] (Supplementary Note 2.4). From sequence data we also inferred the phase of the two diploid Clementine haplotypes, identifying ten crossovers from the meiosis that produced the haploid Clementine, and annotated nominal centromeres as large regions of low recombination (Supplementary Note 10.1, Supplementary Fig. 10, 23). Independently we also sequenced and assembled a draft genome of the (diploid) sweet orange variety 'Ridge Pineapple' by combining deep 454 sequence with light Sanger sampling (Table 1; Supplementary Note 3) and inferred chromosome phasing using the recently reported rough draft genome of a sweet-orange-derived dihaploid[16].

The citrus genome retains substantial segmental synteny (that is, local co-linearity) with other eudicots, although it has experienced extensive large-scale rearrangement on the chromosome scale (Supplementary Note 5). Based on analysis of synteny we propose a specific model for the origin of the citrus genome from the paleo-hexaploid eudicot ancestor[18] through a series of chromosomes fissions and fusions (Supplementary Figs. 11,12). Despite the compactness of the citrus genome, 45% is repetitive, with long-terminal repeat retrotransposons and numerous uncharacterized elements, each making up nearly half of the repetitive content; the remainder comprises DNA transposons and LINEs (Supplementary Note 4.1). We identified ~25,000 protein-coding gene loci in both Clementine and sweet orange by computational methods combined with extensive long-read 454 and Sanger expressed sequence tags (Supplementary Note 4.2).

To investigate the origin of cultivated varieties we sequenced the genomes of four mandarins (including Clementine), two pummelos, and one sour orange, as well as the sweet orange genome reported above (Table 1). (Cultivars derived from *C. medica* (the third purported wild species)*, i.e.,* citrons, limes, and lemons, were not part of this study.) Two distinct types of chloroplast genomes (cpDNA) were readily identified, with mandarins all having an "M" type

and pummelos and oranges sharing a "P" type, with limited variation within each cpDNA type (Table 1; Supplementary Note 6.2), consistent with prior studies of mitochondrial markers [19]. Citrus nuclear genomes tell a more complex story. We find that while the sequenced pummelos are evidently genotypes from the sexual *C. maxima* species with minimal introgression of other species, all the mandarin-type citrus we sequenced show substantial admixture with pummelo and therefore cannot simply be selections from an ancestral *C. reticulata* population. The sweet and sour oranges are also hybrids of varying complexity, with pummelo-type chloroplast genomes in both cases.

The two diploid pummelos that we sequenced contain three distinct haplotypes, since Low acid (Siamese Sweet) pummelo is the known female parent of Chandler pummelo[20], so that the two pummelos share one haplotype at each locus (Supplementary Note 10.4). Within the two sequenced pummelos and between their non-shared alleles (derived from the other parent of Chandler, *i.e*., Siamese Pink pummelo) modest levels of heterozygosity were observed, with a genome-wide nucleotide heterozygosity of 5.7 heterozygous (het) sites/kb (Fig. 2a). The presence of a second low-heterozygosity peak (~1 het site/kb) in the distribution can be explained by a strong ancient bottleneck in the *C. maxima* population ~100-300 kya (Supplementary Note 9.3). Our reanalysis of three Chinese pummelos reported in Xu *et al*.[16] (including the Wusuan pummelo that we identify as from the same somatic lineage as Siamese Sweet pummelo), shows that both Thai and Chinese pummelos are derived from the same wild population (Supplementary Note 11.1). Only a single short 1.5 Mb segment on chromosome 2 of Chandler shows unusually high heterozygosity that could reflect interspecific introgression. These observations are consistent with pummelo domestication by selection from a wild sexual *C. maxima* population.

To sample a range of mandarin types, we sequenced two "traditional" mandarins without prior suspected admixture: Ponkan, an old and widely grown Asian variety that was presumed to be typical of *C. reticulata*, and Willowleaf, a common Mediterranean variety, as well as two mandarins believed to be hybrids of "traditional" mandarins with other citrus: Clementine, the diploid parent of the haploid reference accession, and W. Murcott (believed to be synonymous with the cultivar also known as Nadorcott and Afourer), widely grown in California and the Mediterranean (Supplementary Note 1). In contrast to pummelos, the "mandarin" accessions we sequenced typically include segments of high nucleotide heterozygosity (~17 het sites/kb, consistent with inter-specific variation) that span tens of cM or Mbp. These highly heterozygous blocks are interspersed with long segments of substantially lower levels of heterozygosity (~5 het sites/kilobase) that are consistent with intra-specific variation and clearly distinct from the higher-heterozygosity blocks (Fig. 2d). In the lower heterozygosity segments, both alleles are often distinct from those observed in the pummelos and presumably derive from *C. reticulata*, which is widely cited as the true species from which cultivated mandarins arose[7]. In contrast, the higher heterozygosity blocks typically carry one allele that matches the pummelos, and one non-pummelo allele, also presumably *C. reticulata*. The presumptive *C. reticulata* alleles are typically common to multiple mandarin accessions, further supporting their identification.

Thus, our surprising conclusion is that "traditional" mandarin types like Ponkan and Willowleaf, are in fact interspecific introgressions of *C. maxima* (pummelo) into *C. reticulata* (wild mandarin). Furthermore, while these traditional mandarins were previously thought to be unrelated, we detect extensive haplotype sharing between them (Supplemental Note 10.2). Since

microsatellite-based population structure analyses of a wide range of *Citrus* genotypes shows mandarins as a defined cluster of genotypes[21], such admixture is likely widespread among mandarin types. Indeed, reanalysis of data from Xu *et al*.[16] in the light of our discovery of interspecific introgression in multiple mandarin types, shows that the traditional Chinese Huanglingmiao mandarin (incorrectly treated by Xu *et al*. as a pure *C. reticulata*) also exhibits previously unsuspected admixture between *C. reticulata* and *C. maxima* (Supplementary Note 11.2).

Although none of our cultivated mandarin genotypes represent pure *C. reticulata*, we can nevertheless extract wild mandarin alleles from our data by comparing the (admixed) cultivated mandarins with each other and the two pure pummelos. By such genome-wide comparisons we identified 1,537,264 putative fixed single nucleotide differences between *C. reticulata* and *C. maxima* (Supplementary Note 7). These diagnostic variants can in turn be used to partition the mandarin, pummelo and orange genomes into segments according to their species ancestry (Fig. 3). The characterization of *C. reticulata* genomic segments from modern mandarins is analogous to the extraction of African haplotypes from Mexican Americans[22] and native American haplotypes from extant ethnic human populations that are admixtures with American, African, and European roots[23].

We can estimate the parameters of a simple population genetic model for the divergence of *C. reticulata* and *C. maxima* from an ancestral south Asian citrus founder population, using a coalescent framework and our collection of fixed interspecific differences and intraspecific variation (Supplementary Note 9). This analysis is consistent with effective population sizes of several hundred thousand trees for *C. maxima* and somewhat fewer for *C. reticulata*, with larger effective population size for pummelos in keeping with their higher heterozygosity. Note that the likely occurrence of apomixis in wild mandarin populations, a trait that seems to be absent in *C. maxima*, may contribute to reducing the effective *C. reticulata* population size relative to the census size. If we assume a per site mutation rate of $\mu \sim 1\text{-}2 \times 10^{-9}$/yr (comparable to that observed in poplar trees[24]) then we can estimate that *C. reticulata* and *C. maxima* diverged ~1.6-3.2 Mya, consistent with the divergence between *Citrus* and the related genus *Poncirus*, which is estimated at 4-9.6 Mya [25]. As noted, the excess of low heterozygosity segments in pummelo is consistent with a substantial population bottleneck several hundred thousand years ago and prior to the separation of Thai and Chinese pummelo lineages (Supplementary Note 9, 11).

Some specific citrus genotypes are generally recognized as "hybrid" varieties. For example, Clementine mandarin (also known as Algerian tangerine) is believed to be a chance seedling from a 'Mediterranean' mandarin (*e.g*., Willowleaf) selected just over a century ago in Algeria[26]. While various male parents have been proposed, serological and molecular studies demonstrated that the Clementine was likely a mandarin × sweet orange hybrid[6, 17, 27]. We confirm this hypothesis at the sequence level by definitively identifying a Willowleaf and sweet orange allele at each Clementine locus; demarcating the recombination breakpoints in the meiosis that produced the haploid Clementine sequence; and determining the Willowleaf and sweet orange haplotypes that contributed to diploid Clementine (Supplementary Note 10.1, Supplementary Fig. 23). Similarly, the W. Murcott mandarin is believed to be a chance zygotic seedling of Murcott tangor, itself a presumed F1 hybrid of sweet orange and an unknown mandarin. Our sequence analysis confirms the suspected grandparent/grandchild relationship between sweet orange and W. Murcott (Supplementary Note 10.2.1). Although the other parent and

grandparent of W. Murcott are not known (but see[28]), a search for these ancestors will be enabled by the other observed alleles.

Sweet orange (*C.* × *sinensis* L. Osbeck) is the citrus type most widely cultivated for fruit and juice and is widely believed to be an interspecific hybrid, but its origin is unknown[4, 6]. Different sweet orange cultivars share the same genomic organization with little sequence variation, having arisen by mutation from the original sweet orange domesticate (see, e.g. Ref. 29). Using our genome-wide catalog of fixed *C. reticulata*/*C. maxima* alleles, we can represent the sweet orange genome as segments of these two parental species or hybrid segments thereof (Supplementary Note 10.2; Fig. 3a), with clear boundaries between different segments types (Fig. 2c). A recently proposed "(P×M)×M" backcross scheme for the derivation of sweet orange from mandarin and pummelo[16], however, is easily ruled out by the presence of clear "P/P" (*i.e.*, *C. maxima*/*C. maxima*) segments in sweet orange, which requires *both* parents to have some pummelo ancestry. (The P/P segment on chromosome 2 has been confirmed by directed resequencing of three genes in this region[30].) Unexpectedly, in our analysis we found that sweet orange shares alleles with Ponkan mandarin across nearly three-quarters of the genome, and many of the same segments are also shared with Willowleaf and Huanglingmiao (Supplementary Note 10.2.1; Supplementary Fig. 25). This leads to the surprising conclusion that these three traditional mandarins, previously considered independent selections, in fact show substantial kinship with each other and an ancestor of sweet orange, suggesting much more limited genetic diversity among the traditional mandarins than previously recognized (Supplementary Note 10.2.2). The nature of the female parent of sweet orange is more difficult to infer, but the distribution of observed heterozygous segments in sweet orange is more readily accounted for if the female parent was itself a pummelo with substantial introgression of wild mandarin. Neither our pummelos, nor the related pummelos of Xu *et al*. show such admixture, which must now be sought across a broader diversity of pummelo (Supplementary Note 10.3, Supplementary Fig. 28).

Finally, Seville or sour orange (also known as *C.* × *aurantium*), which has historically been an important rootstock for citrus and, more familiarly, is used in marmalade and other products, is another traditional cultivar type that is widely regarded as a pummelo-mandarin hybrid. Our genomic analysis shows that sour orange is indeed the direct result of a simple interspecific F1 cross between a pummelo (*C. maxima*) seed parent and a wild mandarin (*C. reticulata*) pollen parent (Supplementary Note 10.4). Surprisingly in light of our discovery of widespread pummelo admixture among traditional mandarins, no such admixture is found in the *C. reticulata* parent of sour orange, but the specific parental genotypes remain unknown. Sour orange may have arisen as a natural hybrid of two wild *Citrus* species, and persisted by virtue of its reproduction through apomixis, followed by deliberate human cultivation and distribution. We found no detectable recent relationship between sweet and sour orange.

Among cultivars traditionally classified as "mandarins", however, we found another surprise. Our analysis of the genome of a presumed "wild mandarin" from Mangshan, China[16] (CMS) shows (a) a chloroplast genome that is distinct from both *C. reticulata* and *C. maxima* (Fig. 4a); (b) limited heterozygosity (Fig. 4b), again uniformly distributed across the genome, and no segments of pummelo or mandarin ancestry, indicating no admixture; (c) ~2% homozygous differences from both *C. reticulata* and *C. maxima* uniformly across the genome, a rate comparable to the divergence between *C. maxima* and *C. reticulata* (Fig 4b). At the level of

nucleotide diversity, CMS is as diverged from C. *maxima* and *C. reticulata* as they are from each other (Fig. 4b) and is clearly separated from pummelos, oranges and mandarins by principal coordinate analysis (Fig. 4c, Supplementary Note 11.3). By all these measures, we find that Mangshan "mandarin" is unrelated to the other cultivated mandarins discussed above (including Huanglingmiao mandarin). We therefore propose that despite its morphology Mangshan "mandarin" represents a distinct species from *C. reticulata*, supporting the nomenclature *C. mangshanensis* [31].

Discussion

Our genomic analyses clarify some of the murky early history of citrus domestication. The nuclear and chloroplast genomes of cultivated pummelos are consistent with the identification of pummelos as a single citrus species, *C. maxima*. In contrast, the nuclear genomes of sequenced "mandarin" type cultivars all contain substantial admixture of *C. maxima*, despite the similarity of mandarin chloroplast sequences. Our results thus show that the various conventional citrus taxonomies that associate mandarin citrus types with the ancestral citrus species *C. reticulata* are too simplistic. It is particularly surprising that even the traditional mandarin types with no prior suspicion of relatedness or admixture such as Ponkan, Willowleaf, and Huanglingmiao mandarin show substantial haplotype sharing and all include introgressed pummelo segments. A supposed "wild mandarin" from Mangshan, China, turns out to represent a distinct taxon only distantly related to *C. reticulata*, based on analysis of its nuclear and chloroplast genomes. (In Xu *et al.*'s analysis of sweet orange ancestry[16], Mangshan "mandarin" Clementine, and Huanglingmiao were used to represent *C. reticulata.* Our discovery of substantial pummelo admixture in Clementine and Huanglingmiao, and the distinctness of Mangshan "mandarin" from *C. reticulata,* further invalidates their conclusions.)

Remarkably, even in the absence of a pure type specimen for *C. reticulata* we can characterize the genome of this wild mandarin progenitor species from genome-wide comparative analysis of admixed descendants[22]. Our collection of 1,537,264 SNPs (Supplementary File 1) that differentiate *C. reticulata* from *C. maxima* can be used to guide the search for pure *C. reticulata* mandarin types (or recognize other cryptic species) among the hundreds of known cultivars and other germplasm accessions. Small-fruited mandarins that are less desirable for fresh consumption based on appearance, flavor, texture and aroma may be considered likely candidates. With the discovery that *C. mangshanensis* is a distinct group the possibility of additional undescribed wild citrus species must also be considered.

The prevalence of interspecific admixture in cultivated citrus suggests that either early in domestication or in a natural hybrid zone prior to domestication, *C. reticulata* and *C. maxima* interbreeding occurred. Given the typical size of the hybrid blocks, only a few generations of introgression occurred prior to the selection of attractive cultivars, which were then propagated asexually by apomictic or vegetative means, perhaps in southern China[32]. Our analysis of sweet orange and sour orange shows that these ancient and widely cultivated genotypes are pummelo-mandarin admixtures that are unrelated to each other, despite some degree of phenotypic similarity[33]. The discovery that sour orange is a simple F1 hybrid of *C. maxima* and *C. reticulata* implies that pure *C. reticulata* individuals were part of the breeding germplasm at the origin of sour orange. Remarkably, we found that extant Ponkan, Willowleaf, and Huanglingmiao mandarins are related to each other and to the male parent of sweet orange. Although the female

parent of sweet orange remains unknown, it cannot have been a pure pummelo (though it had pummelo cytoplasm, based on cpDNA and mtDNA[19]). Its identity is constrained by the high proportion of hybrid P/M segments in sweet orange, which can be naturally explained if the female parent of sweet orange were (PxM)xP.

Like many other agricultural enterprises, the global citrus industry relies substantially on large-scale monoculture which makes it particularly challenging to meet consumer demand for greater product diversity while trying to incorporate tolerance and/or resistance to biotic and potentially catastrophic abiotic stresses[34]. Advances in citrus genomics[35, 36] should soon allow the identification of the somatic mutations that, with their ancient genetic backgrounds, underlie the diversity of citrus color, flavor, and aroma in modern cultivars. Our analysis of the relationships between cultivated citrus and the ancestral species from which they were derived emphasizes the limited ancestral germplasm that contributed to the commercially-important cultivar types like sweet orange, and highlights the opportunities for the creation of new combinations of the ancestral citrus types with novel fruit quality traits or even the re-creation of sweet orange with improved disease resistance *via* sexual hybridization, beyond the current approaches based on somatic mutations and genetic engineering.

Online Methods

Haploid *C.* x *clementina* 'Clemenules' sequencing and assembly

A total of 4.6M Sanger reads (including 469k fosmid end and 73k BAC end reads), were obtained from an induced haploid plant *C.* × *clementina* 'Clemenules', assembled with Arachne and integrated with a genetic map producing chromosome-scale pseudo-molecules (nearly 97% of ESTs aligned to the genome) (Supplementary Note 2).

C. x *sinensis* genome sequencing and assembly

A total of 16.5 Gb sequence (36M 454 reads and 750k Sanger PE reads) was generated from *C.* × *sinensis* 'Ridge pineapple' and assembled with Newbler (Supplementary Note 3).

Annotation of repeats and genes in citrus genome assemblies

Repeat analysis was performed separately in the Clementine and sweet orange genomes. The method used RepeatModeler to find novel repeats in the genome sequence. Repeat sequences from this analysis were and RepeatMasker and PASA combined ESTs (1.6M for clementine; 6.5M for sweet orange) with Fgenesh+, exonerate and GenomeScan gene predictions to generate gene models (Supplementary Note 4).

Evolutionary comparisons with other plant genomes

Evolutionary comparisons to plant genomes used ortholog assignment to generate chromosome to chromosome relationships within and between genomes and predict ancestral genome structures (Supplementary Note 5).

Analysis of resequencing datasets

Illumina shotgun sequence reads from eight accessions (17x-110x depth; Table 1) were mapped to the haploid Clementine reference using bwa, and single nucleotide variants were identified using samtools and in-house scripts (Supplementary Note 6). Heterozygosity in diploid accessions was estimated in windows of 100-500 kb by dividing the number of confidently inferred heterozygous single nucleotide variant ("het") sites by the number of eligible sites in the window at which confident variant calls could be made, based on depth and alignment quality (Supplementary Note 6).

Identification of two ancestral species (*C. maxima vs. C. reticulata* alleles) and admixture analysis

Diagnostic alleles for the two ancestral citrus species, *C. maxima* and *C. reticulata,* were derived from a comparative analysis of two pummelos and two traditional mandarin types, and were used to study the admixture patterns in the sequenced cultivars (Supplementary Note 7,8).

Population genetic analysis and simulations

Population genetic analysis of the two citrus species and demographic inference were based on coalescent simulations conducted using MaCS (Supplementary Note 9).

Analysis of relatedness in citrus

Parentage and relatedness analysis for Clementine and Mangshan mandarin (CMS) made use of homozygous SNPs in each diploid genome relative to the haploid Clementine reference as well as to the inferred second haplotype of Clementine (Supplementary Note 10.1, 11.3). In the same way, the haploid sweet orange assembly was used for identifying shared haplotypes with sweet orange (Supplementary Note 10.2.1,10.5). A modified identical-by-state (IBS) method was used for haplotype sharing analysis among mandarins and other citrus pairs (Supplementary Note 10.2.2,10.4).

Acknowledgments

National Science and Technology Institute of Genomics for Citrus Breeding (FAPESP and CNPq) and Brazilian Agricultural Research Corporation (Embrapa) (MAT, JF-A, MMach, JJ,JG,JS) and Embrapa-Monsanto Agreement (JF-A). Agence nationale de la recherche (ANR) grant CITRUSSEQ PCS-08-GENO (XP,MR,PO); PB, AL and MB was supported in part by the US National Institute of Health grant 5R01HG00783 to M.B.; (Prometeo/2008/121) from the Generalitat Valenciana, Spain and by a grant (AGL2011-26490) from the Ministry of Economy and Innovation-Fondo Europeo de Desarrollo Regional (FEDER) (PA,LN); Conselleria de Agricultura, Pesca, Alimentación y Água from the Generalitat Valenciana (JP,DR); Ministerio de Economia e Innovación grants PSE-060000-2009-8 and IPT-010000-2010-43 and by the companies integrated in the Citruseq-Citrusgenn consortium, Anecoop S. Coop., Eurosemillas S.A., Fundación Ruralcaja Valencia, GCM Variedades Vegetales A.I.E, Investigación Citrícola Castellón,S.A and Source Citrus Genesis – SNFL (LE,JVM-S,VI,AH-O,MT); Florida Citrus Production Research Advisory Council (FCPRAC), the Florida Department of Agriculture and Consumer Services (Grant # 013646), the Florida Department of Citrus (FDOC), and the Citrus Research and Development Foundation, Inc. (Grant #71), on behalf of the Florida citrus growers

(FG,CC,WGF); Project Citrustart from Ministero delle Politiche Agricole Alimentari e Forestali; Project IT-Citrus Genomics PON_01623 from MIUR, Programma Operativo Nazionale "Ricerca e Competitività" 2007-2013 (MMorg,SS,FC,CDF,SPin,AZ). The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Author Contributions

GW Developed and applied methods to analyze citrus genetic diversity, population history, and ancestry; SP genome annotation and initial analysis of genetic diversity; JJ sequence assembly and map integration of haploid Clementine reference; JS, FM Analysis of synteny and genome evolution.; UH Analysis of population history and ancestry; KL, JP, AC, KJ Dideoxy shotgun sequencing and analysis of haploid Clementine reference; SS, SPin, AZ, CDF, XP, MRu Analysis of sequencing and resequencing data, and repetitive sequence annotation and analysis.; FC Sanger and Illumina sequencing; AL, PB, MB sweet orange gene model predictions; CC, WGF 454 sequencing of sweet orange and Siamese Sweet pummelo; CC contributions to sweet orange transcriptome, annotation, and the strategic rationale for comparative analyses; PA, LN haploid Clementine DNA; JP, DR haploid Clementine transcriptome; JT, FRT, LHE, JVM-S, VI, AH-O, MT generation of BAC clones of the haploid Clementine and provided genome sequences of diploid Clementine and Willowleaf mandarins; BD, CK, MMohi, TH, KF Sweet orange 454 transcriptome, and genome sequencing and assembly; MAMach and MAT Ponkan shotgun sequence; MRo W. Murcott shotgun sequence; MMorg Chandler pummelo, Seville sour orange shotgun sequence; GR, JF-A, FQ, LN, MRo Project coordination; DSR, FG, GW, SP wrote the paper with substantial input from MT, PO, MM, OJ, Mro; FG, DSR, OJ, PO, MMach, MMorg, MT, JSch, PW Project coordination and scientific leadership.

Figure Legends

Fig. 1. A selection of mandarin, pummelo, and orange fruits, including cultivars sequenced in this study.

Pummelos (numbered 1, 2 in outline, on left) are large trees that produce very large fruit, with white, pink, or red flesh color (2) and yellow or pink rinds. Most cultivars have large leaves having petioles with prominent wings. Apomictic reproduction is absent and most selections are self-incompatible. Mandarins (3 - 7) are smaller trees bearing smaller fruit, with orange flesh (9, 11) and rind color. Mandarins have both apomictic and zygotic reproduction and some are self-compatible. Oranges (8, 10) are generally intermediate in tree and fruit size, flesh (10) and rind color is commonly orange, and apomictic reproduction is always present. (The sour orange shown (12) is immature.)

Fig. 2. Nucleotide diversity distribution in citrus.

(**a**) Nucleotide heterozygosity distribution computed in overlapping 100 kb windows (with 5 kb step size) across the low acid (LAP) and Chandler (CHP) pummelo genomes and between the non-shared haplotypes of this parent-child pair (LAP/CHP) is shown. The peak at ~6 heterozygous sites/kb in all three pairwise comparisons represents the characteristic nucleotide diversity of the species *C. maxima*; the peak near ~1 heterozygous site/kb reflects a bottleneck in

the ancestral *C. maxima* population after divergence from *C. reticulata* (Supplementary Note 9.3). **(b)** Nucleotide heterozgosity distribution computed in overlapping 500 kb windows (with 5 kb step size) in Ponkan (PKM, solid line) and Willowleaf (WLM, dashed line) mandarins. Genomic segments are designated M/M, M/P or P/P based on a set of 1,537,264 SNPs that differentiate *C. reticulata* (M) from C. *maxima* (P). Both mandarins contain admixed segments from *C. maxima* introgression (M/P) as well as M/M segments, and these are plotted and normalized separately for easy comparison. **(c)** Nucleotide heterozgosity distribution computed in overlapping windows of 500 kb (5 kb offsets) for sweet orange (SWO) and sour orange (SSO). The three different genotypes of the SWO genome (M/M, P/P and M/P), and the SSO genotype M/P are normalized and plotted separately. **(d)** Nucleotide heterozygosity for the traditional Willowleaf mandarin (WLM) plotted along chromosome 6, computed in overlapping windows of 200 kb (with 100 kb step size). This chromosome shows an example of the clear discontinuity in single nucleotide variant heterozygosity levels between ~5/kb in the M/M segment (red bar) and ~17/kb in the M/P segment (blue bar).

Fig. 3. Admixture patterns and nucleotide diversity in cultivated citrus.

For each of the three groups of sequenced citrus, variation in nucleotide diversity (averaged over 500 kb windows with step size 250 kb) is shown across the genome for one representative cultivar above genotype maps (horizontal bars: green = *C. maxima/C. maxima*; blue = *C. maxima/C. reticulata*; red = *C. reticulata/C. reticulata*; grey=unknown; the 9 chromosomes are numbered at the top). **(a)** SWO nucleotide diversity with genotype maps for SWO and SSO. Note the *C. maxima/C. maxima* genotype (green segments present on chromosomes 2 and 8) in SWO. **(b)** WLM nucleotide diversity and genotype maps for three traditional mandarins (PKM, WLM, Huanglingmiao (HLM)) and three recent mandarin types (CLM, WMM, HCR). For the haploid Clementine reference sequence (HCR), red and green segments indicate *C. reticulata* and *C. maxima* haplotypes, respectively. All five mandarin types show pummelo introgressions (blue or green segments). **(c)** LAP nucleotide diversity and genotype maps for two pummelos (LAP, CHP).

Fig. 4: Mangshan mandarin is a species distinct from *C. maxima* and *C. reticulata*.

(a) Midpoint-rooted neighbor-joining phylogenetic tree of Citrus chloroplast genomes. (b) The frequency distributions of the pairwise sequence divergences (across 100 kb windows) between CMS and *C. maxima* (green), CMS and *C. reticulata* (red), *C. reticulata* and *C. maxima* (light blue), as well as the distinctly lower CMS intrinsic nucleotide diversity (dashed blue). (c) The first two coordinates of principal coordinate analysis (PCo) of the citrus nuclear genomes, based on pairwise distances and the metric multidimensional scaling. The *C. maxima* - *C. reticulata* axis (PCo1, 47.5% variance) separates pummelos (green) from mandarins (red), with oranges (blue) lying in between; PCo2 (19.6% of variance) separates CMS from the others.

Tables

Table 1. Sequenced cultivars and proportions derived from the ancestral species *C. reticulata* and *C. maxima*.

Three letter abbreviations as used throughout this work and common systematic designation are shown. Sequence depth reported as count of aligned reads to reference, after removal of duplicate reads. Chloroplast genome type inferred from shotgun reads aligning to the sweet orange chloroplast genome[37], with M indicating mandarin type and P indicating pummelo type. Diploid nuclear genotype proportions refer to fraction of genome in megabases using the HCR physical map (proportions of unknown genotype are not shown but can be inferred by subtracting the three genotype proportions from 100%)**.** The last two columns show proportions of *C. maxima* and *C. reticulata* haplotypes, and are derived from the three genotype proportions. max. = *C. maxima*; ret. = *C.reticulata*. *Ponkan mandarin is widely assumed to represent *C. reticulata*, but as shown here it has substantial admixture from *C. maxima*.

Cultivar	Abbr.	Common designation	Sequence generated	Cp type	ret./ ret.	ret./ max.	max./ max.	ret.	max.
Haploid Clementine	HCR	*C.* x *clementina*	7x Sanger	M	n/a	n/a	n/a	89%	11%
Clementine mandarin	CLM	*C.* x *clementina*	110x Illumina	M	58%	42%	0%	79%	21%
Ponkan mandarin	PKM	*C. reticulata**	55x Illumina	M	85%	14%	0.7%	92%	8%
Willowleaf mandarin	WLM	*C.* x *deliciosa*	110x Illumina	M	91%	8.8%	0%	95%	4.4%
W. Murcott mandarin	WMM	*C. reticulata* x *C.* x *sinensis*	25x Illumina	M	69%	30%	0.4%	85%	15%
Chandler pummelo	CHP	*C. maxima*	22x Illumina	P	0%	0.4%	99.6%	0.2%	99.8%
Low acid pummelo	LAP	*C. maxima*	17x Illumina	P	0%	0%	100%	0%	100%
Sweet orange	SWO	*C.* x *sinensis*	80x Illumina	P	14%	82%	3%	55%	44%
Seville sour orange	SSO	*C.* x *aurantium*	36x Illumina	P	0%	98%	0%	49%	49%

References

1. Bové, J. HUANGLONGBING: A DESTRUCTIVE, NEWLY-EMERGING, CENTURY-OLD DISEASE OF CITRUS. *J. Plant Path.* **88**, 7-37 (2006).

2. Reuther, W., Webber, H.J. & Batchelor, L.D. (eds.) The Citrus Industry, Vol. 1, Edn. 1. (University of California, Division of Agricultural Sciences, Berkeley; 1967).

3. Spiegel-Roy, P. & Goldschmidt, E.E. Biology of citrus. (Cambridge University Press, Cambridge ; New York; 1996).

4. Scora, R.W. On the history and origin of Citrus. *Bull. Torrey Botanical Club* **102**, 369-375 (1975).

5. Barrett, H.C. & Rhodes, A.M. A numerical taxonomic study of affinity relationships in cultivated citrus and its close relatives. *Syst. Biol.* **1**, 105-136 (1976).

6. Nicolosi, E. et al. Citrus phylogeny and genetic origin of important species as investigated by molecular markers. *TAG Theoretical and Applied Genetics* **100**, 1155-1166 (2000).

7. Swingle, W.T. & Reece, H.C. in The Citrus Industry, Vol. 1, Edn. 2nd edition. (eds. W. Reuther, H.J. Webber & L.D. Batchelor) 190-430 (University of California Press, Berkeley; 1967).

8. Tanaka, T. Fundamental discussion of Citrus classification. *Studia Citrologica* **14**, 1-6 (1977).

9. Moore, G.A. Oranges and lemons: clues to the taxonomy of Citrus from molecular markers. *Trends Genet* **17**, 536-540 (2001).

10. Cornille, A. et al. New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genet* **8**, e1002703 (2012).

11. Myles, S. et al. Genetic structure and domestication history of the grape. *Proc Natl Acad Sci U S A* **108**, 3530-3535 (2011).

12. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497-501 (2012).

13. Hufford, M.B. et al. Comparative population genomics of maize domestication and improvement. *Nat Genet* **44**, 808-811 (2012).

14. Morrell, P.L., Buckler, E.S. & Ross-Ibarra, J. Crop genomics: advances and applications. *Nat Rev Genet* **13**, 85-96 (2011).

15. Aleza, P. et al. Recovery and characterization of a Citrus clementina Hort. ex Tan. 'Clemenules' haploid plant selected to establish the reference whole Citrus genome sequence. *BMC Plant Biol* **9**, 110 (2009).

16. Xu, Q. et al. The draft genome of sweet orange (Citrus sinensis). *Nat Genet* **45**, 59-66 (2013).

17. Ollitrault, P. et al. A reference genetic map of C. clementina hort. ex Tan.; citrus evolution inferences from comparative mapping. *BMC genomics* **13**, 593 (2012).

18. Salse, J. In silico archeogenomics unveils modern plant genome organisation, regulation and evolution. *Curr Opin Plant Biol* **15**, 122-130 (2012).

19. Froelicher, Y. et al. New universal mitochondrial PCR markers reveal new information on maternal citrus phylogeny. *Tree Genetics & Genomes* **7**, 49-61 (2011).

20. Cameron, J.W.a.S., R K Chandler – an early-ripening hybrid pummelo derived from a low-acid parent. *Hilgardia* **30**, 359-364 (1961).

21. Barkley, N.A., Roose, M.L., Krueger, R.R. & Federici, C.T. Assessing genetic diversity and population structure in a citrus germplasm collection utilizing simple sequence repeat markers (SSRs). *Theor Appl Genet* **112**, 1519-1531 (2006).

22. Johnson, N.A. et al. Ancestral components of admixed genomes in a Mexican cohort. *PLoS Genet* **7**, e1002410 (2011).

23. Bustamante, C.D., Burchard, E.G. & De la Vega, F.M. Genomics for the world. *Nature* **475**, 163-165 (2011).

24. Tuskan, G.A. et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313**, 1596-1604 (2006).

25. Pfeil, B.E. & Crisp, M.D. The age and biogeography of Citrus and the orange subfamily (Rutaceae: Aurantioideae) in Australasia and New Caledonia. *Am J Bot* **95**, 1621-1631 (2008).

26. Trabut, J.L. L'hybridation des Citrus: une nouvelle tangérine "la Clémentine". *Revue Horticole* **10**, 232-234 (1902).

27. Samaan, L.G. Studies on the origin of Clementine tangerine (Citrus reticulata Blanco). *Euphytica* **31**, 167-173 (1982).

28. Luro, F. et al. in Eleventh International Citrus Congress (Wuhan, China; 2008).

29. Novelli, V.M., Cristofani, M., Souza, A.A. & Machado, M.A. Development and characterization of polymorphic microsatellite markers for the sweet orange (Citrus sinensis L. Osbeck). *Genetics and Molecular Biology* **29**, 90-96 (2006).

30. Garcia-Lor, A. et al. A nuclear phylogenetic analysis: SNPs, indels and SSRs deliver new insights into the relationships in the 'true citrus fruit trees' group (Citrinae, Rutaceae) and the origin of cultivated species. *Ann Bot* **111**, 1-19 (2013).

31. Liu, G.F., He, S.W. & Li, W.B. Two new species of citrus in China. *Acta Botanica Yunnanica* **12**, 287-289 (1990).

32. Gmitter, F.G. & Hu, X. The possible role of Yunnan, China, in the origin of contemporary citrus species (Rutaceae). *Economic Botany* **44**, 267-277 (1990).

33. Morton, J.F. Fruits of Warm Climates. (Florida Flair Books, Miami, Florida, USA; 1987).

34. Gottwald, T.R. Current epidemiological understanding of citrus Huanglongbing. *Annu Rev Phytopathol* **48**, 119-139 (2010).

35. Talon, M. & Gmitter, F.G., Jr. Citrus genomics. *Int J Plant Genomics* **2008**, 1-17 (2008).

36. Gmitter, F.G. et al. Citrus genomics. *Tree Genetics & Genomes* **8**, 611-626 (2012).

37. Bausher, M.G., Singh, N.D., Lee, S.B., Jansen, R.K. & Daniell, H. The complete chloroplast genome sequence of Citrus sinensis (L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. *BMC Plant Biol* **6**, 21 (2006).

Fig. 1

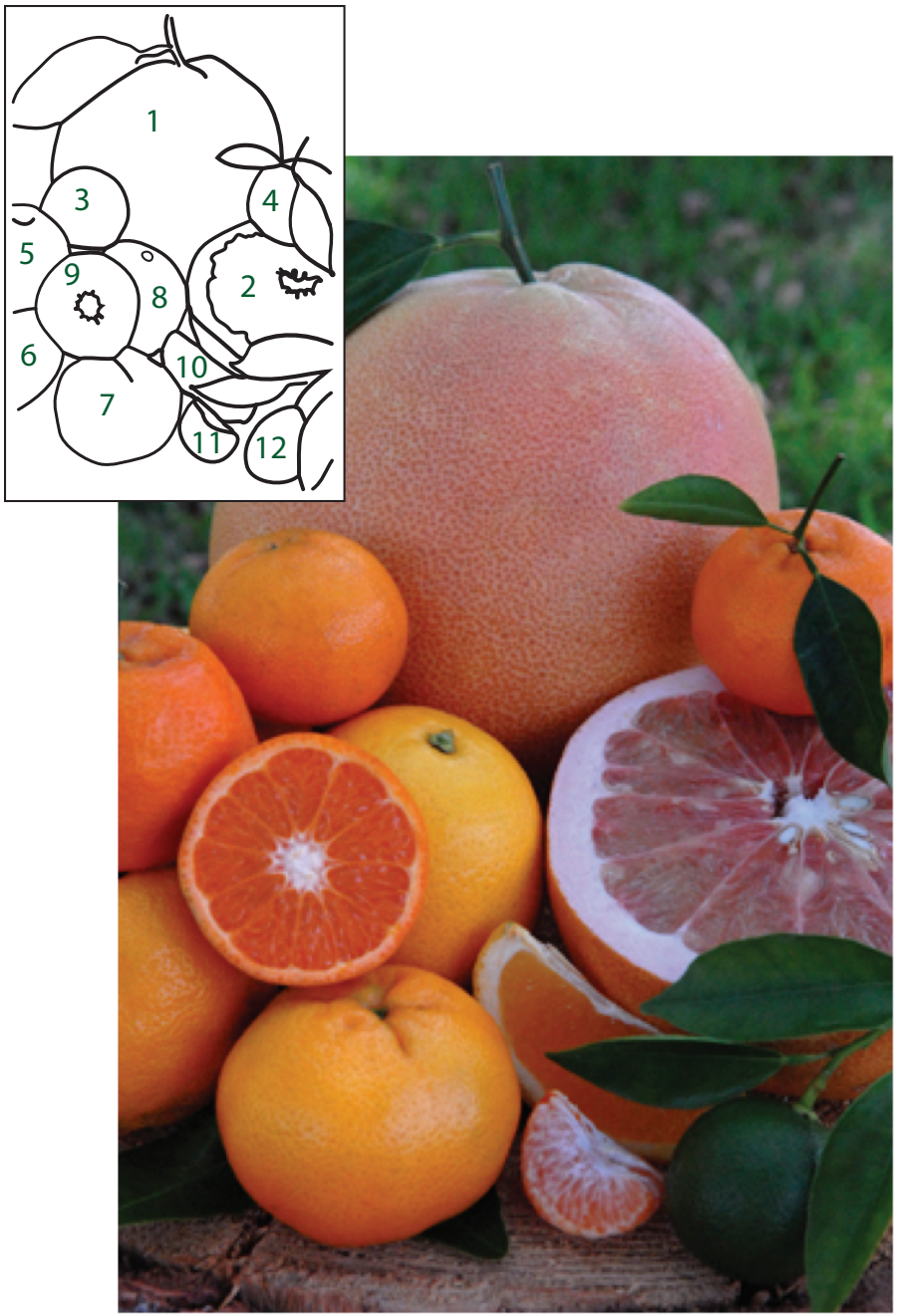


Fig. 2

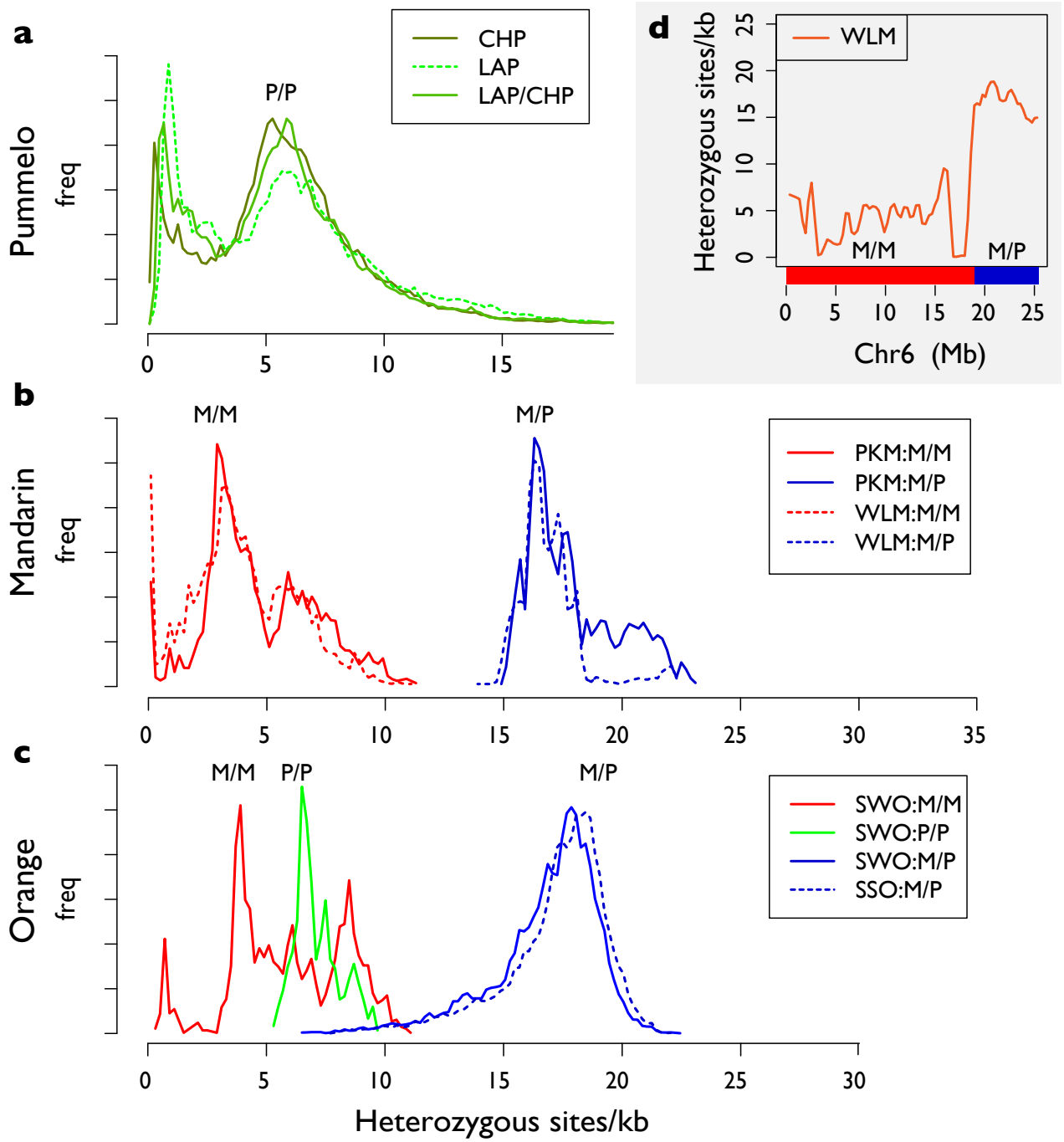

a
Pummelo
freq
CHP
LAP
LAP/CHP
P/P
0
5
10
15
b
Mandarin
freq
M/M
M/P
PKM:M/M
PKM:M/P
WLM:M/M
WLM:M/P
0
5
10
15
20
25
30
35
c
Orange
freq
M/M
P/P
M/P
SWO:M/M
SWO:P/P
SWO:M/P
SSO:M/P
0
5
10
15
20
25
30
Heterozygous sites/kb
d
Heterozygous sites/kb
WLM
M/M
M/P
0
5
10
15
20
25
Chr6 (Mb)

Fig. 3

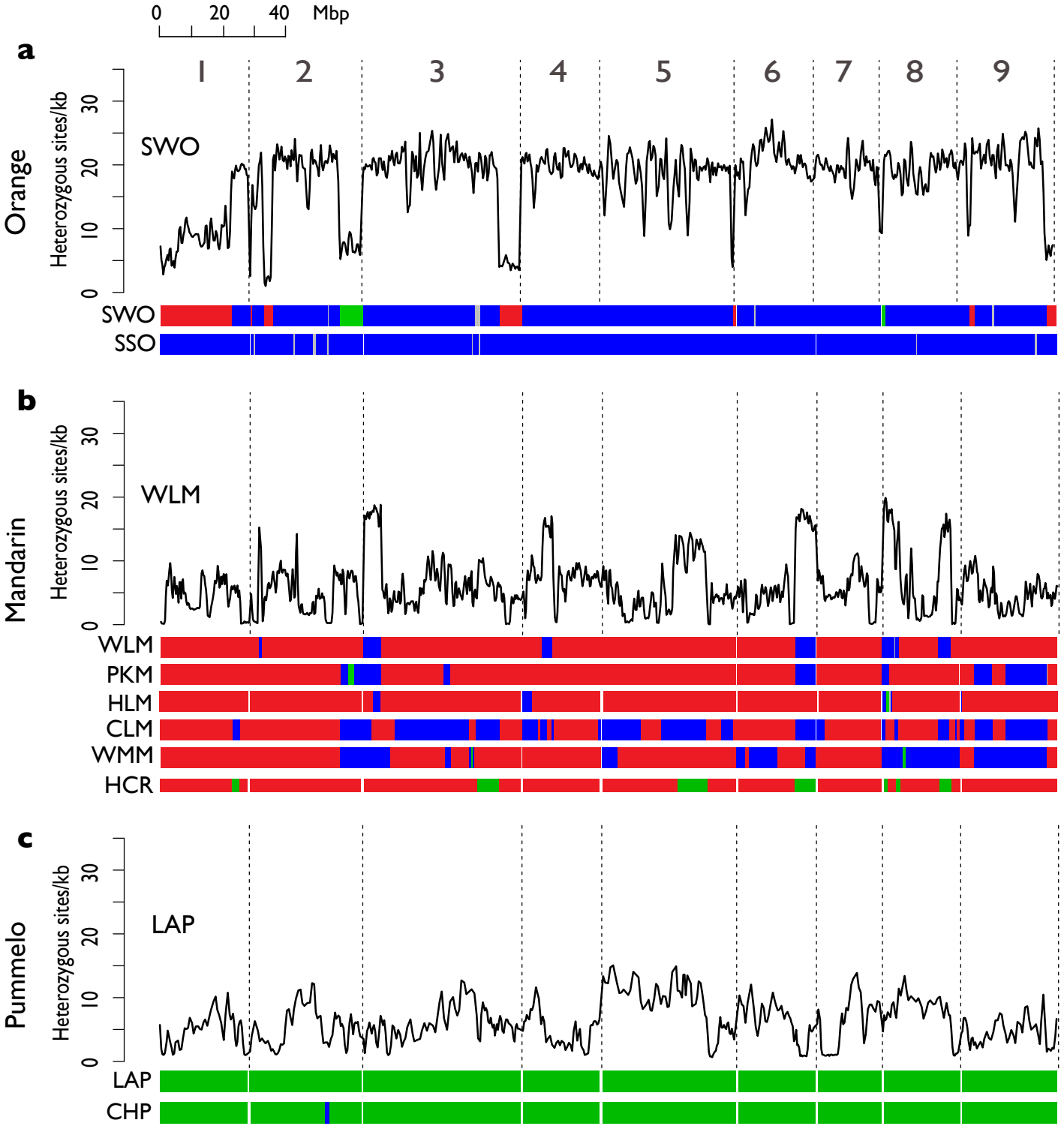
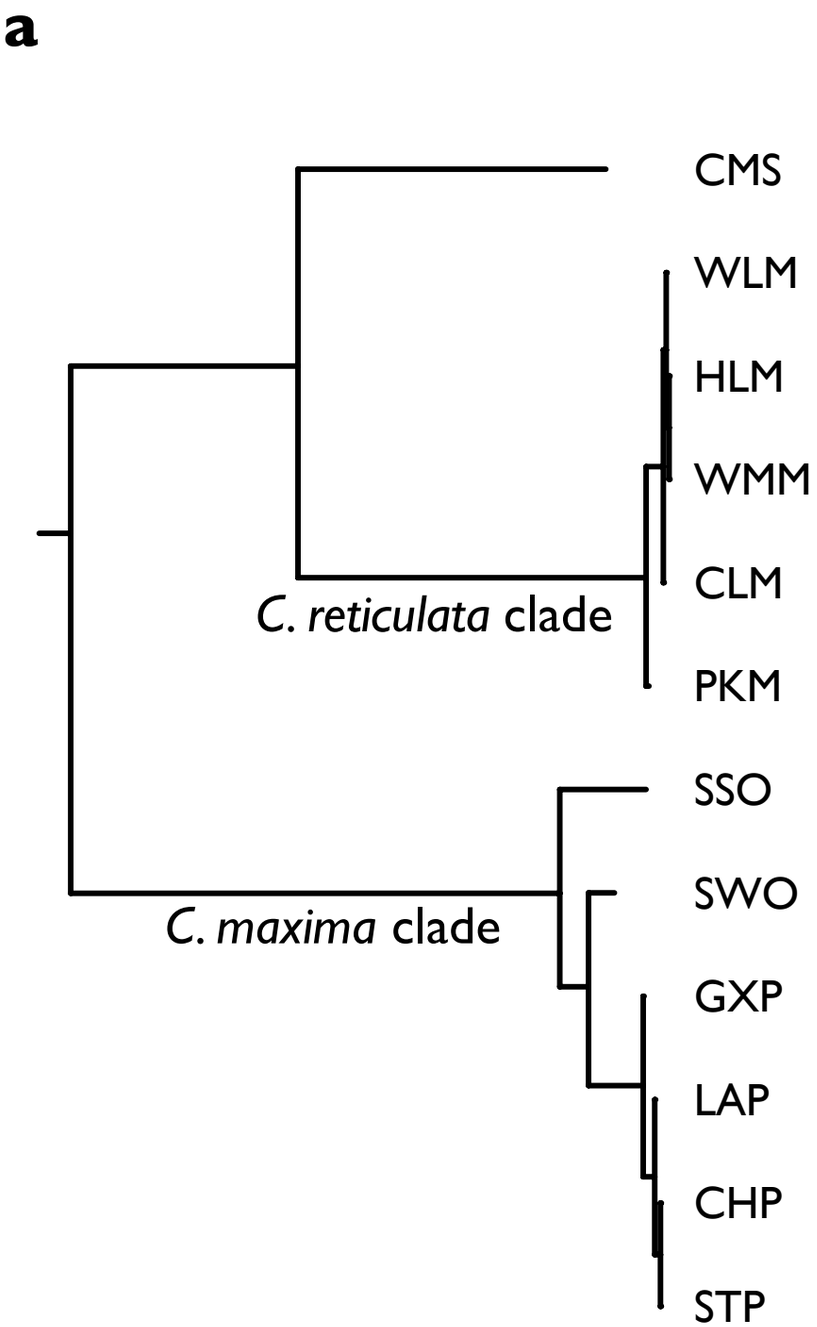
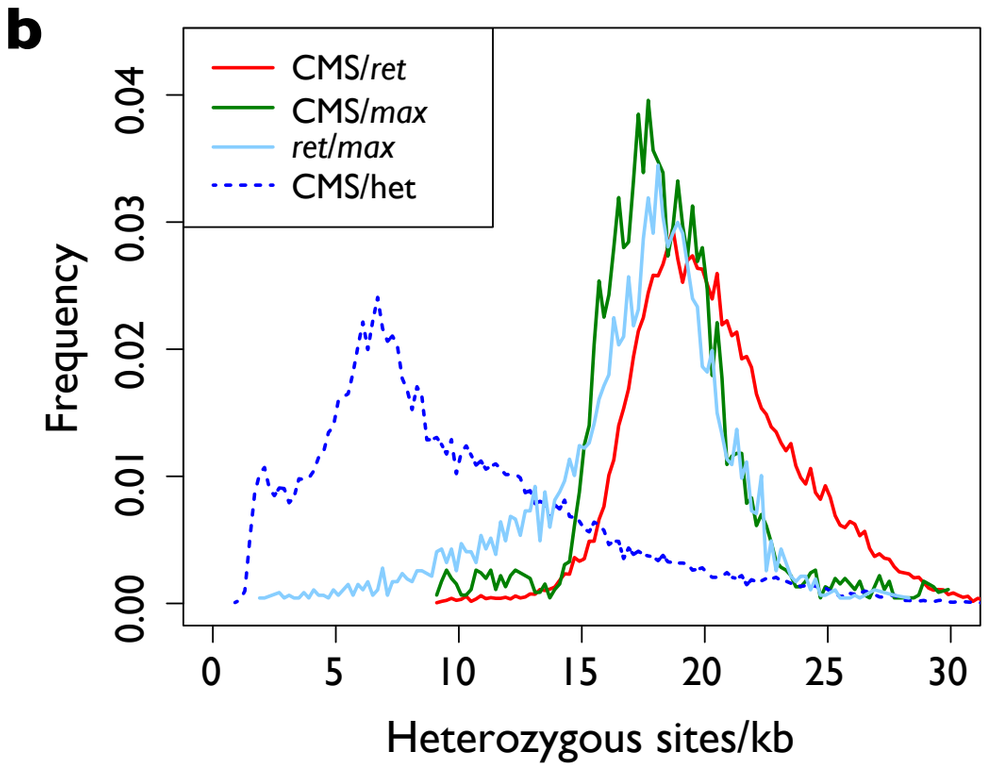
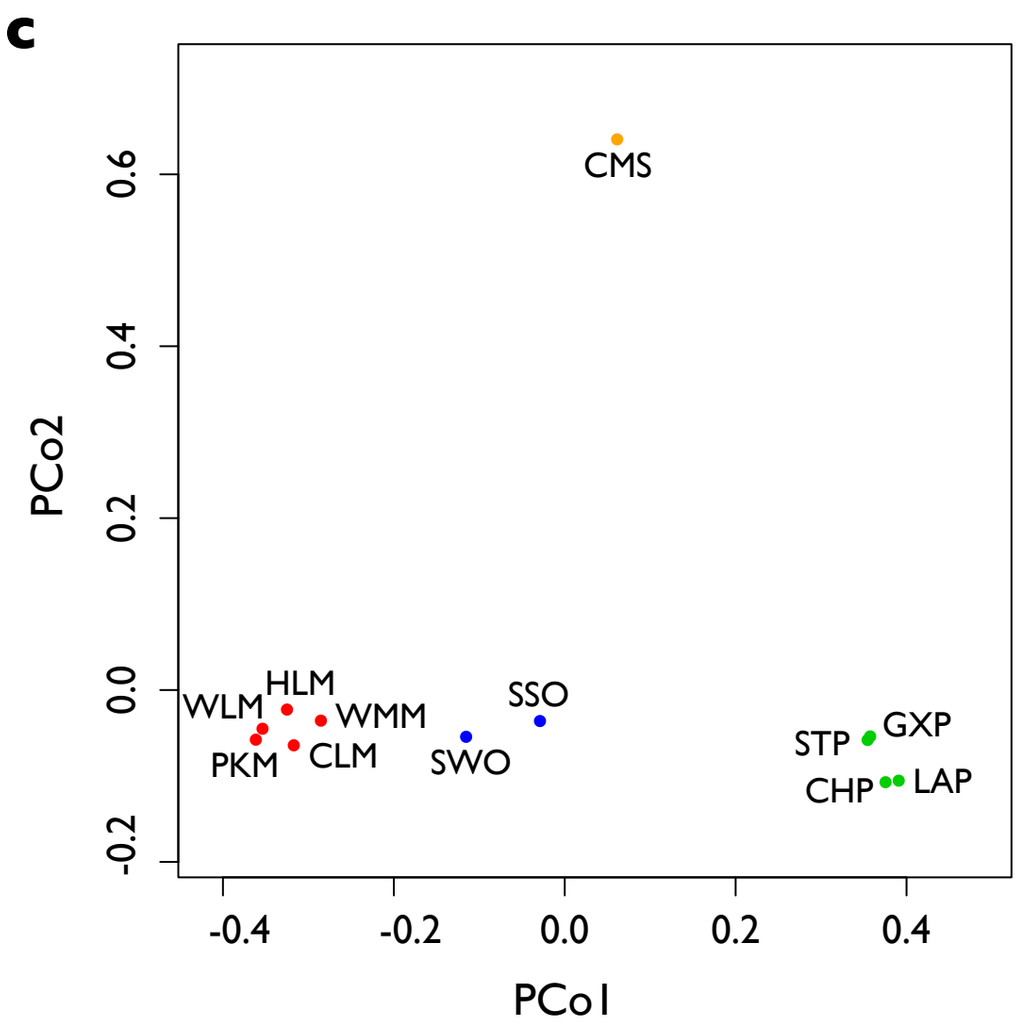

0
20
40
Mbp
a
1
2
3
4
5
6
7
8
9
Orange
Heterozygous sites/kb
0
10
20
30
SWO
SWO
SSO
b
Mandarin
Heterozygous sites/kb
WLM
WLM
PKM
HLM
CLM
WMM
HCR
c
Pummelo
Heterozygous sites/kb
LAP
LAP
CHP

Fig. 4


a
CMS
WLM
HLM
WMM
CLM
PKM
C. reticulata clade
SSO
SWO
GXP
LAP
CHP
STP
C. maxima clade


b
CMS/ret
CMS/max
ret/max
CMS/het
Frequency
0.00
0.01
0.02
0.03
0.04
0
5
10
15
20
25
30
Heterozygous sites/kb


c
CMS
HLM
WLM
WMM
SSO
PKM
CLM
SWO
STP
GXP
CHP
LAP
PCo2
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.2
0.0
0.2
0.4
PCo1

Complex history of admixture during citrus domestication revealed by genome analysis

Supplementary Text

Table of Contents

Supplementary Note 1: Background information on citrus taxonomy and sequenced cultivars

Citrus and related genera are generally considered to have originated and diversified genetically in an area that extends from northeast India and Myanmar northwards into southwest China, through southeast Asia and the Malay archipelago, and southwards to Australia[1, 2]. The earliest recorded references to citrus fruit come from China more than 4,000 years ago[3], though domestication and the distribution of edible citrus types undoubtedly occurred substantially earlier in prehistoric times. Citrus slowly spread from Asia throughout the world, following ancient land and sea trade routes.

Various taxonomic systems have been used to describe the diversity of citrus forms and species: Swingle recognized 16 distinct species in 2 subgenera[4], while Tanaka expanded the list to 162 species[5]. More contemporary work has suggested that there are three true biological citrus species, and possibly a fourth, that have contributed to the origins of the commonly known and major cultivated types of citrus fruit [6-8]. These species are citron (*C. medica* L.), pummelo (*C. maxima* (Burm.) Merrill), mandarin (*C. reticulata* Blanco), and *C. micrantha* Wester. We show here that the cultivated mandarins are not simply selections from a single wild species, but are in fact admixtures of two ancestral wild species. We also provide evidence from genome analysis that Mangshan mandarin represents a distinct species from *C. reticulata* and *C. maxima*.

The so-called "secondary species" such as sweet orange (*C.* × *sinensis* L. Osb.), grapefruit (*C.* × *paradisi* Macf.), lemon (*C.* × *limon* (L.) Burm. f.) and lime (*C.* × *aurantifolia* (Christm.) Swingle) presumably arose from serial introgressive hybridizations of two or more of these four true species. For clarity we represent these secondary species herein with the symbol "×" in the systematic name (e.g. *C.* × *clementina*), and note that these secondary species are not true sexual species in the usual sense, but are instead derived from a single progenitor genome by asexual means, and differ only by accumulated mutations.

There is abundant documentation of the origin of grapefruit as a hybrid between pummelo and sweet orange and the subsequent diversification and proliferation of grapefruit cultivars from a common ancestor that was introduced into Florida in the early 1800s (Ref. [9]). It has long been assumed that the generation of sweet orange and its diversification into contemporary cultivars is a similar story to that of grapefruit, but no historical documentation is available. Of particular note in this context, pummelo and mandarin are presumed to be the true species contributing to sweet orange.

The diversification of cultivars within other groups (sweet orange, Clementine and Satsuma mandarins, etc.) has likewise come from accumulated somatic mutations in the lineages derived from the ancestral varietal genotypes. These mutations have been discovered either as sport limbs on trees or among nucellar

embryo-derived seedling progeny. There are few known places in the world where the wild ancestors of citrus can be found, and the identities and pedigrees of the presumptive ancestral individuals of most modern cultivated citrus varieties have been lost in undocumented antiquity.

Although the ancestral species are interfertile, as are their hybrids, the degree of phenotypic and sequence differentiation that they display is typical of biological species in other groups. It is likely that the ancestral species were once geographically isolated but were brought together by human intervention, leading to extensive spontaneous hybridization and the origin of the secondary species. To some extent, the proliferation of species names in the various commonly used taxonomic systems derives from a desire to retain species status for the commercially important cultivar groups. A further complication is the occurrence of a type of apomixis, nucellar embryony, among some of the mandarins and many secondary species. Seeds from these apomictic types are frequently polyembryonic, bearing primarily nucellar (and occasionally zygotic) embryos. In such taxa, the phenotype is "true breeding" as expected for a biologically valid species. Pummelos and citrons are not known to have nucellar embryony, although there has been some anecdotal speculation that citron may at times produce nucellar embryos, despite the monoembryony observed in the seeds.

Supplementary Table 1.1: Sequenced Citrus accessions.
Abbreviations in the last column, Mono = monoembryonic, seeds containing zygotic embryos; Poly = polyembryonic, seeds typically containing nucellar embryos via apomixis. Sequencing contributions: (1) International Citrus Genome Consortium (Genoscope, France; DOE Joint Genome Institute, USA; IGA, Italy; in collaboration with and through support of the ICGC members from Brazil, France, Italy, Spain and the USA), (2) Spain (IVIA, CNAG), (3) Brazil (CCSMIA), (4) UCR, (5) IGA, (6) UF, (7) Roche Life Sciences/454.

Cultivar	Abbr.	Seq. source	Other common names	University of California Riverside Citrus Variety Collection (http://www.citrusvariety.ucr.edu/index.html)	Alternate systematic name(s)	Repro-ductive mode
Haploid Clemenules Clementine (reference genome)	HCR	(1)				n/a
Clemenules Clementine mandarin	CLM	(2)	Clementina de Nules	*C. clementina* hort ex. Tanaka	*C.* × *clementina*	*mono*
Ponkan mandarin	PKM	(2,3)	Chinese Honey	*C. reticulata* Blanco	*C. reticulata*	*poly*
Willowleaf mandarin	WLM	(2)	Mediterranean mandarin	*C. deliciosa* Ten.; also referred to as *C. reticulata* Blanco	*Swingle: C. reticulata; Tanaka: C.* × *deliciosa*	*poly*
W. Murcott mandarin	WMM	(4)	Likely equivalent to Nadorcott, Afourer	*C. reticulata* Blanco		*poly*
Chandler pummelo	CHP	(5)	Hybrid of Siamese Sweet (acidless) pummelo (seed parent) and Siamese Pink pummelo	*C. maxima* (Burm.) Merrill	*C. maxima*	*mono*
Low acid pummelo	LAP	(6)	Siamese sweet pummelo. Siamese acidless pummelo. Seed (i.e., female) parent of Chandler.	*C. maxima* (Burm.) Merrill	*C. maxima*	*mono*
Sweet orange ('Ridge Pineapple')	SWO	(1,2,6,7)	Seedy cultivar type *C.* × *sinensis*	*C. sinensis* (L.) Osbeck	*C.* × *sinensis*	*poly*
Seville or sour orange	SSO	(5)	Sevillano	*C. aurantium* L.	*C.* × *aurantium*	*poly*

The history of the sequenced accessions is given below.

Clementine (CLM). The diploid clone of Clemenules (CLM) is a bud sport of Fina, which is probably the original Clementine brought to Spain from Algeria; it has been maintained in the IVIA Germplasm Bank as IVIA-022. Clemenules is the most important Clementine variety in Spain. Clementine itself is believed to be derived from a cross between Mediterranean mandarin and sweet orange; this concept was first proposed by Samaan [10] who indicated that the seed parent of Clementine was most likely the 'Baladi' mandarin (synonymous for the Mediterranean or Willowleaf mandarin in Egypt), and a most likely pollen parent was sweet orange, based on serological analysis of pollen grain proteins, among five putative parental types tested.

The **Haploid Clementine (HCR)** used for the reference genome is from a haploid tree derived from a single ovule of Clemenules Clementine after induced gynogenesis [11], and so represents a single haplotype produced by one round of meiotic recombination of diploid Clemenules Clementine. It has been maintained in the IVIA Germplasm Bank as accession number IVIA-638.

Ponkan (PKM) is the most widely grown mandarin in the world, with China, India (where it is known as Nagpur suntara), and Brazil being the countries where it is the dominant mandarin cultivar [12]. We sequenced a typical Brazilian type that was introduced into the germplasm collection of Instituto Agronomico de Campinas in 1930s (accession number IAC-06018) as well as a Spanish variety obtained from a commercial nursery. It therefore has no accession number.

Willowleaf (WLM) is a mandarin of unknown origin that has been grown in the Mediterranean basin under various names since the 1800s. The Willowleaf mandarin clone that was sequenced was acquired from a commercial nursery, therefore there is no specific accession number associated with it.

W. Murcott (WMM) is a mandarin believed to have originated from a zygotic seedling of Murcott, and with pollen parentage uncertain. It was imported to California from Morocco in 1985, where it is also known as Nadorcott or Afourer (University of California, Riverside, Citrus Variety Collection, http://www.citrusvariety.ucr.edu/citrus/wmurcott.html). Murcott itself is presumed to be a tangor, i.e., a cross between a tangerine (mandarin) and a sweet orange [13].

Sour orange (SSO) probably originated in northeastern India or adjacent areas and was one of the first citrus fruits brought to Europe. Arabs are thought to have carried it to Arabia in the 9th century. It was reported to be growing in Sicily in 1002 AD and it was cultivated in Seville, Spain at the end of the 12th century[14]. Seville sour orange is a type typically used for marmalade production. The selection of sour orange from which the genome sequence was derived is known as Santa Marina 1, and the source tree is part of the germplasm collection at the Pallazelli farm of the CRA-ACM. Santa Marina 1 is a typical sour orange and is widely used as a rootstock for citrus trees in Italy.

The **sweet orange (SWO)** cultivar that was sequenced, the Ridge Pineapple, was originally selected as a nematode resistant rootstock and released in Florida by Harry Ford (University of Florida Citrus Experiment Station) in 1964. This clone was originally selected and used by the ICGC because it is assumed that, being seedy compared to many other commercial orange cultivars, it should have a lower likelihood of the chromosomal rearrangements that are typically associated with reduced seed numbers in several citrus types. The specific tree from which the sequence was derived is in the Florida Department of Agriculture and Consumer Services Foundation Block under accession number SPB-602. This is the same clone from which a BAC library was developed.

The **Low-Acid pummelo (LAP)** (more properly known as the Siamese Sweet pummelo) genome that was sequenced came from a sample tree in the UF-Citrus Research and Education Center germplasm collection in Florida; the budwood was originally supplied to the UF-CREC from the USDA-ARS collection, by H. C. Barrett. This accession was first introduced in the USA by the USDA, and is also held in the Citrus Variety Collection at UC-Riverside as accession number CES 2240. It is the maternal parent of Chandler pummelo, the other pummelo accession sequenced.

The specific clone of **Chandler pummelo (CHP)** that was sequenced was purchased from a commercial citrus nursery, therefore there is no specific accession number associated with it.

Supplementary Note 2: Haploid *C.* × *clementina* 'Clemenules' sequencing and assembly

Supplementary Note 2.1: Source material for haploid Clementine (*C.* × *clementina* 'Clemenules')

DNA for sequencing was extracted from leaves collected from plants growing in a temperature-controlled greenhouse that were vegetatively propagated from the original haploid plant. This original haploid plant was obtained by *in situ* parthenogenesis of clementine Clemenules induced by irradiated pollen of Fortune mandarin, followed by direct embryo germination *in vitro*. The hemizygosity of this plant was confirmed by analysis of 238 SSRs markers. The haploid plant was included in the Citrus Germplasm Bank of the Instituto Valenciano de Investigaciones

Supplementary Note 2.2: Shotgun sequencing of haploid Clementine.

Paired-end reads from shotgun libraries were collected with standard Sanger sequencing protocols on ABI 3730XL capillary sequencing machines at Genoscope in Evry, France; Department of Energy Joint Genome Institute in Walnut Creek, California; Institute for Genomic Applications in Udine, Italy; and the HudsonAlpha Institute, Huntsville, Alabama. All data has been deposited in the NCBI Trace Archive (SEQ_LIB_ID='Citrus clementina') and the assembly and annotation has been deposited at NCBI Genome database under the Accession AMZM00000000; BioProject ID PRJNA47957. Shotgun data are summarized in Supplementary Table 2.1, including insert size and standard deviation as measured from placement of read pairs on the assembly.

Supplementary Table 2.1: Shotgun sequencing summary for haploid *C.* × *clementina*.

Genomic libraries included in the haploid *C.* × *clementina* genome assembly and their respective assembled sequence coverage levels in the final release. Sequencing center abbreviations are GS Genoscope, IGA Institute of Applied Genomics, HA HudsonAlpha, JGI Joint Genome Institute.

Library ID	Avg insert size ± std. dev. (base pairs)	Number of reads sequenced	Assembled Sequence Coverage (x)	Sequencing source (center)
AEU0AAC	2,273 ± 70	316,711	0.481	GS
AEU0AAE	2,453 ± 79	662,991	1.010	GS
AEU0AAA	2,952 ± 190	249,569	0.279	GS
CL1A	2,993 ± 352	397,077	0.654	IGA
ORA	2,992 ± 545	230,400	0.340	HA
ORB	3,070 ± 576	192,000	0.310	HA
GYSX	6,897 ± 673	414,048	0.690	JGI
GPUY	6,867 ± 665	29,952	0.050	JGI
GXXX	6,942 ± 677	341,376	0.600	JGI
AEU0ABB	8,447 ± 1793	686,158	1.087	GS
AEU0AAB	10,845 ± 825	101,230	0.135	GS
AEU0ABA	11,902 ± 1,663	414,923	0.664	GS
GXIX	38,762 ± 3,750	469,344	0.560	JGI
CRETE	101,415 ± 26,114	36,640	0.055	IGA
CRETB	120,757 ± 31,739	18,771	0.0275	IGA
CRETH	142,032 ± 51,908	18,070	0.0275	IGA
Total		**4,579,260**	**6.97**	

Supplementary Table 2.2: Raw shotgun sequence assembly summary for haploid *C.* × *clementina*.

Summary of the raw haploid Clementine whole genome shotgun assembly as produced directly from the ARACHNE assembler, prior to contaminant screening and map integration. The table shows total contigs and total assembled base pairs for each set of scaffolds greater than the specified size.

Scaffolds longer than … (bp)	No. of Scaffolds	No. of contigs	Total Scaffold Size	Non-gap bp	% Non-gap bp
5,000,000	23	3,722	226,930,946	223,759,276	98.60%
2,500,000	34	4,451	264,254,615	260,507,467	98.58%
1,000,000	45	4,826	282,941,158	278,832,978	98.55%
500,000	50	4,915	287,038,503	282,842,522	98.54%
250,000	53	4,944	288,205,900	283,976,109	98.53%
100,000	66	5,094	290,432,044	285,849,032	98.42%
50,000	93	5,359	292,203,691	287,363,244	98.34%
25,000	179	5,977	295,128,813	289,952,134	98.25%
10,000	525	7,358	300,348,022	294,333,068	98.00%
5,000	1,211	8,938	304,956,949	298,378,296	97.84%
2,500	2,415	11,145	309,481,095	302,408,280	97.71%
1,000	2,638	11,492	309,906,778	302,738,072	97.69%
0	2,931	11,785	310,048,644	302,879,938	97.69%

A total 4,579,260 reads as summarized in Supplementary Table 2.1 were assembled using a modified version of Arachne v.200710161 (Ref. [15]) with parameters maxcliq1=100, correct1_passes=0 and BINGE_AND_PURGE=True. This produced 2,931 scaffold sequences, with L50 of 6.8 Mb, 66 scaffolds larger than 100 kb, and total assembled size of 302.9 Mb (310.0Mb including Ns). Raw assembly statistics are shown in Supplementary Table 2.2.

Supplementary Note 2.4: Genetic map integration and chromosome-scale pseudomolecule construction

The citrus genetic map[16] was used to identify false joins in the initial assembly. Scaffolds were broken if they contained a putative false join coincident with an area of low BAC/fosmid coverage. A total of 5 breaks were identified and broken, resulting in 2,936 scaffolds in the broken assembly. Genetic markers were aligned to the broken assembly using two methods. First, SSR markers were placed using three successive rounds of e-PCR [17] with N=0, N=1 and N=3. Second, markers with sequence associated with them were placed with BLAT [18] and BLASTN [19] and the best placement (based on alignment identity and marker coverage) was selected to position the marker. A total of 59 scaffolds had markers that aligned to them.

Optimal order and orientation of the broken scaffolds was obtained from the marker positions, and care was taken to properly orient the telomere in the production assembly (identified using the TTTAGGG repeat). Along with the 59 scaffolds containing marker alignments, BAC/fosmid joins were used to incorporate 2 additional scaffolds. Hence, a total of 61 scaffolds were joined using 52 joins to form the production assembly containing 9 chromosomes capturing 283.8 Mb of non-gap sequence (288.6 Mb including gaps).

A subset of 16 scaffolds could not be reliably oriented using marker placements. The orientation of 11 out of 16 un-oriented scaffolds was resolved using an analysis of BAC/Fosmid joins, leaving 5 scaffolds unresolved (Supplementary Table 2.3).

Each map join is denoted with 10,000 Ns. Including gaps, the pseudomolecules contain 288.6 Mb out of 310.0 Mb total assembled sequence (93.1%). After screening for contaminants, the release assembly containing chromosomes and unmapped *C. clementina* scaffolds, is composed of 1,398 scaffolds covering a total of 301.4 Mb with a contig L50 of 118.9 kb and a pseudomolecule L50 of 31.4 Mb.

Plots of the marker placements on the chromosomal pseudomolecules are shown in Supplementary Figures 1-9.

Supplementary Table 2.3: Unoriented scaffolds in haploid Clementine. Summary of the five regions in linkage groups 1 and 9 where the combination of low marker density, poor synteny, and no BAC/Fosmid support prevented the scaffold orientation from being determined.

Linkage Group	Start	End	Length (bp)
1	6,793,809	6,796,892	3,083
1	6,806,892	9,262,838	2,455,946
1	13,619,782	15,214,087	1,594,305
9	10,984,017	14,462,003	3,477,986
9	14,472,003	14,492,897	20,984

Supplementary Note 2.5: Screening and final assembly release.

We classified the remaining scaffolds in various bins depending on sequence content. We identified contamination using megablast against GenBank NR [20] and BLASTP [19] against a set of known microbial proteins. We classified additional scaffolds as unanchored rDNA (7), mitochondrion (11), chloroplast (27), low base percentage (26), and repetitive (1064). We also removed 351 scaffolds that were less than 1kb in sequence length. The resulting final statistics are shown in Supplementary Table 2.4. The genome sequence is publicly available at http://www.phytozome.net.

Supplementary Note 2.6: Assembly Completeness.

Based on similarity searches with 114,618 citrus ESTs obtained from GenBank [20], it was estimated that at least 97.4% of available expressed gene loci were included in the 9 chromosome assemblies. The ESTs that were not found were screened against GenBank and over half of them were identified as prokaryotic rDNA.

Supplementary Table 2.4. Comparision of the haploid *C.* × *clementina*, *C. sinensis* double haploid (Xu et al. [21]) and *C.* × *sinensis* assemblies.
This table compares the released assemblies of haploid *C.* × *clementina*; double haploid *C. sinensis* (downloaded from http://citrus.hzau.edu.cn/orange) and diploid *C.* × *sinensis*.

	Haploid *C.* × *clementina* raw assembly (this study)	Haploid *C.* × *clementina* v1.0 pseudomolecules this study)	Double haploid *C. sinensis* pseudomolecules (Ref. [21])	*C.* × *sinensis* v1.0 "Ridge Pineapple" (this study)
Assembly total size (Mb)	310.0	301.4	327.8	319.2
Total number of scaffolds	2,931	1,398	4,811	12,574
Total L50 (Mb)	6.8	31.4	1.8	0.2505
Longest scaffold (Mb)	30.5	30.5	8.4	5.93
Anchored assembly size (Mb)	n/a	288.6 (95.98 %)	239.0 (72.9%)	n/a
Number of anchored scaffolds	n/a	48	160	n/a
Anchored scaffold (chromosomal pseudomolecule) L50 (Mb)	n/a	31.4	28.8	n/a
Longest pseudomolecule	n/a	51.05	36.15	n/a
Total number of contigs	11,785	8,692	17,140	53,536
Contig L50 (kb)	115.9	118.9	51.0	6.6
Longest contig (Mb)	1.23	1.23	0.323	0.119
Total size of contigs (Mb)	302.9 (2.3% gaps)	295.2 (2.1% gaps)	301.0 (8.2 % gaps)	252.2 (20.9% gaps)
Number of anchored contigs	n/a	4,955	7,839	n/a
Anchored contig L50 (kb)	n/a	122.8	57.8	n/a
Anchored contig assembly size (Mb)	n/a	283.8 (93.7%)	223.9 (74.3%)	n/a
%GC	35.0	35.0	34.1	34.6

Supplementary Note 3: *C.* × *sinensis* genome sequencing and assembly

Supplementary Note 3.1: Source material for diploid sweet orange (*C.* × *sinensis* 'Ridge Pineapple')

DNA for sequencing was extracted from leaves that had been collected from a mature tree of the 'Ridge Pineapple' sweet orange. This tree was growing in a citrus orchard managed by the Citrus Budwood Registration Bureau (CBRB), Division of Plant Industry (DPI), Florida Department of Agriculture and Consumer Services (FDACS), in Dundee, Florida. This orchard was established by FDACS-DPI-CBRB as a source for seeds of citrus rootstock cultivars for the Florida citrus nursery industry; therefore the original source tree was certified as true to type.

Supplementary Note 3.2: Shotgun sequencing of diploid sweet orange.

Genome sequence was generated on Sanger and 454 platforms. Approximately 1.2× depth of shotgun coverage was produced using paired-end Sanger sequencing. Sequencing reads were collected with standard Sanger sequencing protocols on ABI 3730XL capillary sequencing machines at the Department of Energy Joint Genome Institute in Walnut Creek, California. All data has been be deposited in the NCBI Trace Archive (SEQ_LIB_ID='CITRUS SINENSIS' and CENTER_NAME='JGI'). Shotgun data are summarized in Supplementary Table 2.1.

Supplementary Note 3.2.1: Libraries, read lengths, quality, estimated coverage.

To ensure that coverage of different parts of the genome was as even as possible (because this decreases the number of gaps caused by uneven sampling), 14 different single-end shotgun libraries were prepared (Supplementary Table 3.4) as well as several paired-end libraries with pair distances of 3kb and 8kb (Supplementary Table 3.4) for sequencing on the 454 platform. A total of 51.5× of 454 sequencing data were generated. The raw sequence data were screened against the organelle sequences (chloroplast and mitochondria) to generate the best possible genomic assembly without any contaminant sequence. The assembly was generated using 454 GS *de novo* Assembler version 2.3 ('Newbler').

Supplementary Table 3.4: Shotgun sequencing summary for diploid *C.* × *sinensis*.

Statistics are given for the amount of sequence obtained from various libraries by the sequencing centers involved in generating sweet orange sequence data. (Abbreviations: seq., sequence; cov. coverage; ave. average, HQ high quality)

Sequencing centre	Library count or name	Platform	Raw Reads (M)	Raw Seq. (Gbp)	Ave. raw read length (bp)	Raw cov.	HQ reads (M)	HQ seq (Gbp)	Ave HQ read length (bp)	HQ cov.
U Florida	4	454 Shotgun FLX	9.76	2.71	277.7	8.5×	9.80	2.27	231.9	7.1×
Roche	10	454 Single end shotgun Titanium	6.77	4.08	602.4	12.8×	6.77	2.27	335.0	7.1×
	3	454 nominally 3kb paired end (mapped insert size 1.8 ± 0.7 kb)	11.29	5.20	460.1	16.2×	11.29	3.54	313.2	11.1× (22.8× clone cov.)
	8	454 nominally 8kb paired end (mapped insert size 5.7 ± 1.1 kb)	8.12	3.84	472.8	12.0×	8.12	2.57	316.5	8.1× (44× clone cov.)
JGI	BUZX	Sanger PE (estimated 3.4 kb insert) nominally 3kb ± 0.5 kb	0.0076	0.0085	1,108	0.03×	0.0066	0.0074	701	0.02×
	BUZY	Sanger PE (estimated 6.6 kb insert) nominally 8kb ± 2.8 kb	0.503	0.425	846	1.3×	0.453	0.376	700	1.2×
	BUZZ	Sanger PE (estimated 34.6 kb insert size) nominally 35 ±5 kb	0.244	0.216	887	0.7×	0.193	0.163	680	0.5×
Total	28		36.69	16.48		51.5×	36.63	47.83		35.1×

Supplementary Table 3.5: Details of *C.* × *sinensis* 3kb insert library sequencing by 454 Life Sciences.
The amount of sequence generated in 3kb paired-end library sequencing is shown.

	HQ Reads	Linker+	pUC	Left	Right	Linker-	Uniqueness	TP
Lib1 Run1	593,317	68%	0%	138	142	199	55.52%	223,999
Lib1 Run2	1,114,551	71%	0%	132	133	210	55.52%	439,347
Lib1 Run3	1,361,004	78%	0%	145	148	264	55.52%	589,391
Lib1 Run4	1,404,995	74%	0%	142	144	235	55.52%	577,239
Lib3 Run1	1,242,103	74%	0%	171	177	287	78.47%	721,262
Lib3 Run2	1,149,899	46%	0%	133	145	243	78.47%	415,070
Lib3 Run3	1,193,550	74%	0%	177	180	275	78.47%	693,068
Lib2 Run1	1,117,943	64%	0%	162	167	269	61.76%	441,883
Lib2 Run2	1,085,321	52%	0%	138	146	217	61.76%	348,553
Lib2 Run3	1,028,411	60%	0%	150	154	252	61.76%	381,088
Total	11,291,094	69%	0%	148	153	245	64%	4,830,900

Supplementary Table 3.6: Details of *C.* × *sinensis* 8kb paired end libraries and WGS reads from 454 Life Sciences.
The amount of sequence generated in 8kb paired-end library sequencing is shown.

	HQ Reads	Linker+	pUC	Left	Right	Linker-	Uniqueness	TP
Lib1-A	118,733	46%	2%	113	115	176	63.65%	34,764
Lib1-B	695,894	62%	2%	162	168	275	32.67%	140,956
Lib2-A	607,494	62%	2%	164	167	284	36.64%	138,003
Lib2-B	651,306	63%	2%	166	169	283	35.71%	146,526
Lib3-A	610,063	59%	3%	158	162	275	38.00%	136,776
Lib3-B	613,102	60%	3%	161	165	279	36.62%	134,711
Lib4-A	517,913	52%	3%	144	147	246	43.09%	116,048
Lib4-B	486,265	50%	3%	138	141	238	44.43%	108,024
Lib5-A	546,603	60%	0%	153	158	231	87.61%	287,327
Lib5-B	531,881	54%	0%	143	147	209	90.67%	260,419
Lib6-A	548,434	69%	0%	164	168	254	88.01%	333,047
Lib6-B	568,346	70%	0%	166	171	255	88.32%	351,374
Lib7-A	545,807	67%	0%	145	149	245	89.31%	326,598
Lib7-B	565,478	73%	0%	160	163	270	88.65%	365,946
Lib8-A	508,357	56%	0%	139	142	213	89.09%	253,621
Lib8-B	4,222	30%	1%	78	80	129	99.12%	1,255
Total	8,119,898	58%	1%	147	151	241	57.12%	3,135,396

Supplementary Note 3.3: Shotgun sequence assembly of diploid *C.* × *sinensis*

The v.1 sweet orange assembly was produced using the 454 GS *de novo* Assembler version 2.3 ('Newbler'). Total scaffold length of the assembly is 319 Mb (of which 20.9% is gaps) spread over 12,574 scaffolds. Half the genome is accounted for by 236 scaffolds that are at least 251 kb long (N50/L50) (Supplementary Table 2.4). Each read was quality-trimmed by default and any trimmed read that is shorter than 50 bp long was discarded. The following options were used:

-large was used to enable large and complex genome assembly mode.

-het was used to enable "heterozygous" mode which specifies that the project's read data is from a diploid or non-inbred organism. This prompts Newbler to adjust the algorithms it uses to reflect an increase in the expected variability in sequence identity.

-scaffold organizes the contigs into scaffolds using paired-end information to order and orient the contigs and to approximate the distance between contigs.

The genome sequence is publicly available at Phytozome (http://www.phytozome.net).

Supplementary Table 3.7: Contig summary information for *C.* × *sinensis* assembly.
Statistics for scaffold and contig lengths from the Newbler assembly are shown.

Minimum Scaffold Length	Number of Scaffolds	Number of Contigs	Total Scaffold Length (bp)	Total Contig Length (bp)	Scaffold Contig Coverage
All	12,574	53,536	319,231,331	252,507,433	79.10%
1 kb	12,574	53,536	319,231,331	252,507,433	79.10%
2.5 kb	8,960	49,819	311,209,056	244,574,591	78.59%
5 kb	5,002	44,967	298,329,566	232,422,498	77.91%
10 kb	3,076	41,197	283,990,788	221,082,681	77.85%
25 kb	1,425	35,755	258,491,908	201,650,975	78.01%
50 kb	960	32,789	241,967,478	189,099,986	78.15%
100 kb	592	28,584	215,831,105	170,625,214	79.05%
250 kb	237	20,077	159,873,107	129,025,662	80.71%
500 kb	90	13,172	109,432,220	90,535,652	82.73%
1 Mb	32	8,044	69,841,201	59,025,356	84.51%
2.5 Mb	7	2,991	28,345,257	24,593,787	86.77%
5 Mb	2	1,029	11,071,338	9,895,318	89.38%

Supplementary Note 3.4: Paired End Library Span Estimation.

Paired-end sequencing was implemented in circularized libraries as described in [see 'Methods' tab at http://454.com/applications/whole-genome-sequencing/index.asp], so that the 5' and 3' ends of a single pyrosequencing read are derived from opposite ends of a DNA fragment. Estimates of the distance spanned by paired-end reads in a library are made when at least 8 consistent mate pairs are found that align to the same contig or scaffold. Both halves of a paired-end read must align to the same contig with the expected directionality (the read halves 3' ends point toward each other, after reverse-complementation of the left half). Summary statistics for the distance between mated pairs are kept for each library. As additional scaffolds are formed, additional useful paired-end reads become available and the library span is re-estimated. Paired-end reads whose halves are too far away from the mean of the distribution and those whose halves do not have the expected relative orientation are excluded from the span distance calculation. The estimate is less robust when either little paired-end information for a library is available or when very few contigs are significantly longer than the actual library span (in the latter case, the estimated span may be significantly lower than the actual span).

Supplementary Note 4: Annotation of citrus genome assemblies

The haploid Clementine and diploid sweet orange genome assemblies were annotated using the JGI plant genome annotation pipeline [22]. Before protein-coding genes were predicted, repetitive content was analyzed as described below.

Supplementary Note 4.1: Analysis of repetitive content in *Citrus* genome assemblies.

Repetitive sequences were identified in separate analyses in the assembly of *C.* × *sinensis* and *C.* × *clementina* with the *de novo* repeat finding algorithm RepeatModeler v1.0.5 (http://www.repeatmasker.org/RepeatModeler.html). For both analyses, repetitive sequences less than 500nt long were removed, and the remainder was annotated with predicted protein domains with Pfam [23] and Panther [24]. Sequences that had been annotated with a protein domain not associated with transposable elements or other repeat sequences were removed from the library. (A manual review step removed additional protein coding gene-associated sequences from the *C.* × *sinensis* repeat library). The resulting repeat library was used to lower case mask the assembly with RepeatMasker v. Open-3.0 (http://repeatmasker.org). To quantify the masking of each genome, the GFF format file output by RepeatMasker was analyzed to generate a table summarizing masked nts by repeat family (Supplementary Table 4.1). To run RepeatMasker efficiently, the input assemblies were broken into 500kb segments with 1kb overlap. There is potentially a small amount of sequence that could be counted twice in the repeat masking by family statistics.

Type I retrotransposon and Type II DNA transposon abundances across the nine chromosomes of haploid Clementine assembly show an inverse relationship with gene content, whereas simple repeats do not (Supplementary Fig. 10).

Supplementary Table 4.1: Repeat families in citrus genome assemblies in this work.

The amount of sequence (b.p.) represented by different repeat types in the *C.* × *clementina* and *C.* × *sinensis* assemblies is shown. Note that since repeat libraries were constructed independently, some rare repeats in one assembly are not described in the other (reported in table as 0.00%)

Repeat Type	*C.* × *clementina*	%	*C.* × *sinensis*	%
DNA	59,897	0.02%	221,550	0.07%
DNA/En-Spm	1,700,186	0.56%	945,074	0.30%
DNA/Harbinger	293,208	0.10%	274,724	0.09%
DNA/MuDR	2,633,432	0.87%	2,038,664	0.64%
DNA/Pogo	0	0.00%	110,654	0.03%
DNA/Sola	107,924	0.04%	0	0.00%
DNA/Tc1	0	0.00%	229,459	0.07%
DNA/TcMar-Pogo	254,408	0.08%	78,950	0.02%
DNA/TcMar-Tc1	199,070	0.07%	0	0.00%
DNA/hAT	0	0.00%	224,918	0.07%
DNA/hAT-Ac	2,644,452	0.88%	1,190,014	0.37%
DNA/hAT-Tip100	224,166	0.07%	76,913	0.02%
LINE/L1	3,597,664	1.19%	4,495,118	1.41%
LINE/L2	110,719	0.04%	0	0.00%
LINE/R1	32,290	0.01%	0	0.00%
LTR	22,507	0.01%	456,081	0.14%
LTR/Caulimovirus	0	0.00%	694,846	0.22%
LTR/Copia	23,756,816	7.88%	22,521,281	7.05%
LTR/ERVK	174,218	0.06%	0	0.00%
LTR/Gypsy	36,192,528	12.01%	26,310,282	8.24%
RC/Helitron	0	0.00%	130,548	0.04%
Low Complexity	4,855,980	1.61%	4,462,997	1.40%
putative SINE	84,327	0.03%	0	0.00%
SUBTEL_sa	107	0.00%	0	0.00%
Simple Repeat	1,507,593	0.50%	1,992,080	0.62%
Unknown	56,131,599	18.62%	32,745,816	10.26%
rRNA	70,037	0.02%	0	0.00%
Total	**134,653,128**	**44.67%**	**99,199,969**	**31.06%**

Supplementary Note 4.2: Protein-coding gene annotation of Citrus

Supplementary Note 4.2.1: Haploid *C.* × *clementina* gene model annotation

Protein-coding genes were predicted with a pipeline that combines expressed sequence tag (EST), homology, and de novo prediction methods [22].

We obtained 770,602 EST sequences from LifeSequencing from the diploid Clementine var. Nules that is the parent of the haploid reference. To these, we added 210,567 *C.* × *sinensis* and 118,365 *C.* × *clementina* ESTs downloaded from GenBank, 58,656 EST assemblies that had been generated from sweet orange 454 ESTs assembled with Newbler and 401,708 454 EST reads from LifeSequencing to make a total of 1,559,898 ESTs. These were aligned to the Clementine genome (requiring 95% sequence identity and 50% coverage of the input sequence) and further assembled with PASA [25] to generate 76,372 EST assemblies.

We aligned predicted protein sequences from Arabidopsis (v. TAIR8); peach (JGI v. 1.0) and grapevine (Genoscope 12× 05/10/10) to the softmasked Clementine v1.0 assembly (see above) with gapped BLASTX [19] and generated putative protein-coding gene loci from regions with EST assemblies and/or protein homology, extending to include overlap where necessary.

Gene predictions were generated from putative loci with FGenesH+ (Ref. [26]), exonerate [27] (with option -model protein2genome) and GenomeScan [28]. The gene prediction at each locus with the highest amount of support from EST assemblies and protein homology was chosen to be improved using evidence from the EST assemblies with a second round of PASA. Gene models with homology to repeats were removed.

This produced an annotation at each of 24,533 protein coding loci, with 9,396 alternative splice forms, making a total of 33,929 predicted transcripts. Protein coding gene content varies across the chromosomes, with high levels corresponding to repeat-poor regions away from the expected positions of telomeres (Supplementary Fig. 10).

Supplementary Table 4.2: Summary of protein-coding annotation of Citrus assemblies compared to representative eudicots.
Statistics are shown for the eudicot genomes used as protein homology inputs for gene prediction in the citrus genomes. EST overlap with gene models for other eudicots were not determined and those cells are greyed out in the table.

Description	*C.* × *clementina* v1.0	*C.* × *sinensis* v1.1	*A. thaliana* TAIR 8	*G. max* v1	*V. vinifera* 12x	*P. persica* v1.0
Primary transcripts (loci)	24,533	25,379	27,014	46,367	26,346	27,864
Alternate transcripts	9,396	20,768	5,601	9,420	0	837
Total transcripts	33,929	46,147	32,615	55,787	26,346	28,701
For primary transcripts (longest at locus)						
Average number of exons	5.3	4.9	5.3	6.0	6.2	5.0
Median exon length	156	153	155	142	143	151
Median intron length (bp)	171	166	99	185	212	164
Median gene length (bp)	2,467	2,103	1,893	2,906	3,572	2,076
Gene models with EST overlap	19,422	15,755				
EST support over 100% of their lengths	8,684	10,985				
EST support over 95% of their lengths	13,592	12,037				
EST support over 90% of their lengths	14,206	12,508				
EST support over 75% of their lengths	15,234	13,369				
EST support over 50% of their lengths	16,963	14,234				
Transcripts with Pfam annotation	16,986	17,457	18,264	34,065	15,751	18,275
Transcripts with panther annotation	14,015	14,924	14,449	26,960	14,124	14,702
Transcripts with KOG annotation	10,807	9,515	12,095	20,601	9,412	11,109
Transcripts with KEGG Orthology annotation	3,710	3,524	3,369	6,950	3,692	3,925
Transcripts with E.C. number annotation	2,005	1,918	3,100	3,698	2,020	2,107

Supplementary Note 4.2.2: Diploid *C.* × *sinensis* gene model annotation

To annotate the sweet orange genome with expressed sequences from its own genotype, we generated 5,935,974 454 EST sequences from 17 different cDNA samples/conditions (Supplementary Table 4.2) on the 454 platform at Roche. In total, we generated 2,248,334,318 bp, with median EST length 419 bp.

These EST reads were filtered to remove rDNA and chloroplast sequences (35.38% of total), leaving 3,835,882 putative transcript fragments. To these, we added 549,116 sequences from GenBank (downloaded March 3rd, 2010) to make a total of 4,384,998 input sweet orange ESTs.

We aligned this set of sweet orange EST sequences to the genome and assembled them using the PASA pipeline [25], which produced 85,463 EST assemblies.

Predicted protein sequences from rice (partially non-redundant set of predicted protein sequences from TIGR v. 5), Arabidopsis (TAIR version 8) and grapevine (Genoscope v. 12× from 05/10/10) were collected from external sources, together with predicted soybean protein sequences generated in the JGI annotation pipeline. These diverse angiosperm sequences were aligned to the soft-masked genome (see above) using gapped BLASTX [19]. Regions on the genome where there was a protein alignment and/or overlap with an EST assembly generated in the previous step were considered to be putative protein-coding gene loci in subsequent gene prediction step as follows. These loci were extended by 1kb in each direction and submitted to FgenesH (provided by Asaf Salamov at JGI), along with related angiosperm peptides and/or ORFs from the overlapping EST assemblies. In a separate gene prediction effort, hybrid gene predictions that integrate EST information with *ab initio* predictions were generated with GeneMark-ES+ Ref. [29]. These two sets of predictions were integrated with the EST assemblies by picking the predicted model at each locus that has the best support from homology and EST evidence and then using PASA [25] to improve agreement between gene predictions and EST assemblies. The results were filtered to remove genes with over 20% of their coding sequence overlapping genomic regions annotated as repetitive.

This pipeline predicted 25,376 protein-coding loci, each with a primary transcript. An additional 20,771 alternative transcripts were predicted, generating a total of 46,147 transcripts that comprise the 'orange1.1' annotation version. 16,318 primary transcripts have EST support over at least 50% of their length. Two-fifths of the primary transcripts (10,813) have EST support over 100% of their length.

Supplementary Table 4.2: *C.* × *sinensis* mRNA samples for cDNA sequencing on 454 platform.
This table shows the conditions/treatment and the plant tissue from which mRNA was extracted for cDNA sequencing.

Sample #	Treatment	Plant material
1	4°C for 48 hours	Germinated seeds/plants
2	Freezing -20°C for 48 hours	Germinated seeds/plants
3	Darkness for 48 hours	Germinated seeds/plants
4	Sodium chloride 0.5M for 48 hours	Germinated seeds/plants
5	100 % O_2 for 24 hours	Plants
6	100 % $C0_2$ for 24 hours	Plants
7	Ethylene 50-55 ppm 4 hours	Plants
8	37°C for 48 hours	Germinated seeds/plants
9	Canker for 8 days	Plants
10	Diaprepes larvae for 7 days	Plants
11	Psyllids for 10 days	Plants
12	Mechanical cutting for 24 hours	Plants
13	Salicylic acid 0.5 mM for 24 hours	Germinated seeds/plants
14	Under water for 24 hours	Germinated seeds/plants
15	pH 9 solution for 24 hours	Germinated seeds/plants
16	None	Germinated seeds/plants
17	None	Leaves from germinated seeds from the plant used for genome sequencing

Supplementary Note 5: Comparisons with other plant genomes and evolutionary analysis.

Supplementary Note 5.1: Identification of paralogous segments in the Citrus genome.

We characterized genome duplications by aligning the 24,533 Clementine genes to themselves with stringent alignment criteria and statistical validation as described previously [30]. Clementine was used to represent Citrus in this analysis because of our high quality chromosomal assembly. We identified and characterized seven large genomic blocks that are involved in three-to-three paralogous relationships, with a divergence corresponding to the paleo-hexaploidization event 100-130 million years ago [31]. These blocks cover 34% of the Clementine genome (in 226 pairwise paralogous segmental relationships) and involve the following chromosome to chromosome relationships (Supplementary Fig. 11a):

- c3 (blue),
- c1-c7 (turquoise),
- c3-c5-c9 (purple),
- c2-c4-c6-c8 (orange),
- c5-c7-c8 (green),
- c2-c3-c4-c6 (yellow),
- c1-c2 (red).

Supplementary Note 5.2: Dating of paralogous segments.

Sequence divergence as well as speciation event dating analysis was based on the rate of nonsynonymous (Ka) vs. synonymous (Ks) substitutions calculated with MEGA3 Ref. [32]. The average substitution rate, $r = 6.5 \times 10^{-9}$ substitutions per synonymous site per year for grasses is classically used to calibrate the ages of the considered gene [33, 34]. The time since gene insertion is then estimated using the formula MYA (millions of years ago) = Ks /2r, where $r = 6.5 \times 10^{-9}$ and most orthologs have a Ks value of ~0.9 (Supplementary Fig. 11b). This gives a divergence time for grape and citrus of ~70 million years ago. Alternatively, if a substitution rate of 1.2×10^{-9} is used, as estimated for poplar [35, 36], the time scale of divergence is increased by a factor of approximately 6.

Supplementary Note 5.3: Synteny analysis.

To analyze conserved synteny between Citrus and other eudicots, we first compared Citrus with grape, since grape is known to have preserved the ancestral eudicot hexaploid chromosome organization more than other sequenced species. 4,862 orthologous relationships were identified relative to the 7 eudicot proto-chromosomes (Supplementary Fig. 11c,12). These segments cover 69% of the *C.* × *clementina* genome.

The following chromosome-to-chromosome relationships have been established (**c** for *C.* × *clementina* and **g** for grape as chromosome nomenclature, following [31]):

- c1/g11-g4-g19-g17,
- c2/g12-g14-g1-g6-g10,
- c3/g7-g17-g2-g15-g5-g16-g18-g8-g19-g4,
- c4/g13-g1-g17-g2-g12-g16,
- c5/g18-g14-g12-g2-g4-g6-g7,
- c6/g8-g14-g4-g2,
- c7/g9-g7-g4,
- c8/g3-g6-g8,
- c9/g5-g8-g1.

Supplementay Note 5.4: Evolutionary scenario.

We integrated independent analyses of the duplications within and between six diverse sequenced eudicot genome – Citrus, Arabidopsis [37], Cacao [38], soybean [39], Populus [35], and grapevine [31] — to produce a coherent scenario for the evolution of these genomes from their common hexaploid ancestor. Seven ancestral linkage groups could be identified on modern chromosomes, in agreement with the seven ancestral chromosomal groups previously proposed in eudicots.

This pattern is found on the following chromosome pair combinations in *Citrus* compared to the 7 ancestral linkage groups reported in grape:

- g1-g14- g17 / c2-c3-c4-c6,
- g2-g15-g12-g16 / c3,
- g3-g4-g7-g18 / c5-c7-c8,
- g4-g9-g11 / c1-c7,
- g5-g7-g14 / c3-c5-c9,
- g6-g8-g13 / c2-c4-c6-c8,
- g10-g12-19 / c1-c2.

Here we represent chromosome "x" of grapevine as "gx" and chromosome "y" of Citrus as "cy".

Based on the ancestral paleo-hexaploidization reported for the eudicots, we propose an evolutionary scenario that has shaped the 9 *Citrus* chromosomes from the 7 chromosomes eudicot ancestor and more precisely to the 21 paleo-hexaploid intermediate (Supplementary Fig. 11c).

To reach the actual modern 9 chromosomes *Citrus* ancestor structure from the 21 chromosomes intermediate ancestor, we require 23 fissions and 35 fusions. The actual *Citrus* genome would descend from one additional round of whole genome polyploidization that corresponds to the gamma event [31] (Supplementary Fig. 11b-c).

Supplementary Note 5.5: Calculation of four-fold degenerate transversion frequency

Methods closely followed those of the Populus genome paper [35]. Briefly, Segments were found by locating blocks of BLASTP hits with significance 1E-18 or better with fewer than 5 intervening genes between such hits. The 4DTv distance between orthologous genes on these segments is plotted as a histogram (Supplementary Fig. 13).

Supplementary Note 6: Analysis of resequencing datasets

In addition to sequencing the haploid Clementine mandarin reference genome and the diploid sweet orange as described above, we used the Illumina platform to resequence eight diploid citrus accessions, achieving shotgun depths from 17-110× (Table 1, main text). Here, we describe our protocols for identifying heterozygous and homozygous single nucleotide variants (SNVs) relative to the

high quality haploid Clementine reference, and for quantifying rates of heterozygosity and homozygous differences relative to the reference in the resequenced individuals.

For reliable SNV calling, we did not consider genomic regions that have low mapping quality (low confidence that shotgun reads are correctly mapped) or low depth (which provides limited support for alternate alleles). Instead, SNVs are called from "eligible" (*i.e.*, callable) sites, which are defined as sites covered by sufficient but not excessive number of high quality bases (phred score 20 or more) from reads with mapping scores at least 25. The lower and upper bounds on the read depths for an eligible site depend on the genome wide depth-of-coverage distribution, and a rule of thumb is given below (Supplementary Note 6.1). **Our subsequent analyses always refer to eligible sites**. For example, nucleotide heterozygosity of an individual can be measured in het sites/kb, meaning number of heterozygous sites per 1,000 eligible sites. Similarly, a sliding window of 100 kb contains 100,000 eligible sites.

The error rates in SNV calling are estimated by simulation as well as by comparison to array data (see Supplementary Notes 6.1.1 and 6.1.2).

Supplementary Note 6.1: Read mapping and SNP calling

Illumina paired end reads were mapped to the haploid Clementine reference (HCR) using 'bwa aln –n 8 –q 15' and 'bwa sampe' (Ref. [40]). Reads with mapping score below 25 were discarded, and duplicate reads were removed with an in-house script. Mpileup files were then generated using 'samtools mpileup -BA' (Ref. [41]). Single nucleotide variants (SNVs) were called from mpileup files using an in-house python script based on the following criteria:

1) *Read coverage*. To exclude bases with unusually high or low read coverage relative to the genome-wide per base read depth distribution, we only considered sites with coverage between a lower cutoff (around half the peak coverage) and upper cutoff (twice the peak coverage). For reliable SNV calling, we required coverage of at least four reads.
2) *Base quality*. To reduce false positives in SNV calling, we only considered bases in aligned sequence reads with phred quality score 20 or more in our SNP calling algorithm.
3) *Calling SNVs*. For a diploid genome, we called a heterozygous SNV when two alleles were present at a site with each allele supported by at least two reads. We further assumed that the non-reference allele frequency within the individual (in mapped reads) follows a binomial distribution with probability p (default p = ½), and call a "het" only if the allele frequency does not reside in the tails of the distribution (*i.e.*, 1% probability of seeing an allele at least as extreme as the observed frequency). This is done to avoid false positives due to mapping and base calling errors. A "homozygous SNV" relative to the reference haploid sequence is called when the non-reference allele frequency exceeds 0.9, allowing for mis-mapping and base call errors.

Supplementary Note 6.1.1: Estimation of false positive and false negative variant call rates.

Calling variants relative to a reference sequence using computational analysis of resequencing data is susceptible to both false positive and false negative calls. False negatives are variants that are present in the sequenced genotype but are missed by the computational analysis; false positives are variants that are called by computation when they are not present in the true genotype. False positive and negative rates are influenced by intrinsic sequencing errors, errors in aligning short reads to the reference genome, and low coverage of variant alleles, which depends in turn on sequencing depth and uniformity.

To place bounds on the total false positive and false negative rates in our study, we created a synthetic reference sequence with known variation, and tested our ability to detect this known variation. We introduced artificial substitutions and indels into the haploid Clementine reference (HCR) and analyzed our diploid Clementine (CLM) Illumina resequencing data relative to this synthetic mutant. Since the position and nature of the substitutions are known, we can estimate (1) the false negative rate, which is the rate at which true single nucleotide variants are missed, and (2) the false positive rate as the rate at which single nucleotide variants are called when they are not in fact present (assuming that the reference genome is highly accurate – errors in the reference genome will also be counted as false positives). The false positive rate is measured per non-variant "eligible" site, where "eligible" sites are those at which our protocols make a genotype call and the meaning of non-variants is explained below. The false negative rate is measured per predicted variant (*i.e.*, artificial substitution in the mutant reference sequence). We did not call indels in our analysis.

The synthetic reference genome (denoted "HCMut", for haploid Clementine mutant) was produced by introducing single nucleotide substitutions at a rate of 2% (for chromosomes 1-3), 1% (chromosomes 4-6), and 0.5% (chromosomes 7-9) following the corresponding Poisson distribution. In a similar manner, indels of size 1-3bp were introduced with rates equal to 10% of the corresponding base substitution rates, with single nucleotide indels being the dominant form. As a control for indel effects, no indels were generated on chromosomes 1, 4, and 7. Variants were not introduced less than 10 bp from an adjacent variant. We used variable substitution rates to reflect the different levels of heterozygosity observed in citrus, ranging from the inter-specific sequence divergence around 2% to typical within species sequence variation around 0.5%. The substitutions and indels introduced by this procedure should show up as homozygous non-HCMut variant calls when the CLM reads are aligned to the mutated HCMut sequence.

To assess the dependence of SNV call error rates on read depth, we subsampled the diploid Clementine mandarin dataset at three depths of coverage: 100×, 34×, and 17×. At each depth of coverage, analysis was performed as follows:

1) As a control, reads were first mapped to HCR (haploid Clementine reference) and SNVs were called. This was done for three reasons. 1) to identify candidate assembly errors in HCR showing up as homozygous non-HCR variants. These sites were subsequently masked to ensure a reliable error rate estimate of SNV calls. 2) to discover heterozygous positions (hets) in the diploid Clementine, which were subsequently masked to estimate false positive rate for calling hets when the same reads are aligned to HCMut. 3) to define 'HCR-non-variant' sites as homozygous reference sites relative to HCR. Only these sites are used for error rate estimate in the later steps.

2) To estimate SNV call errors, the same reads were mapped to HCMut (synthetic "mutant" haploid Clementine) and SNVs were called with our pipeline. These calls are made for eligible, 'HCR-non-variant' sites only, both to minimize the effect of assembly errors and to estimate false positive rate for calling hets when they are expected to be absent thanks to the masking in step 1). Note that the indels introduced in generating HCMut change the length of the HCMut sequence relative to HCR, so a mapping between the HCR and HCMut coordinates was generated.

3) Error rates were computed by comparing the set of introduced nucleotide substitutions and SNVs discovered from reads mapped to HCMut. We consider only eligible sites with the range of read depths given in Supplementary Tables 6.1 and 6.2. The false positive rate is the probability of mis-calling a non-variant as a variant, either as a heterozygous site (FP/het) or homozygous non-reference site (FP/hom). A non-variant is not only an "HCR-non-variant" but also a site without artificial base substitution or an indel. The false negative rate is the probability of mis-calling an artificial base substitution in HCMut (showing up as homozygous non-reference in the reads), either as heterozygous (FN/het) or as homozygous reference (FN/hom).

Several observations were made following this analysis: 1) We noted a slight decrease in SNV call error rates in the absence of introduced indels (chromosomes 1,4 and 7). For this reason, we used the six chromosomes with introduced indels to estimate the error rates more conservatively. 2) We did not observe a striking correlation between SNV call error rate and sequence divergence rate. 3) The SNV call error rates decrease strikingly with increasing read depth of coverage. 4) The error rate for heterozygous SNV calls is notably larger than that for homozygous calls. This might be due to the fact that alignment errors often lead to false hets rather than homozygous genotypes. A detailed summary of the analysis is given in Supplementary Tables 6.1 and 6.2.

It is worth noting that the SNV call error rates are low at the three depths of coverage (100×, 34× , 17×). For example, at 17×, the false positive rate (FPR) in miscalling a non-variant site as a het is 0.05%, and the FPR for miscalling a non-variant site as homozygous-non-reference is three orders of magnitude lower, at 0.00004%. By contrast, the false negative rate (FNR) for miscalling a

homozygous-non-reference site as heterozygous site is 0.3%, and the FNR for miscalling a homozygous-non-reference site as nonvariant (i.e. homozygous reference) is two orders of magnitude lower at 0.003%. Our high depth of coverage resequencing data (17× to 110×) thus allows us to make reliable SNV calls, and avoid biases in nucleotide heterozygosity estimate associated with low-coverage sequencing[42].

Supplementary Table 6.1. False positive rate (FPR) in SNV calls based on a synthetic reference sequence and the Clementine resequencing dataset.
At each depth of coverage and from a set of pre-determined non-variant sites, FPR is computed based on the mis-called heterozygous and homozygous non-ref sites separately. The total FPR in miscalling a non-variant as a variant site is the sum of the two mis-called types, and is estimated at 2.0×10^{-5}, 9.8×10^{-5}, and 5.3×10^{-4} at 100×, 34× and 17× respectively. FP/het=false positive heterozygous calls, FP/hom=false positive homozygous non-reference calls. The range of allowable read depths is shown in the column headers (top row).

		100× (depth 60-180)		34× (depth 20-70)		17× (depth 10-30)	
Observed	Type	# Sites	Proportion	# Sites	Proportion	# Sites	Proportion
Homozygous reference	True Negative	141,710,197	1.00	143,781,266	1.00	135,215,672	1.00
Heterozygous	FP/het	2,864	2.02×10^{-5}	14,026	9.75×10^{-5}	71,883	5.31×10^{-4}
Homozygous non-reference	FP/hom	28	1.98×10^{-7}	32	2.23×10^{-7}	50	3.70×10^{-7}

Supplementary Table 6.2. False negative rate (FNR) in SNV calls based on a synthetic reference sequence and the Clementine resequencing dataset.
At each depth of coverage and from a set of introduced SNVs in the reference sequence, FNR is computed based on the mis-called heterozygous and homozygous reference calls separately. The total FNR as miscalling/missing a homozygous non-reference variant is the sum of the two mis-called types, and is estimated at 1.0×10^{-3}, 1.9×10^{-3} and 3.5×10^{-3} at 100×, 34× and 17× respectively. FN/het=false negative heterozygous calls, FN/hom=false negative homozygous reference (non-variant) calls. The range of allowable read depths is shown in the column headers (top row).

		100× (depth 60-180)		34× (depth 20-70)		17× (depth 10-30)	
Observed	Type	# Sites	Proportion	# Sites	Proportion	# Sites	Proportion
Homozygous non-reference	True Positive	1,759,644	9.99×10^{-1}	1,784,063	9.98×10^{-1}	1,680,501	9.97×10^{-1}
Heterozygous	FN/het	1,736	9.86×10^{-4}	3,418	1.91×10^{-3}	5,815	3.45×10^{-3}
Homozygous reference	FN/hom	54	3.07×10^{-5}	51	2.85×10^{-5}	55	3.26×10^{-5}

Supplementary Note 6.1.2: Validation of SNV calls by GoldenGate assay

To validate our computational SNV predictions, we compared the SNV calls for the diploid Clementine with SNVs assayed by an Illumina GoldenGate array[43, 44]. In the common set of 339 genotype calls, only one discrepancy was found. The concordance rate is thus 338/339 = 99.7%. This high concordance rate is

consistent with the low false negative rate (~10^{-3}) estimated in the previous section.

Supplementary Note 6.2: Assignment of cpDNA

Since chloroplast genomes are typically inherited maternally [45, 46], determining the number of phylogenetically distinct cpDNA types among the citrus cultivars should inform us about the ancestral progenitor species from which the cultivars are derived. Unlike the nuclear genome, no inter-specific hybrid cpDNA can be produced if maternal inheritance is strictly observed. Conversely, the phylogenetic grouping of cpDNA can reveal the maternal inheritance pattern among the citrus cultivars.

We examined the relationship between the chloroplast sequences from the 8 citrus cultivars: Low acid pummelo (LAP), Chandler pummelo (CHP), Ponkan mandarin (PKM), Willowleaf mandarin (WLM), sweet orange (SWO), Seville sour orange (SSO), Clementine mandarin (CLM), and W. Murcott mandarin (WMM).

To determine which shotgun sequence reads were derived from the chloroplast genome, short reads from each of the 8 diploid citrus shotgun datasets were mapped to the chloroplast genome sequence of Ridge Pineapple sweet orange[47] using bwa [40] as described above. After filtering low mapping quality (maq<25) and duplicate reads, SNVs were called with our SNV caller as described above. Since no heterozygosity is expected, a cpDNA SNV was called when the non-reference allele was supported by 90% or more of the reads for a given accession. Pairwise mismatches among the 8 cultivars were computed based on SNVs called from a common set of 92.6k sites.

The 8 citrus cpDNA sequences appear in 2 clusters in a neighbor-joining phylogenetic tree based on a distance matrix generated from pairwise mismatches (Supplementary Fig. 14). The first "mandarin" cluster includes WLM, PKM, CLM and WMM, with zero mismatches between WLM, CLM, and WMM. There are only 2 mismatches between these three mandarins and PKM, the fourth member of the cluster. These cpDNA haplotypes were defined as being *C. reticulata*, following the conventional taxonomic designation.

The second "pummelo" cluster contains LAP, CHP, SWO and SSO, with no mismatches between the two pummelos LAP and CHP, as expected from their mother-child relationship. These cpDNA types were identified with *C. maxima*, indicating that sweet and sour orange have maternal pummelo ancestry. Pairwise mismatches within the cluster are 29 (LAP/SWO), 56 (LAP/SSO), and 45 (SWO/SSO).

Across the entire chloroplast genome, the number of mismatches between the "mandarin" and "pummelo" clusters are in the range 320-332, or ~3.5/kb. This rate corresponds to ~1/6 of the rate of nuclear sequence divergence between *C. maxima* and *C. reticulata* (see Supplementary Note 9.1), which is consistent with

the chloroplast:nuclear molecular clock rate ratio of 3:16 (Ref. [48]). The clear separation of the cpDNA into two clusters is consistent with the proposition that these cultivars were descended from two ancestral species, *C. maxima* and *C. reticulata*.

Supplementary Note 7: Identification of two ancestral species (*C. maxima vs. C. reticulata* alleles)

Diagnostic SNVs that differentiate between *C. maxima* and *C. reticulata* were derived from analysis of the diploid genomes of two pummelos (LAP, CHP) and two ancient mandarin types without a previously suspected history of admixture (PKM, WLM). Here our goal is to identify polymorphic sites across these genomes that are candidate fixed differences between the two progenitor species. This analysis is complicated by the observation that the mandarins do in fact contain previously unsuspected pummelo introgression, as described in the main text (Fig. 3b). Nevertheless, careful analysis of these admixed regions allows us to identify candidate fixed differences across the entire genome.

Analysis of heterozygosity in the two pummelo genomes suggests that they are derived from a homogeneous sexual population of *C. maxima* except for a short segment ~25 Mbp from the start of chromosome 2 of CHP that has unusually high heterozygosity (Supplementary Fig. 15). Supporting evidence for the hypothesis that LAP and CHP represent nearly pure *C. maxima* comes from the following: (1) in both LAP and CHP, nucleotide heterozygosity (π) shows a peak at ~6 het sites/kb (Fig. 2a) and (2) nucleotide heterozygosity between non shared haplotypes of the two pummelos has a peak at approximately the same value of π (~6 het sites/kb) (Fig. 2a). Note that there are three haplotypes in the two pummelos due to their parent/offspring relationship (see Supplementary Note 10.4), and (3) all three distributions of heterozygosity (Fig. 2a) do not show a higher, inter-specific peak, like the one that can been seen in the SWO and SSO nucleotide heterozygosity distributions (Fig. 2c). Note that the lower peak (~1 het site/kb) in the pummelo heterozygosity histogram (Fig. 2a) could be due to a population bottleneck in the *C. maxima* species (see Supplementary Note 9.2).

In contrast, both "traditional" mandarins, PKM and WLM exhibit two distinct features in their nucleotide heterozygosity distribution: one averaging ~6 het sites/kb and the other ~17 het sites/kb, (Fig. 2b, main text). These two regimes of nucleotide heterozygosity are organized along the genome as distinct blocks with sharp boundaries (Supplementary Fig. 16). The regions of lower heterozygosity generally represent diploid segments of wild mandarin, that is, *C. reticulata*. Regions of higher heterozygosity are interpreted as hybrid segments in which one *C. reticulata* haplotype is paired with an alternate haplotype from a distinct species. Comparison of the alleles at these excess heterozygous sites with LAP/CHP identifies the other species as pummelo (*C. maxima*). The existence of these hybrid regions indicates inter-specific introgression of *C. maxima* into a presumptive *C. reticulata* background.

In the absence of a pure *C. reticulata* genome, *C. maxima* and *C. reticulata* specific alleles can be inferred based on sites segregating in the four genomes but fixed in the two pummelos, where at any such site, the *C. maxima* allele is the allele fixed in the two pummelos and the *C. reticulata* allele is the second allele present in PKM/WLM, respectively. False signals can arise, for example, from novel mutations present in only one haploid sequence of PKM or WLM, and they make site-specific admixture analysis unreliable. Nevertheless, sliding-window-based analysis that uses a large number of markers gives more reliable results (see next section).

A refined set of diagnostic markers can be obtained based on both the local nucleotide diversity levels of PKM and WLM and the divergence between one of the two pummelos and either PKM or WLM (Supplementary Fig. 17). The basic idea is to distinguish between candidate regions with inter-specific admixture from those without. For comparing two diploid genomes, we adopt the definition of F_{st} used by Hudson *et al.* [49] and Keinan *et al.* [50] and denoted by D (for divergence) to avoid possible confusion with the concept of F_{st} as applied to populations:

$$D = 1 - 0.25 \times (\pi_1 + \pi_2)/ \pi_{12} ,$$

where the numerator of the second term is the average nucleotide diversity (*i.e.*, heterozygosity) within each diploid, and π_{12} is the average nucleotide divergence between the two diploids. Here we define the *nucleotide divergence* between two diploids as the probability that randomly chosen alleles from each are different. For example, when comparing two identical diploids with heterozygosity $\pi_1 = \pi_2 = \pi$, the between-individual divergence is $\pi_{12} = \pi/2$, since half of the time, different alleles are chosen from each individual. (The other half of the time the same allele is chosen from each, which does not contribute to the divergence measure.)

To build intuition, note that for identical twins, since $\pi_{12} = \pi/2$, we have D = 0. For parent-child pairs, one haplotype is always shared, and (assuming both parent and child have the same heterozygosity $\pi_1 = \pi_2 = \pi$), we have $\pi_{12} = 3/4\ \pi$, so that D = 1/3. For unrelated diploid individuals drawn from the same homogenous population, $\pi_{12} = \pi$, and D = ½. In contrast, two individuals sampled from two distinct gene pools are expected to have values of D ranging between 0.5 and 1, depending on the degree of divergence of the two populations.

For two highly diverged populations like *C. maxima* and *C. reticulata*, the value of D is close to 1 for regions without admixture. When one of the two individuals is admixed, D is close to 0.5. For regions where both individuals are admixed, D is small and approaches zero. D can be computed in a genomic window containing many variable sites to provide a local measure of relatedness across the genome.

For the four citrus genomes (the two pummelos LAP, CHP and the two "traditional mandarins" WLM, PKM), high D values (~0.9, Supplementary Fig.

17) for all four pummelo/mandarin pairwise comparisons and low intrinsic nucleotide heterozygosity (~5 het sites/kb, Supplementary Figs. 15, 16) characterize most genomic regions, and correspond to two highly diverged species without introgression for those regions. For these chromosomal segments, diagnostic *C. maxima* and *C. reticulata* alleles can be more reliably identified from sites having two alleles separately fixed in the two pummelos and the two "traditional mandarins" (*i.e.*, PKM and WLM) respectively.

Other regions are characterized by high pummelo/PKM D values (~0.9) and low intrinsic nucleotide heterozygosity for PKM, but significantly lower pummelo/WLM D values (~0.5) and much higher nucleotide heterozygosity for WLM (e.g. ~18-22 Mbp along chromosome 8). These regions are consistent with PKM as a pure (diploid) *C. reticulata* and WLM as a *C. reticulata*/*C. maxima* inter-specific hybrid. A refined set of diagnostic *C. maxima* and *C. reticulata* alleles for these genomic regions can be identified from sites having two alleles separately fixed in the two pummelos and PKM. A similar analysis can be done for regions where both D and heterozygosity are consistent with PKM being a hybrid and WLM being pure *C. reticulata* (e.g. ~15-28Mbp along chromosome 9)

Finally, there are regions (*e.g.*, the right end of chromosome 6) where both PKM and WLM are characterized by a sharp rise in nucleotide diversity and sudden drop in pummelo/mandarin D (Supplementary Figs. 16,17). These features are consistent with both of the "traditional mandarins" being inter-specific *C. reticulata/C. maxima* hybrids for those regions. In this case, inference about diagnostic alleles can be made based on sites having one allele fixed in the two pummelos (the *C. maxima* allele) and two alleles present in both PKM and WLM, with the second (non-pummelo) allele identified as the putative *C. reticulata* allele.

Thus, a clearly defined set of diagnostic *C. maxima*/*C. reticulata* alleles can be obtained based on the nucleotide heterozygosity of PKM and WLM and on the pummelo/mandarin D values. They can be divided into five categories:

1) alleles separately fixed in LAP/CHP and PKM/WLM (*e.g.*, chromosomes 1, 5, 7)
2) alleles separately fixed in LAP/CHP and PKM (for segments where WLM is highly heterozygous and pummelo/WLM D~0.5, e.g., ~18-22 Mbp along chromosome 8)
3) alleles separately fixed in LAP/CHP and WLM (for regions where PKM is highly heterozygous and pummelo/PKM D~0.5, e.g. most of chromosome 9)
4) alleles only fixed in LAP/CHP but segregating in both PKM and CLM (for segments where both PKM and WLM are highly heterozygous and pummelo/mandarin D~0.5. see *e.g.*, left ends of chromosomes 3 and 8, right end of chromosome 6). In these cases, the non-*C. maxima* allele is considered to be a *C. reticulata* allele.
5) For a short segment at ~25 Mbp from the start of chromosome 2, CHP is

highly heterozygous (Supplementary Fig. 15) and might not be a pure *C. maxima*. Therefore CHP was not used to derive diagnostic SNPs on this chromosome and diagnostic SNPs were based on LAP/PKM and LAP/WLM as follows:

a. 0-5 Mb: alleles separately fixed between LAP and PKM (WLM contains highly heterozygous segments in this region and is unlikely to be pure *C. reticulata*)
b. 5 Mb-end of chromosome: alleles separately fixed between LAP and WLM (PKM contains highly heterozygous segments in this region and is unlikely to be pure *C. reticulata*).

In this way, we obtain 1,537,264 diagnostic SNVs that we can use to differentiate *C. maxima* and *C. reticulata* (below, often abbreviated as *C. max* and *C. ret* respectively). These can be found in Supplementary File 1.

Supplementary Note 8: Admixture in the citrus genomes

Using a sliding window of 2,000 diagnostic SNVs, the likelihood of each of the three genotypes (*C. ret./C. ret., C. max./C. max., C. max./C. ret.*) can be estimated for each window. The genotype with over 50% support among the 2,000 SNVs in a window is considered the genotype for the window. If no genotype has more than 50% support, the corresponding genomic segment is considered of unknown genotype.

Based on the set of diagnostic alleles as defined above, LAP has 100% support for the *C. max./C. max.* genotype across its genome, as do 8 of the 9 CHP chromosomes. One short (~1.3Mb) segment ~ 25 Mbp from the beginning of chromosome 2 of CHP shows 76% support for a *C. max/C. ret* genotype based on *C. maxima* and *C. reticulata* alleles in LAP and WLM respectively. These and other results of the admixture analysis are shown in Figure 3, Supplementary Figure 18, main Table 1, Supplementary Table 8.1.

From this analysis, we conclude that all "mandarin" types that we sequenced include some admixture of *C. maxima* introgression, since no pure *C. reticulata* genotype was found among the 8 sequenced citrus cultivars. In particular, both PKM and WLM have inter-specific hybrid genotypes for three chromosome segments (*i.e.*, end of chromosome 6, and the beginning of chromosomes 3 and 8, see main Fig. 3, Supplementary Fig. 16). The identical locations of these hybrid segments between PKM and WLM raise the possibility that these cultivars share ancient ancestry. PKM also contains a *C. max./C. max.* segment near the end of chromosome 2 (Fig. 3, Supplementary Table 8.1). In an analysis of genome composition in which we calculated genomic fractions in genetic map units with respect to the reference genetic map of citrus [16], we found that among the 8 diploid citrus genomes, WLM and PKM contain the most *C. ret./C. ret.* genotype at 88% and 86% respectively, followed by CLM (76%) and WMM (74%). The sweet orange genome is characterized by 75% *C. max./C. ret.*, 20% *C. ret./C. ret.*,

and 5% *C. max.*/*C. max*. In contrast, the sour orange genome consists of at least 99% *C. max.*/*C. ret* (Supplementary Table 8.1).

In summary, except for the two pummelo genomes, admixture is prevalent in cultivated citrus, including, notably, “traditional” mandarin types that were previously thought to be derived purely from wild *C. reticulata*, without suspected admixture.

Supplementary Table 8.1. Size and proportions of admixed regions in citrus cultivars.
For the eight diploid genomes, sizes and corresponding proportions of the three genotypes (M/M, P/P, M/P) are given both in physical base pairs (Mbp) and genetic map length (cM). For the haploid Clementine reference (HCR), proportions of the two haplotypes (M and P) are given. M=*C. reticulata* P=*C. maxima*

Cultivar	Genotype	Distance (Mbp)	Proportion (Mbp)	Distance (cM)	Proportion (cM)
HCR	unknown	0	0	0	0
	M	254.9	0.89	959	0.90
	P	31.7	0.11	108	0.10
WLM	unknown	0.5	0.002	4	0.004
	M/M	261.0	0.91	938	0.88
	P/P	0	0	0	0
	M/P	25.1	0.09	125	0.12
PKM	unknown	0.2	0.001	3	0.003
	M/M	243.4	0.85	922	0.86
	P/P	2.0	0.007	11	0.01
	M/P	41.0	0.14	131	0.12
CLM	unknown	0	0	0	0
	M/M	165.7	0.58	812	0.76
	P/P	0	0	0	0
	M/P	121.0	0.42	255	0.24
WMM	unknown	1.2	0.004	1.0	0.001
	M/M	199.0	0.69	790	0.74
	P/P	0	0	0	0
	M/P	86.5	0.30	277	0.26
SSO	unknown	4.3	0.015	7.1	0.007
	M/M	0	0	0	0
	P/P	0	0	0	0
	M/P	282.6	0.98	1061	0.99
SWO	unknown	3.3	0.01	3.6	0.003
	M/M	38.9	0.14	213	0.20
	P/P	8.9	0.03	55	0.05
	M/P	235.7	0.82	797	0.75
LAP	unknown	0	0	0	0
	M/M	0	0	0	0
	P/P	286.6	1.00	1067	1.00
	M/P	0	0	0	0
CHP	unknown	0	0	0	0
	M/M	0	0	0	0
	P/P	285.3	0.996	1059	0.99
	M/P	1.3	0.004	8	0.008

Supplementary Note 9: Population genetic analysis and simulations

The divergence time between the ancestral populations of *C. maxima* and *C. reticulata* can be estimated using the non-admixed regions of the nuclear genome.

Supplementary Note 9.1: *C. maxima* and *C. reticulata* divergence time estimate from nuclear genomes

The joint genotype frequencies of two diploid genomes can be characterized by four parameters, namely the frequencies of the following genotypes: AA|BB, AB|AA, AA|AB, AB|AB, where A and B denote two different alleles, and the genotypes of the two diploid individuals are separated by '|'. We do not distinguish between ancestral and derived alleles, so that AB|AA simply means that the first individual is heterozygous and the second is homozygous.

These joint genotype frequencies can be fitted using the simplest population genetic model describing the divergence of two populations (Supplementary Fig. 19). This "pants model" is specified by four parameters:

- the population divergence time T;
- the effective population sizes of the two extant populations N_{max} (*C. maxima*) and N_{ret} (*C. reticulata*);
- the ancestral Citrus effective population size N.

In the absence of a pure *C. reticulata* genome among the sequenced individuals, we used LAP and WLM to obtain the four joint genotype frequencies, but excluded segments in WLM that were identified as admixed described in the previous section. Regions of unusually low nucleotide heterozygosity (<1 het site/kb) within WLM suggest very recent shared ancestry and thus also excluded from consideration. Other pairs of accessions can also be used, but results are similar. The observed paired genotype frequencies are as follows (the first and second genotypes refer to *C. maxima* and *C. reticulata* respectively):

AA|BB = 1.49%

AB|AA = 0.81%

AA|AB = 0.47%

AB|AB = 0.026%

Coalescent simulations were performed using MaCS[51]. The best-fit "pants model" parameters are:

$N_{max}/N = 0.31$

$N_{ret}/N = 0.19$

$T/N = 0.71$

$2N\mu = 0.0079$

where μ is the nucleotide substitution rate per base per generation, and T is measured in units of generations.

Assuming that citrus has a similar nucleotide substitution rate as poplar[35, 36], we used

$$\mu = (1\text{-}2) \times 10^{-9} \text{ /bp/yr}$$

to estimate the divergence time of *C. maxima* and *C. reticulata* as

$$T = 1.4\text{-}2.8 \text{ Mya.}$$

To crudely estimate the effective population sizes, we used a generation time of 5 years to obtain

$$N = (4.0\text{-}7.9) \times 10^5,$$

$$N_{max} = (1.2\text{-}2.4) \times 10^5,$$

$$N_{ret} = (0.7\text{-}1.5) \times 10^5.$$

The larger effective population size of pummelo is consistent with the higher level of standing variation in pummelo relative to mandarin.

Supplementary Note 9.2: A bottleneck in the *C. maxima* population

The presence of a second peak at ~1 het site/kb in the density spectrum of the pummelo nucleotide heterozygosity in addition to the main peak at ~6 het sites/kb (Fig. 2a) suggests a more complex demographic history than that described by the simple model with constant effective population size described above (Supplementary Note 9.1). Using LAP (the parent of CHP) as an example we show below that this could have been caused by a severe ancient population bottleneck in pummelo.

To estimate the time and strength of the bottleneck, we used a 3-epoch model with piecewise constant effective population size (Supplementary Fig. 20a). In this model, the pummelo population started with size N_{max} and experienced a bottleneck from time T_2 to T_1 during which the effective populations size was N_b. For simplicity we assume that the population recovered to its original size after the bottleneck.

Coalescent simulations were carried out using MaCS [51] with variable recombination rate modeled as an array of recombination hotspots, each with a strength 30 cM/Mbp and size 1 kb, as well as a between-hotspot recombination rate of 0.01 cM/Mbp. Hotspots were distributed to recover the global properties of the Citrus genetic map. A reasonable fit for the observed LAP heterozygosity spectrum is shown in Supplementary Figure 20b for 10 kb sliding windows. The bottleneck parameters were estimated by MaCS (Ref. [51]) as follows:

$$T_1/(4N_{max}) = 0.014$$

$$T_2/(4N_{max}) = 0.054$$

$N_b/N_{max} = 0.18$

The strength of the bottleneck depends on $(T_2-T_1)/N_b$, and can be measured in terms of the inbreeding coefficient[52, 53].

$F=1 - \exp(-(T_2-T_1)/(2N_b))=0.36$

In comparison to the out-of-Africa population bottleneck for anatomically modern humans [F~0.175 (Ref. [54, 55])], pummelo may have undergone a much more severe population crash and recovery.

With such a strong bottleneck, how will the divergence time of *C. maxima* and *C. reticulata* be affected? For this, we now turn to a more realistic model than the simplest pants model.

Supplementary Note 9.3: A more realistic model for the estimate of the divergence time between *C. maxima* and *C. reticulata*

We can use the pummelo bottleneck parameter values from the last section to revisit the citrus speciation time estimate, with the improved "pants model" that incorporates a *C. maxima* bottleneck (Supplementary Fig. 21).

The 4 variable parameters of the model (N, N_{max} ,N_{ret}, T) can be estimated by fitting the 4 joint-genotype frequencies of LAP and WLM as before. The fit from coalescent simulations in the infinite sites model gives

$N_{max}/N = 0.54$

$N_{ret}/N = 0.20$

$T/N = 0.78$

$2N\mu = 0.0077$

Assuming as before $\mu = (1\text{-}2) \times 10^{-9}$/bp/yr, the *C. maxima* and *C. reticulata* divergence time is

T = 1.5-3.0 Mya.

For the same mutation rate, the new estimate is ~7% older than the estimate without a bottleneck.

The other parameters can be estimated assuming a generation time of 5 years:

$N = (3.9\text{-}7.7) \times 10^5$

$N_{max} = (2.1\text{-}4.2) \times 10^5$

$N_{ret} = (0.8\text{-}1.6) \times 10^5$

The most notable change in the presence of the bottleneck is the *C. maxima* population size prior to and after the bottleneck, with an increase of 75%.

The time and population size associated with the bottleneck can be similarly obtained:

$N_b = (3.8\text{-}7.5) \times 10^4$

$T_1 = 60\text{-}120$ kya

$T_2 = 230\text{-}450$ kya

This shows that the pummelo population was reduced by 80% to around 57,000 trees during the bottleneck, which lasted 170 - 330 kyr. Compared with species divergence time, the occurrence of the bottleneck is recent and its duration short (about one tenth of the speciation time). This might explain why the species divergence time estimate is not much affected by the bottleneck.

To examine the sensitivity of the demographic parameters to the assumption of an infinite sites model, we also conducted coalescent simulations with an in-house script allowing for parallel mutations. With the bottleneck parameter values from last section, the best fit of the observed joint genotype frequencies of LAP/WLM yields very similar results to that of the infinite sites model. In particular, the divergence time of *C. maxima* and *C. reticulata* was found to be T = 1.6-3.2 Mya, a few percent larger than found with the infinite sites model.

We note that the above estimates of *C. maxima* and *C. reticulata* divergence time are in line with paleontological and phylogenetic findings that put the divergence time between *Citru*s and its sister genus *Poncirus* at 4.0-9.6 Mya [56].

Supplementary Note 10: Analysis of relatedness in citrus.

The direct inference of parent-child relationships between two diploid individuals is possible when phased haplotypes are available for one of them, as is the case with the haploid Clementine reference (HCR) described above, and the haploid sweet orange assembly (RefSO) [21]. By mapping reads from the second individual to a haploid reference, we can identify genomic segments that are shared between two genomes based on the absence of "homozygous SNVs" in the diploid relative to the haploid reference.

This concept is illustrated in Supplementary Figure 22, where three homologous sequence segments coalesce in two steps to reach their common ancestor A. If the reference sequence is R and the two haplotypes of the diploid are h_1 and h_2, the first coalescence event with common ancestor B can involve R and h_1, R and h_2, or h_1 and h_2 (Supplementary Fig. 22). In the Newick notation, these 3 genealogies are ((R, h_1), h_2), ((R, h_2), h_1), and (R, (h_1, h_2)) respectively.

For topology ((R, h_1), h_2) where R and h1 coalesce first, homozygous non-reference SNVs observed in the diploid (h_1, h_2) genome directly measure the number of base substitutions in the R-B branch (Supplementary Fig. 22a). In the infinite sites model, where each substitution occurs at a different site, the number of homozygous SNVs is about half of the nucleotide diversity (i.e. mismatch rate) between R and h_1. As an example, assume the reference R=*C. reticulata*, and the diploid is a hybrid of h_1 = *C. reticulata* and h_2 = *C. maxima*, the rate at which homozygous non-reference SNVs occur in the diploid h_1/h_2 is about half of the (within-species) *C. reticulata* nucleotide heterozygosity. When R and h_1 share an ancestor in the immediate past (*e.g.*, shared haplotype between offspring and parent), the R-B branch length approaches zero and no homozygous non-reference SNVs will be present in the diploid (h_1, h_2) compared to the reference, R. The reverse is also true and is the basis of our inference of a shared haplotype based on the absence of homozygous non-reference SNVs. The same reasoning holds for topology ((R, h_2), h_1) (Supplementary Fig. 22b).

For the third topology (R, (h_1, h_2)) where h_1 and h_2 coalesce first, homozygous non-reference SNVs in the diploid (h_1, h_2) measure the number of base substitutions on the two branches B-A and A-R (Supplementary Fig. 22c), and correlate with the genetic distance between R and the diploid. As an example, if R is a *C. maxima* haplotype and h_1 and h_2 are *C. reticulata* haplotypes, homozygous non-reference SNVs in the diploid (h_1, h_2) approximate the inter-species divergence.

In summary, homozygous non-reference SNVs in a diploid genome compared to a reference haplotype can reflect intra- or inter-specific nucleotide diversity, but the absence of homozygous non-reference SNVs in certain regions implies shared sequence segments between the haploid reference and the diploid genome.

Because a parent and its offspring share haplotypes across the genome, one should not observe any homozygous non-reference SNVs across the genome of the parent relative to one haploid sequence of the offspring (and *vice versa*).

Note that this method of shared sequence detection between two diploid individuals (including the inference of parent/offspring relationship) is independent of population structure and admixture, and does not require the knowledge of allele frequencies in the population. These allele frequencies are required by existing human kinship inference methods (e.g., PLINK [57], KING [58], REAP [59]). Instead, our method makes use of the availability of the reference sequence of one of the two individuals being compared.

Supplementary Note 10.1: Origin of Clementine mandarin and high degree of inbreeding within Clementine

Clementine is commonly described as having arisen from a cross between a traditional Mediterranean-type mandarin (exemplified by Willowleaf) and a sweet orange[8, 60]. In our analysis, we do not assume this, but instead try to infer the parentage of Clementine by sequence analysis.

We mapped the short reads of Clementine mandarin (CLM), sweet orange (SWO) and Willowleaf mandarin (WLM) to the haploid Clementine reference (HCR) as described above. We used the CLM heterozygous SNVs and HCR sequence to infer the second haploid sequence of CLM (HCA; where the A stands for "alternative" haplotype). This allowed us to compare SWO and WLM to the two haploid genomes of Clementine separately. The homozygous non-reference SNVs relative to HCR and HCA are plotted separately for SWO (Supplementary Fig. 23a) and WLM (Supplementary Fig. 23b) in overlapping windows of 500,000 callable sites, with step size 250 kb.

We find that the SWO homozygous non-reference SNV rates with respect to HCR and/or HCA are uniformly low across the genome (Supplementary Fig. 23a), suggesting parent/offspring kinship between SWO and CLM.

Similarly, the low genome-wide homozygous SNV rate relative to HCR and/or HCA in WLM (Supplementary Fig. 23b) provides evidence for parent/offspring relationship between WLM and CLM.

Since CLM arose in the past hundred years or so, it has a much younger history than WLM and SWO. Based on the match between CLM and WLM cpDNA, we conclude that WLM is the female parent of CLM and SWO is the male parent of CLM. Inference of the parentage of CLM confirms, at nucleotide resolution, earlier studies utilizing a limited number of markers [8, 44, 60].

We derived a parentage map of the haploid Clementine reference, and the positions of crossovers that occurred in the formation of HCR, by comparing the homozygous SNV rates of SWO and WLM along the HCR chromosomes. For this purpose, we identify shared haplotypes between HCR and SWO/WLM using a cutoff of 0.02% for the homozygous SNV rate. For example, an HCR segment is of SWO origin if the homozygous SNV rate in the SWO reads is below 0.02% throughout the segment.

Some HCR segments, however, seem to share their haplotypes with *both* SWO and WLM, as *both* SWO and WLM have very low (<0.02%) homozygous SNP rate relative to those segments. As we will show below, this is mostly due to a high degree of inbreeding in the CLM diploid genome. For these ambiguous segments, we can assign their parental origin according to their adjacent segments. In this way, we are seeking a minimal-recombination reconstruction of the HCR parentage map. Very rarely, the two neighboring segments of an ambiguous HCR sequence have different parental origins (e.g. near the end of Chr. 3), and this type of ambiguity can be resolved by the parentage of the second CLM haploid sequence (HCA).

The parentage map for HCR is shown in Supplementary Figure 23c. We identify 10 crossovers in the generation of HCR. This is consistent with the total genetic map length of 11 Morgans [16]. For this minimal-crossover reconstruction of the CLM parentage map, the proportions of SWO and WLM in the haploid Clementine reference sequence are 19% and 80% respectively when measured in

physical bases, or 29% and 70% respectively in genetic map units. About 1% of the HCR genome cannot be assigned by our methods and has unknown parental origin.

To examine the degree of inbreeding in the CLM diploid genome, we computed the nucleotide heterozygosity in 100kb windows and identified regions of low heterozygosity (*i.e.*, regions that are homozygous) (Supplementary Fig. 24). About 19% of the CLM genome (or 10% of the total genetic map distance) has nucleotide heterozygosity below 0.02% and results from haplotypes being shared between the two haploid sequences of CLM. These identical-by-descent (IBD) segments account for most of the regions in HCR that have ambiguous parental origin, and they reveal an unexpectedly high degree of genetic relatedness between the two parents of CLM, namely WLM and SWO (see Supplementary Note 10.2.1).

Supplementary Note 10.2: Haplotype sharing analysis

Supplementary Note 10.2.1: Haplotype sharing between sweet orange and mandarins

To detect shared genomic segments between SWO and other citrus genomes, we made use of the haploid SWO assembly [21] and inferred the second sweet orange haplotype based on heterozygous SNVs in the SWO shotgun reads. Other citrus genomes were then compared to the two SWO haploid sequences separately to identify shared haplotypes based on the absence of homozygous SNVs as was described for Clementine in the previous section.

Surprisingly, all three traditional mandarins (PKM, WLM, and the recently sequenced Chinese mandarin HLM (see Supplementary Note 11)) share extensive haplotypes with SWO, with PKM showing the highest relatedness. As in the last section, we use a cutoff of 0.02% on the homozygous SNV rate in 100kb windows to define haplotypes shared with SWO. Proportions of the SWO genome sharing haplotypes with PKM, WLM, and HLM are 76%, 34% and 38% respectively for the 9 assembled SWO chromosomes measured in Mb. These are shown in Supplementary Figure 25, which also shows common haplotype-sharing regions for all three mandarins. This surprising result suggests shared ancestry between SWO and the traditional mandarins, and highlights the limited genetic diversity for the cultivated mandarins in our study, most of which were previously thought to be unrelated.

The homozygous segments observed in the Clementine genome (see Supplementary Note 10.1) confirm the genetic relatedness of its two parents (SWO and WLM), which we have computed directly (see above).

Citrus breeding records (University of California, Riverside, Citrus Variety Collection, http://www.citrusvariety.ucr.edu/citrus/wmurcott.html) indicate that Murcott mandarin is a parent of W. Murcott mandarin (WMM), and there is general consensus that sweet orange is a parent of Murcott mandarin [13].

Haplotype sharing analysis finds that 34% of the SWO genome is shared with WMM, consistent with the suspected grandparent/grandchild relationship.

In contrast, similar analyses found no relatedness between SWO and Seville Sour Orange (SSO), or between SWO and the four pummelos (Low acid pummelo (LAP), Chandler (CHP), and the two recently sequenced Chinese Guanxi (GXP) and Shatian (STP) pummelos); see Supplementary Note 11).

The high proportion of haplotype sharing between Ponkan mandarin and SWO (~76% of genome) indicates a close relationship between the two. Since a parent/offspring pair would share haplotypes throughout the genome (*i.e.*, 100%), PKM and SWO are not simply parent and child. Though haplotype sharing between a grandparent and a grandchild has a mean of 50%, simulations show the 95% confidence interval is 28-72% sharing, assuming both the parents and grandparents are unrelated. It thus seems possible that PKM and SWO could be related as a grandparent/grandchild pair with possible haplotype sharing among the parents. Alternately, PKM and SWO share one or more common ancestors.

In an independent test of the relatedness between PKM and SWO, 168 SSR markers [61-63] distributed across all 9 chromosomes were examined for allele sharing between PKM and SWO. Of these markers, 155 (92%) have matching alleles between PKM and SWO, further supporting their close relationship.

Supplementary Note 10.2.2: Haplotype sharing among three traditional mandarins

To calculate the proportions of haplotypes shared between the traditional mandarins, we made use of the identical-by-state (IBS) measure. For two diploid genomes, the joint genotypes can be classified into 4 types: AA|BB (no shared alleles or IBS0), AB|AA and AA|AB (one shared allele or IBS1), and AB|AB (two shared alleles or IBS2). It has been shown [64] that the identical-by-state ratio (IBSR) can be written thus

$$IBSR = IBS2/(IBS2+IBS0)$$

IBSR is also independent of population allele frequencies and has a mean of ⅔ for two unrelated individuals from the same homogenous population. If two individuals share haplotypes over a genomic segment, IBS0 is zero and IBSR=1 for that segment. Thus IBSR values close to 1 can reveal relatedness between two diploid genomes.

Identical genotypes shared between two diploids can be separately inferred as a special case of haplotype sharing, when the distance measure D (see Supplementary Note 7 for definition) becomes zero for certain sequence segments. This corresponds to IBS0=IBS1=0 and only IBS2 is non-zero. The total haplotype sharing proportion is defined as a simple sum of both genotype and haplotype sharing. Thus both identical twins and a parent/offspring pair share haplotypes across 100% of the genome.

Supplemental Note 10.2.3: Development of an IBS2+ method for calculating haplotype sharing

In the presence of interspecific admixture the above IBSR measure breaks down when some genomic segments are interspecific hybrid in both diploids. For example, the sweet and sour orange genomes are mostly hybrids of *C. maxima* and *C. reticulata* and are unrelated based on homozygous SNV rate (see Supplementary Note 10.5). The predominant joint genotype is thus IBS2, with the number of IBS0 sites decreasing exponentially with time since speciation. The value of IBSR is approximately 1 even though the two oranges are not related. In other words, one cannot use IBSR to detect haplotype sharing for *C.ret/C.max* hybrid regions of the genome. In this case, we made use of the common shared haplotypes with SWO when comparing the hybrid segments of two mandarins. This provides a lower bound on shared haplotypes as the two mandarin genomes can share haplotypes beyond those shared with SWO. In practice, the lower bound gives a good approximation to the true haplotype sharing as the hybrid proportion between two mandarin genomes is usually small. We refer to this modified approach as the "IBS2+" method.

High coverage Illumina reads for the three traditional mandarins, Ponkan (PKM), Willowleaf (WLM) and the recently sequenced Chinese mandarin Huanglingmiao (HLM, Supplementary Note 11) were mapped to the haploid Clementine reference, and pairwise IBSR and D were computed in 100 kb windows. Admixture analysis (Supplementary Note 8) was used to determine the hybrid segments for a pair of mandarins. For these hybrid regions, haplotype sharing between a pair of mandarins is estimated using their common shared haplotypes with SWO based on homozygous SNV rate relative to the haploid SWO reference sequence and the inferred second haploid SWO sequence. To translate from the SWO coordinate system to the haploid Clementine coordinate, we generated a correspondence map or dotplot (Supplementary Fig. 26).

As an example, Supplementary Fig. 27 shows haplotype sharing between Ponkan (PKM) and Willowleaf (WLM) mandarins in 100 kb windows, with IBSR > 0.99 and D < 0.01 for haplotype and genotype sharing respectively, to account for errors in SNV calls and cumulative somatic mutations for not-too-distant shared ancestries (up to ten thousand years). The hybrid proportions of the genome are 5.4% (molecular distance) and 9.1% (genetic map distance). Haplotype sharing accounts for 56% (molecular distance) and 46% (map distance) of the genome. Consistent with the significant proportions of common haplotype sharing with SWO, the genetic relatedness of PKM and WLM is surprisingly high, comparable to the amount of haplotype sharing of a grandparent/grandchild pair.

We calculated the haplotype sharing proportions among three traditional mandarins (Supplementary Table 10.1). As an independent test of the IBS method, we also computed haplotype sharing for the parent/offspring pair WLM/CLM (see Supplementary Note 10.1). We show estimates of the common haplotypes shared with SWO; our modified IBS method (IBS2+) described above that makes special consideration of the hybrid segments and total haplotype

sharing due to the small proportion of hybrid segments. As noted above, our IBS2+ estimate is very close to the hybrid value (Supplementary Table 10.1).

A significant portion of the total haplotype sharing for all four pairs of mandarins is due to common haplotypes shared with SWO (Supplementary Table 10.1). Together with the high-degree of haplotype sharing between the traditional mandarins, this is further evidence for recent shared ancestry between SWO and the three traditional mandarins, as has been indicated by the notable amount of haplotype sharing between SWO and the mandarins (Supplementary Note 10.2.1). Furthermore, if the three traditional mandarins represent random samplings of the existing mandarins, this analysis suggests that cultivated mandarins have surprisingly limited genetic diversity.

Supplementary Table 10.1. Haplotype sharing among the three traditional mandarins and between the parent/offspring pair WLM/CLM.
Haplotype sharing is shown as the proportion of physical map distance (Mb prop.) as well as genetic map distance (cM prop.). The column "common w/ SWO" shows SWO haplotypes shared with both mandarins, as estimated by requiring homozygous SNV rate $< 2 \times 10^{-4}$ relative to the haploid SWO reference and the second haploid SWO sequence. The column "IBS2+" estimates total haplotype sharing based on common shared haplotypes with SWO for the hybrid segments and IBSR for other regions. The last column lists the hybrid proportions.

		Common w/ SWO	IBS2+	Hybrid
PKM/WLM	Mb prop.	0.276	0.558	0.0542
	cM prop.	0.204	0.459	0.0908
PKM/HLM	Mb prop.	0.264	0.554	0.0130
	cM prop.	0.204	0.467	0.0366
WLM/HLM	Mb prop.	0.137	0.348	0.0145
	cM prop.	0.110	0.353	0.0415
WLM/CLM	Mb prop.	0.279	0.976	0.0585
	cM prop	0.209	0.978	0.0509

Supplementary Note 10.3: Origin of Sweet Orange

Based on the admixture pattern of sweet orange (SWO) and its hybrid (*C. maxima/C. reticulata*) proportion, we can make some inferences regarding the origin of SWO. First, to account for the M/M and P/P segments of SWO (where M=*C. reticulata* P=*C. maxima*), both parents must be an admixed pummelo (*C. maxima*) or mandarin (*C. reticulata*). Thus admixture analysis allows us to rule out two models previously proposed for SWO origin, either as the direct F1 hybrid *C. maxima* × *C. reticulata* [8, 65] or as the backcross (*C. maxima* × *C. reticulata*) × *C. reticulata* [21]. We note that the presence of P/P segment on SWO chromosome 2 (Supplementary Note 8) has recently been confirmed by directed sequencing of three genes[66].

In SWO, 75% of the total genetic map distance comprises hybrid P/M segments. This proportion restricts the possible models of SWO origin. Although the genomic coordinates of the M/M, M/P and P/P segments within PKM and SWO are consistent with a parent-offspring relationship between SWO and PKM, this is ruled out because haplotype sharing is less than 100%. Thus, a hypothetical mandarin-like parent of SWO (which we denote PKX) could have identical M/M, P/M and P/P segment boundaries to those found in PKM and would also have 100% haplotype sharing with SWO (necessarily implying sequence variation relative to PKM). Based on the cpDNA inheritance pattern in citrus, this mandarin-like parent would be the male parent of SWO.

We used this hypothetical male parent, PKX, in simulations to investigate the constraint on models of SWO origin imposed by the proportion of P/M segments in the SWO genome. Since SWO nuclear DNA has a high proportion of *C. maxima* alleles, we are left with two simple models for SWO,

Model A (F1 hybrid maternal parent): SWO = (*C. maxima* × *C. reticulata*) × PKX

Model B (Backcross maternal parent): SWO = ((*C. maxima* × *C. reticulata*) × *C. maxima*) × PKX.

Note that in Model B, there are two other possible schemes for the female parent of SWO, namely, *C. maxima* × (*C. reticulata* × *C. maxima*) and *C. maxima* × (*C. maxima* × *C. reticulata*), and we don't distinguish among the three versions.

To estimate the relative likelihood of the two models, we performed 100,000 simulations for each model using both the admixture patterns of SWO and PKM and the genetic map of the 9 chromosomes. Distribution of the different genotype proportions (of the total genetic map distance) can then be calculated and compared to the observed values in SWO. Here we focus on the inter-specific hybrid proportion in the SWO genome, with an observed value of 0.75.

In Model A, the female parent of SWO is assumed to be an F1 hybrid between *C. maxima* and *C. reticulata*. Distribution of the hybrid proportion in SWO from the simulation is shown in Supplementary Figure 28, with the observed value shown as a vertical dashed red line. The mean of the distribution is 0.50, and the standard deviation is 0.11. The observed value resides in the tail of the distribution (one-sided test, p=0.01).

In Model B, on the other hand, the female parent of SWO contains a higher fraction of *C. maxima* alleles than in model A (the mean values are 0.75 and 0.50 for model B and A, respectively), which in turn can result in a higher hybrid proportion in SWO. Distribution of the SWO hybrid proportion is shown in Supplementary Figure 28, with a mean 0.71 and standard deviation 0.09. The observed value (0.75, Supplementary Table 8.1) is located near the mode of the distribution and within 1 standard deviation uncertainty (Supplementary Fig. 28). Furthermore, the observed proportion of pure *C. maxima* genotype in SWO (0.05,

Supplementary Table 8.1) also lies near the mode of the corresponding distribution (mean=0.054, S.D.=0.026). The same is true for the proportion of pure *C. reticulata* genotype in SWO (observed=0.20, Supplementary Table 8.1; simulation mean=0.23, S.D.=0.09).

Thus we conclude that based on the typical fractions of P/M, M/M and P/P segments found in simulated crosses, Model B is a far more likely model for the origin of sweet orange. We note, however, that sweet orange was likely a human selection from perhaps hundreds or thousands of sampled wild hybrids or populations derived from human activity. Without an understanding of this selection process, we cannot definitively rule out Model A. In any event, the high P/M proportion in sweet orange can restrict the models of SWO origin that can be proposed. In particular, it is likely that the male parent of SWO is a Ponkan-like mandarin (PKX), and the female parent more likely originated from a backcross (model B) instead of a direct hybridization of *C. maxima* and *C. reticulata* (model A).

Finally, we note that while the Seville sour orange (SSO) is an F_1 cross between *C. maxima* and *C. reticulata* (as shown below), it is not a parent of sweet orange (Supplementary Note 10.5).

Supplementary Note 10.4: Parent/offspring relationship between Low acid and Chandler pummelos

Low acid pummelo (LAP) has been reported as the seed parent of Chandler pummelo (CHP) (Ref [67]). To test this relationship, we computed allele sharing between LAP and CHP using two IBS measures: zero allele sharing (IBS0) counts discordant homozygous SNPs (*i.e.,* LAP|CHP joint genotype=AA|BB) and two-allele sharing (IBS2) counts concordant heterozygous SNPs (joint genotype=AB|AB). The distribution of zero and two allele sharing frequencies per polymorphic site in 500kb windows is shown in Supplementary Figure 29. The genome wide absence of IBS0 sites is consistent with a parent/offspring relationship between LAP and CHP.

Supplementary Note 10.5: Seville sour orange is an F_1 hybrid of *C. maxima* and *C. reticulata*, and is not related to sweet orange

From the admixture analysis, the genotype of Seville sour orange (SSO) is a *C. maxima/C. reticulata* hybrid across the entire genome, with a characteristic inter-specific nucleotide heterozygosity of ~1.7% (Supplementary Fig. 30). This implies that SSO originated from an F1 cross between a *C. maxima* individual and a *C. reticulata* individual.

To examine the genetic relatedness between sour orange and sweet orange, SSO homozygous SNVs with respect to the haploid SWO reference [21] (RefSO) were computed (Supplementary Fig. 30). The genome-wide homozygous SNV rate is ~0.35%. This implies that SSO and SWO are not related. Since the observed homozygous SNV rate is about half the within-species nucleotide diversity, one

can further conclude that SSO and SWO are as unrelated as two haplotypes randomly chosen from the same species (Supplementary Fig. 22a).

Overall, we derived five pedigree relationships between citrus cultivars in this work (Supplementary Table 10.2).

Supplementary Table 10.2: Summary of principal citrus relationships. The principal relationships derived in this study are summarized.

Accession	Pedigree
Clementine	willowleaf × sweet orange
W. murcott	grandchild of sweet orange
Sour orange	(unknown pummelo) × (unknown mandarin)
Sweet orange	(pummelo-mandarin hybrid or backcross with pummelo) × (admixed mandarin related to Ponkan)
Chandler pummelo	child of low acid pummelo

Supplementary Note 11: Analysis of Chinese citrus genomes

To test our hypothesis that there is widespread pummelo introgression into cultivated mandarin types and to examine the genetic diversity of citrus, we analyzed deep sequence data from three pummelos and three mandarins[21] using the set of diagnostic SNV alleles that distinguish *C. maxima* and *C. reticulata*.

One of the six Chinese citrus genomes[21] has an identical genotype to Clementine mandarin; a second is identical to Low-acid (Siamese Sweet) pummelo from our cultivar collection. Our analysis therefore focuses on the four other genome sequences: two Chinese pummelos (Guanxi pummelo (GXP) and Shatian pummelo (STP)), Huanglingmiao mandarin (HLM), and "Mangshan mandarin" (CMS). In citrus taxonomies that assign distinct species names to specific mandarins (e.g., *C. clementina* for Clementine mandarin) the binomial *C. mangshanensis* has been suggested for "Mangshan mandarin," and we adopt this notation based on the demonstration below that CMS is highly divergent from other sequenced *Citrus*, including other "mandarins."

Supplementary Note 11.1: The two Chinese pummelos represent *C. maxima* without inter-specific admixture

Based on the set of diagnostic *C. maxima* and *C. reticulata* alleles described above (see Supplementary Note 7), both Guanxi pummelo (GXP) and Shatian pummelo (STP) show a pure *C. maxima* genotype like the two Siamese pummelos in our collection (LAP and CHP).

To look for close kinship and population structure among the four pummelos, we calculated pairwise distances between individuals, D, using the formula given in Supplementary Note 7. There are a few scenarios for which D is expected to have particular values: (1) a pair with identical genotypes is expected to have D = 0; (2)

parent/offspring pairs are expected to have D = 1/3; (3) unrelated diploid individuals drawn from the same homogenous population are expected to have D = ½. In contrast, two individuals sampled from two distinct gene pools are expected to have values of D between 0.5 and 1, depending on the degree of divergence of the two sub-populations.

Using this measure we find that the two Chinese pummelos (GXP and STP) are not close relatives (D = 0.47), and furthermore that they bear no kinship with the two Siamese pummelos (mean D = 0.49). On the other hand, our two Siamese pummelos are a parent/offspring pair with D = 0.33.

The heterozygosity levels of both Chinese pummelos are similar to those of the Siamese pummelos (~ 6 het sites/kb) based on their common mapped sites. Furthermore, the average nucleotide divergence between Chinese and Siamese pummelos is very close to the average nucleotide diversity of the four pummelos. This suggests limited divergence between the Chinese and Thai pummelo populations.

The absence of SNVs in the chloroplast genomes of the four pummelos, is also consistent with a recent common ancestor for the cpDNA sequences.

Finally, the distribution of heterozygosity in 10 kb windows (Supplementary Fig. 31) shows that all four pummelos are characterized by similar intra-species scale of ~5-6 het sites/kb, and that all show the secondary peak around 1 het sites/kb. This is consistent with the existence of an ancient population bottleneck in *C. maxima*, prior to the separation of the Chinese and Thai populations.

Supplementary Note 11.2: The genome of Huanglingmiao mandarin shows pummelo introgression

Huanglingmiao mandarin (HLM) is another "traditional" mandarin type without previously suspected inter-specific introgression. Indeed, Xu *et al.* used HLM and "Mangshan mandarin" to compile their catalog of nominal mandarin-specific variants; this would be inappropriate if HLM shows evidence of introgression with pummelo, as we report here, and even more inappropriate given the uniqueness of "Mangshan mandarin" relative to other traditional mandarins also reported here.

Using our set of diagnostic *C. maxima* and *C. reticulata* alleles (see Supplementary Note 7), the HLM genome shows pummelo introgression on chromosomes 3, 4 and 8 (Supplementary Fig. 32). More specifically, most of the introgressed regions have hybrid *C.maxima/C.reticulata* genotypes, with the exception of one short segment on chromosome 8 with a pure pummelo genotype (*i.e., C. maxima/C.maxima*) (Supplementary Fig. 32). As expected, these inter-specific hybrid segments manifest much higher nucleotide heterozygosity (~2%) than the rest of the genome. Furthermore, we find HLM shares haplotypes with PKM and WLM, indicating recent shared ancestry. This observation lends

further support to our proposal that pummelo introgression is widespread among cultivated mandarins.

Supplementary Note 11.3: *C. mangshanensis* represents a distinct species from *C. maxima* and *C. reticulata*

Previous studies have considered "Mangshan mandarin" to be a wild mandarin[68, 69], although there were earlier suggestions that it should be considered a "transitional" wild species called *C. mangshanensis* (CMS) [70].

We used our diagnostic set of *C. maxima* and *C. reticulata* single nucleotide differences derived from traditional mandarins (see Supplementary Note 7) to analyze the CMS ("Mangshan mandarin") genome. CMS is clearly not a wild *C. reticulata*, since it does not contain these diagnostic *C. reticulata* SNVs. It also does not resemble other sequenced mandarins since it does not contain any of the three expected genotypes *ret/ret, ret/max, or max/max*. Phylogenetic analysis of the CMS chloroplast genome together with 11 *C. maxima* and *C. reticulata* chloroplast genomes was performed using neighbor-joining in the PHYLIP package [71] and also shows that CMS is a third distinct type (main Fig. 4a) that is deeply diverged from *C. reticulata* cpDNAs

To quantify the divergence of CMS from *C. maxima* and *C. reticulata*, we calculated the rate of homozygous differences between it and our reference haploid Clementine sequence (HCR), and looked at its variation relative to both the *C. maxima* and *C. reticulata* segments of HCR. We observed ~2% divergence rates (~20 SNVs/kb) between CMS and both *C. maxima* and *C. reticulata* (main Fig. 4b), which is similar to the divergence between *C. maxima* and *C. reticulata*. The same divergence rate is observed between CMS and the other three available haploid sequences – the second haploid sequence of the Clementine diploid, the dihaploid sweet orange reference, and the alternate sweet orange haplotype. Thus the pairwise divergences between *C. mangshanensis*, *C. maxima* and *C. reticulata* are all comparable.

Based on these differences, we suggest that CMS should be considered a distinct species, *C. mangshanensis* on the same footing as *C. maxima* and *C. reticulata*. That CMS represents a pure member of this species with no detectable admixture is further supported by the observation that, the intrinsic nucleotide heterozygosity of CMS along the chromosomes shows no sharp transitions between different levels. The magnitude of this nucleotide heterozygosity (~6 het sites/kb) is comparable to that found in *C. maxima* and *C. reticulata*, again consistent with the variation observed within other *Citrus* species.

To visualize the relatedness and diversity of the different citrus genotypes, we generated a map of 12 citrus genomes based on principal coordinate analysis (PCo) of pairwise distances (D) (Supplementary Note 7) using metric multidimensional scaling (main Fig. 4c). The first principal coordinate (PCo1) separates the pummelos and the mandarins, with the two oranges lying between them. The second principal coordinate (PCo2) separates *C. mangshanensis* from

the pummelos, mandarins and oranges. As with pairwise fixed differences, this map clearly shows that *C. mangshanensis* is distinct from the pummelos, the mandarins, and the oranges.

Supplementary Note 11.4: Revisiting the hypothesis of Xu et al. [21] for the origin of sweet orange: alternative analysis and conclusions.

In their report of a dihaploid sweet orange genome assembly, Xu et al. [21] proposed a specific model for the origin of sweet orange, namely that sweet orange arose from a cross between a mandarin and a pummelo-mandarin F1 hybrid, *i.e.*, "(PxM)xM." Their proposal has several key differences from the models for sweet orange we discussed in Supplementary note 10.3. Several discrepancies with our data, and our analysis of their data, allow us to reconsider their proposal at multiple levels.

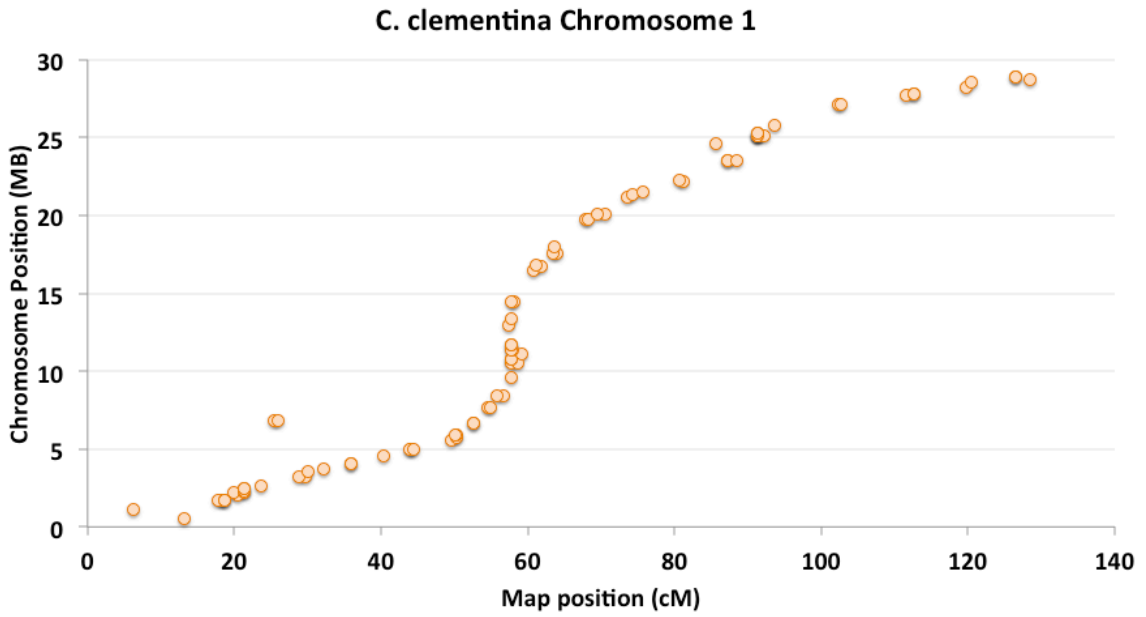
From a technical perspective, Xu *et al.* designate mandarin ("M") and pummelo ("P") alleles based on the contrast between the three "mandarins" they sequenced (*i.e.*, Mangshan, Clementine, and Huanglingmiao) and three pummelos. Their analysis treats these "mandarins" as selections from a single species, *C. reticulata*. As we have shown, however, (1) Clementine and Huanglingmiao mandarin both contain admixed segments of pummelo, and so do not faithfully represent *C. reticulata* alleles uniformly across the genome, and (2) Mangshan "mandarin" is a distinct third species, *C. mangshanensis*, and cannot be used to identify "M" segments. Thus the inputs to their analysis of "M" and "P" haplotypes are problematic.

Furthermore, sweet orange contains segments with the *C. max*/*C. max* (i.e., "P/P") genotype (Supplementary Note 10.3). In the simple (PxM)xM scheme proposed by Xu *et al.*, only P/M and M/M segments can occur in sweet orange. The detection of a P/P segment has been confirmed by independent observations [66] and rules out a simple (PxM)xM scheme unless the M pollen parent was admixed, which contradicts the assumptions of Xu et al.'s analysis.

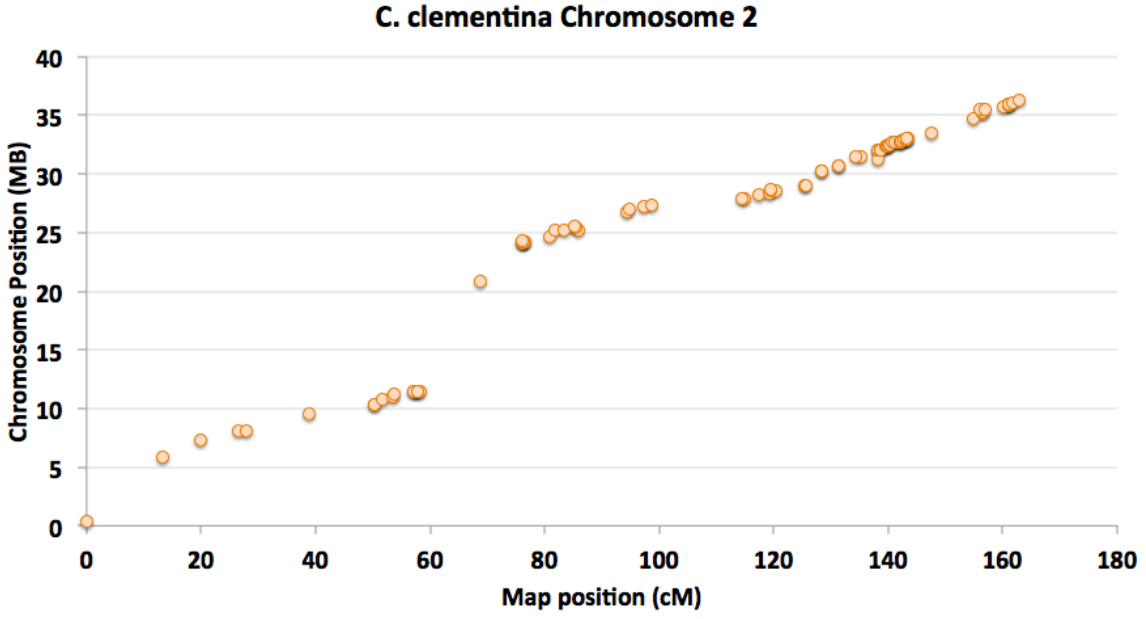
Finally, if one parent of SWO is mandarin-like, the high proportion (~75% map distance) of P/M hybrid segments in SWO places strong constraints on the second parent. In particular, it is much more likely that the second parent is (PxM)xP than PxM (see Supplementary Note 10.3). Thus the (PxM)xM scheme proposed by Xu *et al.* fails in two ways: it cannot explain the P/P segments in SWO, and also cannot account for the high proportion of P/M segments in SWO.

Supplementary Figures

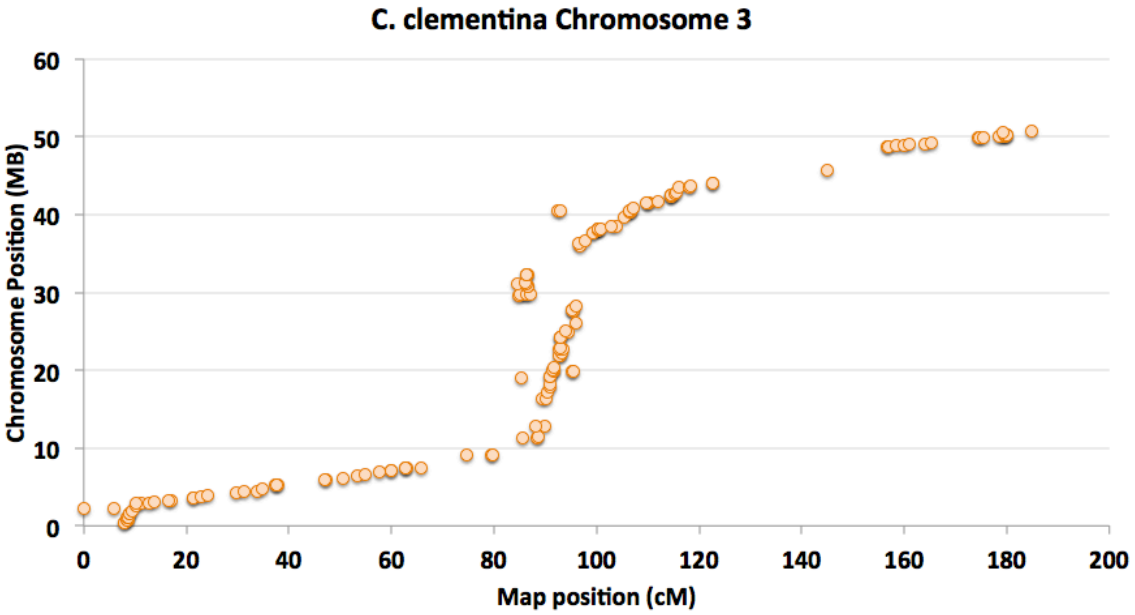
Supplementary Figure 1: Correspondence between genetic[16] and physical map for Haploid Clementine Chromosome 1



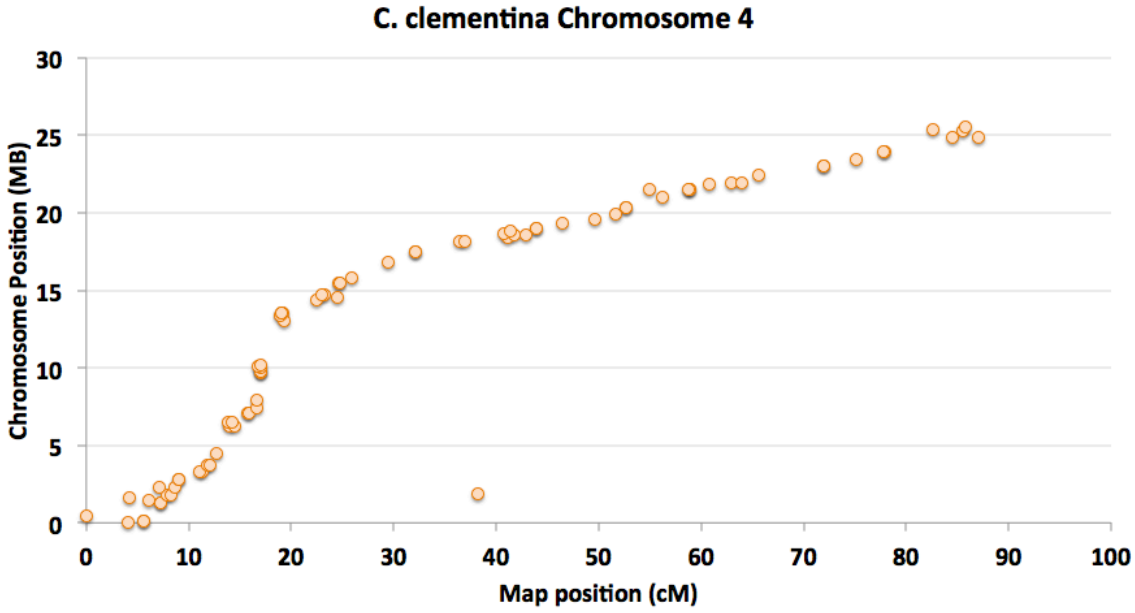
Supplementary Figure 2: Correspondence between genetic and physical map for Haploid Clementine Chromosome 2



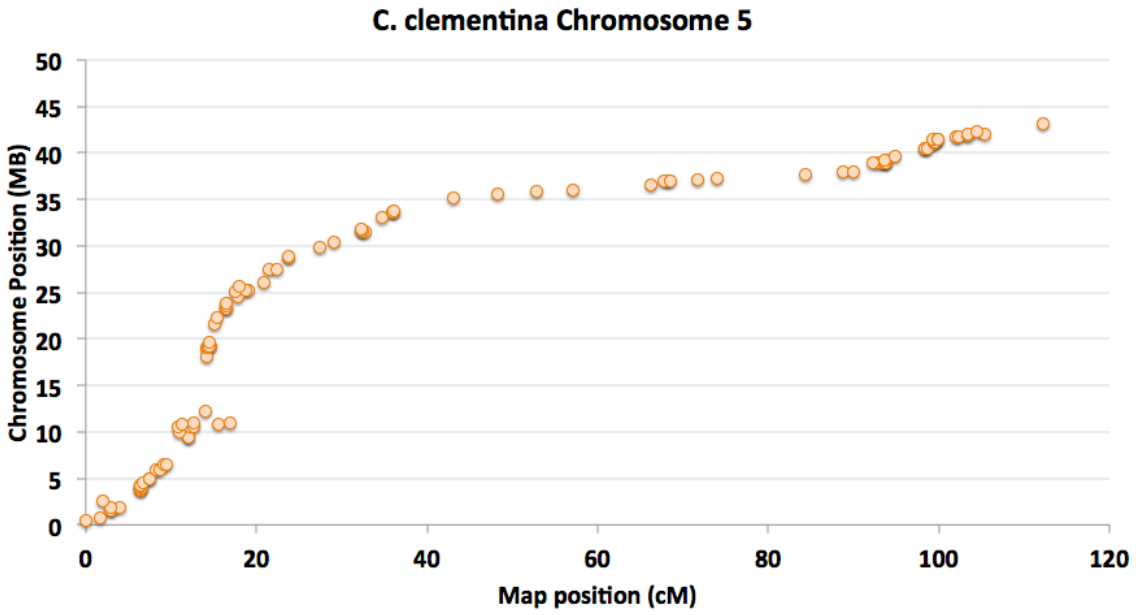
Supplementary Figure 3: Correspondence between genetic and physical map for Haploid Clementine Chromosome 3



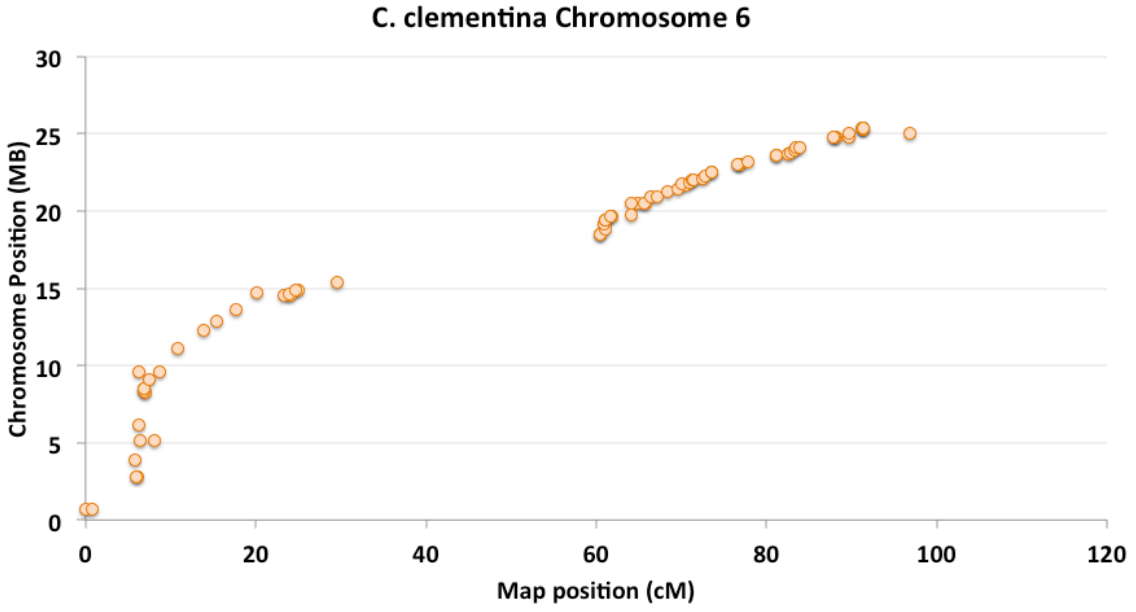
Supplementary Figure 4: Correspondence between genetic and physical map for Haploid Clementine Chromosome 4



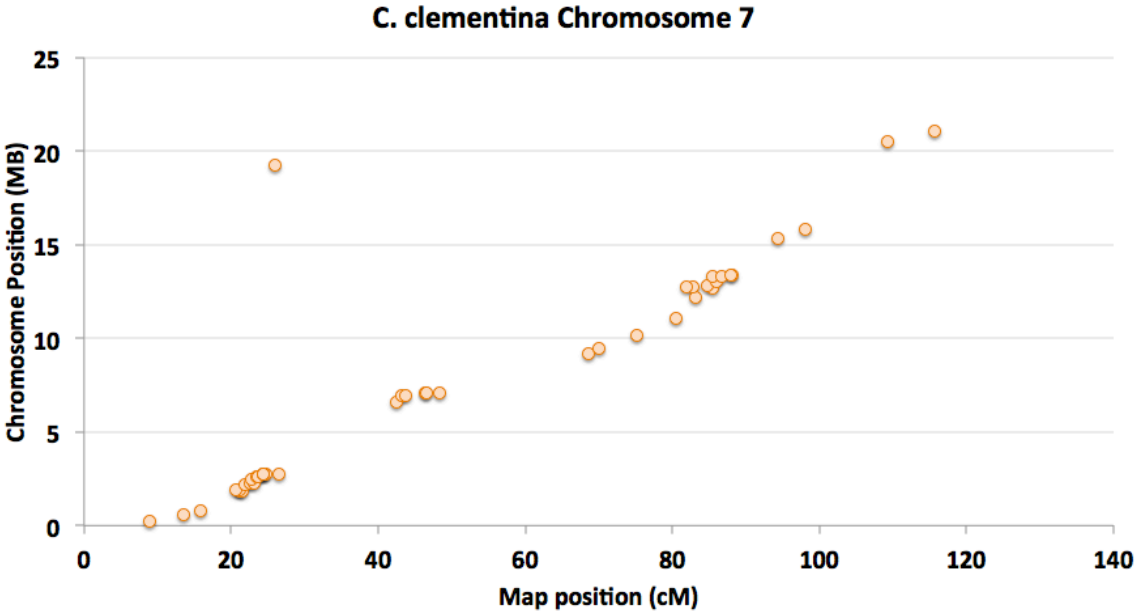
Supplementary Figure 5: Correspondence between genetic and physical map for Haploid Clementine Chromosome 5



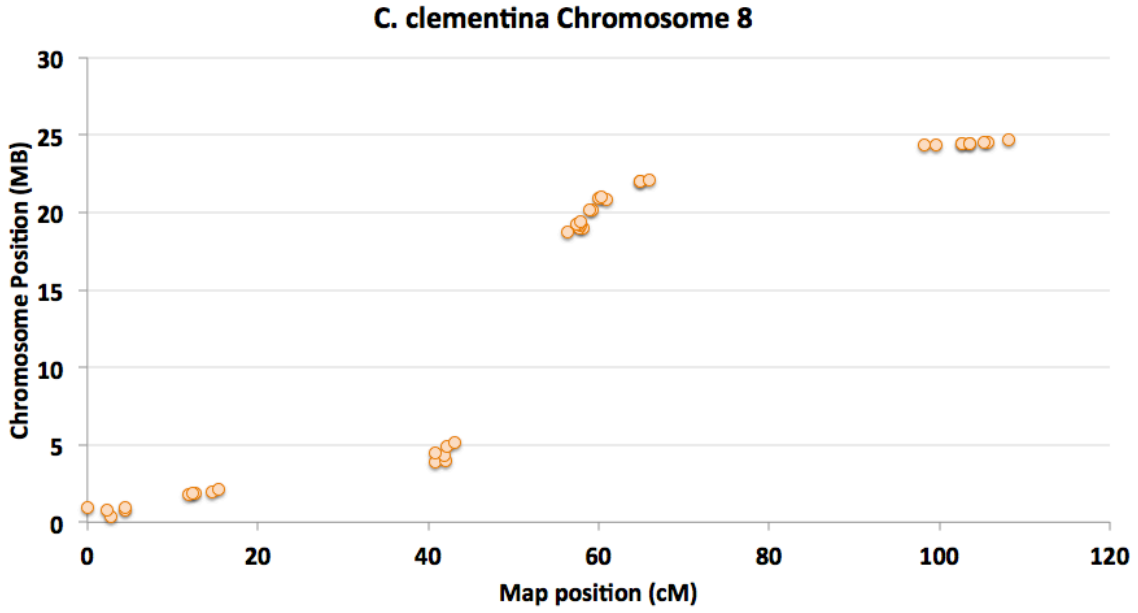
Supplementary Figure 6: Correspondence between genetic and physical map for Haploid Clementine Chromosome 6



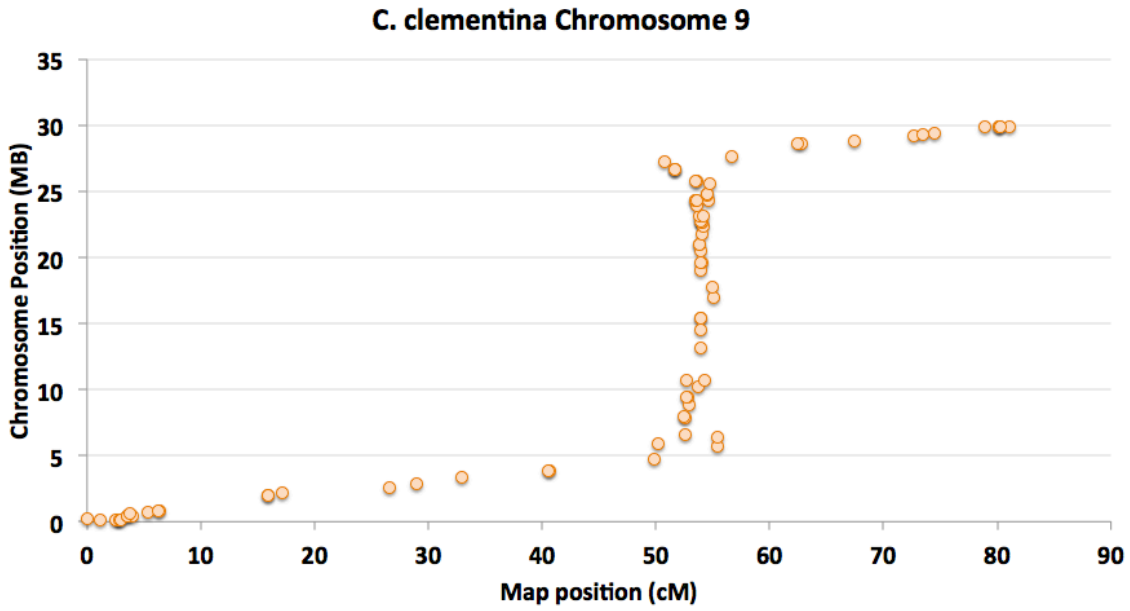
Supplementary Figure 7: Correspondence between genetic and physical map for Haploid Clementine Chromosome 7



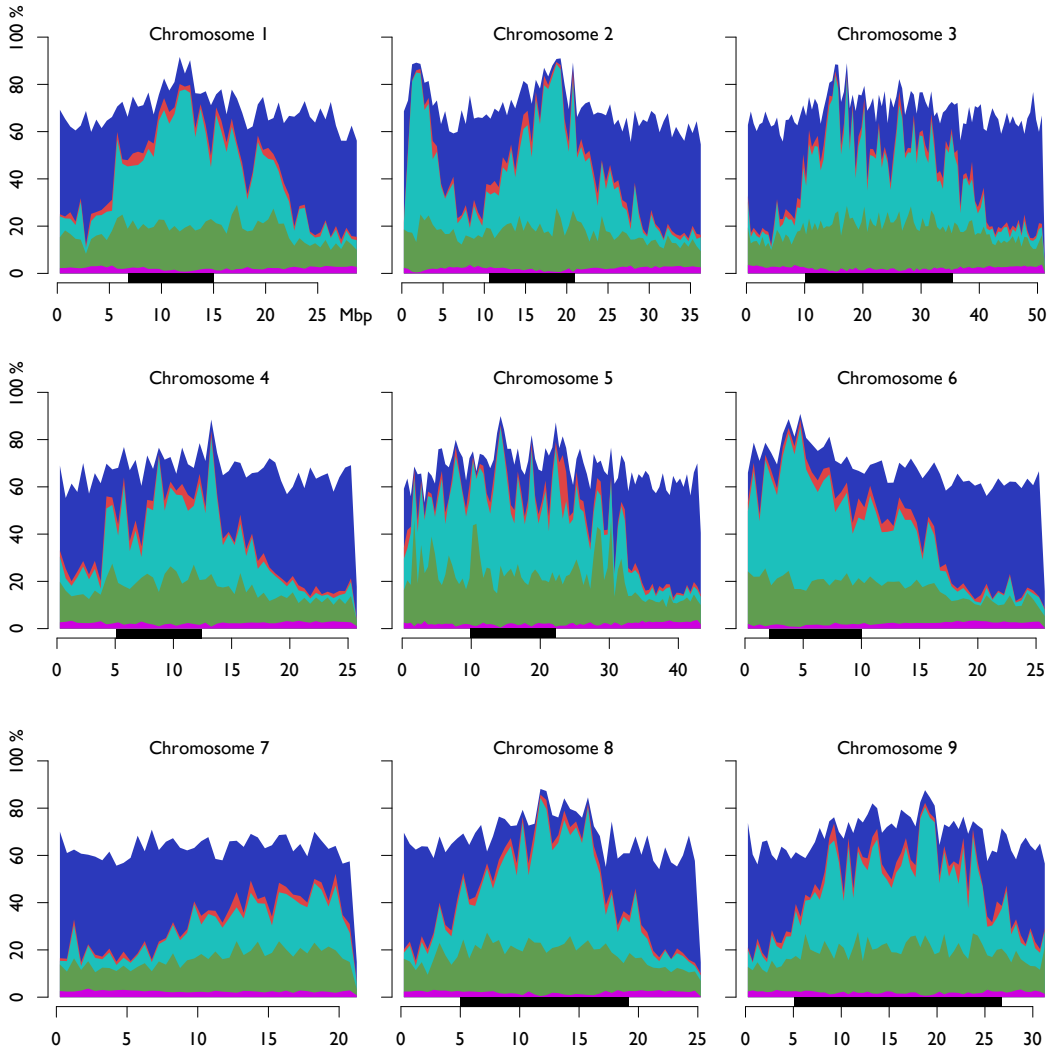
Supplementary Figure 8: Correspondence between genetic and physical map for Haploid Clementine Chromosome 8



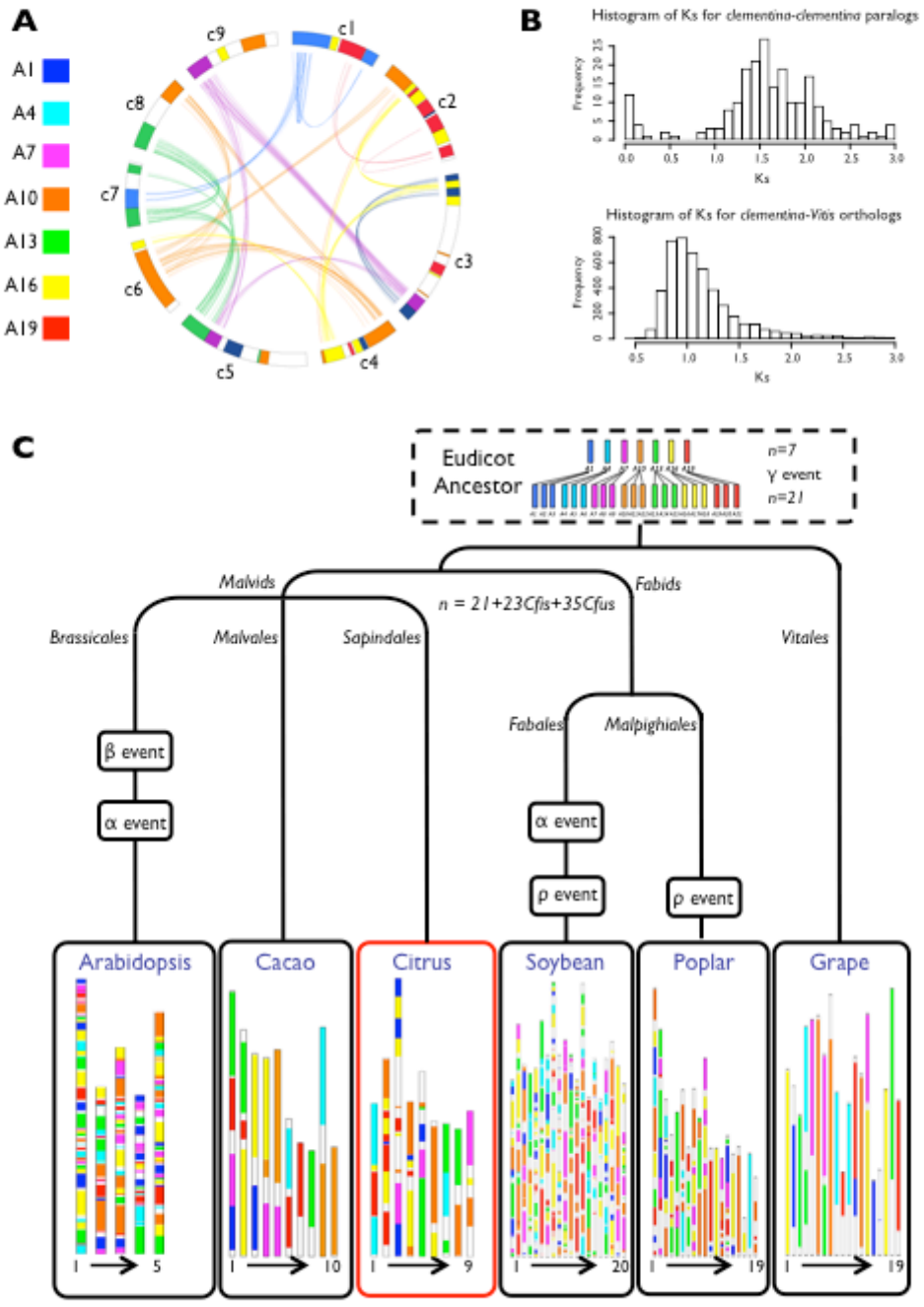
Supplementary Figure 9: Correspondence between genetic and physical map for Haploid Clementine Chromosome 9



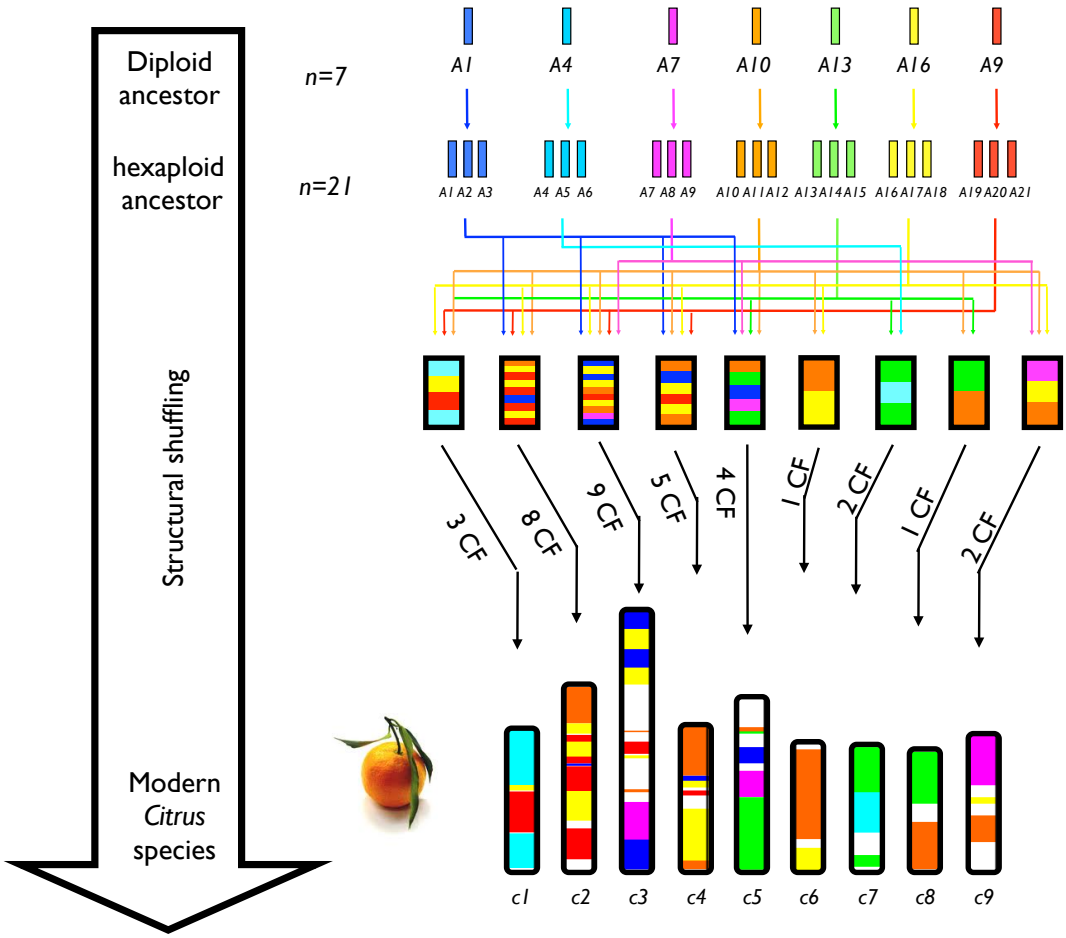
Supplementary Figure 10: Genomic landscape of Clementine.
The percent of the assembly that consists of protein-coding genes (blue), DNA transposons (red), Retrotransposons (cyan), uncharacterized complex repeats (green) and simple repeats (magenta) is plotted for non-overlapping 500kb windows. Regions of approximately zero genetic map distance, presumably corresponding to centromeres are indicated by black bars.



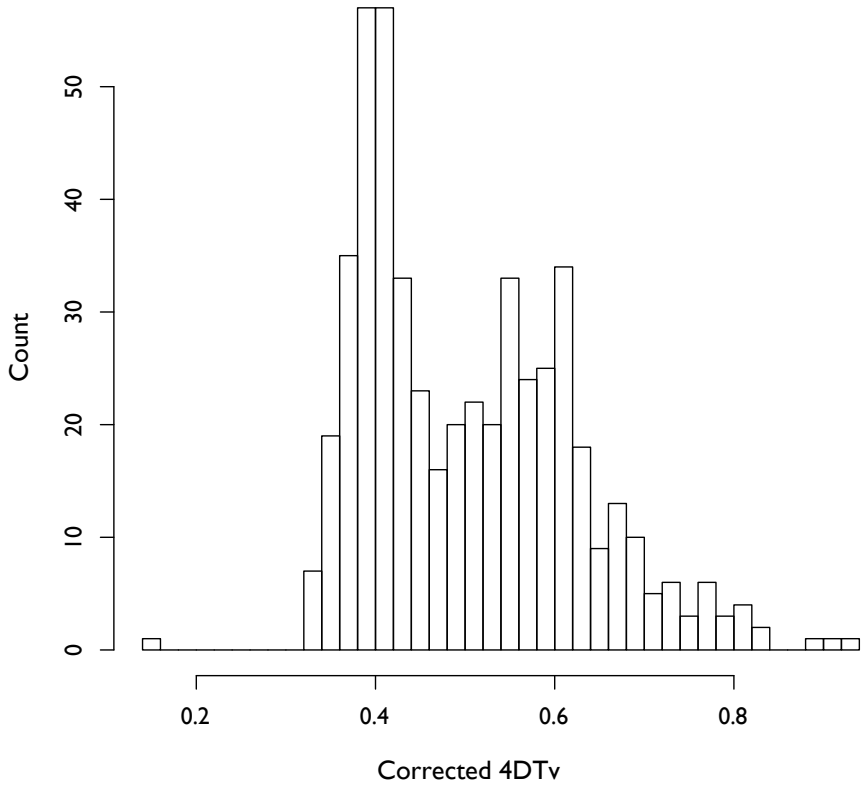
Supplementary Figure 11: Citrus genome synteny and duplication pattern and evolutionary history
(A) Schematic representation of the paralogous pairs identified within the citrus (c1 to c9) genome. Each line represents a duplicated gene. The different colours reflect the origin from the seven ancestral chromosomes (A1, A4, A7, A10, A13, A16, and A19). (B) K_s distributions (expressed in MYA on x-axis) of Citrus paralogs and Citrus-Grape orthologs. (C) Evolutionary scenario of the *Sapindales* family from the Eudicot ancestor.



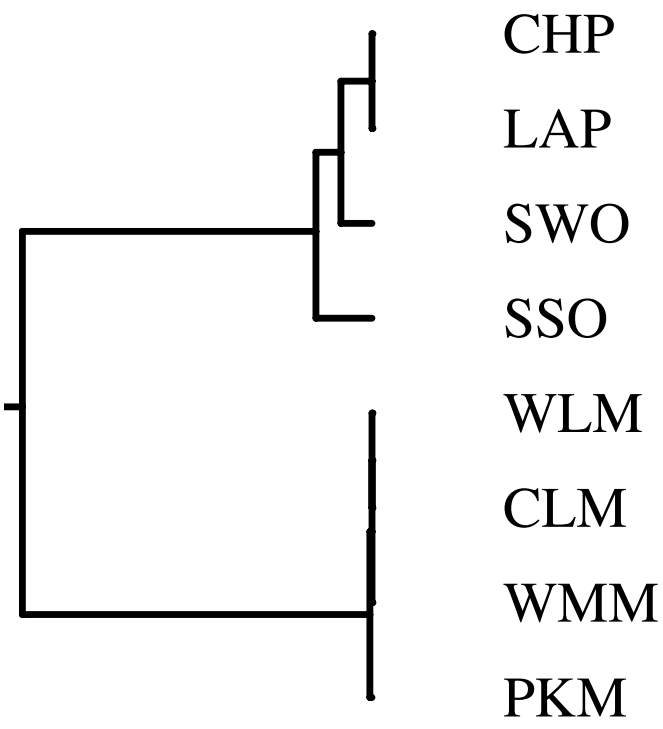
Supplementary Figure 12: Ancestry of citrus



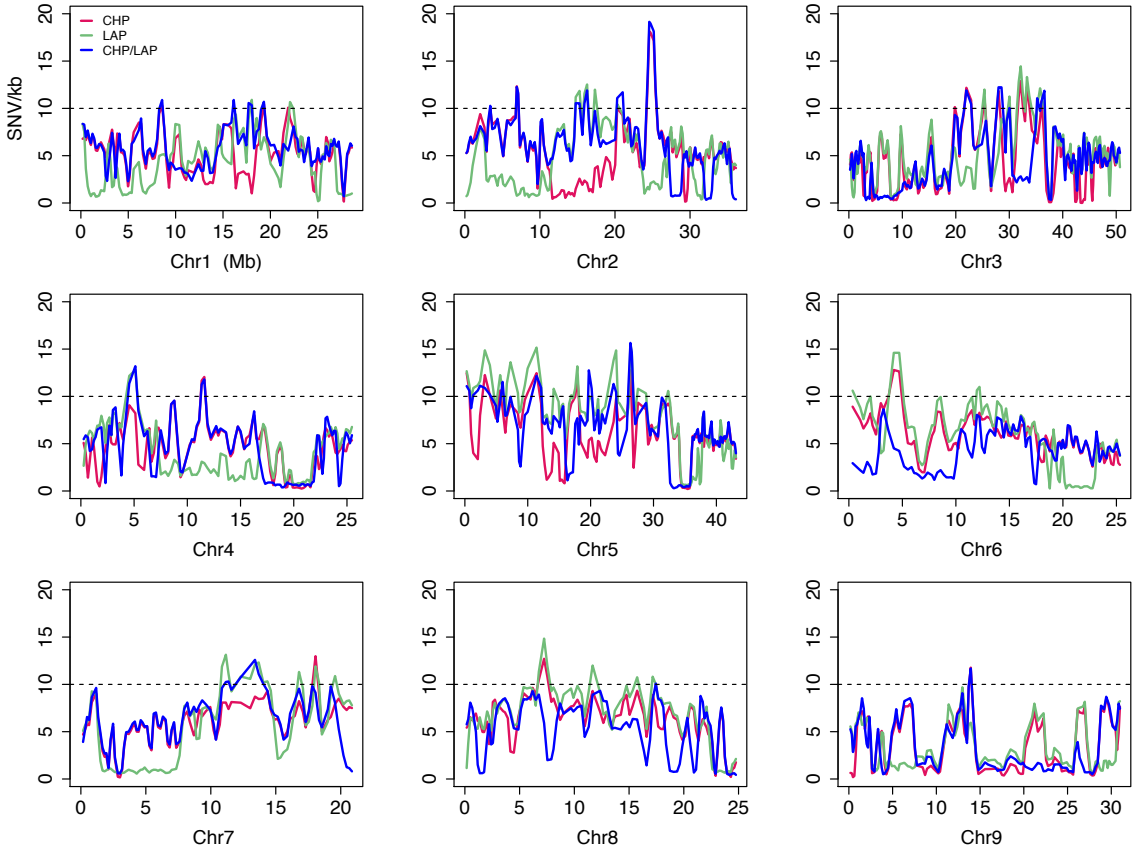
Supplementary Figure 13: Histogram of 4DTv between *C.* × *clementina* and grapevine.



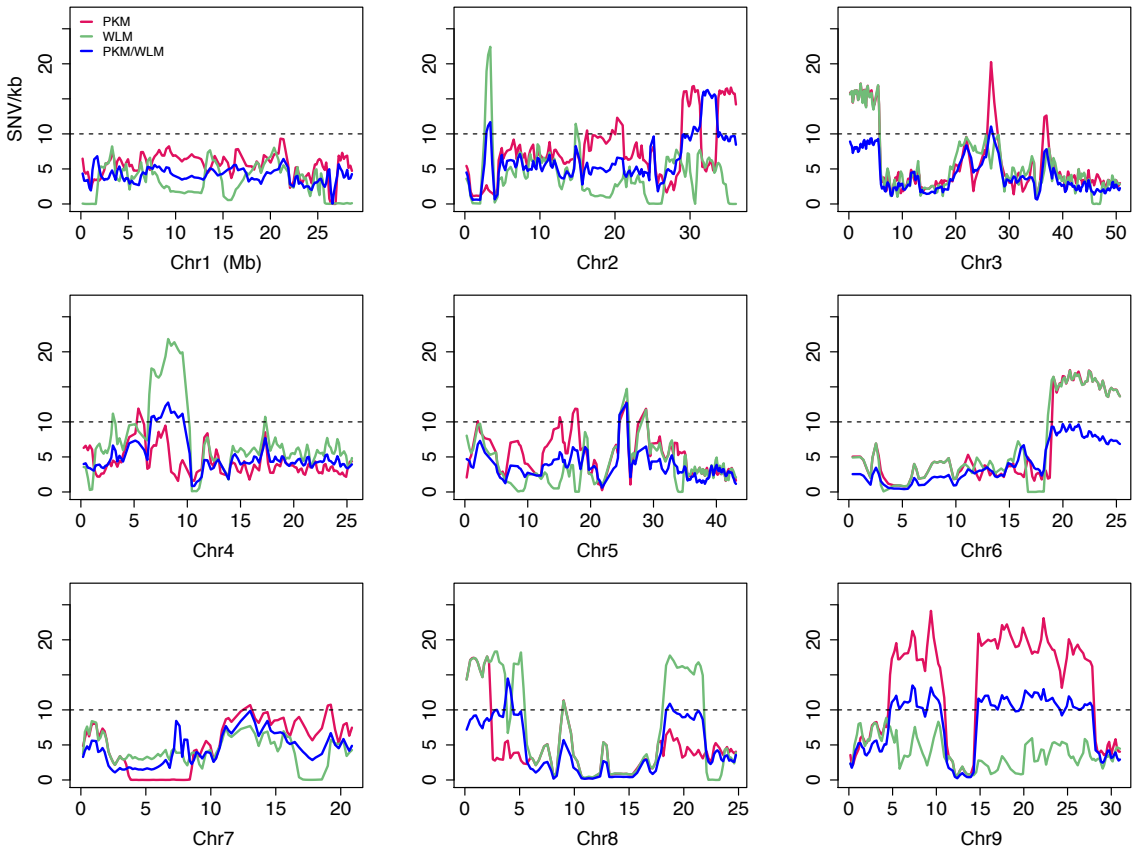
Supplementary Figure 14: Tree of chloroplast genome sequences
A neighbor-joining tree was generated from a distance matrix of pairwise mismatches among the 8 citrus cpDNA sequences. The deep split of the two clusters corresponds to two highly-diverged ancestral species of *C. maxima* and *C. reticulata*.



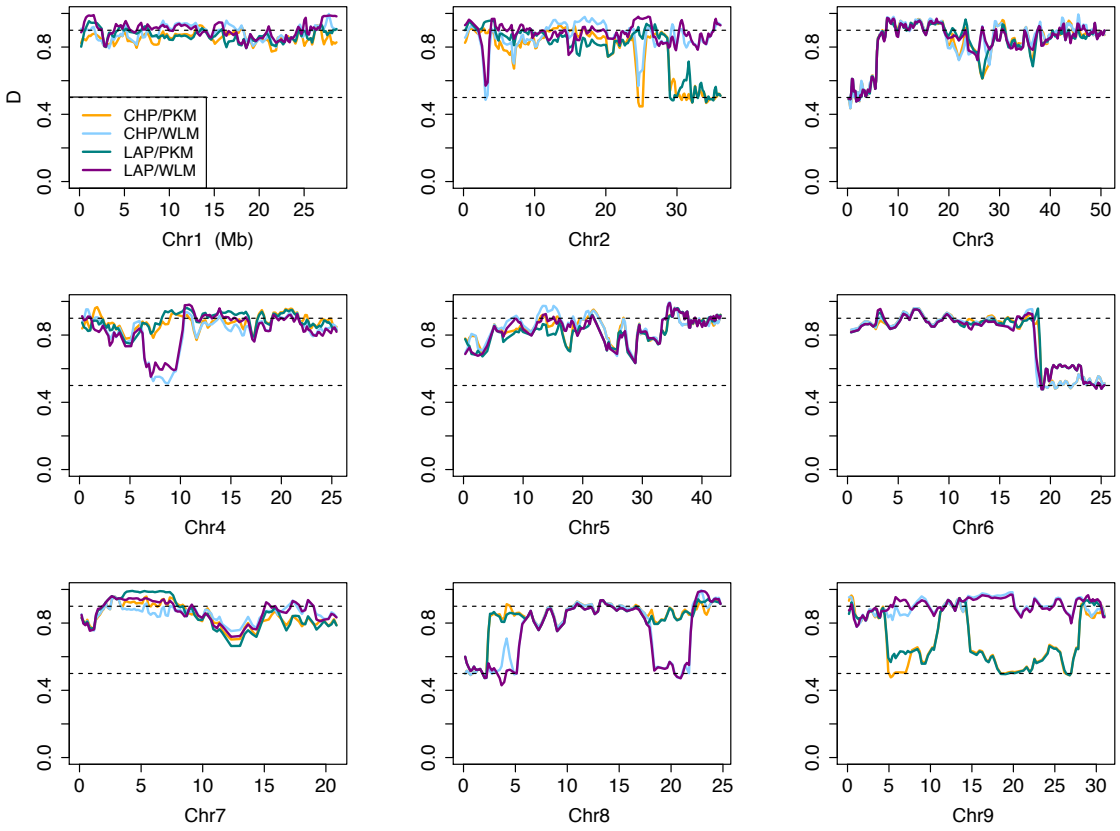
Supplementary Figure 15: Nucleotide diversity in pummelos.
Nucleotide diversity within and between two pummelos (200kb windows; 100kb step size). Low acid pummelo heterozygosity (LAP), Chandler pummelo heterozygosity (CHP), and the inferred heterozygosity between the non-shared haplotypes of LAP and CHP (LAP/CHP) are plotted. In computing nucleotide diversity between LAP and CHP, we have taken into account the fact that LAP and CHP share one haploid sequence because LAP is a parent of CHP. The dashed horizontal line marks 1% heterozygosity (10 SNVs/kb).



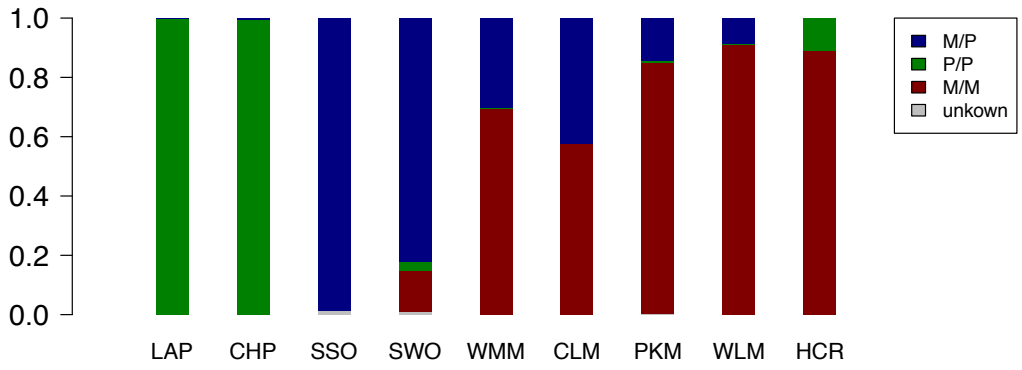
Supplementary Figure 16: Nucleotide diversity in mandarins.
Nucleotide diversity within and between Ponkan mandarin (PKM) and Willowleaf mandarin (WLM) (200kb windows; 100kb step size). The between-PKM-and-WLM diversity (WLM/PKM) measures the probability that two randomly chosen alleles (one from PKM and one from WLM) are different. The dashed line marks 1% diversity (10 SNVs/kb).



Supplementary Figure 17: Pummelo-mandarin divergence.
Plotted in 200kb windows are distances between one of the two pummelos (LAP, CHP) and either WLM or PKM (see Supplementary Note 7 for definition of D). High D values (around the dashed line at 0.9) for most of the genome characterize two highly divergent species without admixture in those regions. Intermediate D values (around the dashed line at 0.5) could arise, for example, when one of the two genomes compared consists of inter-specific hybrid segments. Note the difference between PKM and WLM.

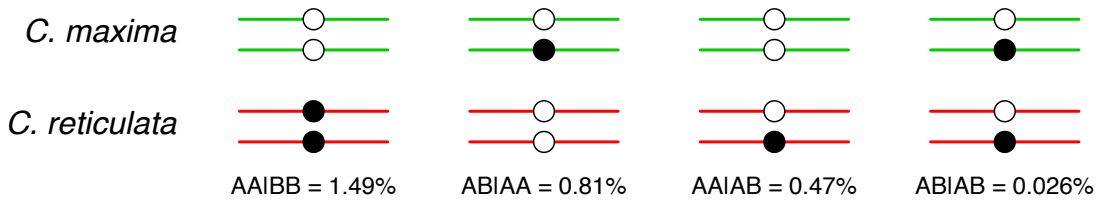


Supplementary Figure 18: Proportions of admixture in citrus.
The fraction of each genotype is based on segment sizes in Mb. The two ancestral species are denoted by M=*C.reticulata* and *P=C.maxima*.

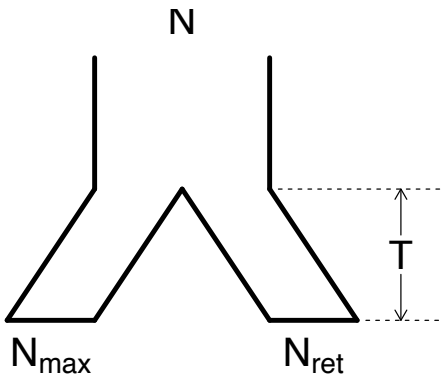


Supplementary Figure 19: Divergence of *C. maxima* and *C. reticulata*.
(**a**) The four joint genotype frequencies for a *C. maxima* and a *C. reticulata*. Numbers below the figure are based on the Low acid pummelo (LAP) and Willowleaf mandarin (WLM) genomes, to the exclusion of admixed regions. (**b**) The four-parameter "pants model" (where the parameters are population divergence time, T; the effective population sizes of the two extant populations N_{max} (*C. maxima*) and N_{ret} (*C. reticulata*) and the ancestral Citrus effective population size N) can be used to describe the divergence of two populations from a common ancestor by fitting the four joint genotype frequencies.

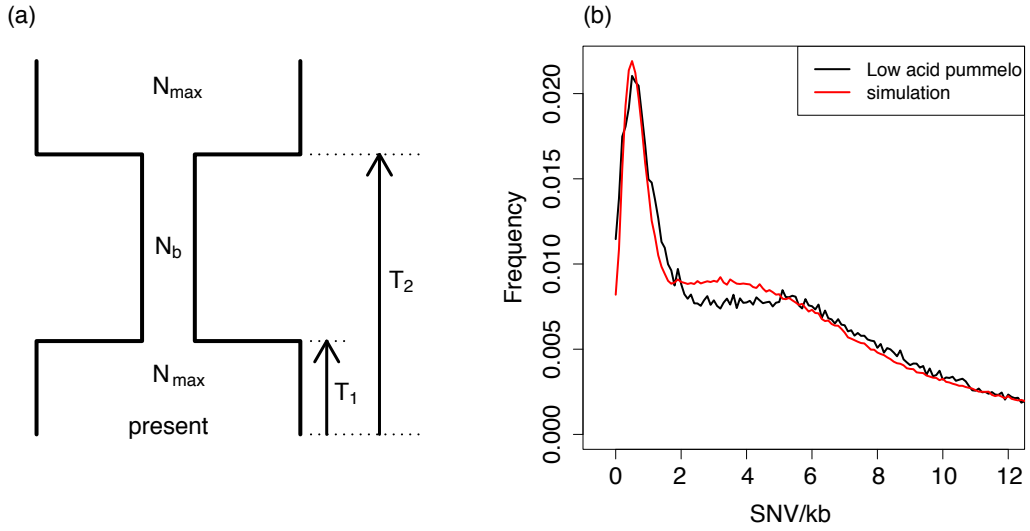
(**a**)



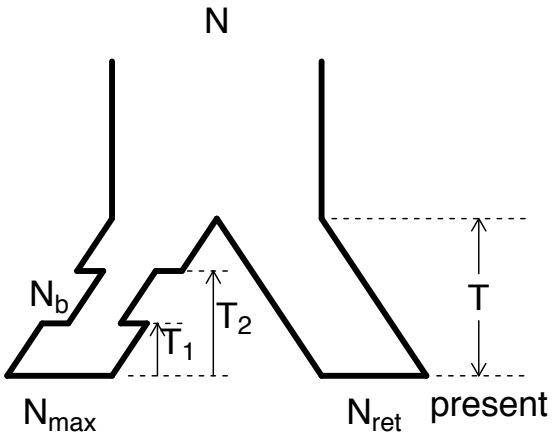
(**b**)



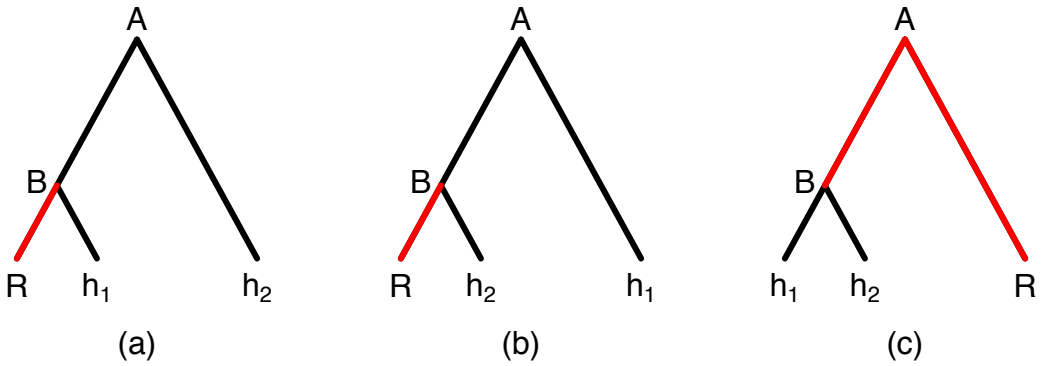
Supplementary Fig 20: A severe bottleneck in the ancient *C. maxima* population.
(a) A three-epoch model with piecewise constant population sizes for *C. maxima* (b) Model fit (simulation) to the heterozygosity spectrum of Low acid pummelo (LAP) in 10 kb windows. The pronounced peak around 1 heterozygous site per kb could be attributed to a strong bottleneck in the *C. maxima* population.



Supplementary Figure 21: A more realistic model for the divergence of *C. maxima* and *C. reticulata*.
In this model, the *C. reticulata* species maintained a constant effective population size N_{ret} since its divergence from *C. maxima*. However, the *C. maxima* population went through a severe bottleneck starting at T_2 and ending at T_1, with effective population sizes N_b and N_{max} during and after the bottleneck respectively.

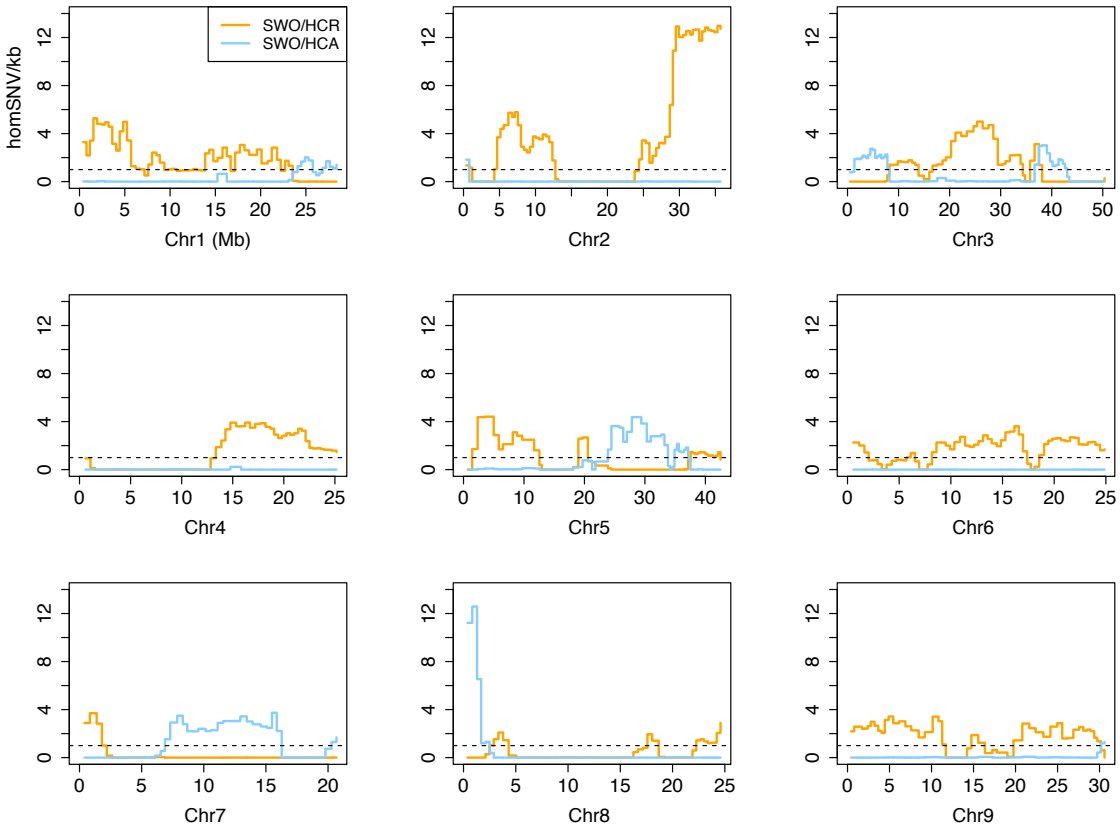


Supplementary Figure 22: Homozygous SNPs of a diploid with respect to a reference sequence.
The three possible genealogies of 3 orthologous sequences are shown: the haploid reference (R) and the two haplotypes (h_1, h_2) of a related diploid. Homozygous SNVs in the diploid with respect to the reference R arose through mutations along the red branches.

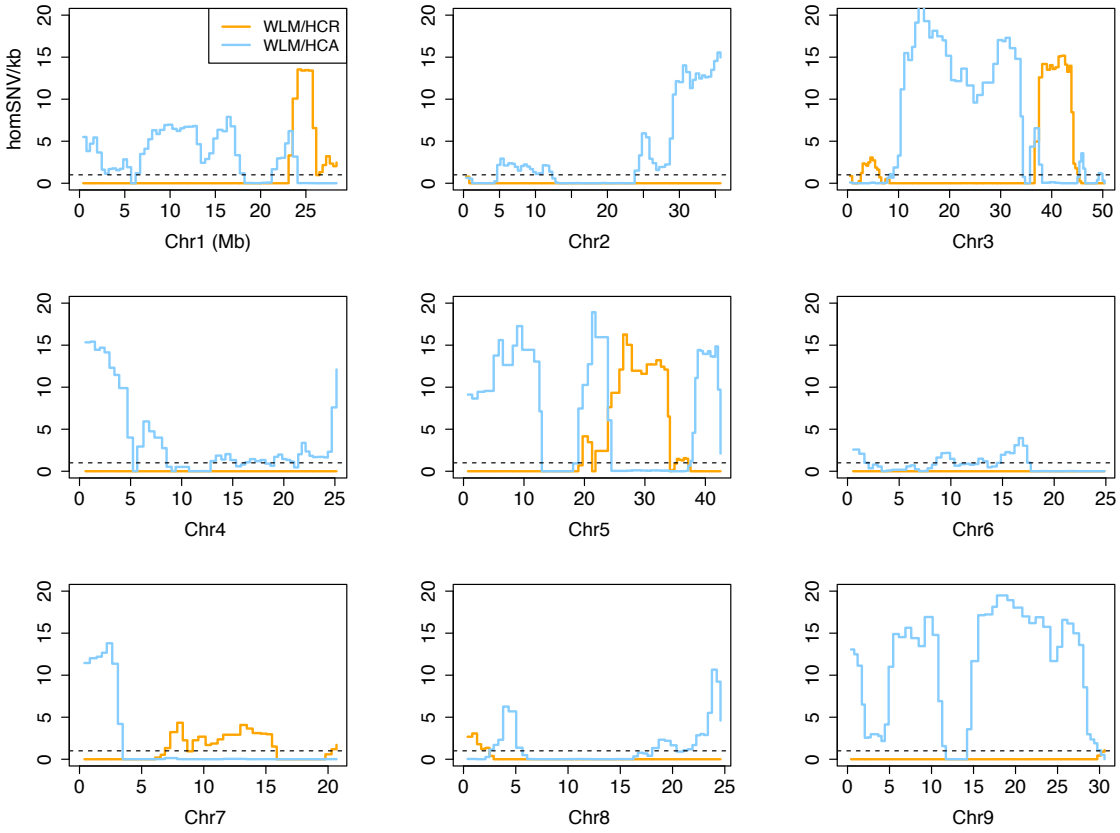


Supplementary Figure 23a: Homozygous SNV rates in sweet orange compared to Clementine

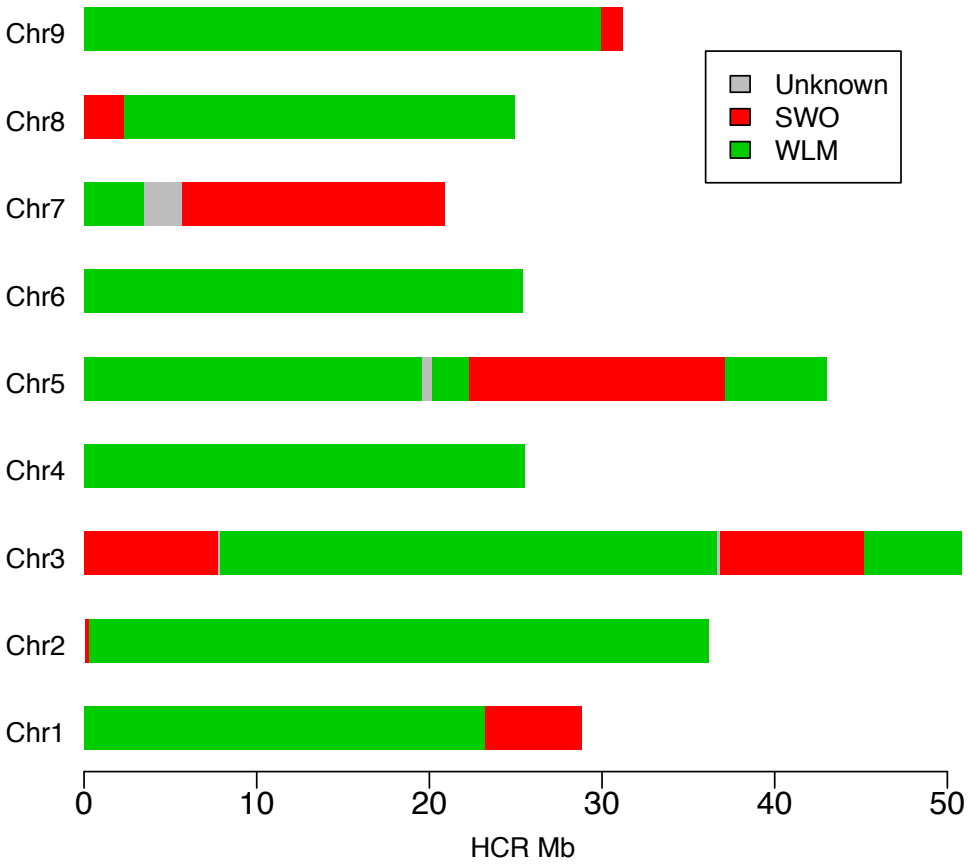
The rates of homozygous SNVs in sweet orange (SWO) relative to the haploid Clementine reference (HCR) and the complementary haploid Clementine (HCA) are plotted (500kb windows; step size 250kb).



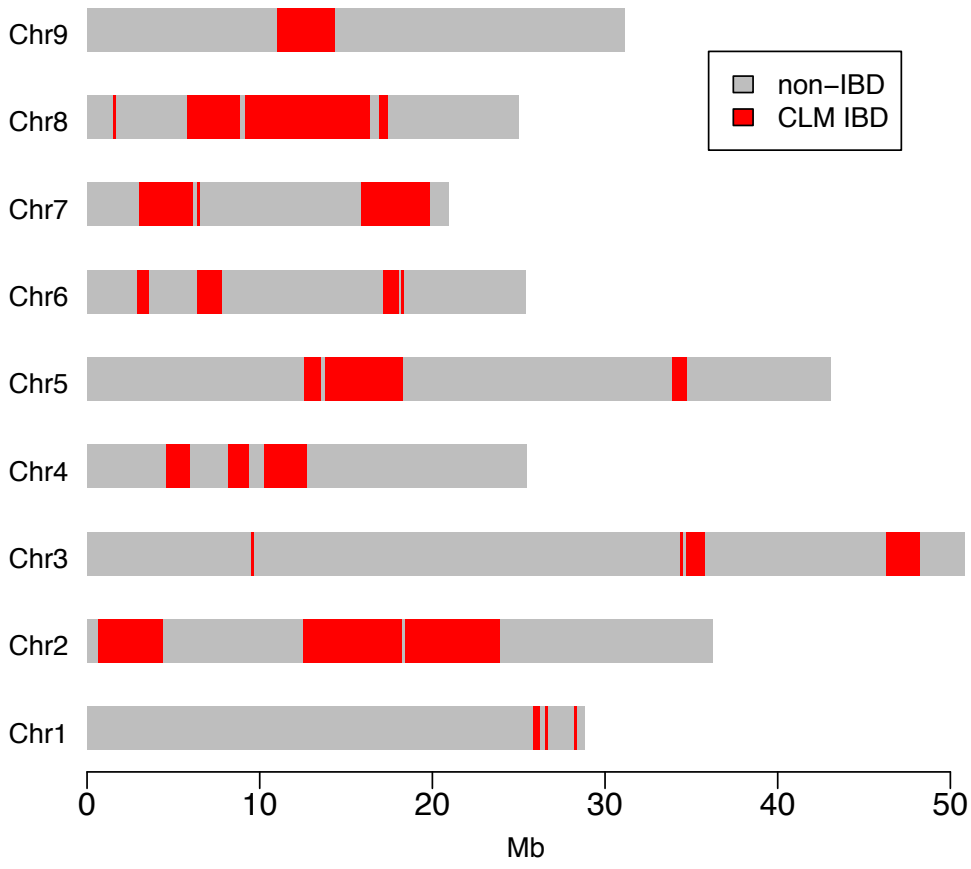
Supplementary Figure 23b: Homozygous SNV rates in mandarin compared to Clementine.
The rates of homozygous SNVs in Willowleaf mandarin (WLM) relative to the haploid Clementine reference (HCR) and the complementary haploid Clementine (HCA) are plotted (500kb windows; step size 250kb).



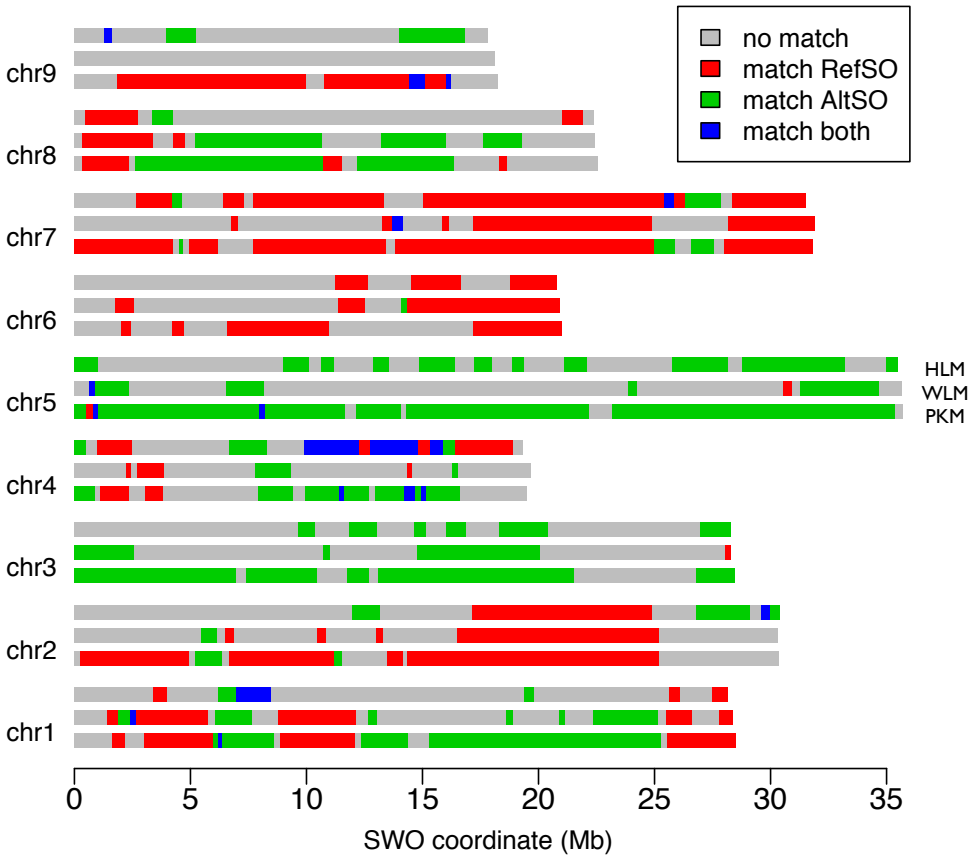
Supplementary Figure 23c: Parentage of haploid Clementine mandarin.
The parental origin of the haploid Clementine reference sequence (HCR) is shown based on a minimal-crossover inference model (see Supplementary Note 10.1). Haplotype sharing between HCR and SWO or WLM is inferred when the homozygous SNV rate in SWO or WLM with respect to HCR falls below 0.02% for a 100kb window, otherwise the HCR segment is labeled 'unknown'.



Supplementary Figure 24. The Clementine mandarin (CLM) genome shows a high degree of inbreeding.
Segments in which the nucleotide diversity between the two haploid sequences (calculated in 100kb windows) falls below 0.02% were considered to be identical-by-descent in the diploid CLM. These segments (red) make up 19% of the genome.

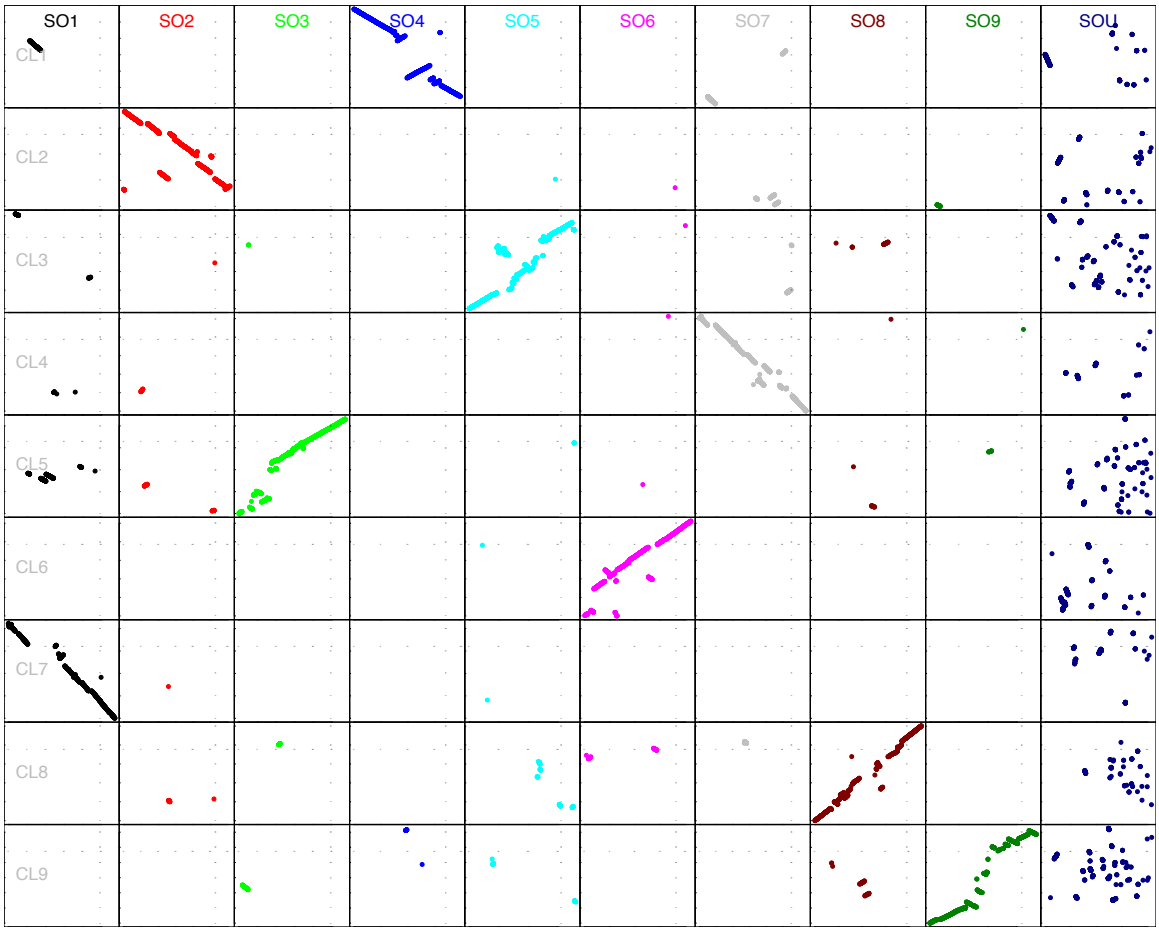


Supplementary Figure 25: Haplotype sharing between sweet orange and mandarins.
Haplotype sharing between sweet orange (SWO) and three traditional mandarins [Ponkan mandarin (PKM), Willowleaf (WLM) and Huanglingmiao (HLM)]. Haplotype sharing is calculated separately for the haploid sweet orange assembly [21] (RefSO, red) and the inferred, second haploid sequence of sweet orange (AltSO, green). Note that this figure uses chromosome nomenclature and orientation from Xu *et al.*[21]. For the translation between Xu *et al.* and our notation based on the published Citrus linkage map[16] see Supplementary Figure 26.

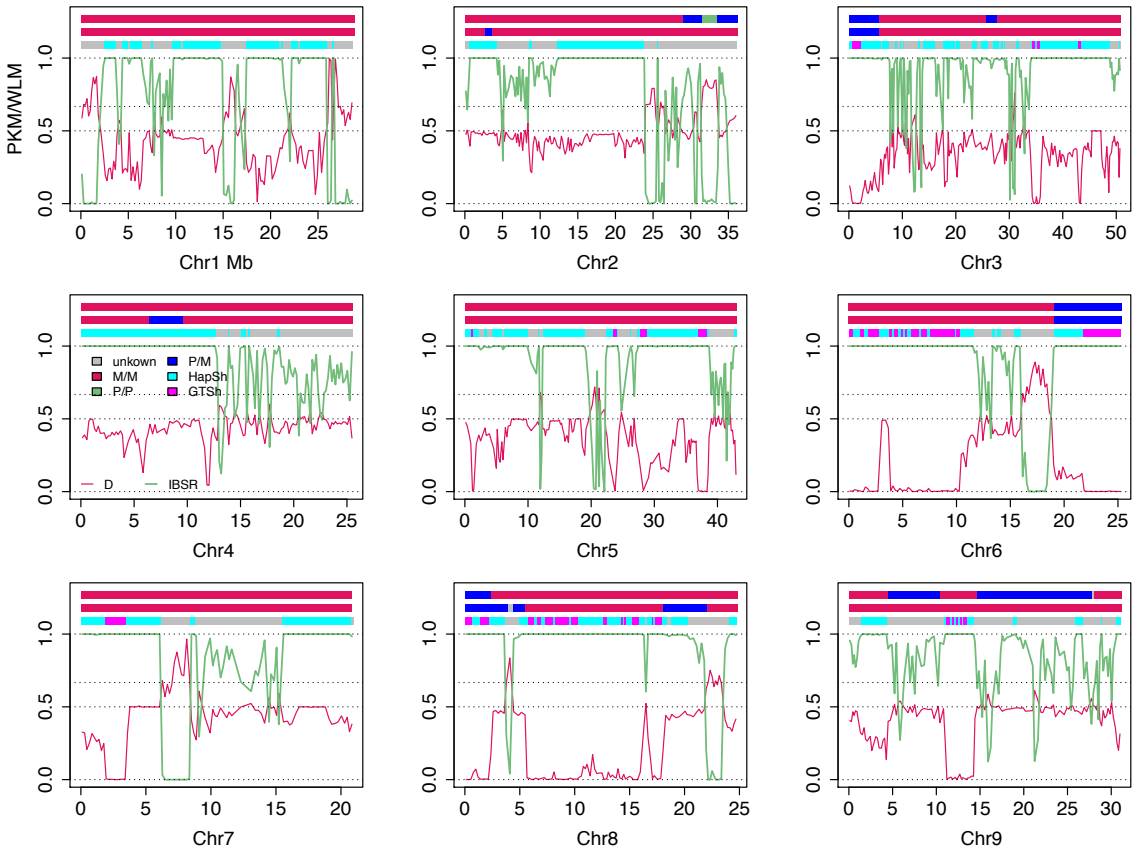


Supplementary Figure 26: Comparison of sweet orange and Clementine assemblies.

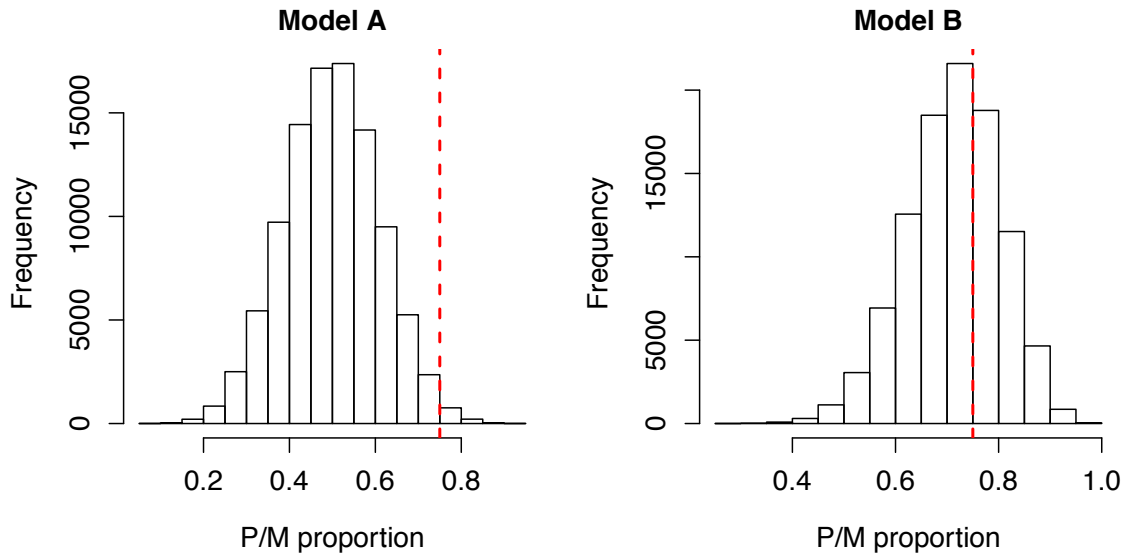
Correspondence of chromosomal regions between the haploid sweet orange assembly [21] and the haploid Clementine assembly. The 9 rows refer to the 9 chromosomes of the haploid Clementine reference (CL1-9). The 10 columns correspond to the assembled 9 chromosomes (SO1-9) and the unanchored scaffolds (SOU) of the haploid sweet orange. Dots in this final column correspond to sequences that are placed on to chromosomes in our Clementine assembly but are unmapped in Xu et al. Note also scattered discrepancies between the two assemblies.



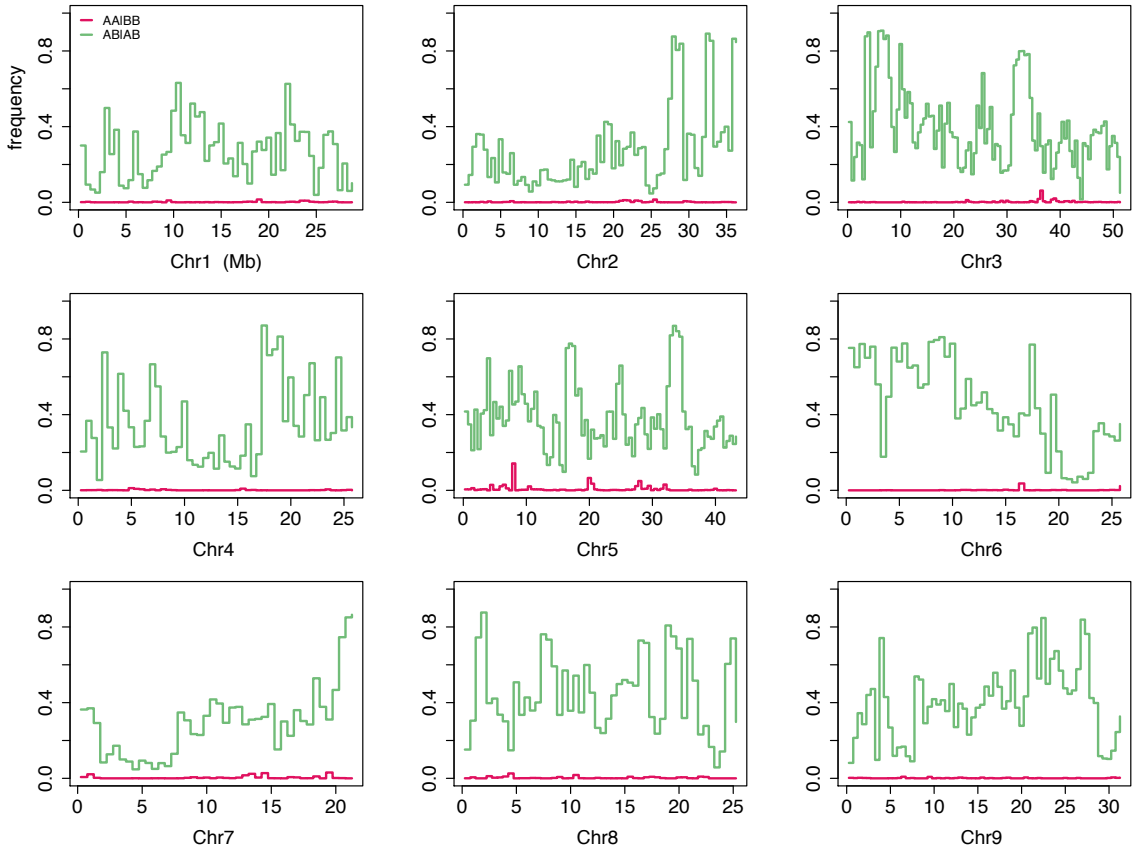
Supplementary Figure 27: Haplotype sharing between Ponkan and Willowleaf mandarins.
A matching haplotype is called for each 100 kb window when IBSR = IBS2/(IBS2+IBS0) > 0.99, except for hybrid *C.max*/*C.ret* regions where common shared haplotypes with sweet orange are used to estimate haplotype sharing between the two mandarins. The top and middle bars depict the admixture patterns for PKM and WLM respectively *(M=C. reticulata, P=C. maxima)*, and the lower bar shows haplotype sharing between PKM and WLM across the genome (cyan=shared haplotype, magenta=identical genotype, gray=unrelated). The red and green curves plot the distance D and IBSR values. The coordinates are relative to the haploid Clementine reference sequence. See Supplementary note 10.2.2 for details.



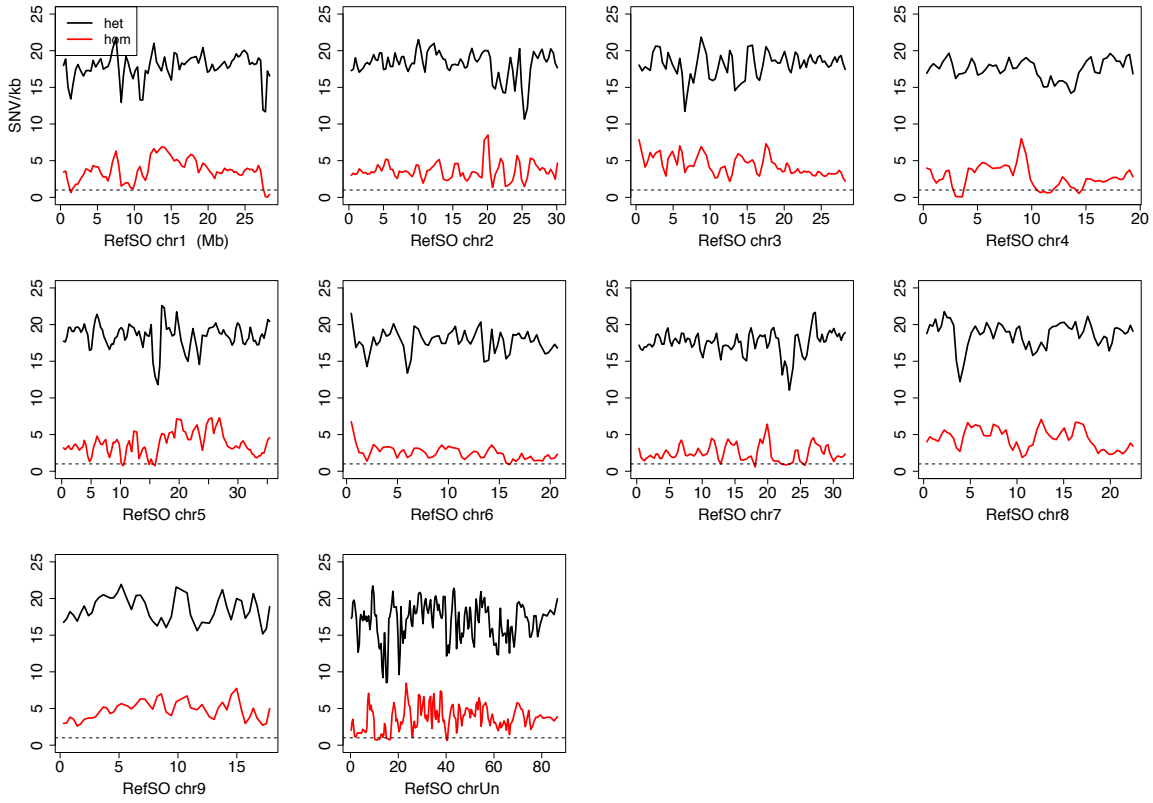
Supplementary Figure 28: Models of hybridization in sweet orange.
Two models for the hybrid P/M proportion in sweet orange are presented. Model A (left): SWO=(*C.max.* × *C.ret.*) × PKX (Female parent is an F1 hybrid) and Model B (right): SWO=((*C.max.* × *C.ret.*) × *C.max.*) × PKX (Female parent is a backcross). For both models, the distributions of inter-specific hybrid proportion in SWO based on simulations are shown, with a red dashed line indicating the observed value. Denoting one parent of SWO as an admixed mandarin (PKX), the other parent is highly constrained: model B is much more likely than model A. See Supplementary Note 10.3 for details. (M=*C. reticulata* P=*C. maxima*)



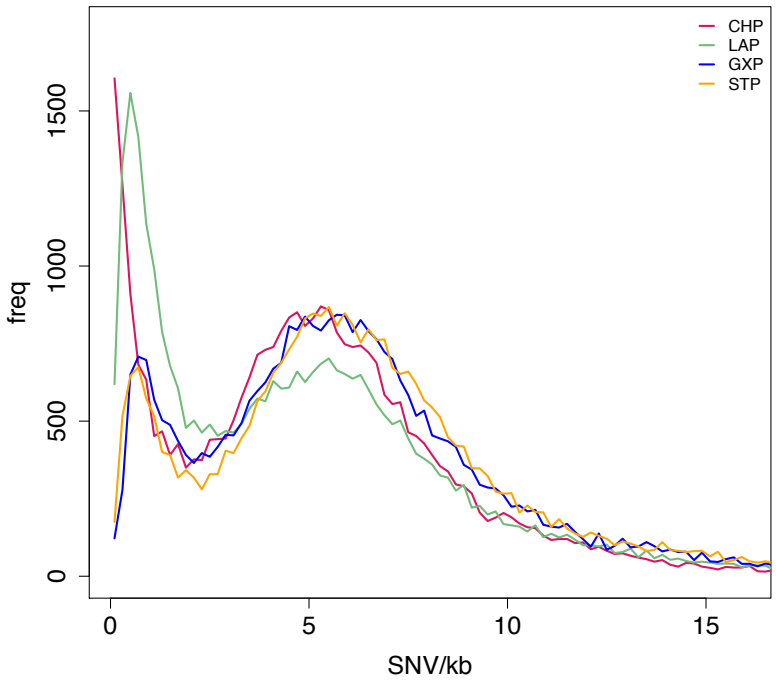
Supplementary Figure 29: Allele sharing between Low-acid pummelo (LAP) and Chandler pummelo (CHP).
Plotted in non-overlapping 500kb windows are frequencies of zero and two allele sharing per polymorphic site in the two pummelos. The LAP/CHP joint genotypes AA|BB and AB|AB correspond to zero and two allele sharing respectively. Data are consistent with parent/offspring relationship between LAP and CHP.



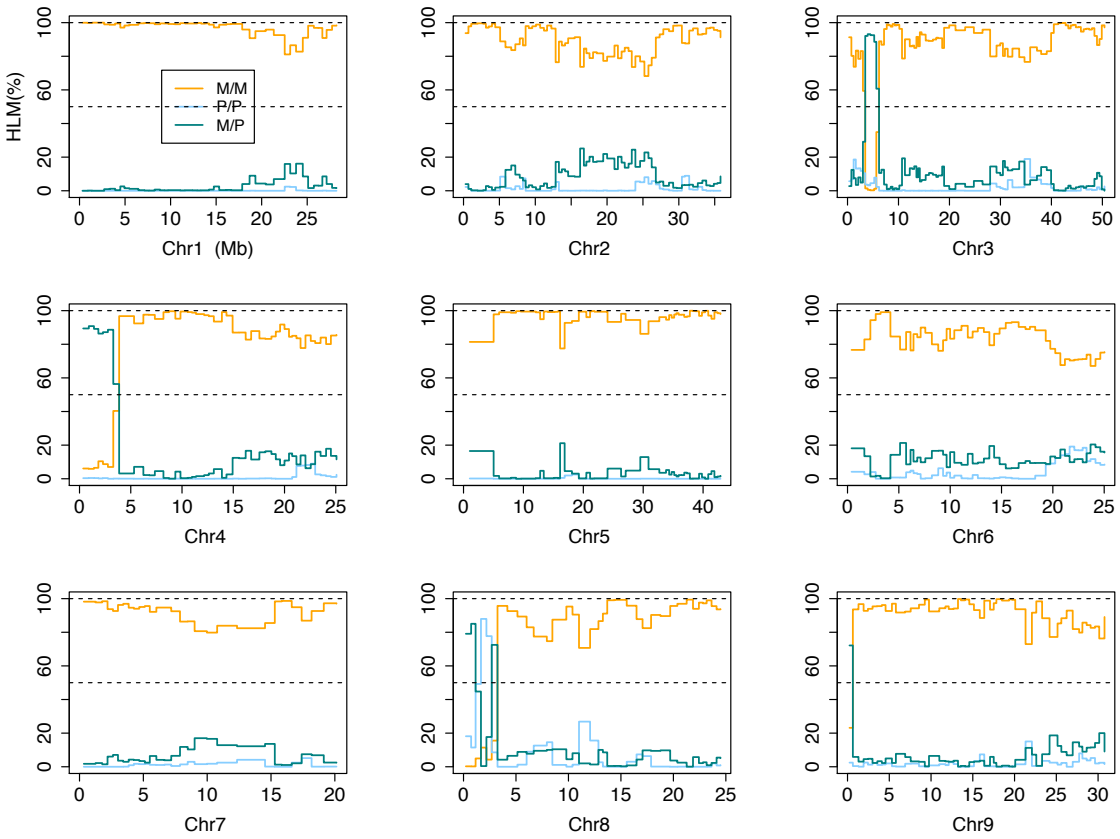
Supplementary Figure 30. Seville sour orange is not related to sweet orange. The nucleotide heterozygosity is plotted (100kb windows, 50 kb offset) for the Seville sour orange (SSO, black). This shows a characteristic interspecific *C. max./C. ret.* divergence of ~2% (~20 het sites/kb in SSO). The homozygous SNV rate (100 kb windows) with respect to the haploid sweet orange assembly (RefSO, red) is also shown. The typical homozygous SNV rate (~3-4 SNV/kb) indicates no genetic relatedness between SSO and sweet orange. The dashed line is at 1 SNV/kb. The chromosomes are numbered and oriented according to the assembly of Xu et al.[21].



Supplementary Figure 31: Nucleotide heterozygosity distribution in pummelos.
The histograms of nucleotide heterozygosity in two Chinese (GXP, STP) and two Siamese (LAP, CHP) pummelos are shown, based on common mapped sites in 10 kb sliding windows. Besides the common dominant peak around 5-6 SNV/kb, all four pummelos also share the secondary peak near 1 SNV/kb, consistent with a possible bottleneck in the ancestral *C. maxima* population prior to the separation of the Chinese and Thai pummelos.



Supplementary Figure 32: Admixture analysis of Huanglingmiao mandarin (HLM).
Along each chromosome the probabilities are shown of three genotypes in windows of 2,000 diagnostic SNPs distinguishing *C. maxima* and *C. reticulata*. The dashed line is at 50%. Note the pummelo admixture segments on chromosomes 3, 4, and 8. (M= *C. reticulata,* P= *C. maxima).*



References

1. Gmitter, F.G. & Hu, X. The possible role of Yunnan, China, in the origin of contemporary citrus species (Rutaceae). *Economic Botany* **44**, 267-277 (1990).

2. Gmitter, F.G., Soneji, J.R. & Rao, M.N. in Breeding Plantation Tree Crops: Temperate Species. (eds. S. Mohan Jain & P.M. Priyadarshan) 105-134 (Springer, New York; 2008).

3. Webber, H.J. in The Citrus Industry, Vol. 1. (eds. W. Reuther, H.J. Webber & L.D. Batchelor) 1-39 (University of California, Division of Agricultural Sciences, Berkeley; 1967).

4. Swingle, W.T. & Reece, H.C. in The Citrus Industry, Vol. 1, Edn. 2nd edition. (eds. W. Reuther, H.J. Webber & L.D. Batchelor) 190-430 (University of California Press, Berkeley; 1967).

5. Tanaka, T. Fundamental discussion of Citrus classification. *Studia Citrologica* **14**, 1-6 (1977).

6. Scora, R.W. On the history and origin of Citrus. *Bull. Torrey Botanical Club* **102**, 369-375 (1975).

7. Barrett, H.C. & Rhodes, A.M. A numerical taxonomic study of affinity relationships in cultivated citrus and its close relatives. *Syst. Biol.* **1**, 105-136 (1976).

8. Nicolosi, E. et al. Citrus phylogeny and genetic origin of important species as investigated by molecular markers. *TAG Theoretical and Applied Genetics* **100**, 1155-1166 (2000).

9. Gmitter, F.G. in Plant Breeding Reviews, Vol. 13. (ed. J. Janick) 345-363 (John Wiley & Sons, Inc., 1996).

10. Samaan, L.G. Studies on the origin of Clementine tangerine (Citrus reticulata Blanco). *Euphytica* **31**, 167-173 (1982).

11. Aleza, P. et al. Recovery and characterization of a Citrus clementina Hort. ex Tan. 'Clemenules' haploid plant selected to establish the reference whole Citrus genome sequence. *BMC Plant Biol* **9**, 110 (2009).

12. Saunt, J. in Citrus varieties of the world 71-72 (Sinclair International Ltd., Norwich, UK; 2000).

13. Hodgson, R.W. in The Citrus Industry, Vol. 1. (eds. W. Reuther, H.J. Webber & L.D. Batchelor) (University of California, Division of Agricultural Sciences., Berkeley; 1967).

14. Morton, J.F. Fruits of Warm Climates. (Florida Flair Books, Miami, Florida, USA; 1987).

15. Jaffe, D.B. et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* **13**, 91-96 (2003).

16. Ollitrault, P. et al. A reference genetic map of C. clementina hort. ex Tan.; citrus evolution inferences from comparative mapping. *BMC genomics* **13**, 593 (2012).

17. Schuler, G.D. Sequence mapping by electronic PCR. *Genome Res* **7**, 541-550 (1997).

18. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).

19. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).

20. Benson, D.A. et al. GenBank. *Nucleic acids research* **40**, D48-53 (2012).

21. Xu, Q. et al. The draft genome of sweet orange (Citrus sinensis). *Nat Genet* **45**, 59-66 (2013).

22. Goodstein, D.M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic acids research* **40**, D1178-1186 (2012).

23. Finn, R.D. et al. The Pfam protein families database. *Nucleic acids research* **36**, D281-288 (2008).

24. Mi, H. et al. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic acids research* **38**, D204-210 (2010).

25. Haas, B.J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research* **31**, 5654-5666 (2003).

26. Solovyev, V., Kosarev, P., Seledsov, I. & Vorobyev, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol* **7 Suppl 1**, S10 11-12 (2006).

27. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

28. Yeh, R.F., Lim, L.P. & Burge, C.B. Computational inference of homologous gene structures in the human genome. *Genome Res* **11**, 803-816 (2001).

29. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic acids research* **33**, 6494-6506 (2005).

30. Salse, J., Abrouk, M., Murat, F., Quraishi, U.M. & Feuillet, C. Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Brief Bioinform* **10**, 619-630 (2009).

31. Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-467 (2007).

32. Kumar, S., Tamura, K. & Nei, M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5**, 150-163 (2004).

33. Gaut, B.S., Morton, B.R., McCaig, B.C. & Clegg, M.T. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proc Natl Acad Sci U S A* **93**, 10274-10279 (1996).

34. SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. & Bennetzen, J.L. The paleontology of intergene retrotransposons of maize. *Nat Genet* **20**, 43-45 (1998).

35. Tuskan, G.A. et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313**, 1596-1604 (2006).

36. Buschiazzo, E., Ritland, C., Bohlmann, J. & Ritland, K. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evol Biol* **12**, 8 (2012).

37. Intiative, T.A.G. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).

38. Argout, X. et al. The genome of Theobroma cacao. *Nat Genet* **43**, 101-108 (2011).

39. Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178-183 (2010).

40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

41. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

42. Lynch, M. Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol Biol Evol* **25**, 2409-2419 (2008).

43. Terol, J., Naranjo, M.A., Ollitrault, P. & Talon, M. Development of genomic resources for Citrus clementina: characterization of three deep-coverage BAC libraries and analysis of 46,000 BAC end sequences. *BMC Genomics* **9**, 423 (2008).

44. Ollitrault, P. et al. SNP mining in C. clementina BAC end sequences; transferability in the Citrus genus (Rutaceae), phylogenetic inferences and perspectives for genetic mapping. *BMC Genomics* **13**, 13 (2012).

45. Sakamoto, W., Miyagishima, S.Y. & Jarvis, P. Chloroplast biogenesis: control of plastid development, protein import, division and inheritance. *Arabidopsis Book* **6**, e0110 (2008).

46. Abkenar, A.A., Isshiki, S. & Tashiro, Y. Maternal inheritance of chloroplast DNA in intergeneric sexual hybrids of true citrus fruit trees revealed by PCR-RFLP analysis. *J. Hort. Sci. and Biotech.* **79**, 360-363 (2004).

47. Bausher, M.G., Singh, N.D., Lee, S.B., Jansen, R.K. & Daniell, H. The complete chloroplast genome sequence of Citrus sinensis (L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. *BMC Plant Biol* **6**, 21 (2006).

48. Drouin, G., Daoud, H. & Xia, J. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Molecular phylogenetics and evolution* **49**, 827-831 (2008).

49. Hudson, R.R., Slatkin, M. & Maddison, W.P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583-589 (1992).

50. Keinan, A., Mullikin, J.C., Patterson, N. & Reich, D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* **39**, 1251-1255 (2007).

51. Chen, G.K., Marjoram, P. & Wall, J.D. Fast and flexible simulation of DNA sequence data. *Genome research* **19**, 136-142 (2009).

52. Crow, J.F. & Kimura, M. An introduction to population genetics theory. (Harper & Row, New York; 1970).

53. Blum, M.G. & Jakobsson, M. Deep divergences of human gene trees and models of human origins. *Mol Biol Evol* **28**, 889-898 (2011).

54. Akey, J.M. et al. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS biology* **2**, e286 (2004).

55. Voight, B.F. et al. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 18508-18513 (2005).

56. Pfeil, B.E. & Crisp, M.D. The age and biogeography of Citrus and the orange subfamily (Rutaceae: Aurantioideae) in Australasia and New Caledonia. *Am J Bot* **95**, 1621-1631 (2008).

57. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-575 (2007).

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

59. Thornton, T. et al. Estimating kinship in admixed populations. *American journal of human genetics* **91**, 122-138 (2012).

60. Deng, Z., Gentile, A., Nicolosi, E., Continella, G. & Tribulato, E. in Proc Int Soc Citricul, Vol. 2 849-854 (1996).

61. Barkley, N.A., Roose, M.L., Krueger, R.R. & Federici, C.T. Assessing genetic diversity and population structure in a citrus germplasm collection utilizing simple sequence repeat markers (SSRs). *Theor Appl Genet* **112**, 1519-1531 (2006).

62. Roose, M.L. unpublished SSR marker data. (2012).

63. Chen, C. et al. Verification of Mandarin and Pummelo Somatic Hybrids by Expressed Sequence Tag-Simple Sequence Repeat Marker Analysis. *Journal of the American Society for Horticultural Science* **133**, 794-800 (2008).

64. Lee, W.C. Testing the Genetic Relation Between Two Individuals Using a Panel of Frequency-unknown Single Nucleotide Polymorphisms. *Annals of Human Genetics* **67**, 618-619 (2003).

65. Moore, G.A. Oranges and lemons: clues to the taxonomy of Citrus from molecular markers. *Trends Genet* **17**, 536-540 (2001).

66. Garcia-Lor, A. et al. A nuclear phylogenetic analysis: SNPs, indels and SSRs deliver new insights into the relationships in the 'true citrus fruit trees' group (Citrinae, Rutaceae) and the origin of cultivated species. *Ann Bot* **111**, 1-19 (2013).

67. Cameron, J.W.a.S., R K Chandler – an early-ripening hybrid pummelo derived from a low-acid parent. *Hilgardia* **30**, 359-364 (1961).

68. Li, Y.Z., Cheng, Y.J., Yi, H.L. & Deng, X.X. Genetic diversity in mandarin landraces and wild mandarins from China based on nuclear and chloroplast simple sequence repeat markers. *J. Hort. Sci, Biotech.* **81**, 371-378 (2006).

69. Li, Y., Cheng, Y.J., Tao, N. & Deng, X. Phylogenetic Analysis of Mandarin Landraces, Wild Mandarins, and Related Species in China Using Nuclear LEAFY Second Intron and Plastid trnL-trnF Sequence. *J. Amer. Soc. Hort. Sci.* **132**, 796-806 (2007).

70. Li, W.B., He, S.W. & Liu, G.F. A study of citrus in Hunan province by analysis of leaf peroxidase isozyme. *Acta Hort. Sinica* **14**, 153-160 (1987).

71. Felsenstein, J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166 (1989).