doi:10.1038/nature25447

Table of Contents

Supplementary Notes **3**
Supplementary Note 1: Background information on the sequenced accessions **3**
Supplementary Note 1.1 Nomenclature used in this work 8
Supplementary Note 2. Sequencing data summary **8**
Supplementary Note 3. Analysis of species diversity **9**
Supplementary Note 3.1 Variant calls and heterozygosity 9
Supplementary Note 3.2 Runs of homozygosity (ROH) in mandarins 10
Supplementary Note 3.3. Chloroplast genome phylogeny 10
Supplementary Note 4. Identification of progenitor species **11**
Supplementary Note 4.1 Number of ancestral citrus species 11
Supplementary Note 4.2 Multidimensional scaling 13
Supplementary Note 5. Admixture analysis **13**
Supplementary Note 5.1 Local ancestry inference using species informative markers 13
Supplementary Note 5.2 Widespread pummelo admixture among mandarins 14
Supplementary Note 5.3 Oranges, grapefruit, lemon and limes 14
Supplementary Note 5.4 Admixture in Australian limes 15
Supplementary Note 6. Shared haplotypes revealed by inter-specific phasing **15**
Supplementary Note 6.1 Interspecific phasing in citrus 16
Supplementary Note 6.2 Pummelo admixture pattern divides the mandarins into three types 16
Supplementary Note 6.3 Admixture block size and introgression timing 17
Supplementary Note 6.4 Genetic origins of citrus hybrids 18
Supplementary Note 7 Genetic relatedness among citrus accessions **19**
Supplementary Note 7.1 Coefficient of relatedness calculation 19
Supplementary Note 7.2 Ponkan and Huanglingmiao/Kishu mandarins are the direct parents of Dancy and Satsuma respectively 20
Supplementary Note 7.3 Cocktail grapefruit, Wilking, and other mandarin accessions 21
Supplementary Note 7.4 Genetic relatedness network for mandarins, oranges and grapefruits 22
Supplementary Note 8 Nuclear genome phylogenetic reconstruction and citrus speciation dating **22**
Supplementary Note 8.1 Nuclear genome phylogeny and dating 23
Supplementary Note 8.2 Distinct epochs of speciation for Asian and Australian citrus 24
Supplementary Note 8.3 Comparison with chloroplast genome tree 24
Supplementary Note 9. The origin, biogeography, and dispersal of citrus **26**
Supplementary Note 9.1. Biogeography of *Citrus* and related genera. 26
Supplementary Note 9.2. Citrus dating and fossils 27
Supplementary Note 9.3. The center of origin of citrus 28
Supplementary Note 9.4. Citrus rapid radiation and monsoon weakening 29
Supplementary Note 9.5. Citrus dispersal 30
Supplementary Note 9.6. Origin of Australian citrus 30
Supplementary Note 9.7. Tachibana mandarin dispersal. 31
Supplementary Note 10 Pummelo admixture and citrus fruit size and acidity **31**
Supplementary Note 10.1 Pummelo admixture correlates with fruit size 32
Supplementary Note 10.2. Genome scan for citrus acidity/palatability association . 32

Supplementary Note 11 Acknowledgements **34**

Supplementary Tables **35**

Supplementary Table 1. Biogeographic distribution of the genus Citrus in Southeast Asia and Australia. 35

Supplementary Table 2. Accessions of the genus Citrus and related genera studied in this work. 38

Supplementary Table 3. Sequencing statistics of the 30 new genomes reported in this work. 43

Supplementary Table 4. Candidate SNPs associated with citrus acidity/palatability. 44

References **45**

Supplementary Notes

Supplementary Note 1: Background information on the sequenced accessions

There is a general agreement that citrus are native to Southeast Asia (**Supplementary Table 1**; **Extended Data Fig. 1a**). In this work, whole genome sequences from 58 citrus accessions with different geographical origins and two outgroup genera have been analyzed (**Supplementary Table 2**). Twelve of these genomes, including **Huanglingmiao mandarin** (HLM, *C. reticulata* Hort. ex Tanaka); **Ponkan mandarin** (Chinese honey orange, PKM, *C. reticulata* (Blanco, Swingle); **Willowleaf mandarin** (WLM, *C. deliciosa* Ten. Hort. ex Tanaka); **Clementine mandarin** (cv. Clementina de Nules, CLM, *C. x clementina* Hort. ex Tanaka; *C. x reticulata* Swingle); **W. Murcott mandarin** (WMM, *C. reticulata* Blanco); **Low acid pummelo** (Siamese Sweet, LAP, *C. maxima* [(Burm.) Merr], *C. grandis* Swingle, Tanaka); **Chandler pummelo** (CHP); **Guan-xi-mi-you pummelo** (GXP); **Sha-tian-you pummelo** (STP); **Sweet orange** (cv. Washington Navel, SWO, *C. x sinensis* L. [Osbeck]); **Sour orange** (cv. Sevillano, SSO, *C. x aurantium* L.); and **Mangshan wild mandarin** (CMS, *C. mangshanensis)*, were reanalyzed from previous published works[1,2]. Also re-analyzed are 19 recently published accessions[3] (excluding somatic mutants derived from the same base genome), including 15 mandarins, Cocktail grapefruit, Ambersweet orange, and two Chinese sour oranges (see **Supplementary Table 2**). Fourteen of the 19 accessions are of Chinese origin, and their names and physical traits (fruit size and acidity profile) are unavailable.

Listed below are descriptions of the 30 accessions sequenced in the current work, as well as 5 cultivars (Cocktail grapefruit, Ambersweet orange, Wilking, Fallglo, Kiyomi) developed in the United States and Japan sequenced by Wang et. al.[3]. Descriptions are based on our own observations as well as earlier reports[4,5]. The origin of the sequenced plants is presented in parenthesis (IVIA, Instituto Valenciano de Investigaciones Agrarias Citrus Germplasm Bank, Valencia, Spain; SRA, Station de Recherches Agronomiques de San Giuliano, Corse, France; UCR, University of California at Riverside Citrus Variety Collection; and FDACS/DPI, Florida Department of Agriculture and Consumer Services, Division of Plant Industry).

Sun Chu Sha Kat mandarin (SCM) (*C. reticulata* (Blanco), C. *reticulata* var. austera (Swingle), C. *erythrosa* (Tanaka)) is characterized by small flowers, small but narrow leaves and small fruits. These are broader than long, peel color may change from yellow to deep red and taste is acidic or acidic-sweet. It is used as rootstock (UCR-12A-25-12).

Tachibana mandarin (TBM) (*C. tachibana* (Mak.) Tanaka, *C. reticulata* (Blanco)) is thought to be native to Japan and surrounding islands. It develops easy peeling, small fruits of pale-yellow-orange color and acid flavor. Although taste is not completely unpleasant the fruit is not palatable. No commercial interest (UCR-12B-30-13).

Sunki mandarin (sour mandarin, suanju) (SNK) (*C. sunki* (Hayata, Hort. ex Tanaka, *C. reticulata* (Blanco)) produces easy peeling, very acidic small fruits, of an attractive orange color. Its fruits are not palatable and the plants are used as rootstocks (IVIA-239).

Cleopatra mandarin (CLP) (*C. reshni* (Hort. ex Tanaka), *C. reticulata* (Blanco)) is considered to be native to India. It produces unpalatable, small and very acidic fruits. It is widely used as a salt tolerant rootstock and also as an attractive ornamental because of the deep red color of the peel (SRA-948).

Changsha mandarin (CSM) (*Citrus reticulata* (Blanco)) produces small, juicy, puffy, brilliant orange-red and seedy fruit. The taste is sweet or acidic-sweet. The tree is rather tolerant to frost and yields heavy crops. It is also grown as an ornamental (UCR-12B-23-07).

Kishu mandarin (a.k.a. Kinokuni mandarin) (HSH) (*C. kinokuni* Hort. Ex Tanaka). The seeded form of this small tangerine grows in southern China and also in Japan, where it was introduced. We sequenced the seedless mutant known in Japan as Mukakukishu; sweet, juicy, and easy to peel, it is appreciated because of its pleasant taste and wonderful aroma. Whole genome sequence comparison shows that it has the same base genotype (*i.e.*, is a somatic mutant of) Huanglingmiao[1] mandarin (UCR-12B-39-13).

Satsuma (unshiu) mandarin, cv. Owari (UNS) (*C. unshiu* [(Mak.) Marc]; *C. reticulata* (Swingle)) is a commercial midseason, sterile and parthenocarpic, easy peeling mandarin. Satsumas are a group of commercial varieties with relatively high tolerance to low winter temperatures (IVIA-175)

Dancy mandarin, Dancy tangerine (DNC) (*C. tangerina* (Tanaka), *C. reticulata* (Swingle)) is an easy peeling commercial late harvesting variety of excellent color and good size and perdurability on the tree. Originated in 1867 from a chance seedling (IVIA-437).

King mandarin (KNG) (*C. nobilis* (Lour.), *C. reticulata* (Swingle)) is thought to be a natural tangor, *i.e.*, a hybrid between mandarin and orange, that originated in Vietnam. However this conventional wisdom is evidently wrong, as our whole genome sequence analysis shows that sweet orange is not a direct parent of King mandarin. Fruits have had much commercial interest since they are large in size, develop good flavor when ripe and are of late harvest (IVIA-477).

Rangpur lime (LMA) (*C. x limonia* (Osbeck)) produces non-commercial small and very acidic fruits of orange color. It is mainly used as both rootstock and ornamental plant (SRA-777).

Red rough lemon (RRL) (*C x jambhiri* (Lush)) probably originated in the Himalayan foothills in India. It was thought to be a natural hybrid between citron and lemon. However, we find by whole genome sequence comparison that it originated from an F1 cross *C. reticulata* x *C. medica*. Fruit is acidic, of medium size, with the surface typically deeply pitted and a lemon-yellow to brownish-orange color. It has been used as a rootstock (FDACS/DPI Budwood Registration Bureau ID# 08103.03).

Grapefruit, cv. Marsh (PAR) (*C. x paradisi* (Macfadyen)), one of the most extended varieties of grapefruit, originated as a chance seedling around 1860 in Lakeland, Florida. It is a late-ripening, self-incompatible variety that shows long tree storage capability and very good behavior during postharvest (IVIA-176).

Lemon, cv. Eureka (LIM) (*C. x limon* L. (Burm. f.)) is one of the most important commercial varieties around the world. Produces acid fruit throughout the year and has few thorns (SRA-4).

Humpang citron (HUM) (*C. medica* L.) fruit is large, oblong or oval, of green color when growing but generally yellow when ripe. The surface usually is smooth, the rind and the albedo are very thick and the segments are filled with acidic pale greenish pulp-vesicles. Citrons were the first citrus fruit to reach the Mediterranean region and are cold sensitive, monoembryonic, unpalatable and very fragrant (SRA-722).

Mac Veu citron (VEU) (*C. medica* L., *C. lumia* Risso & Poit). Similar to Humpang citron (SRA-760)

Corsican citron, (COR) (*C. medica* L.) is an acidless citron of unknown origin (SRA-613).

Buddha's hand citron var. Sarcodactylus (BUD) (*C. medica* L. (Noot.) Swingle) produces a very characteristic fruit usually without pulp and split into a number of finger-like sections. This fingered citron is well-regarded because of its fragrance for perfuming rooms and clothing. It is also grown as a dwarf plant for ornamental purposes (SRA-640).

Australian desert lime (ADR) (*Eremocitrus glauca* (Lindl.) Swingle, *C. glauca* (Lindl.) Burkill) is native to Australia and produces fruits of sour taste that can be used as condiment. It is drought tolerant and has very few soil requirements (UCR-12B-38-01).

Eremorange, Australian desert lime hybrid (ADL) (*Eremocitrus glauca x Citrus sinensis*) (SRA-871)

Australian finger lime (AFR) (*Microcitrus australasica* (F. Muell.) Swingle, *C. australasica* F. Muell), native to Australia, develops elongated finger-shaped fruits of different colors. Juice vesicles that can be broken down and separate very easily are of sharp acid flavor. It is used as a food seasoning (UCR-18B-16-04)

Australian finger lime (AFL) is an accession that we find has Australian round lime admixture. BC2 backcross. (SRA-1002)

Australian round lime (ARR) (*Microcitrus australis*, (Swingle), *C. australis* (A. Cunn. ex Mudie)) native to Australia produces rounded green fruit although at full maturity they become yellow. The pulp has low cohesive juice vesicles as the Australian finger lime. It is used as a food seasoning (UCR-18A-32-01).

Australian round lime (ARL). As above (IVIA-313).

Kumquat, Nagami (FOR) (*Fortunella margarita* (Lour.) Swingle) produces small and elliptical orange fruits that are mostly used as a food seasoning. Since trees are small and show slow, cold-tolerant growth it is also used as an ornamental. It produces fertile hybrids when crossed with species of the genus *Citrus* (IVIA-38).

Calamondin (CAL) (*C. madurensis* (Lour.)) that grows in China and the Philippines, produces very small and sour fruits without commercial relevance and the plant use is primarily ornamental, except in some cultures where they are widely used as a condiment (IVIA-135).

Mexican lime (MXL) (*C. aurantifolia* (Christm.) Swingle) is native of the Indo-Malayan region and our analysis confirms that it is a natural hybrid between micrantha and citron. Trees are very sensitive to cold and fruits are small, of a greenish-yellow color, with high acidity, much juice and a very distinctive aroma. It is used as a food seasoning (SRA-140).

Micrantha, Biasong (MIC) (*C. micrantha* (Wester)) it is thought to be native of the Southeast of the Philippines. It produces small, bitter and inedible fruit with a skin comparatively thick and broadly winged leaves (SRA-1114).

Ichang papeda (ICH) (*C. ichangensis* (Swingle) produces inedible fruits that release aroma reminiscent of lemons. This species is mainly used as rootstock because of its cold and drought tolerance characteristics (SRA-687).

Trifoliate orange, Poncirus Pomeroy, (PON) (*Poncirus trifoliata* (L.) Raf.) shows trifoliate leaves and deciduous behavior, two dominant characters that are not present in citrus. The tree also has high resistance to cold. Its fruit has no commercial value and the plant is commonly used as rootstock like its hybrids, especially the citranges, Carrizo and Troyer (SRA-1074).

Chinese box orange (SVR) (*Severinia buxifolia* (Poir.) Tenore) is native to China and grows as a compact tree or a small shrub. Among the trees related to citrus is the hardiest one. It produces small fruits that have no commercial value and it is used as an ornamental species (IVIA-147).

Ambersweet orange [SO5], [(An unnamed hybrid of Clementine mandarin x Orlando tangelo) x unnamed midseason sweet orange seedling], is a variety released by the USDA because of its resemblance to sweet orange, early maturity and deep flesh color. It is the only such hybrid ever legally designated as a "sweet orange", so that its juice could be used to blend with true sweet orange juice, according to juice industry regulations in Florida. All other known sweet oranges are derived only by somatic mutation, not by sexual hybridization, so Ambersweet is not a true sweet orange. (Sequence from Wang et. al.[3] with accession code A20).

Cocktail grapefruit [GF0] [Hybrid of Siamese Sweet pummelo x Frua mandarin], not a true grapefruit. It was developed by the University of California, Riverside in the mid-20th century. As is the case with sweet oranges, true grapefruit all are descended as somatic mutations from an original hybrid form resulting from a hybridization event between unknown pummelo and sweet orange parents, not by hybridization. (Sequence from Wang et. al.[3] with accession code 14J).

Fallglo [M21] [Hybrid of Bower mandarin (Clementine mandarin x Orlando tangelo) x Temple tangor, a presumed mandarin-sweet orange hybrid of unknown parentage], a seeded, early maturing and large fruited mandarin hybrid, developed by the USDA and produced primarily in Florida, USA. (Sequence from Wang et. al.[3] with accession code QH117).

Kiyomi [M20] [Hybrid of Miyagawa-wase satsuma mandarin x Trovita sweet orange], developed by the Okitsu Branch Fruit Research Station, now known as the Okitsu Citrus Research Station, National Institute of Fruit Tree Science. This is a large fruited juicy tangor, with aroma closely resembling sweet orange, and is seedless in the absence of cross pollination. It produces abundant monoembryonic (zygotic) seeds when cross pollinated and has been used as a scion breeding parent in Japan and elsewhere. (Sequence from Wang et. al.[3] with accession code KYM).

Wilking [M19] [Hybrid of King mandarin [KNG] x Willowleaf mandarin], developed by the University of California, Riverside in 1915. Fruit are small in size, quite fragrant and richly aromatic. Because it produces monoembryonic (zygotic) seeds, it has been used in breeding programs, but not grown commercially to any great extent. (Sequence from Wang et. al.[3] with accession code WLK).

Supplementary Note 1.1 Nomenclature used in this work

Sour oranges. We reserve the name "sour orange" (*C. aurantium*) to refer to the genome from which cultivar Seville and other somatic mutants are derived. It is the maternal parent of lemon (*C. limon*). The two sour oranges from South China[3] (accessions CBSC and ZGSC) represent two different genomes both unrelated to sour orange (*C. aurantium*).

Sweet orange. There is one true sweet orange (*C. sinensis*) from which many somatic mutants are derived, including Washington navel and blood orange. The Ambersweet orange is a mandarin x sweet orange hybrid, and not a true sweet orange, as noted above.

Grapefruits. The name "grapefruit" is used to refer to the true grapefruit (*C. paradisi*), which includes cultivar Marsh that we sequenced and other somatic mutants. It is a hybrid between a pummelo and sweet orange. The Cocktail grapefruit is not a true grapefruit, as noted above.

Lemons. We use "lemon" (*C. limon*) to refer to the cultivar Eureka that we sequenced and related somatic mutants. Its seed parent is sour orange and pollen parent is an unknown citron. Red rough lemon that we sequenced is a *C. reticulata x C. medica* hybrid, and is not a true lemon.

Supplementary Note 2. Sequencing data summary

Thirty citrus accessions were newly sequenced for this study. They came from four different sources – IVIA, SRA, UCR, and UF. The ID numbers for each accession are given above in **Supplementary Note 1** and the sequencing statistics are listed in **Supplementary Table 3**.

IVIA and SRA samples (22 accessions). Libraries were constructed using the Illumina TruSeq DNA Sample Prep standard protocol with some modifications. Briefly, 1 μg of high molecular weight genomic DNA was fragmented with a Covaris sonication device. Thereafter, DNA fragments were end-repaired and A-tailed. Adapters were then ligated via a 3′ thymine overhang. Finally, ligated fragments were amplified by PCR (10 cycles). Libraries insert sizes ranged from 400 to 500 bp. The library was applied to an Illumina flowcell for cluster generation. Sequencing was performed on a HiSeq2000 instrument using 100 bp paired-end reads. Primary analysis of the data included quality control on the Illumina RTA sequence analysis pipeline.

UCR samples (7 accessions). DNA was isolated from trees in the University of California Riverside (UCR) Citrus Variety Collection. DNA was prepared using a slightly modified CTAB protocol[6]. DNA libraries were prepared using the NEBNext Ultra Low Input kit in the UCR Institute for Integrative Genome Biology core facility (IIGB) with fragmentation by sonication or Covaris. Average insert sizes, including adapters, were about 325 bp. 100-bp paired ends were sequenced in 3-sample multiplex on an Illumina HiSeq2500 at IIGB.

UF/FDACS/DPI sample (red rough lemon). Unexpanded young leaves of red rough lemon were used for extracting nuclear DNA according to the methods of Carrier et al[7]. The purity and quantity of the DNA were determined using a Qubit 2.0 fluorometer (Thermo Fisher Scientific). Nuclear DNA was randomly sheared and DNA fragments about 500 bp were gel purified. Illumina pair-end DNA libraries were constructed according to the manufacturer's instructions (Illumina Inc.), and sequenced on Illumina HiSeq2000 in BGI Americas Corporation.

Supplementary Note 3. Analysis of species diversity

Supplementary Note 3.1 Variant calls and heterozygosity

For each accession, Illumina paired end reads are mapped to the haploid Clementine reference sequence[2] and the chloroplast genome sequence of sweet orange[8] using "bwa mem (v0.7.12-r1039)"[9], and PCR duplicates are removed using "Picard MarkDuplicates (v1.139)" [http://broadinstitute.github.io/picard/]. We used the HaplotypeCaller of GenomeAnalysisTK-3.4-46[10] to get a set of preliminary variant calls. Only bi-allelic SNPs are used in this analysis. A final set of high quality variant calls are obtained using the following filters:

Read mapping quality score >= 25, read base quality score >= 30, read depth between half and twice the genome wide average for each sample. As done previously[2], allele balance for heterozygous SNPs in each sample is achieved with a binomial filter to exclude 5% of calls in the tails of the binomial distribution with probability 0.5 for sampling the alternative allele.

The heterozygosity of the sequenced accessions is shown in **Fig. 1b** with representative accessions from each species listed in **Extended Data Table 1**. The citrus interspecific hybrids have heterozygosity ~1.5-2.2%, whereas the intraspecies variations are ~0.3-0.6%, except for citrons (nucleotide diversity ~0.1%).

Supplementary Note 3.2 Runs of homozygosity (ROH) in mandarins

Of the sequenced accessions, mandarins are a heterogeneous group with a wide range of nucleotide diversity due to varying degree of pummelo admixture. Some mandarins also show runs of homozygosity (ROH) as a result of haplotype sharing between the parents (**Extended Data Fig. 4a**). To determine the degree of inbreeding (*i.e.*, ROH) in mandarins, we used non-overlapping sliding windows of 200kb and assigned ROH to each window if the heterozygosity for that window is below $2x10^{-4}$. Satsuma, the Chinese accession BTJ (M16), Dancy, Fallglo, Clementine, Sunki and Kiyomi mandarins have the highest ROH proportions in their genomes.

The above cutoff on heterozygosity is chosen to allow for both SNP call errors and somatic mutations accumulated since the most recent common ancestor of the two haplotypes in ROH. This in turn, implies an upper bound on the false positive error rate for SNP calling of $2x10^{-4}$.

Supplementary Note 3.3. Chloroplast genome phylogeny

PhyML3.1[11] was used to reconstruct a maximum likelihood tree using the chloroplast genome sequences, based on the general time reversal model of nucleotide substitution and 200 bootstrap replicates.

The chloroplast genome phylogeny is shown in **Extended Data Fig.** 1b, using *Severinia* as an outgroup. The phylogenetic position of *Poncirus* is situated near the root of the tree, in agreement with a recent study with a somewhat smaller dataset[12].

Three super clades are manifest from the chloroplast tree. The first one includes Australian species and citrons; the second super clade consists of Ichang papeda, *C. mangshanensis* and mandarins, while the third one comprises micrantha and pummelos.

The *mandarin* clade consists of three subtypes, broadly corresponding to *C. tachibana* and the mainland Asian mandarins that can be further divided into acidic (CLP, SNK and SCM mandarins, RRL and LMA limes, and three pure mandarins of Chinese origin (M01, M02, M04) with unknown acidity profile) and non-acidic (21 mandarins and Ambersweet orange) subtypes. The *C. maxima* clade also contains 3 subtypes: i) sour orange (*C. x aurantium*), ii) sweet orange, and iii) pummelos, grapefruit (*C. x paradisi*), Cocktail grapefruit, and two Chinese sour oranges. By contrast, the 4 citrons of the *C. medica* clade share the same chloroplast subtype.

Supplementary Note 4. Identification of progenitor species

As shown in **Fig. 1b** and **Extended Data Table** 1, the genetic diversity of citrus at the species level is characterized by a much greater inter-specific sequence divergence (1.5-2.2%) than intraspecies variation (0.1-0.6%), with somewhat smaller interspecific divergences among the Australian limes (1.1-1.5%). The degree of divergence between two citrus species can be quantified by the genetic distance between two diploid genomes representative of the two species[2]:

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

where π_1 and π_2 are the nucleotide diversity (i.e. heterozygosity) of the two diploid genomes, and π_{12} is the sequence divergence between the two diploids (i.e. probability that two randomly chosen alleles from the two diploids are different). The value of D ranges from 0 to 1, with monozygotic twins having D=0 and two unrelated individuals from a panmictic population having D=0.5. D approaches 1 for two deeply divergent species[2].

Citrus accessions without inter-specific admixture can be identified based on a combined analysis of genetic distance D and nucleotide diversity, as shown previously[2]. We take Sun Chu Sha Kat mandarin (SCM) and low acid pummelo (LAP) as an example. Sliding window analysis shows that D ~ 0.9 across the genome, and that the nucleotide diversity of each diploid is characterized by intraspecies variation along the chromosomes without abrupt transitions between intraspecies variation and inter-species divergence. We thus conclude that SCM is a pure mandarin and LAP is a pure pummelo. By contrast, admixed accessions show various deviations from this pattern as detailed previously[2].

Supplementary Note 4.1 Number of ancestral citrus species

The genus *Citrus* includes an elusive number of species because its boundaries and species composition have been subjects of controversy during the past century. The long history of citrus cultivation has generated a number of botanical characters exhibiting a considerable degree of variability for this fruit crop. This results in numerous citrus fruit genotypes that are difficult to classify. The two basic taxonomic systems in citrus, proposed by Swingle[13] and Tanaka[14], are so different that they are considered to represent two extremes or visions of the same genus[15]. While Swingle, for instance, identifies three different species of mandarins, Tanaka claims 36. Between these two extremes, other propositions have been added with particular nomenclatures and classifications.

Comparative genomic analysis offers a powerful method to dissect the species composition of our set of samples, to validate the existing taxonomic assignments/systems, and to enumerate the progenitor species from which these samples are derived. Based on the distinct scales of the observed interspecies sequence divergence π_{12} (~1.1-2.2%) and intraspecies nucleotide diversity π (~0.1-

0.6%), we propose a simple rule of thumb for species delimitation: two diploid citrus genomes are from different species if their pairwise sequence divergence π_{12} and individual nucleotide diversity (*i.e.*, heterozygosity) π_i satisfies (1) $\pi_{12} > 1\%$ and (2) $\pi_{12} > \pi_1 + \pi_2$. Equivalently, the second condition can be cast in terms of the distance metric D:

$$D = 1 - 0.25 * (\pi_1 + \pi_2)/\pi_{12} > 1 - 0.25 = 0.75$$

When the two diploid genomes have similar nucleotide diversity, D > 0.75 implies an interspecies divergence that is more than twice the average intraspecies variation. Care must be taken to apply this rule only to diploid genomes after excluding any regions of possible interspecific admixture.

The above criteria for species delimitation reveal 10 progenitor species of citrus as well as two outgroup genera (*Poncirus* and *Severinia*) (**Fig. 1c** and **Extended Data Table** 1). These ten ancestral species include 7 Asian species and 3 Australian species. The seven Asian citrus species are *C. medica* (citrons), *C. maxima* (pummelos), *C. reticulata* (pure mandarins), *C. micrantha, C. ichangensis, Fortunella margarita (Nagami kumquat),* and *C. mangshanensis.* The three Australian species include *Eremocitrus glauca* (Australian desert lime), *Microcitrus australis* (Australian round lime) and *Microcitrus australasica* (Australian finger lime). For each species, at least one pure accession can be identified. Among the 28 sequenced mandarins in particular, only five (Tachibana, Sun Chu Sha Kat, and three unnamed Chinese accessions M01, M02, M04) show no inter-specific admixture.

Our species-level taxonomic assignment mostly comprises a subset of the listed citrus species by Swingle and Tanaka, with three major exceptions. First, we assign *C. mangshanensis* (a wild Mangshan "mandarin"[1] unknown to Swingle and Tanaka) a new species based on genome comparison with *C. maxima* and *C. reticulata*[2]. Second, most mandarins in our collection are not considered as distinct species (as proposed by Tanaka), but are instead described by pummelo admixture into a single common wild mandarin species, *C. reticulata* (**Supplementary Note 5**). Similarly, oranges, grapefruit, lemons and limes have admixed or hybrid genomes and are not assigned their own species. Third, whereas *Citrus tachibana* (TBM) was considered a citrus species by both Tanaka and Swingle, it fails the above "1%" criteria to be assigned a new species. In particular, nuclear genome comparison between TBM and the pure mandarin SCM shows that $\pi_{12} = 0.5\%$ and D=0.6, and suggests that TBM belongs to *C. reticulata*[16,17]. This is consistent with the more recent divergence between TBM and mainland Asian mandarins (**Supplementary Note 8**) and, together with its distinct chloroplast subtype (**Extended Data Fig. 1b**), suggests that it may be more useful to consider TBM as a subspecies of *C. reticulata* arising from allopatric isolation. Lastly, *Poncirus* differs from all known species of citrus in numerous striking characters including the presence of deciduous trifoliate leaves, and is assigned a different genus based on sequence divergence (**Extended Data Table 1**) and the nuclear genome phylogeny (**Fig. 1c and Extended Data Fig. 1c**). This is in line with the general point of view[4,12,18], but in contrast to some recent assignment[19].

Supplementary Note 4.2 Multidimensional scaling

Genetic clustering patterns of the sequenced accessions can be revealed by multidimensional scaling analysis based on pairwise genomic distances D as defined above. We used classical multidimensional scaling as implemented in the R programming language[20] (the cmdscale function) for this analysis. The projection onto the top two principal coordinates (**Fig. 1a**) shows that mandarins, pummelos, and citrons form three distinct clusters with oranges, grapefruit, lemon and limes situated at intermediate positions in accordance with their genetic makeup.

Supplementary Note 5. Admixture analysis

Genome-wide species informative markers (SIM) for the progenitor species can be derived using citrus accessions free of inter-specific admixture. As many cultivated citrus accessions are derived from the three principal species of *C. medica* (citrons), *C. maxima* (pummelos) and *C. reticulata* (pure mandarins), we obtain diagnostic SNPs for these three species using two pure mandarins (TBM and SCM), two citrons (Buddha's hand and Humpang), and three pummelos (Low Acid, Guanximiyou and Shatianyou pummelos). Diagnostic alleles for each species are selected from fixed differences between the target species and the other two species as represented by the 7 accessions. In this way, we obtain 301,817 diagnostic SNPs for *C. medica*, 116,803 for *C. maxima* and 169,963 for *C. reticulata*, with a total of 588,583 SIMs. For the three pummelos used to represent *C. maxima*, we allow up to one sample with missing genotype call. Of the sequenced accessions, 46 are derived from these three progenitor species and their hybrids. Note that species informative markers can also be obtained in the absence of pure samples (*i.e.*, free of inter-specific admixture) based on the patterns of nucleotide diversity and genetic distance D, as demonstrated previously[2].

Supplementary Note 5.1 Local ancestry inference using species informative markers

With this set of high density diagnostic SNPs, interspecific admixture segments in the 46 accessions can be detected using a sliding window of 1000 markers. For each accession, the local ancestry assignment in every window is determined as follows. First, the 1000 markers in the window are divided into three ancestral types corresponding to *C. medica* (C), *C. maxima* (P), and *C. reticulata* (M), respectively. Second, for each marker of a given ancestral type, the copy number (2, 1, or 0) of the marker allele in the target diploid genome is recorded. The allele frequency spectrum (n_2, n_1, n_0) for markers of each ancestral type can then be calculated. Third, the number of haplotypes of a given ancestry (*i.e.*, C, P, or M) in the target genome is inferred to be *argmax* n_i, i.e. the most frequent allele copy number (2, 1, or 0).

Finally, the local ancestry of the diploid genome is determined by the contributing haplotypes from all ancestral species.

As an example, consider the assignment of sweet orange ancestry for a window on chromosome 2 that contains 453, 246 and 301 C, P, and M-type markers respectively. For each C-type marker, sweet orange is homozygous alternate (*i.e.*, non-C-type), indicating that *C. medica* does not contribute to its local ancestry. Among the P-type markers, sweet orange is heterozygous for 242 and homozygous alternate for 4, and we thus infer the presence of one *C. maxima* haplotype. Lastly, of the 301 M-type markers, sweet orange is heterozygous for 300 and homozygous for 1 marker, implying the presence of a *C. reticulata* haplotype. Taken together, we infer that sweet orange is a hybrid P/M at this genomic window.

In the rare case when a window spans the boundary between two segments of different ancestry, the local ancestry assignment for the window can be ambiguous. To detect and quantify such transitional windows, we replace step 3 above by a more stringent condition as follows: assuming $n_i \geq n_j \geq n_k$ (i, j, k are a specific permutation of (0, 1, 2)), the number of haplotypes of a given ancestry (*i.e.*, C, P, or M) in the target diploid genome is inferred to be *i* if $n_i > 2*n_j$. Failing this condition, the local ancestry is assigned "Unknown". For markers near fixation, this condition approximately corresponds to a two-thirds majority rule for windows bridging segments of different ancestry.

The results of the admixture analysis are shown in **Fig. 2a** and **Extended Data Fig. 2a**, with admixture proportions listed in **Extended Data Table 2**. We note in passing that initial attempts to identify admixed segments using existing tools (RFmix[21] for local ancestry inference and beagle[22] for phasing) revealed significant false positive rates likely due to the small sample size of each citrus species. By contrast, the simple method described here takes advantage of the large number of sites with nominally fixed interspecific differences (*i.e.*, homozygous sites within representative accessions that differ between species) and is robust for calling interspecific admixture even with sample sizes as few as one or two per species.

Supplementary Note 5.2 Widespread pummelo admixture among mandarins

Except for five mandarins, pummelo admixture is observed in the rest of the 28 mandarin accessions in our collection (**Extended Data Fig. 2a**). Sixteen mandarin accessions contain small amounts of pummelo admixture (1-10% of genetic map length) in the form of a few short pummelo segments, and are classified as Type 2 mandarins (see **Supplementary Note 6.2** for more details). By contrast, significantly higher proportions of pummelo allele (12-38%) in the form of longer segments are found in seven of the sequenced mandarins that are classified as Type 3 (**Fig. 2a** and **Extended Data Fig. 2a**).

Supplementary Note 5.3 Oranges, grapefruit, lemon and limes

The sequenced citrons and pummelos represent pure species (except for a short segment on chromosome 2 of Chandler pummelo[2]). Three accessions are shown to be citron hybrids, including Rangpur lime and red rough lemon (both *C. reticulata* x *C. medica*). The Eureka lemon genome shows 3-way admixture with alleles from *C. medica* (50%), *C. maxima* (18%) and *C. reticulata* (31%) (estimates are based on genetic map length). We show in the next section that lemon originated from a cross between a sour orange and a citron. Oranges and grapefruits derive their genetic ancestry from the two ancestral species of *C. reticulata* and *C. maxima* (**Extended Data Fig. 2a**). It was shown previously that sour orange (cv. Seville, *C. aurantium*)[2,4,18] arose from an F1 cross (*C. maxima* x *C. reticulata*), and that the origin of sweet orange is more complex[2]. One of the two sour oranges from South China[3], CBSC (BO2), is also an F1 hybrid (*C. maxima* x *C. reticulata*), whereas the second one, ZGSC (BO3) also contains P/P segments.

Applying the above admixture identification method to other progenitor species reveals additional citrus hybrid genotypes. More specifically, Mexican lime (*i.e.*, Key lime) = *C. micrantha* x *C. medica*, as noted previously[18,23]. Similarly, whole genome sequence analysis shows Calamondin = *Fortunella* x *C. reticulata*, in line with earlier suggestions[4,12].

Supplementary Note 5.4 Admixture in Australian limes

The six accessions of Australian citrus include two round limes, two finger limes, one desert lime and an eremorange (hybrid of desert lime x sweet orange). Using a sliding window analysis of nucleotide diversity and pairwise distance D as outlined in **Supplementary Note 4**, we found that the desert lime, both round limes, and one finger lime (AFR) represent pure species, whereas the other finger lime accession in our collection (AFL) has interspecific admixture. To delineate the admixture pattern of AFL, we derived genome-wide marker SNPs for the three Australian species using fixed differences among the pure accessions, as was done above for the three Asian species of *C. medica, C. maxima* and *C. reticulata*. The results show that the finger lime accession AFL contains round lime alleles. The number of admixture segments in AFL suggests that it originated from a BC2 backcross (**Extended Data Fig. 2b**). The absence of homozygous genomic regions in AFL implies possibly three different pure finger limes in its ancestral lineage, e.g. AFL = AF3 x (AF2 x (AF1 x AR1)).

Supplementary Note 6. Shared haplotypes revealed by inter-specific phasing

Genetic relatedness can be detected through shared haplotypes. For inter-specific hybrid genomes and admixed genomic regions, we can use inter-specific phasing to extract species-specific haplotypes.

Supplementary Note 6.1 Interspecific phasing in citrus

We can make use of the strong differentiation between citrus species for inter-specific phasing. As a concrete example, consider phasing a heterozygous SNP in a pummelo(P)/mandarin(M) hybrid segment. We use four pure mandarins (Tachibana, Sun Chu Sha Kat, M01 and M02) to represent *C. reticulata*, and use all four pummelos to represent *C. maxima*. To phase the heterozygous SNP of the P/M hybrid segment, we compare the two alleles at the SNP position to the alleles at the same position in the four pummelos and four mandarins, and restrict to bi-allelic sites only.

If the four pummelos are fixed for one allele and the four mandarins are either fixed for a different allele or heterozygous at the SNP position, we assign the allele of the P/M segment that matches the pummelo allele to a *C. maxima* haplotype and the alternate allele to a *C. reticulata* haplotype. Similarly, if the four representative mandarins are fixed for one allele and the pummelos are either fixed for a different allele or heterozygous at the SNP position, we assign the allele of the P/M segment that matches the mandarin allele to a *C. reticulata* haplotype and the alternate allele to a *C. maxima* haplotype. Occasionally, a heterozygous SNP of the P/M segment is not phased if the representative pummelos and mandarins are invariant for the same allele, or when both are polymorphic at the SNP position.

Shared haplotypes among different accessions can be identified by constructing a tree of the phased haplotypes based on pairwise mismatches. As an example, the 2 Mb region at chromosome 3:3.2-5.2 Mb consists of interspecific P/M hybrid segments for 12 mandarins, sweet orange, and sour orange (**Extended Data Fig. 2a**), and can be phased to extract the *C. maxima* and *C. reticulata* haplotypes. The resulting haplotype tree (**Extended Data Fig. 3b**) shows two major branches corresponding to species-level differentiation. Within the *C. maxima* clade, only two pummelo haplotypes are represented among the 12 mandarins. By contrast, much higher haplotype diversity is observed in the *C. reticulata* clade.

Supplementary Note 6.2 Pummelo admixture pattern divides the mandarins into three types

The above method is used to examine genome-wide shared pummelo haplotypes among the sequenced mandarins, oranges, and grapefruit. One of the surprising observations is that 16 of the sequenced mandarins share either one or two pummelo haplotypes (denoted by P1 and P2) (**Extended Data Fig. 3a** light and dark blue colors) across the genome. Two examples are given in **Extended Data Fig. 3b,** which shows haplotype trees for two chromosome segments. At chr3:3.2-5.2 Mb, nine of the 16 mandarins have pummelo admixture and share a single pummelo haplotype (P1). By contrast, at chr2:31.4-33.4 Mb, pummelo admixture is present in 7 of the 16 mandarins, and two pummelo haplotypes are shared among the 7 admixed mandarins, with Ponkan containing both P1 and P2. For these 16 mandarins, the pummelo admixture segments are often few and short, accounting

for 1-10% of the total genetic map length (**Fig. 2a**).

This admixture pattern can be explained by an ancient (possibly the original) introgression event, possibly involving a single pummelo ancestor, whose haplotypes (P1,P2) have been broken into shorter segments after repeated backcrossing with mandarins. We favor the single introgression origin of P1/P2 over an alternative hypothesis in which the P1/P2 haplotypes were common among a sub-population of pummelos, since (1) we do not observe P1/P2 in any of our sequenced pummelos, and (2) the lengths of P1/P2 segments in cultivated mandarins are typically a few to tens of centimorgans, and it is unlikely that such long segments would remain unbroken by recombination in the pummelo population. We call these 16 accessions Type 2 (or early-admixture) mandarins based on their shared pummelo haplotypes P1 and P2, and the five pure mandarins (TBM, SCM, M01, M02, M04) are then designated as Type 1.

Among the remaining seven mandarins (King, Satsuma, W. Murcott, Clementine, Wilking, Fallglo, Kiyomi), other pummelo haplotypes of larger block size are observed, in addition to the shorter P1 and P2 segments (**Extended Data Fig. 3a**). We call these seven accessions Type 3 (or late-admixture) mandarins, a designation based on the presence of additional (non-P1/P2) pummelo haplotypes and larger pummelo admixture proportions (12-38% of total genetic map length). This implies that the initial pummelo introgression carrying the P1/P2 haplotypes is wide spread among the mandarins, and that late-admixture (Type 3) mandarins are distinguished from early-admixture (Type 2) mandarins by later, additional pummelo introgressions (**Fig. 2b**). In the case of Clementine mandarin, the additional (non-P1/P2) pummelo haplotype comes from its paternal parent the sweet orange.

Thus the phenotypically heterogeneous group of mandarins can be classified into three types depending on the pummelo admixture pattern. In this admixture-based classification framework, breeding between sweet orange and mandarins or between late-admixture (Type 3) mandarins would produce additional late-admixture mandarins (**Fig. 2b**).

Supplementary Note 6.3 Admixture block size and introgression timing

As most commercial/cultivated citrus accessions are clonally propagated (by grafting, or apomixis *via* nucellar polyembryony), dating historical introgression events is not possible as each accession represents a frozen genotype created at certain time point in the past. However, an estimate can be made of the number of sexual generations since the initial introgression based on the number and size of pummelo segments in mandarins.

As the early-admixture (Type 2) mandarins show much less pummelo admixture than late-admixture (Type 3) mandarins and share at most admixture sites a single pummelo haplotype (P1) (**Extended Data Fig. 3a**), they most likely originated from

the earliest pummelo introgression that involved as few as one pummelo tree. We use a repeated backcross model to simulate the number and size of pummelo segments observed in Type-2 mandarins. Recombination is modeled as a Poisson process.

Cleopatra and Sunki represent the least admixed mandarins, each containing only one pummelo segment of approximately 19 cM and 26 cM respectively. Simulations show that this admixture pattern can result from five or six generations of backcross. The other early-admixture mandarins may involve fewer generations of backcross based on their admixture proportion and segment sizes. Thus the initial pummelo introgression into the mandarins could be recent and may or may not have predated citrus domestication.

Supplementary Note 6.4 Genetic origins of citrus hybrids

Interspecific phasing also reveals genetic origins of some hybrid accessions (see **Fig. 2b**).

Lemon (cv. Eureka) and Seville sour orange share either a *C. maxima* or a *C. reticulata* haplotype throughout the genome. They also have identical chloroplast genome sequence. It can thus be concluded that sour orange is the maternal parent of lemon (lemon = sour orange x citron), in agreement with some of the previous genetic studies[12,18,24,25]. The parental citron genotype is not found among the four sequenced citron accessions.

Mexican lime (or Key lime) and micrantha share a *C. micrantha* haplotype throughout the genome. They also have identical chloroplast genome sequence. We thus conclude that the micrantha accession we sequenced is the maternal parent of Mexican lime (Mexican lime=micrantha x citron), consistent with earlier studies[12,18,25].

Both Rangpur lime and red rough lemon are F1 crosses between wild mandarins and citrons (*C. reticulata* x *C. medica*, **Fig. 2b**). However, these two genotypes are not related, and their parents are not found in our collection of sequenced citrons and mandarins. Similarly, Calamondin (*Fortunella* x C. *reticulata*) and sour orange cv. Seville (*C. maxima* x *C. reticulata*) have mandarin paternal parents that are not among the sequenced accessions. One sour orange from South Chin, CBSC (BO2), also arose from an F1 cross (*C. maxima* x *C. reticulata*), but is not related to sour orange cv Seville (*C. aurantium*). These F1 hybrid genotypes may have originated from natural crosses in the wild where pure mandarins and other citrus species coexisted.

The haploid Clementine reference[2] sequence can be used to phase the diploid Clementine genome, which, in turn, can be used to phase the parental sweet orange genome. Haplotype sharing between sweet orange and other citrus accessions can thus be estimated and sweet orange offspring can be identified.

Nuclear genome haplotype sharing analysis, together with chloroplast sequence comparison, shows that sweet orange is the male parent of both grapefruit cv. Marsh and eremorange (an Australian desert lime hybrid). This kinship can be used to phase the genomes of grapefruit and eremorange, and to reveal their genetic origins. In agreement with previous genetic studies[18], we find that grapefruit = *C. maxima* x sweet orange, though the maternal parent is not among the four sequenced pummelos. Similarly, by comparing with the pure Australian desert lime genome (ADR), we establish that eremorange= *E. glauca* x sweet orange.

Two other sweet orange hybrids are also confirmed with whole genome sequence. Kiyomi mandarin is shown to be Satsuma x sweet orange. Ambersweet orange also has sweet orange as the male parent and is closely related to Clementine, in agreement with historical record (**Supplementary Note 1**).

Supplementary Note 7 Genetic relatedness among citrus accessions

Supplementary Note 7.1 Coefficient of relatedness calculation

The genetic relatedness between two diploid individuals can be quantified by the genomic proportions sharing zero, one or two haplotypes that are identical by descent, IBD0, IBD1, IBD2. The coefficient of relatedness is defined by[26]

$$r = 0.5 * IBD1 + IBD2$$

where the genomic proportions are measured in genetic map space, and IBD0 + IBD1 + IBD2 =1. Thus for monozygotic twins, we have IBD0=IBD1=0, IBD2=1, and r=1. For parent-offspring pairs with genetically unrelated parents, IBD0=IBD2=0, IBD1=1, and r = 0.5.

To estimate the coefficient of relatedness between two diploid individuals, we calculate in non-overlapping sliding windows of 200 kb the identical-by-state ratio[2]

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

where IBS2 is the number of sites with joint-genotype AB|AB (sharing two different alleles identical-by-state), and IBS0 is the number of sites without allele sharing (with joint-genotype AA|BB). IBSR is independent of population allele frequencies and has a mean of 2/3 for two unrelated individuals from a panmictic population[27]. With haplotype sharing, IBS0=0 and IBSR=1. The IBD state of haplotype sharing for each window is inferred based on both IBSR and the genomic distance D using conservative cutoffs. If IBSR < 0.95, the genomic window is assigned IBD0. If IBSR>=0.95 and D < 0.05, the window is assigned IBD2. The last case (IBSR>=0.95 and D > 0.05) is inferred as IBD1.

However, genomic regions with inter-specific admixture in both individuals need to be dealt with differently, as the IBSR value is inflated from species-specific alleles and does not reflect shared haplotypes[2]. For such genomic regions, we use inter-specific phasing to separate the two species-specific haplotypes (*C. maxima* and *C. reticulata* for admixed mandarins) for each individual and infer the IBD state by direct haplotype comparison. To allow errors from SNP calling and phasing, we consider two haplotypes identical if the mismatch rate is below $2x10^{-4}$.

Supplementary Note 7.2 Ponkan and Huanglingmiao/Kishu mandarins are the direct parents of Dancy and Satsuma respectively

The method outlined above allows us to find previously unknown kinships among the citrus accessions. In particular, Huanglingmiao and Kishu mandarins are somatic mutants (*i.e.*, share the same base genotype upon which additional somatic mutations have accumulated) and we use the Huanglingmiao sequence[1] to represent them in this study. Beside the known mother/child relationship between Willowleaf and Clementine mandarins[2], we find four other mandarins related as parent/child pairs.

Ponkan and Dancy (both early-admixture mandarins) share at least one haplotype throughout the genome, with r=0.66 (IBD1=0.67, IBD2=0.33). Furthermore, Dancy has a high degree of inbreeding with 17% of its genome showing runs of homozygosity (ROH, **Extended Data Fig. 4a**), as a result of parental haplotype sharing. In comparison, Ponkan has 2% of its genome in ROH. This disparity in ROH, together with the extensive haplotype sharing among the sequenced mandarins as shown below, suggests that Ponkan is the parent of Dancy and that significant haplotype sharing exists between Ponkan and the second (unknown) parent of Dancy. This line of reasoning is evident in the Clementine trio (Clementine = Willowleaf x sweet orange), where the high degree of inbreeding observed in Clementine (17% of genome in ROH) results from haplotype sharing between the parents. We can further infer that the second parent of Dancy is most likely an early-admixture mandarin, with a pummelo segment on chromosome 8 that is inherited by the Dancy mandarin. The parent/child relationship for Ponkan/Dancy was also observed recently[17] based on a limited set of DNA markers, though a reversed kinship was proposed, with Dancy being the parent of Ponkan.

Huanglingmiao/Kishu (early-admixture) and Satsuma (late-admixture) mandarins can be similarly shown to have genome-wide haplotype sharing, with coefficient of relatedness r=0.6. Satsuma also shows the highest degree of inbreeding among the sequenced mandarins, with ~25% of its genome in ROH. It can be thus inferred that Huanglingmiao (or a somatic variant) is a direct parent of Satsuma, and that it is closely related to the other parent of Satsuma.

The second parent of Satsuma mandarin can be further constrained in its genetic makeup. Satsuma mandarin is characterized by a high degree of admixture with

pummelo alleles accounting for ~22% of its genome. As Huanglingmiao/Kishu has very little pummelo admixture (~3%), it is unlikely that the second parent is a mandarin. Neither can it be a pummelo, in order to explain the significant inbreeding (homozygous *C. reticulata*) observed in Satsuma. Instead, we propose that the second parent of Satsuma is an orange-like hybrid (in its genetic composition) with pummelo alleles accounting for about half of its genome.

The above inference of the parentage of Satsuma is similarly confirmed by the recent study based on DNA markers[17], where Huanglingmiao/Kishu and Kunenbo-A (not in our collection) were proposed to be the seed and pollen parents respectively. Interestingly, both the chloroplast type and the nuclear genome allelic composition of Kunenbo-A resemble those of the sweet orange, with estimated 35% *C. maxima* and 64% *C. reticulata* based on Structure analysis of about 100 markers. DNA marker analysis further indicates that Huanglingmiao/Kishu was also the pollen parent of Kunenbo-A[17], revealing Satsuma as product of a backcross with Huanglingmiao/Kishu. This explains the high degree of inbreeding observed in the Satsuma genome.

Supplementary Note 7.3 Cocktail grapefruit, Wilking, and other mandarin accessions

Not considered a true grapefruit, the Cocktail grapefruit [hybrid of Siamese Sweet pummelo x Frua mandarin] is a recent cultivar developed by the University of California, Riverside (**Supplementary Note 1**). Haplotype sharing analysis confirms that Low acid pummelo (a.k.a. Siamese Sweet pummelo) is the seed parent of Cocktail grapefruit. Though sequence for Frua mandarin is not available, the two parents of Frua (King x Dancy) are both sequenced. Comparative sequence analysis using sliding windows shows that haplotype sharing between Cocktail grapefruit and either King or Dancy exists for every window throughout the genome, thus establishing King and Dancy as the two grandparents on the paternal side. Unlike a true grapefruit, the Cocktail grapefruit is not a child of sweet orange.

Wilking mandarin [hybrid of King x Willowleaf] is another cultivar developed by the University of California, Riverside (**Supplementary Note 1**). Haplotype sharing analysis confirms the parentage of Wilking.

Fallglo [hybrid of Bower mandarin x Temple tangor] is developed by the USDA (**Supplementary Note 1**). Although both parents are not sequenced, haplotype sharing analysis shows that Fallglo is closely related to both Clementine (r=0.52) and sweet orange (r=0.43). This is consistent with the presumed parentage of Bower (Clementine x Orlando tangelo) and Temple tangor (mandarin-sweet orange hybrid). Significant inbreeding is also observed in Fallglo (**Extended Data Fig. 4a**), as would be expected since both parents have sweet orange in their ancestry.

Re-analyzing the recently sequenced Chinese mandarins[3] reveals six additional parent/child pairs. In particular, Huanglingmiao/Kishu is inferred to be a direct parent of

M08, and Sun Chu Sha Kat (SCM) is a direct parent of M17. The Chinese mandarin accession M12 is related to three other accessions (M10, M11, M14) as parent/child pairs (**Fig. 3a)**. The last pair is M17/M16. Other closely related non-parent/child accessions are also identified, as summarized in the next subsection and **Fig. 3a.**

Supplementary Note 7.4 Genetic relatedness network for mandarins, oranges and grapefruits

We conducted a pairwise calculation of the coefficient of relatedness among the sequenced accessions. Surprisingly, all 28 sequenced mandarins except Tachibana are genetically related to other mandarins with r>1/8 (*i.e.*, 3rd degree relatedness or equivalent). Even at r>1/4 (2nd degree relatedness or equivalent), 24 of the 28 mandarins remain connected in a relatedness network (**Fig. 3a**).

We also find significant haplotype sharing between sweet orange and all sequenced mandarins (r>0.1) except Tachibana and two pure mandarins from South China (M02 and M04) (**Fig. 3b**, **Extended Data Fig. 4b**). Two late-admixture mandarins (Clementine and Kiyomi) are direct offspring of sweet orange. Among sixteen early-admixture mandarins, Ponkan has the highest genetic affinity to sweet orange (r=0.33), followed closely by Dancy (r=0.3). Of the five pure (Type 1) mandarins, Sun Chu Sha Kat shows strong genetic relatedness to sweet orange (r=0.22). This network of relatedness among mandarins and sweet orange indicates a shared gene pool with predominantly *C. reticulata* founder haplotypes probably resulting from the domestication and human selection process.

By contrast, the *C. reticulata* haplotypes in the inter-specific mandarin hybrids (sour orange, Rangpur lime, red rough lemon, calamondin) are not related to the sequenced mandarins. This implies possible extant wild mandarin genotypes yet to be discovered[28].

Tachibana mandarin (*C. tachibana*), native to Taiwan, Japan and the Ryukyu Islands but absent in Asia mainland, lacks haplotype sharing with the mainland Asian mandarins, in agreement with previous suggestions based on biochemical isoenzyme analyses[16]. Collectively, the data indicate that Tachibana represents a genetic isolate that is highly differentiated from the mainland Asian mandarins due to long-term geographical separation and lack of gene flow. As noted above, its genetic differentiation from other mandarins is consistent with it being considered a subspecies of *C. reticulata* in our classification system.

Supplementary Note 8 Nuclear genome phylogenetic reconstruction and citrus speciation dating

We use representative accessions for each species to study the diversification of the genus citrus, with *Severinia* (Chinese box orange) as an outgroup. To reconstruct the species phylogeny, we use variants in non-repetitive, non-genic regions to minimize selection pressure bias. We also exclude pericentromeric

regions with low-recombination rates to reduce bias from linkage disequilibrium.

Each species is represented by one accession, except *C. reticulata*, which is represented by two pure mandarins (Tachibana and Sun Chu Sha Kat). We include both accessions to study the divergence between Tachibana and mainland Asian mandarins, as revealed by the chloroplast genome divergence (**Extended Data Fig. 1b**) and the nuclear genome genetic clustering pattern (**Fig. 1a**). Representative accessions for the other species include Humpang citron (HUM), Low Acid pummelo (LAP), Australian desert lime (ADR), Australian finger lime (AFR), Australian round lime (ARL), micrantha (MIC), Fortunella (FOR), Ichang papeda (ICH), *C. mangshanensis* (CMS), and *Poncirus* (PON). For each accession, the diploid genome is reduced to a haploid sequence by randomly sampling one allele at each position across the genome (the final phylogeny is insensitive to this random allele sampling). We also require no missing genotype calls among the representative sequences. With these considerations, a set of **362,748** variants derived from 13 nuclear genomes are used for phylogenetic inference.

Supplementary Note 8.1 Nuclear genome phylogeny and dating

We used both Bayesian and maximum likelihood phylogenetic inference methods for independent validation.

Bayesian phylogenetic tree was inferred using Mr.Bayes 3.2[29] under the general time-reversal substitution model and allowing rate variation among sites("lset nst=6 rates=invgamma"). Markov chain Monte Carlo analysis was run for one million generations under default setting (i.e rel-burnin=0.25, 2 runs, 4 chains per run). Convergence of simulation was checked using ESS (estimated sample size), PSRF (potential scale reduction factor), split frequencies, as well as multiple independent runs. The reconstructed nuclear genome phylogeny is highly robust with all nodes having maximum statistical support with marginal posterior probability=1 (**Extended Data Fig. 1c**).

We also used PhyML3.1[11] to reconstruct a maximum likelihood tree under the general time-reversal model of nucleotide substitution with 1000 bootstrap replicates. The reconstructed nuclear genome phylogenetic tree matches the Bayesian tree with all but one node having 100% statistical support (**Extended Data Fig. 1c**).

To date the speciation events, we used a penalized likelihood method[30] to infer the branch-specific evolutionary rates and divergence times based on the Bayesian phylogenetic tree. For time calibration, we used the *Citrus linczangensis* leaf fossil, a specimen from the late Miocene of Yunnan that is recognized as the oldest citrus fossil[31]. Both for simplicity and a better estimate of the uncertainty associated with the relative divergence times (*i.e.*, compared to the citrus crown age), we fix the citrus root age at 8 Ma, the midpoint of the late Miocene (11.6-5.3 Ma), with the

understanding that an overall re-scaling factor ("fudge factor") around one is implied in the speciation time estimates. As will be shown in **Supplementary Note 9**, comparison with multiple geological events suggests that the assumed citrus crown age of 8 Ma is quite reasonable and consistent with earlier estimates[12,19].

The *chronos* program from the R package *ape*[32] was used for speciation time inference, with the parameter setting model='correlated' and lambda=1. The 95% confidence intervals for the speciation times are obtained from 200 bootstrap replicates generated with seqboot of phylip[33]. The resulting chronogram is shown in **Fig. 1c**. With the citrus root age fixed at 8 Ma, the *Poncirus-Citrus* split is estimated at 9.1-9.2 Ma.

Supplementary Note 8.2 Distinct epochs of speciation for Asian and Australian citrus

The nuclear genome chronogram reveals two well-separated phases of species radiation associated with the diversification of Asian and Australian citrus respectively (**Fig. 1c**). With our sampled accessions, the Asian radiation (8-6 Ma) occurred in late Miocene and spans a period of 2 million years. It generated seven citrus species (*C. mangshanensis*, *C. ichangensis*, *C. micrantha*, *C. medica*, *C. maxima*, C. *reticulata* and *Fortunella margarita*) as well as the ancestor species that later diversified into the Australian limes.

The second phase of citrus speciation consists of the diversification of the Australian limes during early Pliocene. It spans the period 4.0-4.5 Ma, separated from the Asian citrus radiation epoch by about 1.5 Myr.

Both the nuclear genome phylogeny and the timing of speciation events point to an Asian origin for citrus, with Australian limes evolved later from an ancient citrus of Asian origin. The chloroplast genome phylogeny, though different from the nuclear genome phylogeny, is also consistent with an Asian origin for citrus. Both reject the previously proposed Australian origin[34].

Within the *C. reticulata* species, *C. tachibana* split from the mainland Asian mandarins during early Pleistocene (~2 Ma) and evolved in Taiwan, Japan and the Ryukyu Islands as a genetic isolate from the mainland Asian gene pool as revealed by the chloroplast phylogeny (**Extended Data Fig. 1b**), nuclear genome clustering (**Fig. 1a**), and haplotype sharing analysis (**Fig. 3a**). For this reason, although Swingle[13] and Tanaka[14] assigned Tachibana its own species (*C. tachibana*), we favor designating it instead as a subspecies of mandarin.

Supplementary Note 8.3 Comparison with chloroplast genome tree

Phylogenetic relationships inferred from the nuclear and chloroplast genomes (**Extended Data Fig. 1**) reveal three major discrepancies.

1) In the nuclear genome tree, *Poncirus* is an outgroup of citrus while in the chloroplast phylogeny *Poncirus* resides near the root of the tree. This observation is compatible with the view that *Poncirus* is likely a descendent of an ancient hybrid of citrus with an unknown parent, as suggested before[12].

2) The chloroplast tree clusters citrons with Australian species. By contrast, in the nuclear phylogeny, citrons cluster with pummelos whereas *Fortunella* clusters with Australian limes.

3) The branching of the three Australian species is not the same. In the chloroplast tree, *Microcitrus australasica* is an outgroup to the clade formed by *Eremocitrus glauca* and *Microcitrus australis,* while in the nuclear tree both *Microcitrus* (*australis* and *australasica*) cluster together with *E. glauca* as an outgroup.

Both nuclear and chloroplast phylogenetic trees are statistically highly supported indicating that inconsistencies are essentially biological and not due to stochastic errors. Incongruences between trees inferred from nuclear and chloroplast DNA are rather common in plants[35] and are explained in terms of convergent evolution, lineage sorting and/or reticulate evolution including horizontal gene transfer, hybrid speciation, introgression and chloroplast capture[36-41] .

In general, as shown in many studies in both plant and animals[42,43] nuclear phylogenies and networks agree with morphology-based taxonomy, while relationships inferred from chloroplast and mitochondria are more correlated with geographic proximity. Our data showed that the citrus nuclear genome phylogeny agrees in general terms, with both major morphological citrus characters as presented in **Extended Data Fig. 6** and also with geographic proximity (**Fig. 1d**). However, the chloroplast tree topology is rather incongruent in terms of fruit characteristics and less accurate regarding geographic distribution.

Concordance between citrus morphological characters and the nuclear genome phylogeny includes the presence of low number of fruit loculi and ovules in the clade containing *Fortunella* and Australian desert limes. **Extended Data Fig. 6** also suggests that mandarins do not share many fruit characteristics with *C. ichangensis* though they group together in the chloroplast genome tree.

The grouping of citrons and pummelos as revealed in the nuclear genome phylogeny has not been reported previously. While the relationship between citrons and pummelos is new, proximity of citrons and Australian limes has been reported in practically all chloroplast-derived trees presented previously. In a recent study[12], we reported chloroplast genes with unusually high number of SNPs that may be under positive pressure. A detailed analysis of the chloroplast SNP set shared exclusively by citrons and Australian limes revealed that more than half of these genes may be under positive selection.

On the other hand, citrons and pummelos share several significant morphological characteristics that are rather peculiar in the genus *Citrus* and certainly absent in the Australian limes. Overall, *C. medica* and *C. maxima* show complex floral vascular anatomies with large flowers and ovaries that contain joined stamens and many loculi producing large fruits with yellow or pale yellow peel, very thick rinds, a higher number of segments and larger columellas (**Extended Data Fig. 6)**. In contrast, Australian limes exhibit small leaves, fruits and flowers with free stamens and fewer loculi. Major differences can also be found between Australian citrus species and citrons/pummelos in other major vegetative traits since Australian limes show characteristic patterns of dimorphic foliage, with coriaceous strongly veined leaves that are not typical in the rest of *Citrus*. Finally, citrons and pummelos have overlapping geographical distribution, as wild genotypes of these two species are mostly located in regions in close proximity, from India, Bhutan, Bangladesh, Myanmar and Indochina to Yunnan (**Supplementary Table 1**).

Supplementary Note 9. The origin, biogeography, and dispersal of citrus

The information reported in this work, *i.e.*, the mapping of the distribution of the 10 pure citrus species identified in this study, the phylogenetic relationship (**Extended Data Fig 1**) and chronogram (**Fig. 1c**) inferred from the analyses of the whole genome sequences together with the recent description of *Citrus linczangensis* from the late Miocene of Lincang[31], leads us to propose that the center of origin of citrus was located in Southeast Himalaya, in a region including the eastern area of Assam, northern Myanmar and western Yunnan.

Supplementary Note 9.1. Biogeography of *Citrus* and related genera.

The genus *Citrus* and related genera (*Fortunella, Poncirus, Eremocitrus, Microcitrus*) are of wide distribution in Southeast Asia through northeastern Australia, New Caledonia, Melanesia and western Polynesia[4].

Wild genotypes of several species of the genus *Citrus* have been reported growing freely around this entire region, although the pattern of distribution varies from species to species. Thus, *C. micrantha* has been reported in the Philippines[44]; *C. ichangensis* in northeastern India, northern Myanmar and central and southwestern China, and Yunnan[4,44]; *C. medica* in northeastern, central and southern India, Bangladesh, Myanmar, Bhutan and Yunnan[44-50]; *C. maxima* in Indochina, Malaysia, Yunnan and Hainan[44,45,50,51]; *Citrus reticulata* (mandarins) are present in a wide area from northeastern India to southern and Southeast China[4,46,52]; Sun Chu Sha Kat mandarin (*Citrus reticulata*) is naturally found in Assam, China and Japan[4] while *Citrus tachibana* is widespread in southern Taiwan, the Ryukyu Islands and southern Japan[53] and *C. mangshanensis* in Hunan, China[54]. In addition, wild hybrids between the species of the genus *Citrus* have also been found mostly in areas and habitats occupied by the parental genotypes. Thus, *C.* x *limonia* (*C. reticulata* x *C.*

medica) has been reported in eastern Guangxi but mostly in southern Tibet[48-50,55] while *Citrus* x *jambhiri* (*C. reticulata* x *C. medica*) has been found growing in a wild condition in India[56]. Sour orange, *C. x aurantium* (*C. maxima* x *C. reticulata*), is found in Nepal, northeast India, Garwal and Sikkim[45,46]. Calamondin (*Fortunella sp* x *Citrus reticulata*) is widely cultivated in the Philippines and also in China.

Regarding the relevant cultivated hybrids such as *C*. x *sinensis* (sweet oranges), the presence of wild trees has been reported in tropical forests of northern Myanmar and in the Khasi Mountains in Assam[57]. However, there are no reports on wild *C*. x *aurantifolia* (limes, *C. micrantha x C. medica*) and *C.* x *limon* (lemons, *C*. x *aurantium* x *C. medica*)[4], although there is a general agreement that limes come from the Southeast Asian archipelago[4] while lemons are native of India and northern Myanmar[4,58].

The genus *Fortunella* includes four species, all of them found south of Yangtze River in provinces such as Yunnan or Guizhou[48,49]. The single *Eremocitrus* species, (*E. glauca*) and 5 of the 6 described species of *Microcitrus*, (including *M. australis* and *M. australasica*) are thought to be native of Queensland and northern New South Wales in eastern Australia[4,59]. Regarding the 2 species of *Poncirus*, *P. trifoliata*, is native to eastern Asia[60,61] while *P. polyandra* was identified in Yunnan Province[62].

Supplementary Note 9.2. Citrus dating and fossils.

Previous estimations of citrus divergence times using molecular analyses have produced different results. Initial analyses based on partial chloroplast sequences estimated the age of citrus to be 22-18 Ma[63] while other more solid studies reported that citrus appeared about 7.0 Ma[19]. Using whole genome chloroplast sequences, we recently dated the emergence of citrus as *ca*. 8.0 Ma[12].

The first report of a citrus fossil (*C. meletensis*) concluded that the specimen was dated in the Pliocene of Europe and therefore outside of its native geographic distribution[64]. While this finding is still not properly contextualized, the recent report on the identification of a new *Citrus* species, *Citrus linczangensis* sp. n., from the late Miocene (11.6-5.3 Ma) of Lincang[31], provides definite evidence for the existence of *Citrus* within the province of Yunnan (China) since ca. 8 Ma.

The fossil leaves of *C. linczangensis*, which do not exactly resemble any particular extant citrus species in all characters, combine a number of features that are present in distinct phylogenetic clades as defined in **Extended Data Fig. 6**. These characteristics include the possession of an articulated, subcordate and broadly winged petiole as usually seen in *C. maxima,* some species of *Fortunella*, *C. micrantha* and in *C. ichangensis*. The fossil also exhibits intramarginal venation as found in *C. reticulata* and *C. aurantifolia* (*C. micrantha x C. medica*) and an entire margin as observed in *C. maxima*. The presence of a long winged petiole has traditionally been regarded as an old trait of ancestral citrus[4].

The specimen, therefore, might well represent a common ancestor of the major citrus groups (papedas, pummelos, mandarins, *Fortunella* and *Micrantha*). This finding implies that southwestern China[50] in Late Miocene was a native habitat of citrus and therefore a potential region of early diversification, supporting the hypothesis of a Southeastern Asian origin of *Citrus*.

Supplementary Note 9.3. The center of origin of citrus

The center of origin of citrus has been matter of dispute during almost a century. The elegant but general pioneer visions of Vavilov[65] identified two centers of origin for citrus, the Indo-Burma center (Assam and Burma; oranges, mandarins and citrons), and the Indo-Malayan center (Indochina and the Malay Archipelago; pummelo). This vision was reformulated by Tanaka,[14] who in subsequent work concluded that the primary center was placed within northeast India and northern Myanmar, from where citrus dispersed to a secondary center located in Indochina and Southeast China. The center of origin of citrus was also suggested to be in northeastern India and in the mountainous parts of southern China[66].

In their masterful review, Swingle and Reece[4] indicated that *Citrus* is native to an extensive barrel-shaped area that has its long axis slanting from the northwest (northeastern India to north-central China) to the Southeast (east-central Australia to New Caledonia). Many other propositions have been advanced in more recent years suggesting e.g. that primary centers of citrus origin were in the southwestern mountains of China[48], in Yunnan and adjacent areas in northern east India, northern Myanmar and southwestern and southern China[50].

The analyses of the native habitats of citrus presented in this work (**Extended Data Fig. 1a**) based on documented reports (**Supplementary Table 1**) upon the presence of wild genotypes growing freely in non-cultivated areas reveals that the triangle limited by Eastern India, Northern Myanmar and Western Yunnan concentrated the highest number of wild citrus genotypes (*C. medica, C. maxima, C. reticulata, C. ichangensis, C. x limonia, C. x sinensis, C. x aurantium, C. x limon*). The identification of *C. linczangensis*[31], on the other hand, provides definite evidence that in Late Miocene western Yunnan was a native habitat of citrus and therefore a potential region of early diversification. Furthermore, the phylogenetic relationship and estimation of divergence times (**Fig. 1c**) indicated that the ancestral citrus experienced a relatively fast radiation giving rise to all major citrus species in a period of about 2 Myr, which is consistent with a unique initial area of diversification. While our proposal offers partial support to some of the previous formulations conferring geographic accuracy to the otherwise broad and vague propositions, overall, it clearly precludes ideas based on primary centers located in Australia or nearby islands[4], Malay Archipelago[67] or Thailand[51].

Supplementary Note 9.4. Citrus rapid radiation and monsoon weakening

Both nuclear (**Extended Data Fig. 1c**) and chloroplast (**Extended Data Fig. 1b**) phylogenies and the estimation of divergence times (**Fig. 1c**) indicated that citrus underwent a rapid radiation during late Miocene (ca. 8.0-6.0 Ma). Rapid biological radiations in this region have been reported for practically all major groups of organisms[68,69] including insects, fishes, crabs, amphibians, reptiles, birds, ferns[70] and plants such as the eudicot genera *Caragana, Rheum, Pedicularis, Saussurea, Rhododendron, Primula, Meconopsis, Rhodiola,* and many lineages of gymnosperms, *i.e.*, the conifer genus *Juniperus*. It is widely accepted that climatic oscillations provoked by monsoonal seasonality were the pivotal factors facilitating speciation and diversification in Southeast Asia. However, contrary to what is generally believed, current evidence clearly indicates that monsoons in this region arose at different times and are certainly unrelated to the Tibetan uplift[71].

Monsoon regimes in South and East Asia were probably established in Early Miocene and experienced an extended period of intensification with strong summer monsoons in the Middle Miocene and reached maximum between 18 and 10 Ma[72]. After this phase of monsoon intensification, a period of monsoon weakening started around 10 Ma in East Asia and around 8 Ma in South East Asia[72,73]. It is accepted that the rapid weakening of the monsoon provoked a sudden and drastic climate transition from wetter conditions to a drier climate with seasonal heavy rains[72,74]. In Southeast Asia this dramatic alteration caused major biota changes including migration of mammals and a conspicuous substitution of evergreen tall tropical trees by tall grasses, a transition that in southeastern Himalaya took place 8.0-7.0 Ma[74,75]. Thus, the rapid radiation of citrus is in accordance with the proposed phase of weakening of the East Asian monsoon that occurred around 8.0 Ma in Southeastern Himalaya[72,73].

Interestingly, the physiological and phenotypic adaptations of current citrus[76,77] appear to still carry the signature of that ancient transition from wetter to dryer conditions, since continental citrus that can be defined as mesophytes exhibit intriguing xerophytic adaptations to cope with periods of water stress. Citrus, for instance, possesses efficient waxy coated leaves and fruit peels to reduce water losses; the fruit also develops individual juice sacs to protect water, and the tree shows low photosynthetic and transpiration rates associated with slower growth patterns. Citrus plants also produce lush foliage, and have high chlorophyll content and shallow root systems, as found in tropical understory bushes with lower light availability and poor organic material soils. Citrus also develop a unique spongy fruit albedo to cushion rapid volume alterations produced by sudden water inputs. Taken together, these observations are also compatible with the assumption that ancestral citrus were native of regions with “tropical monsoon or savanna climates” (according to the Köppen-Geiger climate classification[78]) but evolved in a “humid subtropical climate” with more pronounced dry seasons.

Supplementary Note 9.5. Citrus dispersal

After species radiation during the late Miocene, we propose that the prevalent directions of citrus dispersal from the putative center of origin were: west (citrons), northeast (*C. mangshanensis* and papedas), east (*Fortunella* and mandarins) and southeast (pummelos and micrantha). Support for this proposal comes from studies on citrus biogeography and phylogeny, and on the paleogeography of the region, especially the geological history of Wallacea and Japan. According to this view, all citrus except citrons dispersed in a predominant west to east direction (**Fig. 1d**), as many groups of plants and animals did[68,69,79].

Furthermore, there is compelling evidence that the eastern edge of the Himalaya, the Hengduan range, comprising several mountain subranges in western Chinese provinces, including Yunnan, underwent rapid uplift only after the Miocene, reaching maximum elevations shortly before Late Pliocene, *ca* 3.6 Ma[80,81]. Additional evidence based on the reconstruction of paleovegetation and paleoclimate in the Late Pliocene of west Yunnan indicates that uplift of Gaoligong and Nu Mountains (Hengduan range) and the eastern portion of the Tibetan Plateau (Western Yunnan) must have occurred during or after the late Pliocene[82,83]. Therefore, the paleogeography of the region is compatible with the suggestion that during the period from Late Miocene to Late Pliocene, there was land continuity between current Northern Myanmar and Western Yunnan, i.e. there were not orogenic barriers blocking potential west to east routes allowing citrus dispersal from the center of origin.

Supplementary Note 9.6. Origin of Australian citrus

Diversification of Australian limes (4.0-4.5 Ma, **Fig. 1c**) was not driven by monsoons. Since these species have developed xerophytic structures and are found native in dry environments in north- and south-east Australia, the most logical explanation is that they adapted in Australia. This implies a migration from continental Southeast Asia to Australia in agreement with early plant botanists who suggested that the predominant dispersal direction across the Wallace´s line was west to east[79].

Furthermore, a significant percentage of Australia's northern tropical flora is clearly derived from Southeast Asia, an idea supported by molecular and geographic analyses[84-86] that have concluded that lineages of several angiosperm genera such as *Aglaia, Alocasia, Begonia, Pseuduvaria*, *Neonauclea* and *Uvaria* and some palm lineages displayed dispersal patterns largely consistent with initial diversification in continental Southeast Asia and subsequent dispersal to eastern Malesia or Australia.

There is also a general agreement[87] that most of western Malesia emerged throughout the Cenozoic while islands and lands east of Wallace's Line elevated above sea level only during the late Miocene and Pliocene. Hall[87] has indicated that most of Sumatra and Java, Sulawesi, parts of the Banda Arc, and the Moluccas were

elevated above sea level since 5 Ma while Seram and Timor have both emerged in the last 3 Ma. This regional plate reorganization that occurred in Wallacea in the last few million years[88] apparently provided potential stepping stones allowing plant dispersal to New Guinea and Australia as reported for the genera *Bridelia* which reached Australia *ca*. 2 Ma[89], two different genera of *Cucurbitaceae*, *Benincasa* and *Neoachmandra* that arrived at Australia from Southeast Asia *ca*. 5 and 1 Ma, respectively[90] and *Begonia* that dispersed six independent times from continental Asia and western Malesia to Wallacea and New Guinea dating from the late Miocene to the Pleistocene[84].

Likewise, we propose that the three Australian species studied here had an Asian continental ancestor that dispersed from west to east to reach Australia **(Fig. 1d).** This proposal is compatible with the citrus phylogenetic relationships and the inferred timing of Australian lime diversification (**Fig. 1c**). The proposal is also concordant with the predominant west to east dispersal trend observed in the region and with the spatio-temporal diversification patterns reported in other genera as noted above. Further supporting evidence comes from the paleogeography of the region, especially the geological history of Wallacea which postulates that during Late Miocene onwards extensive land masses and islands emerged providing potential stepping stones and allowing island-hopping dispersal[25]. Lastly, New Guinea was very likely an intermediate stop on the way to Australia, as all the five known *Microcitrus* species found native in eastern Australia are very probably derived from *M. warburgiana,* a species exclusively found in New Guinea[4]. The paleogeographic reconstructions of this epoch also provide additional support for this suggestion[87].

Supplementary Note 9.7. Tachibana mandarin dispersal.

It has been well documented that Taiwan, the Ryukyu archipelago and Japan attained their flora and fauna from adjacent mainland through the emergence of land bridges that occurred mostly during the Pleistocene with the expansion of ice sheets[91,92], the route that probably was also followed by Tachibana during its migration in the early Pleistocene. According to wide palaeoceanographic evidence, the expansion of glaciers led to drastic reductions in the levels of the South China Sea[93], creating land bridges and providing major corridors between the islands. The bridges connecting the mainland with the islands occurred many times throughout the Quaternary[94-96], including the period of Tachibana split from mainland Asian mandarins during the early Pleistocene, an epoch characterized by strong glacial maxima[97].

Supplementary Note 10 Pummelo admixture and citrus fruit size and acidity

Though the domestication process for mandarins and sweet orange is complex, two independent lines of investigation point to a strong connection between citrus domestication and pummelo introgression in the mandarin gene pool. For this

analysis, the recently sequenced Chinese accessions[3] are not included because information on fruit size and acidity profile are not available.

Supplementary Note 10.1 Pummelo admixture correlates with fruit size

Using diameter as a measure of fruit size, we observe a strong correlation (Pearson correlation coefficient = 0.88) between pummelo admixture proportion and the fruit sizes of mandarins, oranges and grapefruit (**Extended Data Fig. 5a**). In particular, the two pure mandarins (TBM, SCM) are smaller than any other mandarins with pummelo admixture.

A simple linear regression fit shows outliers on both sides of the regression line. On the side of small fruits are the four acidic mandarins, which are either admixture free (TBM, SCM) or have small amount of pummelo admixture (CLP, SNK). By contrast, Ponkan mandarin (PKM) appears unusually large given its admixture size. These outliers suggest that certain genomic loci could be more significant than others in fruit size determination.

With the addition of two pummelos (LAP, CHP), the correlation between fruit size and pummelo admixture proportion becomes stronger (Pearson correlation coefficient 0.94). However, a polynomial regression of the 2nd degree provides a better fit than simple linear regression (adjusted R^2=0.92) (**Extended Data Fig. 5b**). Though the four acidic mandarins and Ponkan mandarin remain as outliers, the general trend in fruit size increase is clear from pure to admixed mandarins, and from oranges, grapefruit to pummelos. As larger fruit size, relative to pure mandarins, is a desirable trait, mandarins with pummelo admixture were probably selected and propagated during domestication.

Supplementary Note 10.2. Genome scan for citrus acidity/palatability association

It is very plausible that one of the pivotal drivers of fruit domestication is palatability. In citrus, palatability requires reduced acidity that is dependent upon citric acid accumulation in the vacuole[98]. Thirty-seven citrus accessions with known acidity profile were divided into two groups for a case-control study of fruit acidity/palatability. *Poncirus* and *Severinia* were excluded from this study because they are not true citrus, while fruits from the recently sequenced Chinese accessions[3], eremorange and *C. mangshanensis* were not available for acidity determination. The palatable group (case) was formed by 15 accessions including 9 mandarins, all 4 pummelos, sweet orange and grapefruit (**Supplementary Table 2**). The non-palatable group (control) consists of 4 acidic mandarins (TBM, SCM, CLP, SNK), sour orange, all 4 citrons, lemon and limes, as well as CAL, FOR, ICH, MIC, and 5 Australian limes.

To scan for candidate loci associated with fruit palatability, a genome wide association analysis was performed. Since the sample set was relatively small, GWAS

was used as a preliminary approach to reveal chromosome fragments and loci significantly associated with acidity for further manual evaluation. We used a mixed linear model as implemented in gemma[99] to correct for confounding effects from population stratification and sample relatedness. The quality control for the SNP set followed the criteria of **Supplementary Note 3.1**. In addition, we require a SNP call rate > 90% and minor allele frequency > 5%, resulting in 634,888 common SNPs. We used the standardized relatedness matrix and the likelihood ratio test to estimate P value[99], and applied the conservative Bonferroni correction at significance level of alpha = 0.05 (i.e. $P < 7.9x10^{-8}$).

The GWAS analysis, shown as Manhattan plot in **Extended Data Fig. 5c**, yielded 24 SNPs that exhibited genome-wide statistical significance (**Supplementary Table 4**). All 24 SNPs were manually examined in each of the 37 samples on a gene-by-gene basis as related to allele identity and functional impact. Using *Severinia* and *Poncirus* to define the ancestral state of an allele, palatability is found to be associated with derived alleles, suggesting that the ancestral citrus species were acidic. The results of the manual inspection are summarized in **Extended Data Table 3**. The table shows that SNPs located on chromosomes 2, 5 and 9 as well as three of the four SNPs on chromosome 1 are not discriminatory for acidic accessions versus the four pummelo accessions of the palatable group, though in the GWAS analysis significant allele frequency difference exists between the case and control groups. Based on these considerations, there remain two highly discriminatory loci for the case and control groups.

At the first locus (chr1:23512067; Target of EGR1 Protein 1, TOE1), the non-palatable accessions are mostly homozygous for the ancestral allele (T/T) with a few accessions being heterozygous (C/T). By contrast, all palatable accessions are homozygous in the derived allele (C/C) with the exception of King mandarin (C/T), an edible late variety that shows a marked delay in citric acid degradation.

The second locus spans a nearly 2 Mb region at the beginning of chromosome 8 (chr 8:0.3-2.2 Mb; **Extended Data Fig. 5d**) where pummelo admixture is present among all palatable mandarins but absent in the acidic mandarins, suggesting that the introgression of pummelo genes may have played a role in the domestication of mandarins. This region contains 15 significant and rather discriminatory SNPs. Among several potentially significant genes in this region, Ciclev10028714 and Ciclev10028121 have plausible relevance in the regulation of acidity. Ciclev10028714, a gene coding for mitochondrial NAD^+ Isocitrate Dehydrogenase (NAD^+ IDH) catalyzing the oxidative decarboxylation of isocitrate to α-ketoglutarate, a rate-limiting step in the tricarboxylic acid cycle and therefore in citric acid synthesis[100].

While the SNP specifically pinpointed by GWAS (chr 8: 325527) corresponds to a synonymous substitution, manual inspection of the IDH gene revealed three non-synonymous SNPs (positions chr 8: 324328, 326594, and 326608), two of which are

tri-allelic and did not enter the bi-allelic GWAS analysis (**Extended Data Table 4)**. The SNPs at positions 326594 and 326608 correlate quite well with acidity except that the acidic sour orange shares the same genotypes as the palatable mandarins, an observation that may reflect the polygenic nature of acidity regulation. Besides the IDH gene, two variants located 1kb upstream of a glycosyl hydrolase gene (Ciclev10028121) are completely discriminatory for acidic versus the non-acidic accessions. Gene expression or protein activity of these two genes has recently been associated with acidity in citrus[101,102].

Supplementary Note 11 Acknowledgements

The authors acknowledge financial support from the following institutions and companies: Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, INIA, (Ministerio de Economia, Industria y Competitividad e Innovación, Spain) trough grant # RTA-00071-C06-01 and from the Citruseq-Citrusgenn consortium (Anecoop S. Coop., Eurosemillas S.A., Fundación CajaMar Valencia, GCM Variedades Vegetales A.I.E., Investigación Citrícola Castellón S.A. and Source Citrus Genesis) (MT); Agriculture and Food Research Initiative Grant no 2013-67013-21110 from the USDA National Institute of Food and Agriculture (MLR); Florida Citrus Production Research Advisory Council (FCPRAC), Florida Department of Agriculture and Consumer Services grant No. 013646, Florida Department of Citrus (FDOC) and Citrus Research and Development Foundation grant No. 71, on behalf of the Florida citrus growers (F.G., D.D.). The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231 (GAW,DSR). Help and expertise of Isabel Sanchis and Angel Boix (IVIA) on lab and field tasks are also gratefully acknowledged.

Supplementary Tables

Supplementary Table 1. Biogeographic distribution of the genus Citrus in Southeast Asia and Australia. The occurrence of wild genotypes of pure citrus species, admixtures and relevant interspecific hybrids is reported. Classical taxonomy (Swingle[13] and Tanaka[14]) has been maintained for clarity. Numbers indicate approximate locations in Extended Data Figure 1a.

Genera/Species	Geographical Region*	Report	Phenotype	References
		Pure Citrus Species		
Fortunella sp.	1. Southern China 2. South of Yangtze river	Wild species	Freeze-hardy. Small fruit, acid pulp, edible skin.	Zhang, 1981[48]; He et al. 1984[49]; Zhou, 1990[44]
Australian citrus	3. Eastern Australia	Wild *Eremocitrus glauca* *Microcitrus australasica* *Microcitrus australis*	Two ovules per locule (only *E.*) All tree, xerophytic adaptations, small leaves, fruits and flowers	Swingle and Reece, 1967[4]; Zhou, 1990[44]; Sykes, 1997[51]
Citrus micrantha	4.Philippines	Wild *C. micrantha*	Small acid and bitter fruit. Shrubby tree.	Zhou, 1990[44]; Ghosh, 1997[103]
Citrus ichangensis	5. Northeastern India 6. Northern Myanmar 7. Central China 8. Southwestern China[50]	Wild *C. ichangensis*	Non-edible acid fruit. Cold-resistant.	Zhou, 1990[44]; Gmitter and Hu, 1990[50] Ghosh, 1997[103]; Swingle and Reece, 1967[4]
Citrus medica	9. Northeastern India	Wild species	Large acid fruit, persistent style.	De Candolle, 1883[45]; Tanaka, 1961[46]; Rajput

	10. Central India 11. Southern India 12. Bangladesh. 13. Myanmar 14. Bhutan 15. Yunnan.		Thick peel. Tendency to grow continuously.	and Haribabu, 1985[47]; Zhang, 1981[48]; He et al. 1984[47]; Zhou, 1990[44]; Gmitter and Hu , 1990[50]; Yang et al. 2015[104]
Citrus maxima	16. Indochina 17. Malaysia. 18. Yunnan 19. Hainan	Wild species	Large leaves, flowers and fruits. Thick fruit peel. Sweet or acidic.	De Candolle, 1883[45]; Zhou, 1990[44]; Gmitter and Hu, 1990[50]; Scora 1988[51]
Citrus reticulata	20. Assam 21. China 22. Japan	Sun Chu Sha Kat	Small-fruited, acidic-sweet, seedy deep red, mandarin.	Swingle and Reece, 1967[4]
	23. Taiwan 24. Ryukyu Islands 25. Japan	Tachibana (*C. tachibana*)	Small-fruited, acidic-sweet, seedy yellowish mandarin	Tanaka 1931[53]
Citrus mangshanensis	26. Hunan	Wild *C. mangshanensis*	Similar to cultivated mandarins.	He et al. 1984[49]; 1988[105]
		Citrus Admixtures		
Citrus sp. (mandarins)	27. Northeastern India 28. Southern China 29. Southeast China	Wild species of sp. mandarins	*Reticulata*/*maxima* admixtures. Acidic, acidic-sweet or palatable.	Tanaka 1961[46]; Hodgson, 1967[52]
Citrus sinensis	30. Northern Myanmar 31. Assam 32. Southern China	Probably feral *C. sinensis*	*Maxima/reticulata* admixtures Moderate freeze-hardy. Fruit of medium size. Palatable.	Swingle and Reece, 1967[4]; Cooper, 1990[57];

	33. Indochina			
Citrus limon	34. India 35. Northern Myanmar	Wild specimens not found	*Maxima/reticulata/medica* admixtures. Acidic fruits.	Bonavia, 1888[58]; Swingle and Reece, 1967[4]; Webber et al. 1967[67]
		Interspecific Citrus Hybrids		
Citrus limonia	36. Eastern Guangxi 37. Southern Tibet	Wild species	*Reticulata x medica*. Tolerant to stresses. Tart and acidic fruits.	Zhang, 1981[48]; He et al. 1984[49]; Gmitter and Hu, 1990[50]; Chen, 1997[106]
Citrus jambhiri	38. India	Wild species	*Reticulata x medica*. Tolerant to stresses. Tart and acidic fruits.	Hodgson, 1937[56]
Citrus x aurantium	39. Nepal 40. Northeast India 41. Garwal 42. Sikkim	Wild species	*Maxima x reticulata*. Tolerant to stresses. Tart and acidic fruits.	De Candolle, 1883[45]; Tanaka, 1961[46]
Citrus aurantifolia	43. Southeast Asian archipelago	Wild specimens not found	Micrantha x citron. Freeze-sensitive. Mostly acidic fruits.	Swingle and Reece, 1967[4]
Fortunella x C reticulata	44. China 45. Philippines	Wild specimens not found	*Fortunella* sp x *reticulata* Small fruit, acid pulp, edible skin.	Swingle and Reece, 1967[4]

Supplementary Table 2. Accessions of the genus Citrus and related genera studied in this work. When possible, the Swingle (1943)[13] and Tanaka (1954)[14] taxonomic systems have been followed.

Code	Common name	Species name	Nuclear genome relationship (mat x pat)	Chloroplast Genome	Acidity
•COR	Corsican citron	*C. medica* L.	Pure citron	CI	Acidless
•VEU	Mac Veu citron	*C. medica* L *Citrus lumia Risso & Poit* .	Pure citron	CI	Acidic
•BUD	Buddha's hand citron var. Sarcodactylus	*C. medica* L. (Noot.) Swingle	Pure citron	CI	Acidic
•HUM	Humpang citron	*C. medica* L.	Pure citron	CI	Acidic
•SCM	Sun Chu Sha Kat mandarin	*C. reticulata* (Blanco) C. reticulata var. austera (Swingle) *C. erythrosa* (Tanaka)	Pure mandarin	MA	Acidic
•TBM	Tachibana mandarin	*C. tachibana (Mak.)* Tan *C. reticulata* (*Blanco*)	Pure mandarin	MA	Acidic
•SNK	Sunki mandarin (sour mandarin, suanju)	*C. sunki* (Hayata, Hort. ex Tanaka *C. reticulata* (Blanco)	Reticulata/pummelo admixture (low % pummelo)	MA	Acidic
•CLP	Cleopatra mandarin	*C. reshni* (Hort. ex Tanaka) *C. reticulata* (Blanco)	Reticulata/pummelo admixture (low % pummelo)	MA	Acidic
•CSM	Changsha mandarin	*Citrus reticulata* (*Blanco*)	Reticulata/pummelo admixture	MA	Non-acidic
HLM	Huanglingmiao mandarin	*C.* reticulata* (Hort. ex Tanaka)	Reticulata/pummelo admixture (parent of UNS)	MA	Non-acidic
•KSH	a.k.a Mukakukishu or Kinokuni mandarin	*C. *kinokuni* (Hort. ex Tanaka)	Reticulata/pummelo admixture (mutant of Huanglingmiao)	MA	Non-acidic

•UNS	Satsuma (unshiu) mandarin, cv. Owari	*C. * unshiu* [(Mak.) Marc] *C. *reticulata* (Swingle)	Reticulata/pummelo admixture)	MA	Non-acidic
PKM	Ponkan (Chinese honey orange)	*C. *reticulata* (Blanco, Swingle)	Reticulata/pummelo admixture (parent of DNC)	MA	Non-acidic
•DNC	Dancy mandarin, Dancy tangerine	*C. * tangerina* (Tanaka) *C. *reticulata* (Swingle)	Reticulata/pummelo admixture	MA	Non-acidic
WLM	Willowleaf mandarin	*C. x deliciosa* (Ten. Hort. ex Tanaka)	Reticulata/pummelo admixture (maternal parent of CLM)	MA	Non-acidic
CLM	Clementine mandarin, cv. Clementina de Nules	*C. x clementina* (Hort. ex Tanaka) *C.x reticulata* (Swingle)	Reticulata/pummelo admixture WLM x SWO	MA	Non-acidic
•KNG	King mandarin	*C. * nobilis* (Lour.) *C. *reticulata* (Swingle)	Reticulata/pummelo admixture	MA	Non-acidic
WMM	W. Murcott mandarin	*C. x reticulata* (Blanco)	Reticulata/pummelo admixture	MA	Non-acidic
•LMA	Rangpur lime	*C. x limonia* (Osbeck)	Reticulata x citron hybrid	MA	Acidic
•RRL	Red rough lemon	*C x jambhiri (Lush)*	Reticulata x citron hybrid	MA	Acidic
LAP	Low acid pummelo (Siamese Sweet)	*C. maxima* [(Burm.) Merr]. *C. grandis* (Swingle, Tanaka)	Pure pummelo (maternal parent of CHP)	PU	Acidless
CHP	Chandler pummelo	*C. maxima* [(Burm.) Merr]. *C. grandis* (Swingle, Tanaka)	Pure pummelo (Siamese Sweet (LAP) x Siamese Pink)	PU	Non-acidic
GXP	Guan-xi-mi-you pummelo	*C. maxima* [(Burm.) Merr]. *C. grandis* (Swingle, Tanaka)	Pure pummelo	PU	Non-acidic
STP	Sha-tian-you pummelo	*C. maxima* [(Burm.) Merr]. *C. grandis* (Swingle, Tanaka)	Pure pummelo	PU	Non-acidic
•PAR	Grapefruit, cv. Marsh	*C. x paradisi* (Macfadyen)	Pummelo x SWO	PU	Non-acidic

SWO	Sweet orange, cv. Washington Navel	*C. x sinensis* L. (Osbeck)	Pummelo/reticulata admixture (paternal parent of CLM)	PU	Non-acidic
SSO	Sour orange, cv.Sevillano	*C. x aurantium* L.	Pummelo x reticulata hybrid	PU	Acidic
•LIM	Lemon, cv.Eureka	*C. x limon* L. (Burm. f.)	SSO x citron hybrid	PU	Acidic
•ADR	Australian desert lime	*Eremocitrus glauca (Lindl.) Swingle* *C. * glauca (Lindl.) Burkill* *Eremocitrus glauca (Lindl.) Swingle*	Pure glauca	AU	Acidic
•ADL	Eremorange. Australian desert lime hybrid	N.A.	*Eremocitrus glauca x Citrus sinensis*	AU	---
•AFR	Australian finger lime	*Microcitrus australasica (F.Muell.) Swingle* *C. australasica* F. Muell.	Pure australasica	AU	Acidic
•AFL	Australian finger lime	*Microcitrus australasica (F.Muell.) Swingle* *C. australasica* F. Muell	BC2 backcross	AU	Acidic
•ARR	Australian round lime	*Microcitrus australis Swingle* *C. australis* (A. Cunn. ex Mudie)	Pure australis	AU	Acidic
•ARL	Australian round lime	*Microcitrus australis Swingle* *C. australis* (A. Cunn. ex Mudie)	Pure australis	AU	Acidic
•FOR	Kumquat, Nagami	*Fortunella margarita* (Lour.) Swingle	Pure Fortunella	FO	Acidic
•CAL	Calamondin	*Citrus reticulata* Blanco var, austera *Swingle ? x Fortunella sp. ?* *C. madurensis* (Lour.)	Fortunella x mandarin hybrid	FO	Acidic
•MXL	Mexican lime	*C. * aurantifolia* (Christm.) Swingle	Micrantha x citron hybrid	MC	Acidic
•MIC	Micrantha, Biasong	*C. micrantha* (Wester)	Pure micrantha	MC	Acidic
CMS	Mangshan mandarin	*C. mangshanensis*	Pure mangshan	MS	---

•ICH	Ichang papeda	*C. ichangensis* (Swingle)	Pure ichangensis	IC	Acidic
•PON	Trifoliate orange	Poncirus trifoliata (L.) Raf.	Pure Poncirus	PT	Acidic
•SVR	Chinese box orange	*Severinia buxifolia* (Poir.)Tenore		SV	---
M01	Mandarin from south China (CYY)	*C. reticulata*	Pure mandarin	MA	---
M02	Mandarin from south China (HZ)	*C. reticulata*	Pure mandarin	MA	---
M03	Mandarin from south China (SJ)	*C. *reticulata admixture*	Admixed mandarin	MA	---
M04	Mandarin from south China (SPG)	*C. reticulata*	Pure mandarin	MA	---
M08	Mandarin from Zhejiang province (20H)	*C. *reticulata admixture*	Admixed mandarin	MA	---
M10	Mandarin from south China (HPJ)	*C. *reticulata admixture*	Admixed mandarin	MA	---
M11	Mandarin from south China (YSJ)	*C. *reticulata admixture*	Admixed mandarin	MA	---
M12	Mandarin from south China (NJ)	*C. *reticulata admixture*	Admixed mandarin	MA	---
M14	Mandarin from south China (MSJ)	*C. *reticulata admixture*	Admixed mandarin	MA	---
M15	Mandarin from south China (LYJ)	*C. *reticulata admixture*	Admixed mandarin	MA	---
M16	Mandarin from south China (BTJ)	*C. *reticulata admixture*	Admixed mandarin	MA	---
M17	Mandarin from south China (STJ)	*C. *reticulata admixture*	Admixed mandarin	MA	---

M19	Wilking (WLK)	*C. x reticulata*	King x willowleaf	MA	Non-acidic
M20	Kiyomi (KYM)	*C. x reticulata*	Satsuma x sweet orange	MA	Non-acidic
M21	Fallglo (QH117)	*C. x reticulata*	(Clementine x Orlando) x Temple	MA	Non-acidic
GF0	Cocktail grapefruit (14J)	*not assigned by Swingle or Tanaka.*	Low acid pummelo x Frua mandarin	PU	Non-acidic
SO5	Ambersweet orange (A20)	*not assigned by Swingle or Tanaka*	(Clementine x Orlando) x sweet orange	MA	Non-acidic
BO2	Sour orange from south China (CBSC)	--	Pummelo x pure mandarin	PU	---
BO3	Sour orange from south China (ZGSC)	--	--	PU	---

• = Genomes sequenced in this work; otherwise genomes were reanalyzed from original published sequences[1-3].
X = Hybrid origin previously known; * = Information generated in this work.
Chloroplast types: AU, Australian limes; CI, citron; FO, *Fortunella*; MA, mandarin; MC, Micrantha; MS, *Citrus mangshanensis*; IC, Ichang papeda; PT, *Poncirus*; PU, pummelo; SV, *Severinia*.
Non-acidic: palatable citrus fruit; usually with acid content < 10 g/l when overripe. Acidic: unpalatable citrus fruit; acid content > 20g/l when overripe.
Acidless: applies to Corsican citron and low acid pummelo, two varieties with reduced acid content < 2g/l.
Note that the last 19 accessions (beginning with M01) are from Wang et al.[3] with their codes in parenthesis.

Supplementary Table 3. Sequencing statistics of the 30 new genomes reported in this work. When possible, the Swingle (1943)[13] and Tanaka (1954)[14] taxonomic systems have been followed.

Code	Common name	Species name	Total Reads	Coverage
CLP	Cleopatra mandarin	*C. x reshni* (Hort. ex Tanaka)	378265813	126X
KNG	King mandarin	C. x *nobilis* (Lour)	194708238	65X
DNC	Dancy mandarin	*C. x tangerina* (Tanaka)	180698706	60X
UNS	Satsuma mandarin	*C. x unshiu* (Marc)	188558275	63X
PAR	Grapefruit cv. Marsh	*C. x paradisi* (Macfadyen)	534428187	178X
MXL	Mexican lime	*C. x aurantifolia* (Christm.) Swingle	160670087	53X
LMA	Rangpur lime	*C. x limonia* (Osbeck)	177312373	59X
LIM	Lemon cv. Eureka	*C. x limon* L. (Burm. F.)	341041115	114X
PON	Trifoliate orange	*Poncirus trifoliata* (L.) Raf.	410744971	137X
FOR	Kumquat, Nagami	*Fortunella margarita* (Lour.)	38133267	13X
CAL	Calamondin	*C. x madurensis* (Lour.)	47734506	16X
SVR	Chinese box orange	Severinia buxifolia (Poir)	145141150	48X
MIC	Micrantha, Biasong	*C. micrantha* (Wester)	169407177	56X
COR	Corsican citron	*C. medica* L.	59313741	20X
VEU	Mac Veu citron	*C. medica* L	188510474	63X
SNK	Sunki mandarin	C. *x sunki* (hayata)	189645086	63X
ICH	Ichang papeda	*C. ichangensis* (Swingle)	187957308	17X
BUD	Buddha´s hand citron	*C. medica* L.	182240696	62X
ADL	Eremorange	*Eremocitrus glauca x C. sinensis*	178340972	59X
HUM	Humpang citron	*C. medica L.*	208654674	69X
AFL	Australian finger lime	*Microcitrus australasica* (F Muell.)	203376652	68X
ARL	Australian round lime	*Microcitrus australis* Swingle	205354420	68X
ADR	Australian desert lime	*Eremocitrus glauca*	17607283	9.2X
AFR	Australian finger lime	*Microcitrus australasica* (F Muell.)	74982998	39X
ARR	Australian round lime	*Microcitrus australis* Swingle	48391420	25X
CSM	Changsha mandarin	*C. x reticulata* (*Blanco*)	186961315	46X
RRL	Red rough lemon	*C. x jambhiri (Lush)*	117836928	68X
SCM	Sun Chu Sha Kat	*C. reticulata* (Blanco)	72366584	38X
TBM	Tachibana mandarin	*C. tachibana (Mak)*	61787150	32X
KSH	Kishu mandarin	*C. x kinokuni* (Hort. ex Tanaka)	62761835	33x

Supplementary Table 4. Candidate SNPs associated with citrus acidity/palatability. The association study is based on a case-control GWAS analysis of n=37 accessions with known palatability, with conservative Bonferroni correction ($P=7.9x10^{-8}$) at alpha=0.05 significance level. SNPs in intergenic regions are labeled upstream (ups.) or downstream (downs.) of a gene if they are located within 1 kb from a neighboring gene.

Chr	Position	Ref:Alt	Beta	s.e.(beta)	P	Gene (region)	Annotation
1	415175	C:T	4.07E-01	5.24E-02	1.11E-09	Ciclev10008736 (intron)	COP9 SIGNALOSOME COMPLEX SUBUNIT 5
1	23512067	C:T	4.96E-01	3.55E-02	2.96E-08	Ciclev10007611 (exon)	TARGET OF EGR1 PROTEIN 1 (TOE 1)
1	23679916	C:A	-6.52E-01	9.14E-02	8.31E-09	Ciclev10007740 (intron)	RIBOSOMAL RNA METHYLTRANSFERASE NOP2-RELATED
1	24219222	A:G	-6.52E-01	9.14E-02	8.31E-09	Ciclev10010250 (downs.)	PTHR23155//PTHR23155:SF563 - LEUCINE-RICH REPEAT-CONTAINING PROTEIN
2	15484525	G:T	6.66E-01	1.03E-01	7.37E-08	Ciclev10015371 (intron)	S-ALKYL-THIOHYDROXIMATE LYASE SUR1-RELATED
2	15702160	G:T	5.27E-01	7.81E-02	2.85E-08	Ciclev10014095 (intron)	Phospholipid-translocating ATPase
5	35094706	G:A	3.87E-01	5.95E-02	6.30E-08	Ciclev10000105 (intron)	PTHR15245:SF20 - SYMPLEKIN
5	35098538	G:A	3.88E-01	5.98E-02	6.35E-08	Ciclev10000105 (intron)	PTHR15245:SF20 - SYMPLEKIN
8	325527	A:G	5.78E-01	8.66E-02	3.56E-08	Ciclev10028714 (exon)	NAD+-ISOCITRATE DEHYDROGENASE (IDH)
8	631678	T:C	6.48E-01	7.61E-02	1.13E-10	Ciclev10030330 (ups.)	PTHR31175:SF1 - SAUR-LIKE AUXIN-RESPONSIVE PROTEIN-RELATED
8	927020	C:T	6.61E-01	8.23E-02	5.05E-10	Ciclev10028228 (intron)	PEARLI 4
8	1149577	G:T	4.37E-01	6.72E-02	6.05E-08	Ciclev10028121 (ups.)	PTHR31490:SF3 - GLYCOSYL HYDROLASE FAMILY 10 PROTEIN
8	1149586	C:T	4.67E-01	6.31E-02	3.49E-09	Ciclev10028121 (ups.)	PTHR31490:SF3 - GLYCOSYL HYDROLASE FAMILY 10 PROTEIN
8	1174414	T:A	5.15E-01	6.89E-02	4.44E-09	Ciclev10028271 (ups.)	PTHR22950//PTHR22950:SF242 - AMINO ACID TRANSPORTER
8	1413967	A:G	4.66E-01	7.15E-02	5.81E-08	Ciclev10030436 (intron)	PROLINE IMINOPEPTIDASE
8	1651338	G:A	5.04E-01	7.53E-02	3.39E-08	Ciclev10027661 (exon)	SERINE/THREONINE-PROTEIN KINASE (MTOR)
8	1655701	G:T	5.04E-01	7.53E-02	3.39E-08	Ciclev10027661 (exon)	SERINE/THREONINE-PROTEIN KINASE (MTOR)
8	1722788	T:C	4.42E-01	6.77E-02	5.67E-08	Ciclev10027741 (intron)	PRE-MRNA-PROCESSING PROTEIN PRP40
8	2058824	C:T	4.24E-01	6.52E-02	6.15E-08	Ciclev10027948 (exon)	DNA POLYMERASE KAPPA
8	2060290	T:C	4.76E-01	6.18E-02	1.43E-09	Ciclev10027948 (intron)	DNA POLYMERASE KAPPA
8	2063416	T:C	4.49E-01	6.55E-02	2.02E-08	Ciclev10029201 (3´-UTR)	UNKNOWN
8	2137063	G:C	4.49E-01	6.55E-02	2.02E-08	Ciclev10028638 (intron)	SF18 - ATP-DEPENDENT CLP PROTEASE PROTEOLYTIC SUBUNIT
8	2174360	C:T	5.06E-01	7.00E-02	5.99E-09	Ciclev 10029274 (downs.)	PTHR31304:SF1 - LOB DOMAIN-CONTAINING PROTEIN 39
9	30789594	T:C	3.89E-01	5.06E-02	1.40E-09	Ciclev 10007189 (ups.)	KOG4524 - Uncharacterized conserved protein

References

1 Xu, Q. *et al.* The draft genome of sweet orange (Citrus sinensis). *Nature genetics* **45**, 59-66, doi:10.1038/ng.2472 (2013).

2 Wu, G. A. *et al.* Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nature biotechnology* **32**, 656-662, doi:10.1038/nbt.2906 (2014).

3 Wang, X. *et al.* Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. *Nature genetics* **49**, 765-772 (2017).

4 Swingle, W. T. & Reece, P. C. in *The citrus industry, revised 2nd ed., vol. 1, History, world distribution, botany, and varieties* Vol. 1 (eds W. Reuther, H.J. Webber, & L.D. Batchelor) 190–430 (University of California, Berkeley, California, USA, 1967).

5 Frost, H. B. Seed reproduction: development of gametes and embryos. *The citrus industry* **2**, 290-324 (1968).

6 Webb, D. M. & Knapp, S. J. DNA extraction from a previously recalcitrant plant genus. *Plant Molecular Biology Reporter* **8**, 180-185 (1990).

7 Carrier, G. *et al.* An efficient and rapid protocol for plant nuclear DNA preparation suitable for next generation sequencing methods. *Am J Bot* **98**, e13-15, doi:10.3732/ajb.1000371 (2011).

8 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, doi:10.1186/1471-2229-6-21 (2006).

9 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).

10 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

11 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* **59**, 307-321 (2010).

12 Carbonell-Caballero, J. *et al.* A Phylogenetic Analysis of 34 Chloroplast Genomes Elucidates the Relationships between Wild and Domestic Species within the Genus Citrus. *Molecular biology and evolution* **32**, 2015-2035, doi:10.1093/molbev/msv082 (2015).

13 Swingle, W. The botany of Citrus and its wild relatives of the orange subfamily (Family Rutacae, subfamily Aurnatiodideae). *The citrus industry* **1**, 399 (1943).

14 Tanaka, T. Species problem in Citrus. *Japanese Society for Promotion of Science* (1954).

15 Hodgson, R. W. *Horticultural varieties of citrus*. (Division of Agricultural Sciences, 1967).

16 Hirai, M., Mitsue, S., Kita, K. & Kajiura, I. A survey and isozyme analysis of wild mandarin, tachibana (Citrus tachibana (Mak.) Tanaka) growing in Japan. *Journal of the Japanese Society for Horticultural Science* **59**, 1-7 (1990).

17 Shimizu, T. *et al.* Hybrid Origins of Citrus Varieties Inferred from DNA Marker Analysis of Nuclear and Organelle Genomes. *PloS one* **11**, e0166969 (2016).

18 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).

19 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, doi:10.3732/ajb.0800214 (2008).

20 Team, R. C. R: A language and environment for statistical computing. (2013).

21 Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics* **93**, 278-288 (2013).

22 Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* **81**, 1084-1097 (2007).

23 Scora, R. W. Symposium on Biochemical Systematics, Genetics and Origin of Cultivated Plants .9. History and Origin of Citrus. *B Torrey Bot Club* **102**, 369-375, doi:Doi 10.2307/2484763 (1975).

24 Gulsen, O. & Roose, M. Chloroplast and nuclear genome analysis of the parentage of lemons. *Journal of the American Society for Horticultural science* **126**, 210-215 (2001).

25 Curk, F. *et al.* Phylogenetic origin of limes and lemons revealed by cytoplasmic and nuclear markers. *Ann Bot* **117**, 565-583 (2016).

26 Gillespie, J. H. *Population genetics: a concise guide*. (JHU Press, 2010).

27 Lee, W. C. Testing the genetic relation between two individuals using a panel of frequency-unknown single nucleotide polymorphisms. *Ann Hum Genet* **67**, 618-619, doi:10.1046/j.1529-8817.2003.00063.x (2003).

28 Li, Y., Cheng, Y., 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. *Journal of the American Society for Horticultural Science* **132**, 796-806 (2007).

29 Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology* **61**, 539-542 (2012).

30 Sanderson, M. J. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular biology and evolution* **19**, 101-109 (2002).

31 Xie, S. P., Manchester, S. R., Liu, K. N., Wang, Y. F. & Sun, B. N. Citrus Linczangensis Sp N., a Leaf Fossil of Rutaceae from the Late Miocene of Yunnan, China. *Int J Plant Sci* **174**, 1201-1207, doi:10.1086/671796 (2013).

32 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290 (2004).

33 Felsenstein, J. PHYLIP-phylogeny inference package (version 3.2). *cladistics* **5**, 164-166 (1989).

34 Beattie, G., Holford, P., Mabberley, D., Haigh, A. & Broadbent, P. in *Orlando, Florida, USA: International Conference of Huanglongbing Florida.* 25-57.

35 Wang, W., Li, H. & Chen, Z. Analysis of plastid and nuclear DNA data in plant phylogenetics—evaluation and improvement. *Science China Life Sciences* **57**, 280-286 (2014).

36 Chat, J., Decroocq, S. & Petit, R. J. A one–step organelle capture: gynogenetic kiwifruits with paternal chloroplasts. *Proceedings of the Royal Society of London B: Biological Sciences* **270**, 783-789 (2003).

37 Kim, S.-T. & Donoghue, M. J. Incongruence between cpDNA and nrITS trees indicates extensive hybridization within Eupersicaria (Polygonaceae). *Am J Bot* **95**, 1122-1135 (2008).

38 Escobar, J. S. *et al.* Multigenic phylogeny and analysis of tree incongruences in Triticeae (Poaceae). *BMC Evol Biol* **11**, 1 (2011).

39 Yu, W.-B., Huang, P.-H., Li, D.-Z. & Wang, H. Incongruence between nuclear and chloroplast DNA phylogenies in Pedicularis section Cyathophora (Orobanchaceae). *PLoS One* **8**, e74828 (2013).

40 Wang, Z.-H., Peng, H. & Kilian, N. Molecular phylogeny of the Lactuca alliance (Cichorieae subtribe Lactucinae, Asteraceae) with focus on their Chinese centre of diversity detects potential events of reticulation and chloroplast capture. *PloS one* **8**, e82692 (2013).

41 Krak, K., Caklová, P., Chrtek, J. & Fehrer, J. Reconstruction of phylogenetic relationships in a highly reticulate group with deep coalescence and recent speciation (Hieracium, Asteraceae). *Heredity* **110**, 138-151 (2013).

42 Nauheimer, L., Boyce, P. C. & Renner, S. S. Giant taro and its relatives: a phylogeny of the large genus Alocasia (Araceae) sheds light on Miocene floristic exchange in the Malesian region. *Mol Phylogenet Evol* **63**, 43-51 (2012).

43 Rogers, J. & Gibbs, R. A. Comparative primate genomics: emerging patterns of genome content and dynamics. *Nature reviews. Genetics* **15**, 347-359, doi:10.1038/nrg3707 (2014).

44 Zhou, J. in *Proceedings of the International Citrus Symposium Guangzhou, China, Nov. 5-8, 1990/edited by Huang Bangyan, Yang Qian.* ([Beijing]: International Academic Publishers, c1991.).

45 De Candolle, A. *Origine des plantes cultivées*. Vol. 43 (G. Baillière et cie, 1883).

46 Tanaka, T. Citologia: semi-centennial commemoration papers on citrus studies. *Osaka: Citrologia Supporting Foundation* **114** (1961).

47 Rajput, C. & Hari Babu, R. *Citriculture*. (Kalyani, 1985).

48 Zhang, W. in *Proceedings of the International Society of Citriculture/[International Citrus Congress, November 9-12, 1981, Tokyo,*

Japan; K. Matsumoto, editor]. (Shimizu, Japan: International Society of Citriculture, 1982-1983.).
49 He, S., Liu, G. & Xiang, D. in *Proc Int Soc Citriculture.* 100-103.
50 Gmitter, F. G. & Hu, X. L. The Possible Role of Yunnan, China, in the Origin of Contemporary Citrus Species (Rutaceae). *Econ Bot* **44**, 267-277, doi:Doi 10.1007/Bf02860491 (1990).
51 Scora, R. in *Citriculture: proceedings of the Sixth International Citrus Congress: Middle-East, Tel Aviv, Israel, March 6-11, 1988/scientific editors, R. Goren and K. Mendel, editor, N. Goren.* (Rehovot, Israel: Balaban, c1989.).
52 Hodgson, R. W. & Sciences, U. o. C. D. o. A. *Horticultural varieties of citrus*. (University of California, Division of Agricultural Sciences., 1967).
53 Tanaka, T. The discovery of Citrus tachibana in Formosa, and its scientific and industrial significance. *Studia Citrologia* **5**, 1-20 (1931).
54 Liu, G., He, S. & Li, W. Two new species of Citrus in China. *Acta Botanica Yunnanica* **12**, 287-289 (1990).
55 Chen, Z. in *Proc Citrus Germplasm conservation Workshop.* 85-95.
56 Hodgson, R. The citrus fruits of India. *Calif. Citrogr* **22**, 513-514 (1937).
57 Cooper, W. C. *Odyssey of the Orange in China*. (EO Painter Printing Company, 1990).
58 Bonavia, E. *The cultivated oranges and lemons, etc. of India and Ceylon*. (Applewood Books, 1888).
59 Sykes, S. in *Proceedings citrus germplasm conservation workshop.* 6-7.
60 Zhang, D. & Mabberly, D. Citrus (Rutaceae). *Fl. China* **11**, 90-96 (2008).
61 Fang, J., Wang, Z. & Tang, Z. *Atlas of woody plants in China: distribution and climate*. Vol. 1 (Springer Science & Business Media, 2011).
62 NESOM, G. Citrus trifoliata (Rutaceae): Review of biology and distribution in the USA. *Phytoneuron* **46**, 1-14 (2014).
63 Muellner, A., Vassiliades, D. & Renner, S. Placing Biebersteiniaceae, a herbaceous clade of Sapindales, in a temporal and geographic context. *Plant Systematics and Evolution* **266**, 233-252 (2007).
64 Fischer, T. C. & Butzmann, R. Citrus meletensis (Rutaceae), a new species from the Pliocene of Valdarno (Italy). *Plant systematics and evolution* **210**, 51-55 (1998).
65 Vavilov, N. I. in *Theoretical Basis for Plant Breeding* Vol. 1 17-75 (1935).
66 Tolkowsky, S. *Hesperides. A history of the culture and use of citrus fruits*. (J. Bale Sons & Curnow, London., 1938).
67 Webber, H. J. in *The Citrus Industry* Vol. 1 Ch. 1, 1-39 (Univ. California., 1967).
68 Wen, J., Zhang, J. Q., Nie, Z. L., Zhong, Y. & Sun, H. Evolutionary diversifications of plants on the Qinghai-Tibetan Plateau. *Frontiers in genetics* **5**, 4, doi:10.3389/fgene.2014.00004 (2014).
69 Favre, A. *et al.* The role of the uplift of the Qinghai-Tibetan Plateau for the evolution of Tibetan biotas. *Biol Rev* **90**, 236-253, doi:10.1111/brv.12107 (2015).

70 Wang, L., Schneider, H., Zhang, X.-C. & Xiang, Q.-P. The rise of the Himalaya enforced the diversification of SE Asian ferns by altering the monsoon regimes. *BMC Plant Biol* **12**, 1 (2012).

71 Renner, S. S. Available data point to a 4‐km‐high Tibetan Plateau by 40 Ma, but 100 molecular‐clock papers have linked supposed recent uplift to young node ages. *J Biogeogr* (2016).

72 Clift, P. D., Wan, S. M. & Blusztajn, J. Reconstructing chemical weathering, physical erosion and monsoon intensity since 25 Ma in the northern South China Sea: A review of competing proxies. *Earth-Sci Rev* **130**, 86-102, doi:10.1016/j.earscirev.2014.01.002 (2014).

73 Clift, P. D. *et al.* Correlation of Himalayan exhumation rates and Asian monsoon intensity. *Nature Geoscience* **1**, 875-880 (2008).

74 Valdiya, K. S. Emergence and evolution of Himalaya: reconstructing history in the light of recent studies. *Prog Phys Geog* **26**, 360-399, doi:10.1191/0309133302pp342ra (2002).

75 Valdiya, K. Rising Himalaya: Advent and intensification of monsoon. *CURRENT SCIENCE-BANGALORE*- **76**, 514-524 (1999).

76 Iglesias, D. J. *et al.* Physiology of citrus fruiting. *Brazilian Journal of Plant Physiology* **19**, 333-362 (2007).

77 Tadeo, F. R. *et al.* Molecular physiology of development and quality of citrus. *Advances in Botanical Research* **47**, 147-223 (2008).

78 Peel, M. C., Finlayson, B. L. & McMahon, T. A. Updated world map of the Köppen-Geiger climate classification. *Hydrology and earth system sciences discussions* **4**, 439-473 (2007).

79 Hooker, J. D. *On the Flora of Australia: Its Origin, Affinities, and Distribution, Being an Introductory Essay to the Flora of Tasmania*. (Lovell Reeve, 1859).

80 Golonka, J., Krobicki, M., Pajak, J., Nguyen, V. G. & Zuchiewicz, W. *Global plate tectonics and paleogeography of Southeast Asia*. (2006).

81 Sun, B.-N. *et al.* Reconstructing Neogene vegetation and climates to infer tectonic uplift in western Yunnan, China. *Palaeogeography, Palaeoclimatology, Palaeoecology* **304**, 328-336 (2011).

82 Kou, X.-Y., Ferguson, D. K., Xu, J.-X., Wang, Y.-F. & Li, C.-S. The reconstruction of paleovegetation and paleoclimate in the Late Pliocene of West Yunnan, China. *Climatic Change* **77**, 431-448 (2006).

83 Su, T. *et al.* Post-Pliocene establishment of the present monsoonal climate in SW China: evidence from the late Pliocene Longmen megaflora. *Climate of the Past* **9**, 1911-1920 (2013).

84 Thomas, D. C. *et al.* West to east dispersal and subsequent rapid diversification of the mega-diverse genus Begonia (Begoniaceae) in the Malesian archipelago. *J Biogeogr* **39**, 98-113, doi:10.1111/j.1365-2699.2011.02596.x (2012).

85 Richardson, J. E., Costion, C. M. & Muellner, A. N. in *Biotic Evolution and Environmental Change in Southeast Asia* (ed D.J. et al Gower) Ch. 6, 138-163 (Cambridge University Press, 2012).

86 Thomas, D. C. *et al.* Molecular phylogenetics and historical biogeography of the Meiogyne-Fitzalania clade (Annonaceae): Generic paraphyly and late Miocene-Pliocene diversification in Australasia and the Pacific. *Taxon* **61**, 559-575 (2012).

87 Hall, R. Southeast Asia's changing palaeogeography. *Blumea* **54**, 148-161, doi:10.3767/000651909X475941 (2009).

88 van Welzen, P. C., Slik, J. W. F. & Alahuhta, J. Plant distribution patterns and plate tectonics in Malesia. *Biol. Skr.* **55**, 199-217 (2005).

89 Li, Y., Dressler, S., Zhang, D. & Renner, S. S. More Miocene dispersal between Africa and Asia—the case of Bridelia (Phyllanthaceae). *Systematic Botany* **34**, 521-529 (2009).

90 Schaefer, H., Heibl, C. & Renner, S. S. Gourds afloat: a dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. *Proceedings of the Royal Society of London B: Biological Sciences* **276**, 843-851 (2009).

91 Chiang, T.-Y. & Schaal, B. A. Phylogeography of plants in Taiwan and the Ryukyu Archipelago. *Taxon* **55**, 31-41 (2006).

92 Huang, S.-F. Hypothesizing Origin, Migration Routes and Distribution Patterns of Gymnosperms in Taiwan. *Taiwania* **59**, 139-163 (2014).

93 Wang, P. & Sun, X. Last glacial maximum in China: comparison between land and sea. *Catena* **23**, 341-353 (1994).

94 Dobson, M. & Kawamura, Y. Origin of the Japanese land mammal fauna: allocation of extant species to historically-based categories. *The Quaternary Research* **37**, 385-395 (1998).

95 Voris, H. K. Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. *J Biogeogr* **27**, 1153-1167 (2000).

96 Huang, S.-F. & Lin, T.-P. Migration of Trochodendron aralioides (Trochodendraceae) in Taiwan and its adjacent areas. *Botanical Studies* **47**, 83-88 (2006).

97 Gibbard, P. & Cohen, K. M. Global chronostratigraphical correlation table for the last 2.7 million years. *Episodes* **31**, 243-247 (2008).

98 Baldwin, E. in *Biochemistry of fruit ripening* 107-149 (Springer, 1993).

99 Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* **44**, 821-824, doi:10.1038/ng.2310 (2012).

100 Meléndez-Hevia, E., Waddell, T. G. & Cascante, M. The puzzle of the Krebs citric acid cycle: assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. *Journal of Molecular Evolution* **43**, 293-303 (1996).

101 Guo, L.-X. *et al.* Citrate Accumulation-Related Gene Expression and/or Enzyme Activity Analysis Combined With Metabolomics Provide a Novel Insight for an Orange Mutant. *Scientific Reports* **6** (2016).

102 Huang, D., Zhao, Y., Cao, M., Qiao, L. & Zheng, Z.-L. Integrated systems biology analysis of transcriptomes reveals candidate genes for acidity control in developing fruits of sweet orange (Citrus sinensis L. Osbeck). *Frontiers in plant science* **7** (2016).

103 Ghosh, S. P. in *Proc. Citrus Germplasm Conservation Workshop* 96-105 (Brisbane Australia, 1997).
104 Yang, X. *et al.* Genetic diversity and phylogenetic relationships of citron (Citrus medica L.) and its relatives in southwest China. *Tree Genetics & Genomes* **11**, 1-13 (2015).
105 He, S. W. L., G.; Wenpin, L. Wild mandarin oranges. *China Proc. Int. Soc. Citriculture*, 113-121 (1988).
106 Chen, Z. in *Proc Citrus Germplasm conservation Workshop* (ed Bevington KB Sykes SR, Hailstones D) 85-95 (Brisbane. Australia, 1997).