

Genomics of the origin and evolution of *Citrus*

Guohong Albert Wu¹, Javier Terol², Victoria Ibanez², Antonio López-García², Estela Pérez-Román², Carles Borredá², Concha Domingo², Francisco R. Tadeo², Jose Carbonell-Caballero³, Roberto Alonso³, Franck Curk⁴, Dongliang Du⁵, Patrick Ollitrault⁶, Mikeal L. Roose⁷, Joaquin Dopazo^{3,8}, Frederick G. Gmitter Jr⁵, Daniel S. Rokhsar^{1,9,10} & Manuel Talon²

The genus *Citrus*, comprising some of the most widely cultivated fruit crops worldwide, includes an uncertain number of species. Here we describe ten natural citrus species, using genomic, phylogenetic and biogeographic analyses of 60 accessions representing diverse citrus germ plasms, and propose that citrus diversified during the late Miocene epoch through a rapid southeast Asian radiation that correlates with a marked weakening of the monsoons. A second radiation enabled by migration across the Wallace line gave rise to the Australian limes in the early Pliocene epoch. Further identification and analyses of hybrids and admixed genomes provides insights into the genealogy of major commercial cultivars of citrus. Among mandarins and sweet orange, we find an extensive network of relatedness that illuminates the domestication of these groups. Widespread pummelo admixture among these mandarins and its correlation with fruit size and acidity suggests a plausible role of pummelo introgression in the selection of palatable mandarins. This work provides a new evolutionary framework for the genus *Citrus*.

The genus *Citrus* and related genera (*Fortunella*, *Poncirus*, *Eremocitrus* and *Microcitrus*) belong to the angiosperm subfamily Aurantioideae of the Rutaceae family, which is widely distributed across the monsoon region from west Pakistan to north-central China and south through the East Indian Archipelago to New Guinea and the Bismarck Archipelago, northeastern Australia, New Caledonia, Melanesia and the western Polynesian islands¹. Native habitats of citrus and related genera roughly extend throughout this broad area (Extended Data Fig. 1a and Supplementary Table 1), although the geographical origin, timing and dispersal of citrus species across southeast Asia remain unclear. A major obstacle to resolving these uncertainties is our poor understanding of the genealogy of complex admixture in cultivated citrus, as has recently been shown². Some citrus are clonally propagated apomictically³ through nucellar embryony, that is, the development of non-sexual embryos originating in the maternal nucellar tissue of the ovule, and this natural process may have been co-opted during domestication; grafting is a relatively recent phenomenon⁴. Both modes of clonal propagation have led to the domestication of fixed (desirable) genotypes, including interspecific hybrids, such as oranges, limes, lemons, grapefruits and other types.

Under this scenario, it is not surprising that the current chaotic citrus taxonomy—based on long-standing, conflicting proposals^{5,6}—requires a solid reformulation consistent with a full understanding of the hybrid and/or admixture nature of cultivated citrus species. Here we analyse genome sequences of diverse citrus to characterize the diversity and evolution of citrus at the species level and identify citrus admixtures and interspecific hybrids. We further examine the network of relatedness among mandarins and sweet orange, as well as the pattern of the introgression of pummelos among mandarins for clues to the early stages of citrus domestication.

Diversity and evolution of the genus *Citrus*

To investigate the genetic diversity and evolutionary history of citrus, we analysed the genomes of 58 citrus accessions and two outgroup genera (*Poncirus* and *Severinia*) that were sequenced to high coverage, including recently published sequences^{2,3,7} as well as 30 new genome sequences described here. For our purpose, we do not include accessions related by somatic mutations. These sequences represent a diverse sampling of citrus species, their admixtures and hybrids (Supplementary Tables 2, 3 and Supplementary Notes 1, 2). Our collection includes accessions from eight previously unsequenced and/or unexamined citrus species, such as pure mandarins (*Citrus reticulata*), citron (*Citrus medica*), *Citrus micrantha* (a wild species from within the subgenus *Papeda*), Nagami kumquat (*Fortunella margarita*, also known as *Citrus japonica* var. *margarita*), and *Citrus ichangensis* (also known as *Citrus cavaleriei*; this species is also considered a *Papeda*), as well as three Australian citrus species (Supplementary Notes 3, 4). For each species, we have sequenced one or more pure accessions without interspecific admixture.

Local segmental ancestry of each accession can be delineated for both admixed and hybrid genotypes, based on genome-wide ancestry-informative single-nucleotide polymorphisms (Supplementary Note 5). Comparative genome analysis further identified shared haplotypes among the accessions (Supplementary Notes 6, 7). In particular, we demonstrate the F1 interspecific hybrid nature of Rangpur lime and red rough lemon (two different mandarin–citron hybrids), Mexican lime (a micrantha–citron hybrid) and calamondin (a kumquat–mandarin hybrid), and confirm, using whole-genome sequence data, the origins of grapefruit (a pummelo–sweet orange hybrid), lemon (a sour orange–citron hybrid) and eremorange (a sweet orange and *Eremocitrus glauca* (also known as *Citrus glauca*) hybrid). We also verified the parentage of Cocktail grapefruit, with low-acid pummelo as the seed parent and

¹US Department of Energy Joint Genome Institute, Walnut Creek, California, USA. ²Centro de Genómica, Instituto Valenciano de Investigaciones Agrarias (IVIA), Moncada, Valencia, Spain.

³Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain. ⁴AGAP Research Unit, Institut National de la Recherche Agronomique (INRA), San Giuliano, France. ⁵Citrus Research and Education Center (CREC), Institute of Food and Agricultural Sciences (IFAS), University of Florida, Lake Alfred, Florida, USA. ⁶AGAP Research Unit, Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), Petit-Bourg, Guadeloupe, France. ⁷Department of Botany and Plant Sciences, University of California, Riverside, Riverside, California, USA. ⁸Functional Genomics Node, Spanish National Institute of Bioinformatics (ELIXIR-es) at CIPF, Valencia, Spain. ⁹Department of Molecular and Cell Biology and Center for Integrative Genomics, University of California, Berkeley, Berkeley, California, USA. ¹⁰Molecular Genetics Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan.

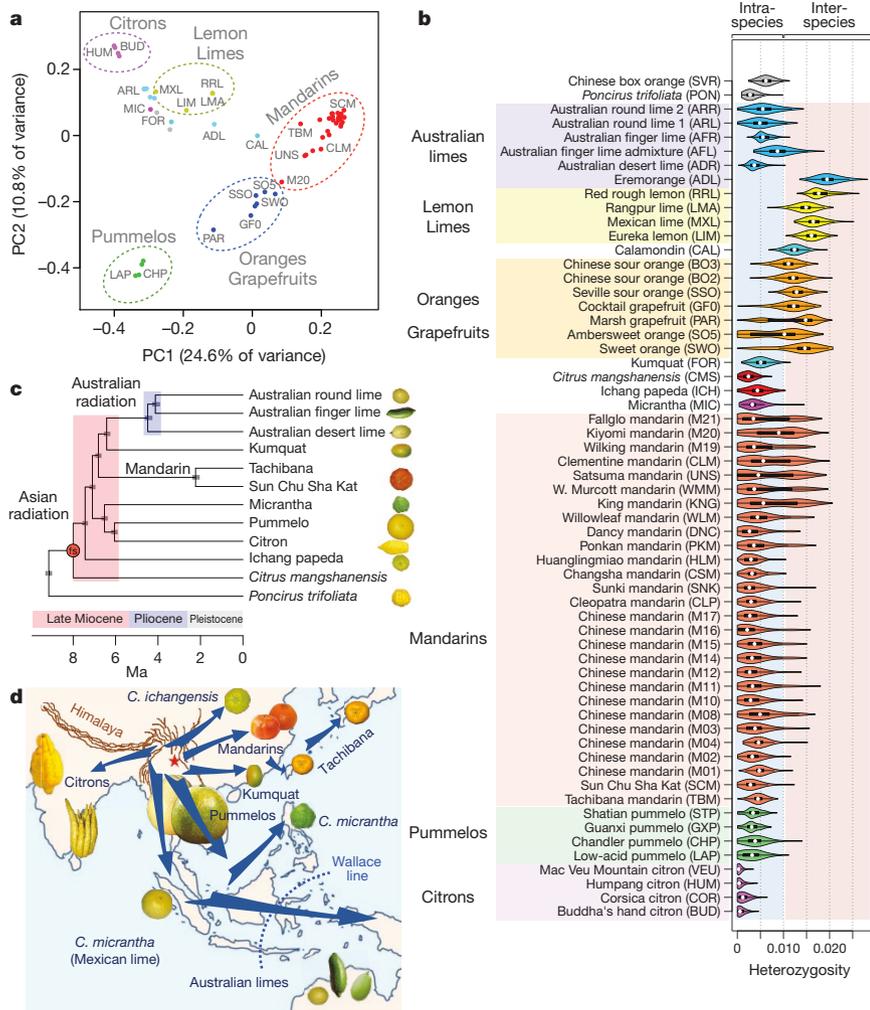


Figure 1 | Genetic structure, heterozygosity and phylogeny of *Citrus* species. **a**, Principal coordinate analysis of 58 citrus accessions based on pairwise nuclear genome distances and metric multidimensional scaling. The first two axes separate the three main citrus groups (citrons, pummelos and mandarins) with interspecific hybrids (oranges, grapefruit, lemon and limes) situated at intermediate positions relative to their parental genotypes. **b**, Violin plots of the heterozygosity distribution in 58 citrus accessions, representing 10 taxonomic groups as well as 2 related genera, *Poncirus* (*Poncirus trifoliata*, also known as *Citrus trifoliata*) and Chinese box orange (*Severinia*). White dot, median; bar limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range. The bimodal separation of intraspecies (light blue) and interspecies (light pink) genetic diversity is manifested among the admixed mandarins and across different genotypes including interspecific hybrids. Three-letter codes are listed in parenthesis with additional descriptions in Supplementary Table 2. **c**, Chronogram of citrus speciation. Two distinct and temporally well-separated phases of species radiation are apparent, with the southeast Asian citrus radiation followed by the Australian citrus diversification. Age calibration is based on the citrus fossil *C. linczangensis*¹⁶ from the Late Miocene (denoted by a filled red circle). The 95% confidence intervals are derived from 200 bootstraps. Bayesian posterior probability is 1.0 for all nodes. **d**, Proposed origin of citrus and ancient dispersal routes. Arrows suggest plausible migration directions of the ancestral citrus species from the centre of origin—the triangle formed by northeastern India, northern Myanmar and northwestern Yunnan. The proposal is compatible with citrus biogeography, phylogenetic relationships, the inferred timing of diversification and the paleogeography of the region, especially the geological history of Wallacea and Japan. The red star marks the fossil location of *C. linczangensis*. Citrus fruit images in **c** and **d** are not drawn to scale.

King and Dancy mandarins as the two grandparents on the paternal side. The origin of the Ambersweet orange is similarly confirmed to be a mandarin–sweet orange hybrid with Clementine as a grandparent. We have previously shown that sour orange (cv. Seville) (*Citrus aurantium*) is a pummelo–mandarin hybrid, and have analysed the more complex origin of sweet orange (*Citrus sinensis*)². Re-analysing sequences from ten cultivars of sweet orange³ shows that they are all derived from the same genome by somatic mutations, and were thus not included in our study.

We identified ten progenitor citrus species (Supplementary Note 4.1) by combining diversity analysis (Extended Data Table 1), multi-dimensional scaling and chloroplast genome phylogeny (Extended Data Fig. 1b). The first two principal coordinates in the multidimensional scaling (Fig. 1a) separate three ancestral (sometimes called ‘fundamental’) *Citrus* species associated with commercially important types^{8,9}—citrons (*C. medica*), mandarins (*C. reticulata*) and pummelos (*Citrus maxima*)—and display lemons, limes, oranges and grapefruits as hybrids involving these three species. The nucleotide diversity distributions (Fig. 1b) show distinct scales for interspecific divergence and intraspecific variation, and reflect the genetic origin of each accession. Hybrid accessions (sour orange, calamondin, lemon and non-Australian limes) with ancestry from two or more citrus species are readily identified on the basis of their higher segmental heterozygosity (1.5–2.4%) relative to intraspecific diversity (0.1–0.6%). Other citrus accessions show bimodal distributions in heterozygosity (sweet orange, grapefruits and some highly heterozygous mandarins) due to interspecific admixture, a process that generally involves complex backcrosses. Among the pure genotypes without interspecific admixture, citrons

show significantly lower intraspecific diversity (around 0.1%) than the other species (0.3–0.6%). The reduced heterozygosity of citrons, a mono-embryonic species, is probably due to the cleistogamy of its flowers¹⁰, a mechanism that promotes pollination and self-fertilization in unopened flower buds, which in turn reduces heterozygosity.

The identification of a set of pure citrus species provides new insights into the phylogeny of citrus, their origins, evolution and dispersal. Citrus phylogeny is controversial^{1,5,6,11,12}, in part owing to the difficulty of identifying pure or wild progenitor species, because of substantial interspecific hybridization that has resulted in several clonally propagated and cultivated accessions. Some authors assign separate binomial species designations to clonally propagated genotypes^{1,6}. Our nuclear genome-based phylogeny, which is derived from 362,748 single-nucleotide polymorphisms in non-genic and non-pericentromeric genomic regions, reveals that citrus species are a monophyletic group and establishes well-defined relationships among its lineages (Fig. 1c and Supplementary Note 8). Notably, the nuclear genome-derived phylogeny differs in detail from the chloroplast-derived phylogeny (Extended Data Fig. 1). This is not unexpected, as chloroplast DNA is a single, non-recombining unit and is unlikely to show perfect lineage sorting during rapid radiation (Supplementary Note 8.3).

The origin of citrus has generally been considered to be in southeast Asia¹, a biodiversity hotspot¹³ with a climate that has been influenced by both east and south Asian monsoons¹⁴ (Supplementary Note 9). Specific regions include the Yunnan province of southwest China¹⁵, Myanmar and northeastern India in the Himalayan foothills¹. A fossil specimen from the late Miocene epoch of Lincang in Yunnan, *Citrus linczangensis*¹⁶, has traits that are characteristic of current major citrus

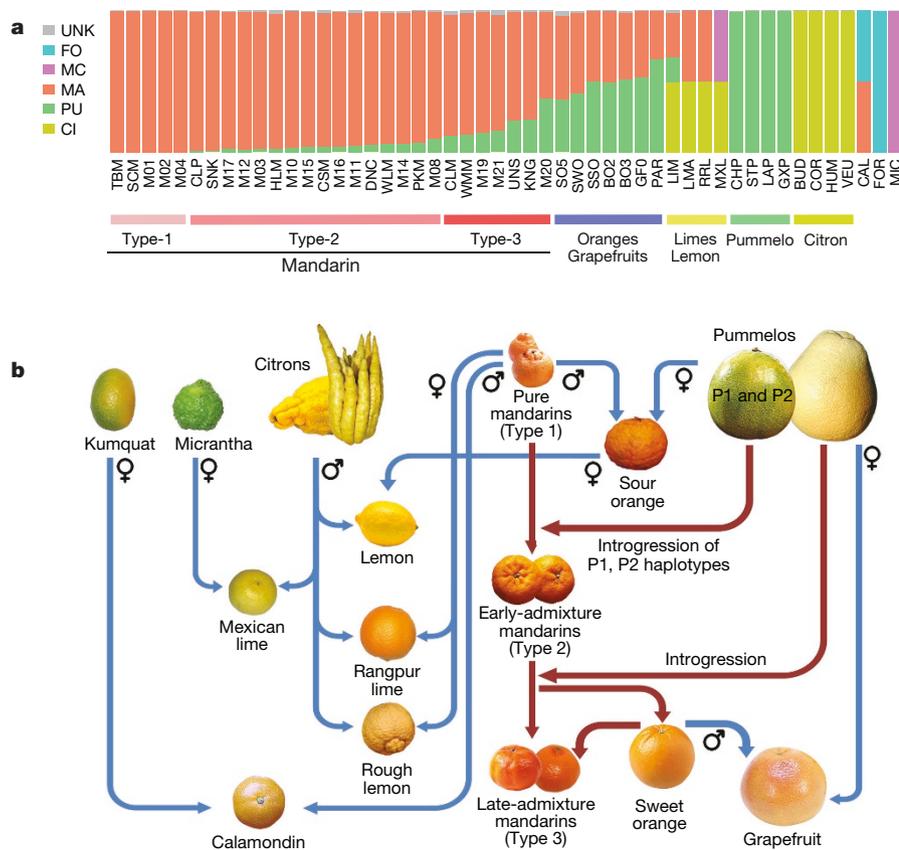


Figure 2 | Admixture proportion and citrus genealogy. **a**, Allelic proportion of five progenitor citrus species in 50 accessions. CI, *C. medica*; FO, *Fortunella*; MA, *C. reticulata*; MC, *C. micrantha*; PU, *C. maxima*; UNK, unknown. The pummelos and citrons represent pure citrus species, whereas in the heterogeneous set of mandarins, the degree of pummelo introgression subdivides the group into pure (type-1) and admixed (type-2 and -3) mandarins. Three-letter code as in Fig. 1, see Supplementary Table 2 for details. **b**, Genealogy of major citrus genotypes. The five progenitor species are shown at the top. Blue lines represent simple crosses between two parental genotypes, whereas red lines represent more complex processes involving multiple individuals, generations and/or backcrosses. Whereas type-1 mandarins are pure species, type-2 (early-admixture) mandarins contain a small amount of pummelo admixture that can be traced back to a common pummelo ancestor (with P1 or P2 haplotypes). Later, additional pummelo introgressions into type-2 mandarins and sweet orange. Further breeding between sweet orange and mandarins or within late-admixture mandarins produced additional modern mandarins. Fruit images are not to scale and represent the most popular citrus types. See Supplementary Note 1.1 for nomenclature usage.

groups, and provides definite evidence for the existence of a common *Citrus* ancestor within the Yunnan province approximately 8 million years ago (Ma).

Our analysis establishes a relatively rapid Asian radiation of citrus species in the late Miocene (6–8 Ma; Fig. 1c, d), a period coincident with an extensive weakening of monsoons and a pronounced climate transition from wet to drier conditions¹⁷. In southeast Asia, this marked climate alteration caused major changes in biota, including the migration of mammals¹⁸ and rapid radiation of various plant lineages^{19,20}. Australian citrus species form a distinct clade that was proposed to be nested with citrons¹², although distinct generic names (*Eremocitrus* and *Microcitrus*) were assigned in botanical classifications by Swingle^{1,5}. Both molecular dating analysis²¹ and our whole-genome phylogenetic analysis do not support an Australian origin for citrus²². Rather, citrus species spread from southeast Asia to Australasia, probably via trans-oceanic dispersals. Our genomic analysis indicates that the Australian radiation occurred during the early Pliocene epoch, around 4 Ma. This is contemporaneous with other west-to-east angiosperm migrations from southeast Asia^{23,24}, presumably taking advantage of the elevation of Malesia and Wallacea in the late Miocene and Pliocene^{25,26} (Supplementary Note 9).

The nuclear and chloroplast genome phylogenies indicate that there are three Australian species in our collection. One of the two Australian finger limes shows clear signs of admixture with round limes (Supplementary Note 5.4). The closest relative to Australian citrus is *Fortunella*, a species that has been reported to grow in the wild in southern China²⁷. Australian citrus species are diverse, and found natively in both dry and rainforest environments in northeast Australia, depending on the species²⁸. Our phylogeny shows that the progenitor citrus probably migrated across the Wallace line, a natural barrier for species dispersal from southeast Asia to Australasia, and later adapted to these diverse climates.

The results also show that the Tachibana mandarin, naturally found in Taiwan, the Ryukyu archipelago and Japan²⁹, split from mainland

Asian mandarins (Fig. 1c, d) during the early Pleistocene (around 2 Ma), a geological epoch with strong glacial maxima³⁰. Tachibana, as did other flora and fauna in the region, very probably arrived in these islands from the adjacent mainland³¹ during the drop in the sea level of the South China Sea and the emergence of land bridges^{32,33}, a process promoted by the expansion of ice sheets that repetitively occurred during glacial maxima (Supplementary Note 9).

Although Tachibana^{5,6} has been assigned its own species (*Citrus tachibana*), sequence analysis reveals that it has a close affinity to *C. reticulata*^{34,35} and does not support its taxonomic position as a separate species (Supplementary Note 4.1). However, both chloroplast genome phylogeny (Extended Data Fig. 1b) and nuclear genome clustering (Fig. 1a) clearly distinguish Tachibana from the mainland Asian mandarins. This suggests that Tachibana should be designated a subspecies of *C. reticulata*. By contrast, the wild Mangshan ‘mandarin’ (*Citrus mangshanensis*)⁷ represents a distinct species, with comparable distances to *C. reticulata*, pummelo and citron² (Extended Data Table 1).

Pattern of pummelo admixture in the mandarins

Using 588,583 ancestry-informative single-nucleotide polymorphisms derived from three species, *C. medica*, *C. maxima* and *C. reticulata*, we delineate the segmental ancestry of 46 citrus accessions (Extended Data Fig. 2 and Supplementary Note 5). Pummelo admixture is found in all but 5 of the 28 sequenced mandarins, and the amount and pattern of pummelo admixture, as identified by phased pummelo haplotypes (Fig. 2a and Supplementary Note 6), suggests the classification of the mandarins into three types.

Type-1 mandarins represent pure *C. reticulata* with no evidence of interspecific admixture and include Tachibana, three unnamed Chinese mandarins (M01, M02, M04)³ and the ancient Chinese cultivar Sun Chu Sha Kat reported here, a small tart mandarin commonly grown in China and Japan, and also found in Assam. This cultivar is likely described in Han Yen-Chih’s AD 1178 monograph ‘Chü Lu’³⁶, which

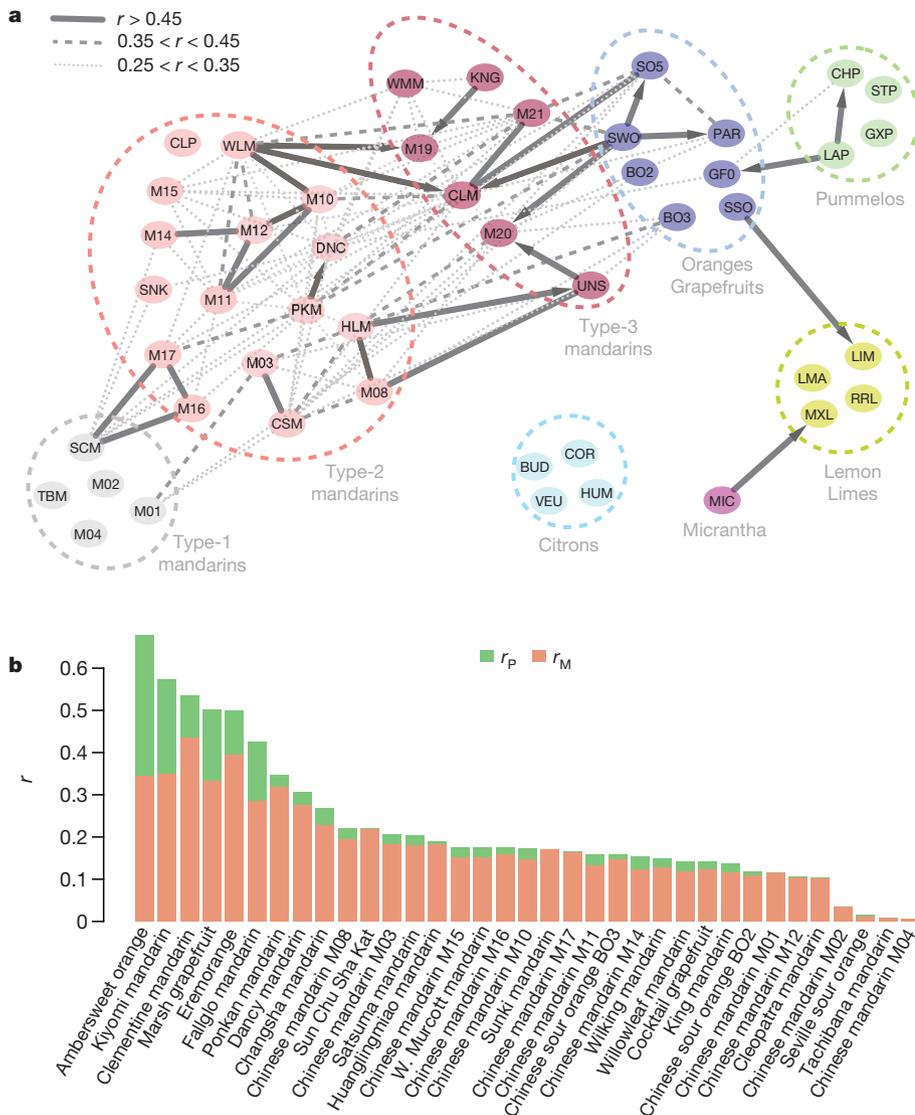


Figure 3 | Citrus relatedness network and haplotype sharing with sweet orange.

a, Genetic relatedness among 48 citrus accessions derived from four progenitor species including citrons, pummelos, pure mandarins and micrantha. Solid lines connect pairs with coefficient of relatedness $r > 0.45$, with parent–child pairs denoted by arrows pointing from parent to child. Dashed and dotted lines correspond to $0.35 < r < 0.45$ and $0.25 < r < 0.35$, respectively. Mandarins are distinguished from other taxonomic groups by an extensive relatedness network, indicating shared haplotypes in the ancestral gene pool. Three-letter code as in Fig. 1, see Supplementary Table 2 for details. **b**, Shown in decreasing order are the values of coefficient of relatedness between sweet orange and other accessions, with *C. maxima* (r_P) and *C. reticulata* (r_M) components in green and light salmon, respectively. There is significant haplotype sharing between sweet orange and all mandarins, except for three of the type-1 mandarins. Five accessions (Clementine and Kiyomi mandarins, eremorange, Marsh grapefruit, and Ambersweet orange) have sweet orange as the male parent.

includes references to citrus cultivated during the reign of Emperor Ta Yu (2205–2197 BC). Sixteen of the twenty-eight mandarins belong to type-2 mandarins, which have a small amount of pummelo admixture (1–10% of the length of the genetic map; Fig. 2a), usually in the form of a few short segments distributed across the genome. Although the lengths and locations of these admixed segments may be distinct in different mandarins, they share one or two common pummelo haplotypes (designated as P1 and P2) (Extended Data Fig. 3). By contrast, the seven remaining mandarins (type-3) contain higher proportions of pummelo alleles (12–38%; Fig. 2a) in longer segments. Although the P1 and P2 pummelo haplotypes are also detectable among type-3 mandarins, other more extensive pummelo haplotypes dominate the pummelo admixture in type-3 mandarins (Fig. 2b and Extended Data Table 2).

These observations suggest that the initial pummelo introgression into the mandarin gene pool may have involved as few as one pummelo tree (carrying both P1 and P2 haplotypes), the contribution of which was diluted by repeated backcrosses with mandarins (Supplementary Note 6.3). The introgressed pummelo haplotypes became widespread and gave rise to type-2 (early-admixture) mandarins (Fig. 2b). We propose that later, additional pummelo introgressions gave rise to type-3 (late-admixture) mandarins and sweet orange, and that some modern type-3 mandarins were derived from hybridizations among existing mandarins and sweet orange. This late-admixture model for type-3 mandarins is consistent with the historical records for Clementine and

Kiyomi (both mandarin–sweet orange hybrids), and for W. Murcott, Wilking and Fallglo (hybrids involving other type-3 mandarins), whereas definitive records for the remaining two late-admixture mandarins (King and Satsuma) are not available.

Domestication of mandarins and sweet orange

Citrus domestication probably began with the identification and asexual propagation of selected, possibly hybrid or admixed individuals, rather than recurrent selection from a breeding population as for annual crops^{37,38}. Additional diversity was obtained by capturing somatic mutations that occur within a relatively few basic genotypes. Therefore, conventional approaches to identifying selective pressures under recurrent breeding³⁹ cannot be applied. We can, however, use genome sequences to infer some features of the early stages of citrus domestication. Here we focus on mandarins, a class of citrus comprising small and easily peeled fruits that are of high commercial value.

All 28 mandarin accessions, except for Tachibana, exhibit an extensive network of relatedness (with a coefficient of relatedness, $r > 1/8$), and all but four mandarins (three of the four are pure or type-1 mandarins) show second degree or higher relatedness ($r > 1/4$) to at least one (mean = 7) other mandarin (Fig. 3a and Supplementary Note 7). By contrast, sequenced pummelos and citrons appear to be independent selections from relatively large populations. In the

absence of historical records for most mandarins, the actual kinships are difficult to infer, owing to extensive haplotype sharing among the ancestors, although some parent–child pairs can be identified. In addition to confirming, using the whole-genome sequence, the parentage of Wilking (King–Willowleaf), Kiyomi (Satsuma–sweet orange) and Fallgo (one grandparent is Clementine), we find parent–child relationships between two pairs of mandarins (Ponkan is a parent of Dancy; Huanglingmiao (a somatic mutant of Kishu) is a parent of Satsuma)³⁴, in addition to the previously established parent–child pair of Willowleaf and Clementine mandarins². Additional parent–child pairs involving the recently sequenced Chinese mandarins³ are also identified (Supplementary Note 7.3). A few cultivar types in this network (Satsuma, Dancy, Clementine, Kiyomi, Fallgo and the Chinese cultivar BTJ mandarins) have marked signs of inbreeding, indicated by runs of homozygosity (Extended Data Fig. 4a) as a result of shared haplotypes between their parents. The high degree of relatedness among mandarins implies extensive sharing of *C. reticulata* haplotypes.

Sweet orange also shows extensive haplotype sharing at the level of $r > 0.1$ with 25 of the 28 sequenced mandarins (except for three pure or type-1 mandarins; Fig. 3b and Extended Data Fig. 4b). Two late-admixture mandarins (Clementine and Kiyomi) are direct offspring of sweet orange. Among the early-admixture (type-2) mandarins, Ponkan shows the highest affinity to sweet orange² with $r \approx 0.36$. Even the pure mandarin, Sun Chu Sha Kat has $r \approx 0.23$, equivalent to second degree relatedness to sweet orange. We can rule out the scenario that sweet orange is the common ancestor of the mandarins, because of a lack of pummelo haplotypes (derived from sweet orange) among the mandarins. Rather, the extensive *C. reticulata* haplotype sharing between sweet orange and mandarins suggests that the mandarin parent of sweet orange was part of an expansive network of relatedness among mandarins.

Because our collection of mandarins represents a diverse set of both ancient and modern varieties, including economically important accessions with mostly unknown parentage, the presence of an extensive relatedness network was not anticipated a priori. The shared *C. reticulata* haplotypes are suggestive of and consistent with signatures of the human selection process, during which mandarins with desirable traits were necessarily maintained through clonal propagation (nucellar polyembryony or grafting). Although one cannot preclude the possibility that the relatedness network was initiated before domestication from a small number of founder trees, human selection of accessions resulting from natural hybridization probably had a key role in the process of domestication that eventually led to the extensive relatedness network observed today. For example, modern mandarins, such as Clementine and W. Murcott, are known to be selections from chance seedlings found in Algeria⁴⁰ and Morocco², at the onset and middle of the last century, respectively.

Pummelo admixture is correlated with fruit size and acidity, suggesting a role for pummelo introgression in citrus domestication. As both fruit size and acidity profile for the most recently sequenced accessions³ are not described, we used 37 citrus accessions in this analysis. We find that the fruit sizes of mandarins, oranges, grapefruit and pummelos show a strong positive correlation (Pearson correlation coefficient $r = 0.88$) with pummelo admixture proportion (Extended Data Fig. 5a, b and Supplementary Note 10.1). In addition to fruit size, a pivotal driver of fruit domestication is palatability, a characteristic that in citrus requires low to moderate levels of acidity. In mandarins, palatability appears to be linked to pummelo introgression at a major locus at the start of chromosome 8 (0.3–2.2 Mb), where all nine known palatable mandarins, but none of the four known acidic mandarins, show pummelo admixture in at least part of the genomic region (Extended Data Fig. 3). This locus is also found to be significant in a genome scan for palatability association (Extended Data Fig. 5c, d and Extended Data Table 3) and contains several potentially relevant genes (Supplementary Note 10.2). Among these genes is a gene encoding

the mitochondrial NAD⁺-dependent isocitrate dehydrogenase (*IDH*) which regulates citric-acid synthesis⁴¹ (Extended Data Table 4).

Our study finds that domesticated citrus fruit crops, such as mandarins and sweet orange, experienced a complex history of admixture, conceptually similar to those well-recognized in annual crops, such as rice⁴² and maize⁴³, and in other fruit trees, such as apple⁴⁴ and grape⁴⁵, for which the current genomic diversity is linked to widespread ancient introgression. Other cultivated citrus groups, the interspecific F1 hybrids in particular, originated from hybridizations of two pure parental species. Several of these involve *C. medica* (citron), including limes and lemons¹⁰. A unique and critical characteristic of the three pivotal species (*C. maxima*, *C. reticulata* and *C. medica*) that gave rise to most cultivated citrus fruits is the occurrence of a complex floral anatomy (Extended Data Fig. 6), thus leading to the development of more complex fruit. Other species were also involved in hybridizations, including *Fortunella* and *C. micrantha*. Distinct from the mandarin lineages, these hybrids are characterized by their acidic fruit, and their selection must have been made on the basis of other characteristics, such as a sweet edible peel and aroma², respectively.

Conclusion

On the basis of genomic, phylogenetic and biogeographic analyses of 60 diverse citrus and related accessions, we propose that the centre of origin of citrus species was the southeast foothills of the Himalayas, in a region that includes the eastern area of Assam, northern Myanmar and western Yunnan. Our analyses suggest that the ancestral citrus species underwent a sudden speciation event during the late Miocene. This radiation coincided with a pronounced transition from wet monsoon conditions to a drier climate, as observed in nearby areas in many other plant and animal lineages. The Australian citrus species and Tachibana, a native Japanese mandarin, split later from mainland citrus during the early Pliocene and Pleistocene, respectively. By distinguishing between pure species, hybrids and admixtures, we could trace the genealogy and genetic origin of the major citrus commercial cultivars. Both the extensive relatedness network among mandarins and sweet orange, and the association of pummelo admixture with desirable fruit traits suggest a complex domestication process.

Our work challenges previous proposals for citrus taxonomy. For example, we find that several named genera (*Fortunella*, *Eremocitrus* and *Microcitrus*) are in fact nested within the citrus clade. These and other distinct clades that we have identified are therefore more appropriately considered species within the genus *Citrus*, on a par with those that formerly were referred to as the three ‘true’ or ‘biological’ species (*C. reticulata*, *C. maxima* and *C. medica*). Additionally, the related genus, *Poncirus*, a subject of continuous controversy since it was originally proposed to be within the genus *Citrus*^{12,46}, is clearly a distinct clade that is separate from *Citrus* based on sequence divergence and whole-genome phylogeny.

In summary, this work presents insights into the origin, evolution and domestication of citrus, and the genealogy of the most important wild and cultivated varieties. Taken together, these findings draw a new evolutionary framework for these fruit crops, a scenario that challenges current taxonomic and phylogenetic thoughts, and points towards a reformulation of the genus *Citrus*.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 November 2016; accepted 10 December 2017.

Published online 7 February 2018.

- Swingle, W. T. & Reece, P. C. in *The Citrus Industry, revised 2nd edn, History, World Distribution, Botany, and Varieties* Vol. 1 (eds Reuther, W. et al.) 190–430 (Univ. California, 1967).

2. Wu, G. A. *et al.* Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat. Biotechnol.* **32**, 656–662 (2014).
3. Wang, X. *et al.* Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. *Nat. Genet.* **49**, 765–772 (2017).
4. Mudge, K., Janick, J., Scofield, S. & Goldschmidt, E. E. A history of grafting. *Hortic. Rev. (Am. Soc. Hortic. Sci.)* **35**, 437–493 (2009).
5. Swingle, W. in *The Citrus Industry, History Botany and Breeding* Vol. 1 (eds Webber, H. J. & Batchelor, L. D.) 129–474 (Univ. California Press, 1943).
6. Tanaka, T. *Species Problem in Citrus* (Japanese Society for Promotion of Science, 1954).
7. Xu, Q. *et al.* The draft genome of sweet orange (*Citrus sinensis*). *Nat. Genet.* **45**, 59–66 (2013).
8. Barrett, H. & Rhodes, A. A numerical taxonomic study of affinity relationships in cultivated *Citrus* and its close relatives. *Syst. Bot.* **1**, 105–136 (1976).
9. Scora, R. W. On the history and origin of *Citrus*. *Bull. Torrey Bot. Club* **102**, 369–375 (1975).
10. Curk, F. *et al.* Phylogenetic origin of limes and lemons revealed by cytoplasmic and nuclear markers. *Ann. Bot.* **117**, 565–583 (2016).
11. Nicolosi, E. *et al.* Citrus phylogeny and genetic origin of important species as investigated by molecular markers. *Theor. Appl. Genet.* **100**, 1155–1166 (2000).
12. Bayer, R. J. *et al.* A molecular phylogeny of the orange subfamily (Rutaceae: Aurantioideae) using nine cpDNA sequences. *Am. J. Bot.* **96**, 668–685 (2009).
13. Jacques, F. M. *et al.* Late Miocene southwestern Chinese floristic diversity shaped by the southeastern uplift of the Tibetan Plateau. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **411**, 208–215 (2014).
14. Jacques, F. M. *et al.* Quantitative reconstruction of the Late Miocene monsoon climates of southwest China: a case study of the Lincang flora from Yunnan Province. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **304**, 318–327 (2011).
15. 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 (1990).
16. Xie, S., Manchester, S. R., Liu, K., Wang, Y. & Sun, B. *Citrus linczangensis* sp. n., a leaf fossil of Rutaceae from the late Miocene of Yunnan, China. *Int. J. Plant Sci.* **174**, 1201–1207 (2013).
17. Clift, P. D., Wan, S. & 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 (2014).
18. Valdiya, K. S. Emergence and evolution of Himalaya: reconstructing history in the light of recent studies. *Prog. Phys. Geogr.* **26**, 360–399 (2002).
19. Wen, J., Zhang, J. Q., Nie, Z. L., Zhong, Y. & Sun, H. Evolutionary diversifications of plants on the Qinghai–Tibetan Plateau. *Front. Genet.* **5**, 4 (2014).
20. Favre, A. *et al.* The role of the uplift of the Qinghai–Tibetan Plateau for the evolution of Tibetan biotas. *Biol. Rev. Camb. Philos. Soc.* **90**, 236–253 (2015).
21. Pfeil, B. E. & Crisp, M. D. The age and biogeography of *Citrus* and the orange subfamily (Rutaceae: Aurantioideae) in Australasia and New Caledonia. *Am. J. Bot.* **95**, 1621–1631 (2008).
22. Beattie, G. A. C., Holford, P., Maberley, D. J., Haigh, A. M. & Broadbent, P. in *On the origins of Citrus, Huanglongbing, Diaphorina citri and Trioza erytraea. International Conference of Huanglongbing* (eds Gottwald, T. R. & Graham, J. H.) 25–57 (Plant Management Network, 2009).
23. 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 (2012).
24. Richardson, J. E., Costion, C. M. & Muellner, A. N. in *Biotic Evolution and Environmental Change in Southeast Asia* Ch. 6 (eds Gower, D. *et al.*) 138–163 (Cambridge Univ. Press, 2012).
25. van Welzen, P. C., Slik, J. W. F. & Alahuhta, J. Plant distribution patterns and plate tectonics in Malesia. *Biol. Skr.* **55**, 199–217 (2005).
26. Hall, R. Southeast Asia's changing palaeogeography. *Blumea* **54**, 148–161 (2009).
27. Zhang, W. *Thirty years achievements in citrus varietal improvement in China in Proc. International Citrus Congress* (ed. Matsumoto, K.) 51–53 (International Society of Citriculture, 1982–1983).
28. Brophy, J. J., Goldsack, R. J. & Forster, P. I. The leaf oils of the Australian species of *Citrus* (Rutaceae). *J. Essent. Oil Res.* **13**, 264–268 (2001).
29. Tanaka, T. The discovery of *Citrus tachibana* in Formosa, and its scientific and industrial significance. *Studia Citologica* **5**, 1–20 (1931).
30. Gibbard, P. & Cohen, K. M. Global chronostratigraphical correlation table for the last 2.7 million years. *Episodes* **31**, 243–247 (2008).
31. Chiang, T.-Y. & Schaal, B. A. Phylogeography of plants in Taiwan and the Ryukyu Archipelago. *Taxon* **55**, 31–41 (2006).
32. Voris, H. K. Maps of Pleistocene sea levels in southeast Asia: shorelines, river systems and time durations. *J. Biogeogr.* **27**, 1153–1167 (2000).
33. Huang, S.-F. Hypothesizing origin, migration routes and distribution patterns of gymnosperms in Taiwan. *Taiwania* **59**, 139–163 (2014).
34. Shimizu, T. *et al.* Hybrid origins of citrus varieties inferred from DNA marker analysis of nuclear and organelle genomes. *PLoS ONE* **11**, e0166969 (2016).
35. Hirai, M., Mitsue, S., Kita, K. & Kajjura, I. A survey and isozyme analysis of wild mandarin, *Tachibana* (*Citrus tachibana* (Mak.) Tanaka) growing in Japan. *J. Jpn. Soc. Hortic. Sci.* **59**, 1–7 (1990).
36. Hagerly, M. J. Han Yen-Chih's Chü lu (monograph on the oranges of Wên-chou, Chekiang). *Toung Pao* **22**, 63–96 (1923).
37. Miller, A. J. & Gross, B. L. From forest to field: perennial fruit crop domestication. *Am. J. Bot.* **98**, 1389–1414 (2011).
38. Gaut, B. S., Díez, C. M. & Morrell, P. L. Genomics and the contrasting dynamics of annual and perennial domestication. *Trends Genet.* **31**, 709–719 (2015).
39. Hamblin, M. T., Buckler, E. S. & Jannink, J. L. Population genetics of genomics-based crop improvement methods. *Trends Genet.* **27**, 98–106 (2011).
40. Trabut, J. L'hybridation des Citrus: une nouvelle tangéline 'la Clémentine'. *Revue Horticole* **10**, 232–234 (1902).
41. 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. *J. Mol. Evol.* **43**, 293–303 (1996).
42. Gross, B. L. & Zhao, Z. Archaeological and genetic insights into the origins of domesticated rice. *Proc. Natl Acad. Sci. USA* **111**, 6190–6197 (2014).
43. Hufford, M. B. *et al.* The genomic signature of crop-wild introgression in maize. *PLoS Genet.* **9**, e1003477 (2013).
44. Cornille, A. *et al.* New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genet.* **8**, e1002703 (2012).
45. Myles, S. *et al.* Genetic structure and domestication history of the grape. *Proc. Natl Acad. Sci. USA* **108**, 3530–3535 (2011).
46. Nesom, G. L. *Citrus trifoliata* (Rutaceae): review of biology and distribution in the USA. *Phytoneuron* **46**, 1–14 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements Please see Supplementary Note 11 for funding information.

Author Contributions M.T., D.S.R. and G.A.W. developed the project and acted as project coordinators and provided scientific leadership; G.A.W. developed methods for admixture analysis and interspecific phasing, and performed comparative genome analysis. J.T., V.I., A.L.-G., E.P.-R., C.B., C.D., F.R.T., J.C.-C., R.A., J.D. and M.T. contributed 26 genomes; J.T., J.C.-C., R.A. and J.D. provided bioinformatics support; J.T. and E.P.-R. contributed to the study of the *IDH* gene; V.I., E.P.-R. and C.B. contributed to the variant analysis of candidate genes using genome-wide association studies; A.L.-G. and C.B. assisted in the biogeographic study; A.L.-G. and F.G.G. contributed to the description of citrus accessions and discriminatory characteristics; P.O. and F.C. contributed to germplasm, admixture analysis and hypothesis on the origin of cultivated citrus species; D.D. and F.G.G. contributed one citrus genome; M.L.R. contributed seven citrus genomes; F.G.G. contributed perspective garnered from more than 35 years of experience working on the genetic improvement of citrus; G.A.W., M.T., D.S.R. and F.G.G. wrote the manuscript; G.A.W. and M.T. contributed the hypothesis on the origin and dispersal of citrus.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to G.A.W. (gwu@lbl.gov), D.S.R. (dsroksar@gmail.com) or M.T. (talon_man@gva.es).

Reviewer Information *Nature* thanks J. Ross-Ibarra, P. Wincker and the other anonymous reviewer(s) for their contribution to the peer review of this work.

 This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

METHODS

Sample collection and sequencing. Whole-genome sequences from a total of 60 accessions were analysed: 58 citrus accessions with different geographical origins and two representative outgroup genera. Twelve of these genomes, including five mandarins, four pummelos, two oranges and a wild Mangshan mandarin (*C. mangshanensis*) were reanalysed from previous works^{2,7}. We also reanalysed 19 genomes from Chinese collections, including 15 unnamed mandarins, 2 Chinese sour oranges, Ambersweet orange and Cocktail grapefruit (a hybrid resembling grapefruit) that have been previously reported³.

The 30 accessions that were newly sequenced came from citrus germ-plasm banks and collections at IVIA, Valencia, Spain; SRA, Corse, France; UCR, Riverside and FDACS/DPI, Florida and included nine mandarins, two limes, one rough lemon, one grapefruit, one lemon, four citrons, one Australian desert lime, one eremorange, two Australian finger limes, two Australian round limes, one kumquat, one calamondin, one micrantha, one Ichang papeda, one trifoliolate orange and one Chinese box orange (Supplementary Note 1).

DNA libraries were constructed using standard protocols with some modifications. Library insert sizes range from 325 to 500 bp. Sequencing was performed on HiSeq2000/2500 instruments using 100-bp paired-end reads. Primary analysis of the data included quality control on the Illumina RTA sequence analysis pipeline (Supplementary Note 2).

Variant calls and *Citrus* species diversity. Illumina paired-end reads were aligned to the haploid Clementine reference sequence² and the sweet orange chloroplast genome assembly⁴⁷ using bwa-mem⁴⁸. PCR duplicates were removed using Picard. Raw variants were called using GATK HaplotypeCaller⁴⁹ with subsequent filtering based on read map quality score, base quality score, read depth and so on (Supplementary Note 3.1).

Interspecific admixtures versus pure citrus species were distinguished based on sliding window analysis of heterozygosity and pairwise genetic distance D (Supplementary Note 4). Genome-wide ancestry informative markers for the progenitor species were derived using pure accessions. Admixture analysis was carried out in sliding windows using ancestry informative markers (Supplementary Notes 5).

***Citrus* relatedness and haplotype sharing.** Interspecific phasing was used to extract admixed haplotypes. Identical-by-descent sharing was calculated for each of the non-overlapping sliding windows across the genome and used to estimate coefficient of relatedness among citrus accessions (Supplementary Notes 6, 7).

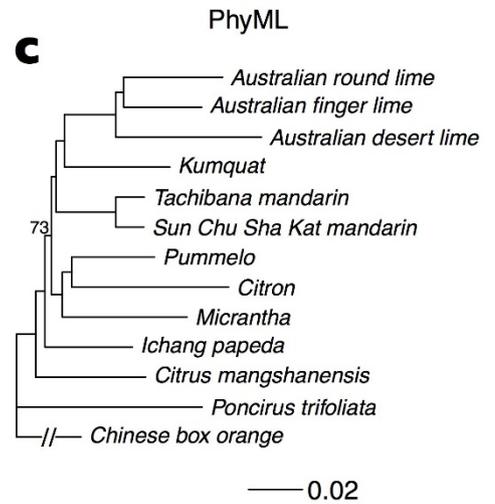
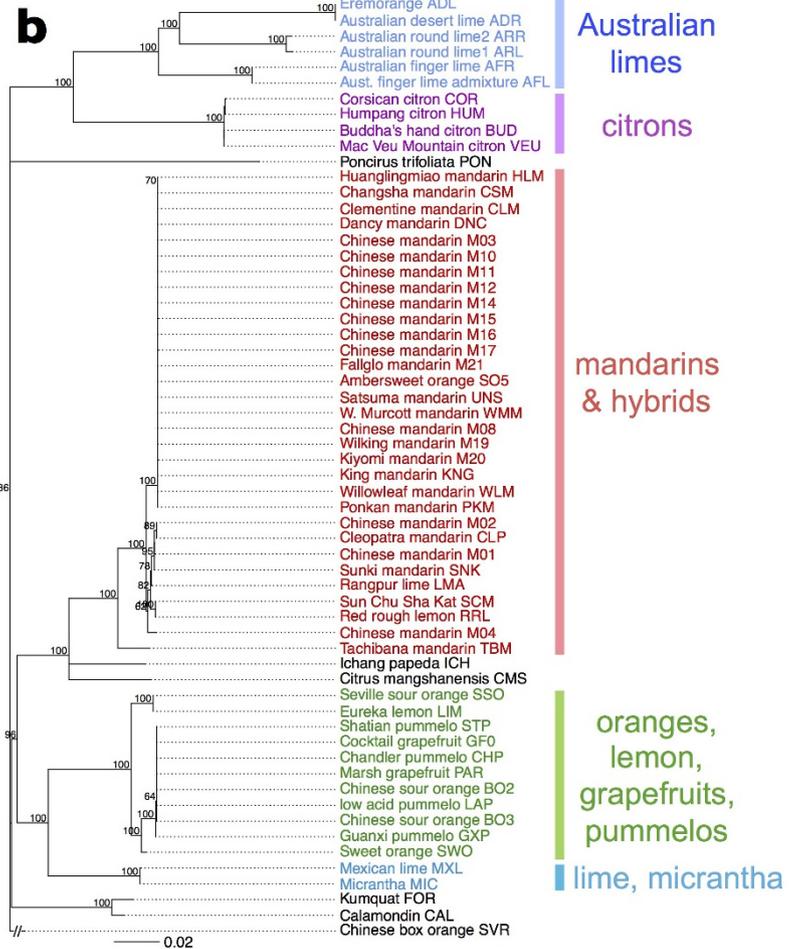
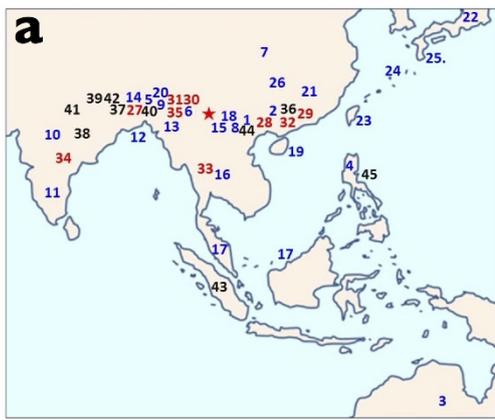
Phylogeny and speciation dating. We used Chinese box orange (genus *Severinia*) as an outgroup. Time calibration is based on the *C. linczangensis*¹⁶ fossil from Lincang, Yunnan, China. MrBayes⁵⁰ was used for whole genome Bayesian

phylogenetic inference, and corroborated with a PhyML⁵¹ reconstructed maximum likelihood tree. A penalized likelihood method⁵² as implemented in APE⁵³ was used to construct the chronogram (Supplementary Note 8).

Genome scan of palatability association. We used a mixed linear model as implemented in gemma⁵⁴ for a case-control study of citrus acidity and palatability with 37 citrus accessions. A conservative Bonferroni correction was used to select significant genomic loci, with subsequent manual examination of each candidate variant in all accessions to identify most discriminatory loci for fruit palatability (Supplementary Note 10).

Data availability. Whole-genome shotgun-sequencing data generated in this study have been deposited at NCBI under BioProject PRJNA414519. Prior resequencing data analysed here can be accessed under BioProject accession numbers PRJNA320985 (mandarins) and PRJNA321100 (oranges), and also under the NCBI Sequence Read Archive accession codes SRX372786 (sour orange), SRX372703 (sweet orange), SRX372702 (low-acid pummelo), SRX372688 (Chandler pummelo), SRX372685 (Willowleaf mandarin), SRX372687 (W. Murcott mandarin), SRX372665 (Ponkan mandarin) and SRX371962 (Clementine mandarin). The Clementine reference sequence used here is available at <https://phytozome.jgi.doe.gov/>.

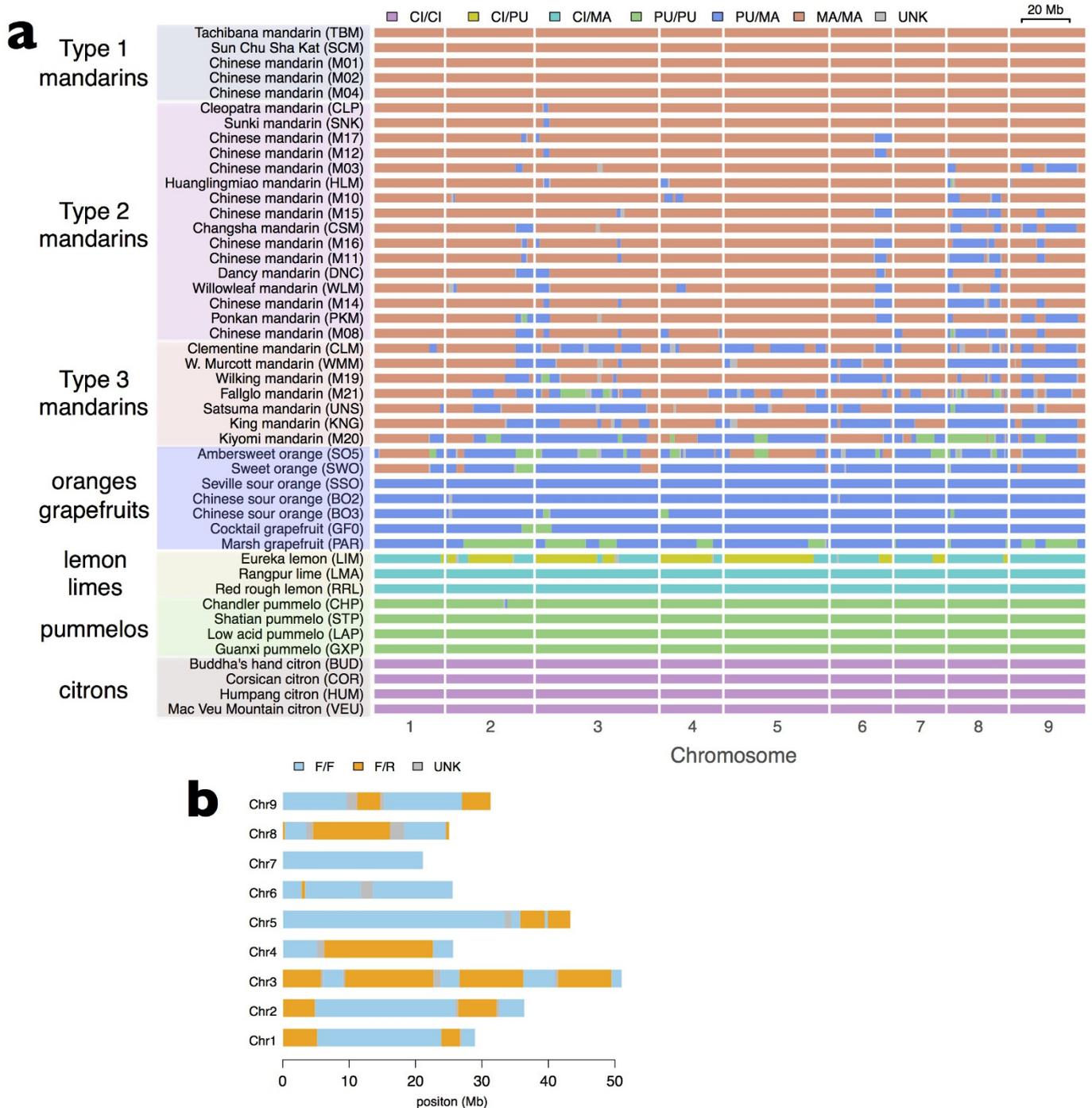
47. Bausher, M. G., Singh, N. D., Lee, S. B., Jansen, R. K. & Daniell, H. The complete chloroplast genome sequence of *Citrus sinensis* (L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. *BMC Plant Biol.* **6**, 21 (2006).
48. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrow-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
49. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
50. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
51. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
52. Sanderson, M. J. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* **19**, 101–109 (2002).
53. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
54. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
55. Frost, H. B. & Soost, R. K. in *The Citrus Industry* Vol. 2 (eds Reuther, W. *et al.*) 290–324 (1968).



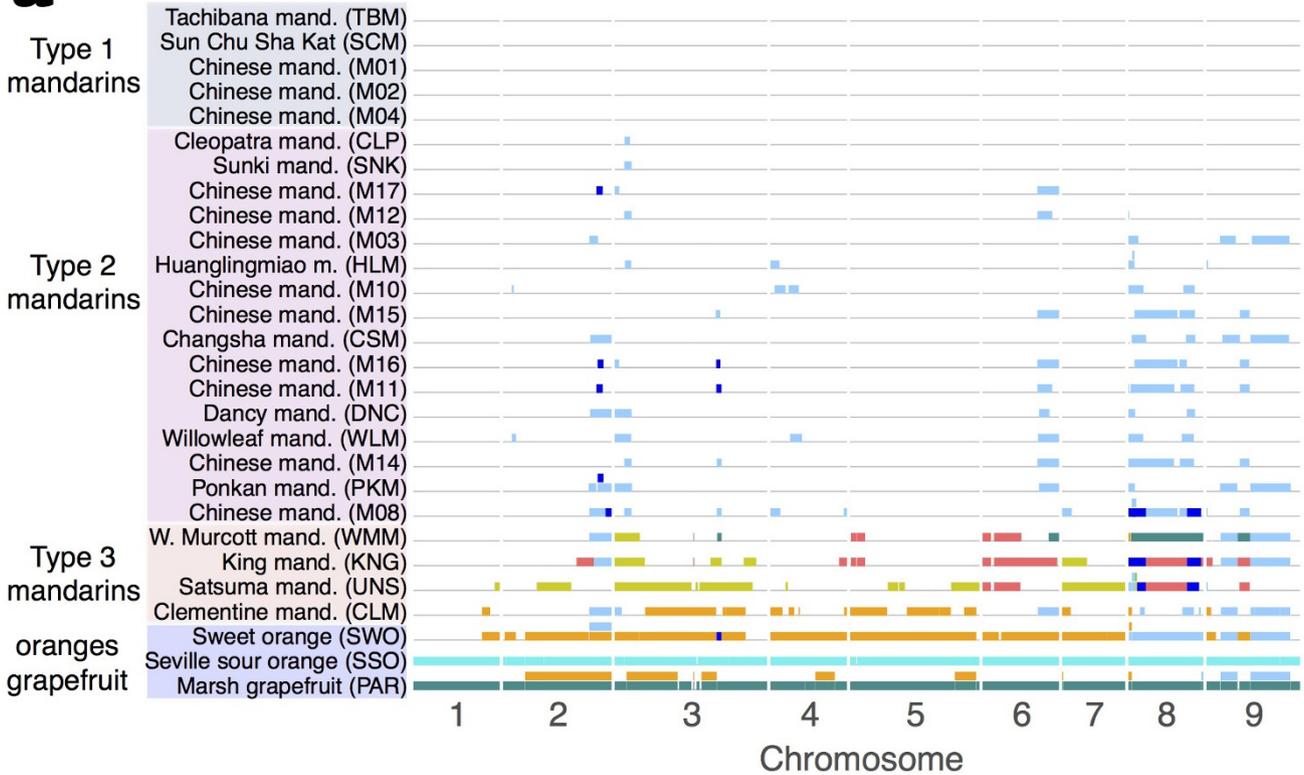
Extended Data Figure 1 | Citrus biogeography and phylogeny.

a. Geographical distribution of the genus *Citrus* in southeast Asia and Australia. Distribution is based on documented reports on the presence of wild genotypes representative of pure citrus species (blue numbers), admixtures (red numbers) and relevant interspecific hybrids (black numbers), growing freely in non-cultivated areas. Numbers are as in Supplementary Table 1. 1, 2, *Fortunella* spp. ; 3, Australian citrus (*E. glauca*; *Microcitrus australasica*; *Microcitrus australis*); 4, *C. micrantha*; 5–8, *C. ichangensis*; 9–15, *C. medica*; 16–19, *C. maxima*; 20–22, *C. reticulata* (Sun Chu Sha Kat); 23–25, *C. tachibana*; 26, *C. mangshanensis*; 27–29, *Citrus* spp. (mandarins); 30–33, *C. sinensis*; 34, 35, *Citrus limon* (probably not truly wild genotypes); 36, 37, *Citrus limonia*; 38, *Citrus jambhiri*; 39–42, *C. aurantium*; 43, *Citrus aurantifolia* (probably not truly wild

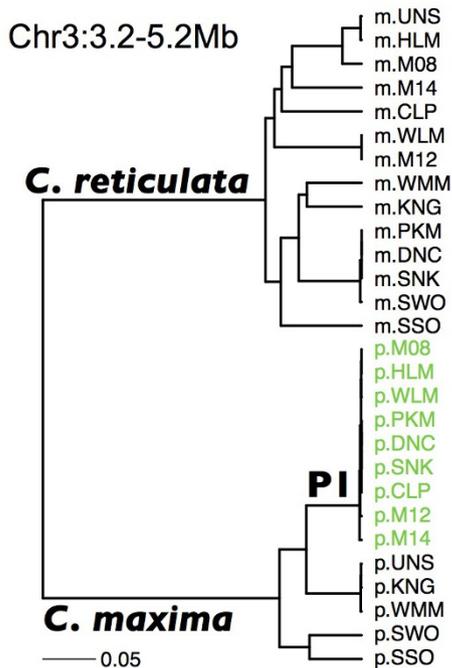
genotypes); 44, 45, *Fortunella* and *C. reticulata* hybrid. The red star indicates the location of the *C. linczangensis* fossil from the Late Miocene of Lincang¹⁶. **b.** Citrus chloroplast genome phylogeny rooted with *Severinia*. The analysis was performed on 58 citrus accessions and 2 outgroup genera, *Poncirus* and *Severinia*. The maximum likelihood tree as inferred from PhyML is shown. Percentage statistical support for the nodes is based on 200 bootstrap replicates. **c.** Citrus nuclear genome phylogeny rooted with *Severinia*. Both Bayesian and maximum likelihood trees yield the same topology with highly supported branches. The maximum likelihood tree reconstructed from PhyML is shown. Branch statistical support is based on 1,000 bootstraps and is shown if it is less than 100%. All branches have posterior probability 1.0 with Bayesian inference using MrBayes (not shown).



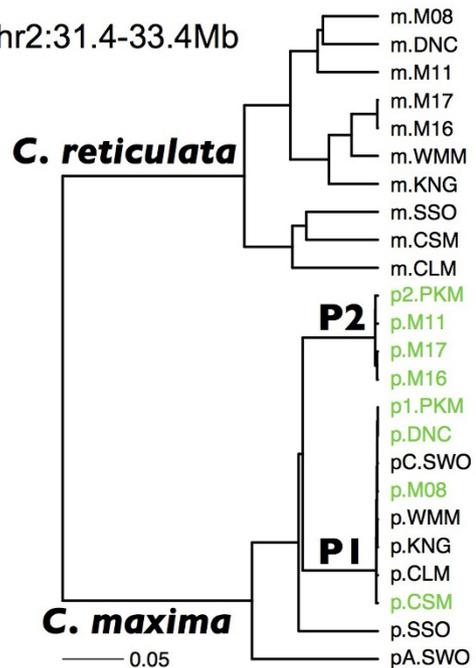
a



b

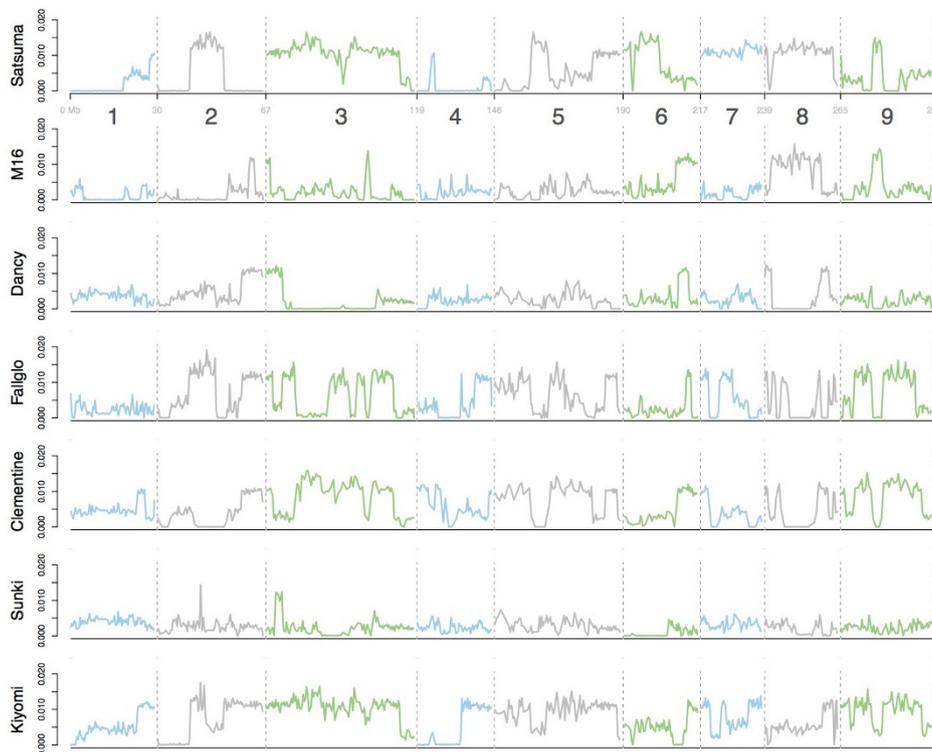
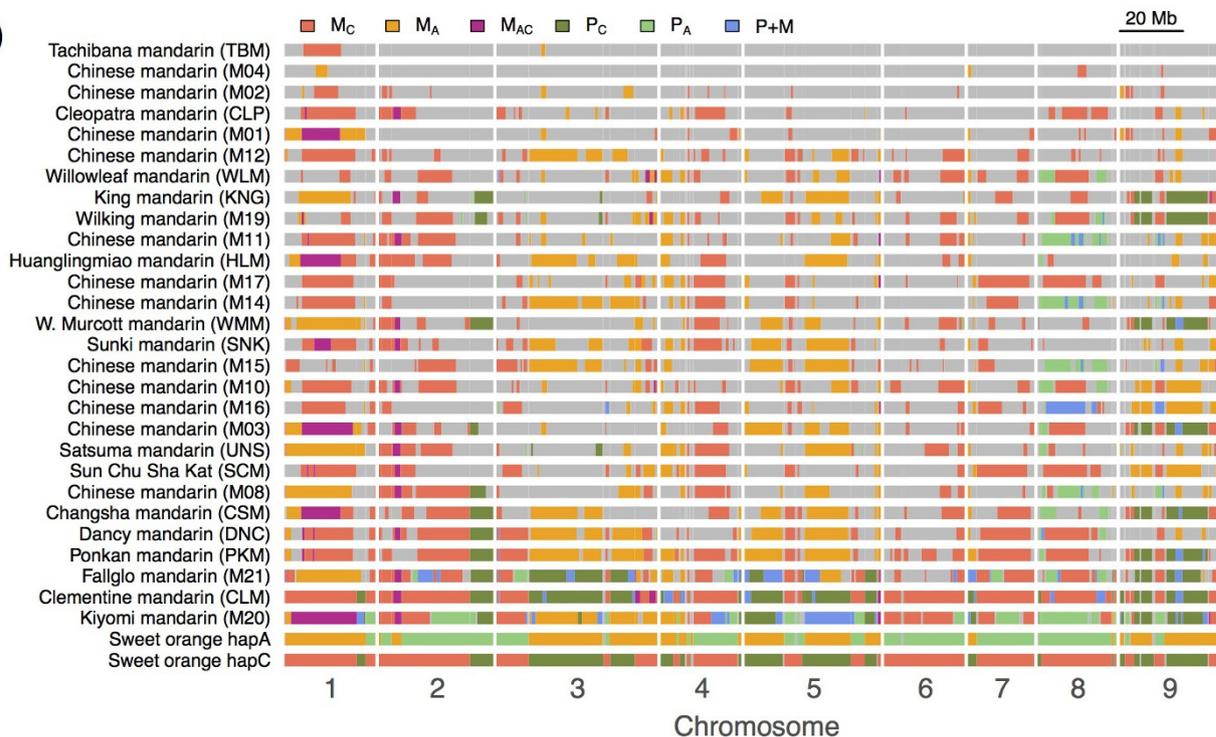


Chr2:31.4-33.4Mb



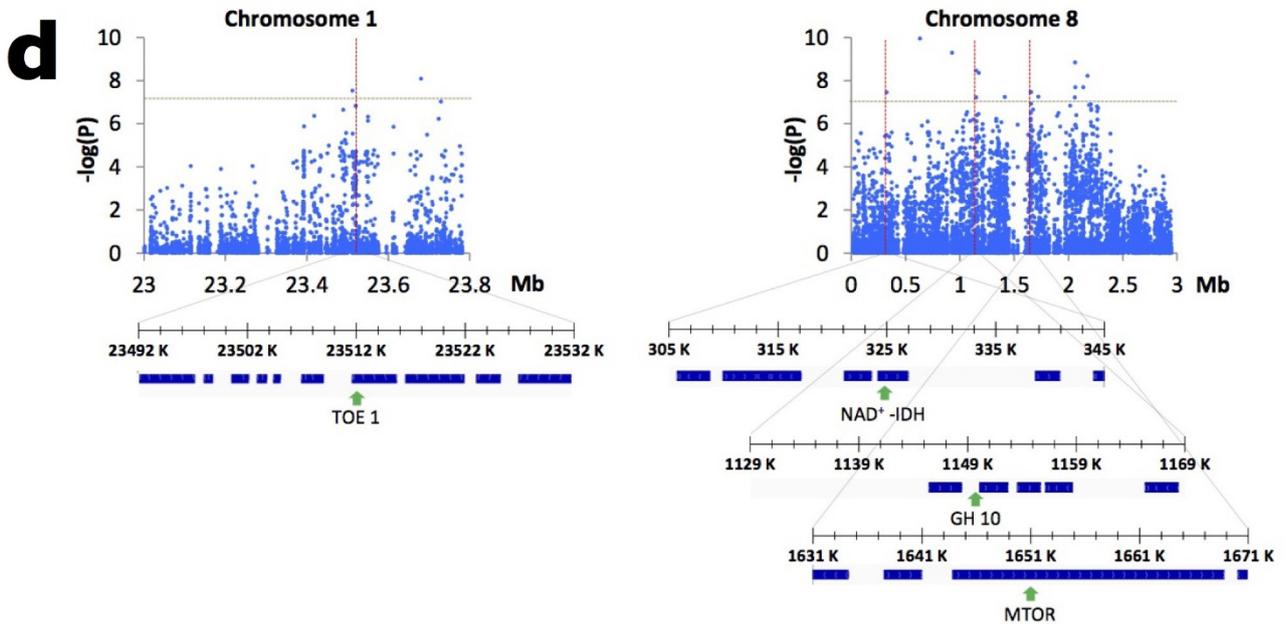
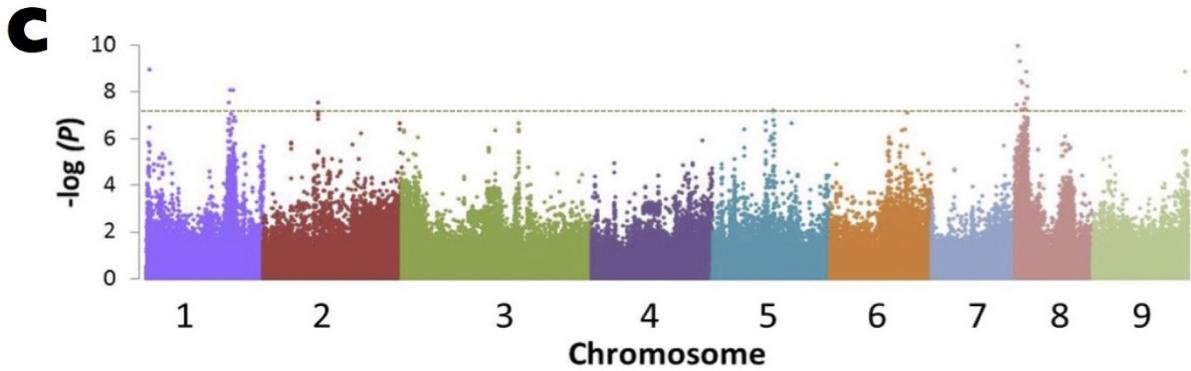
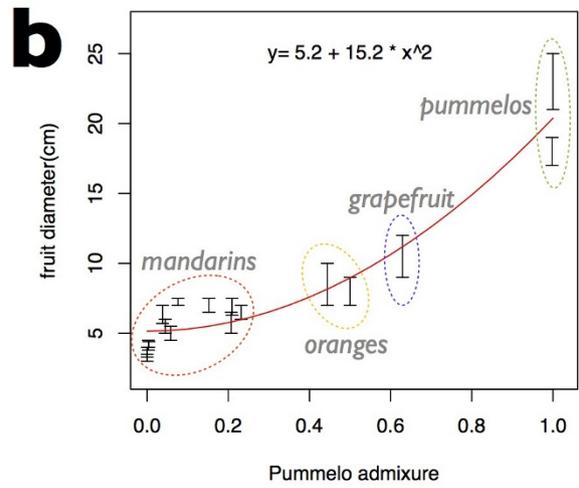
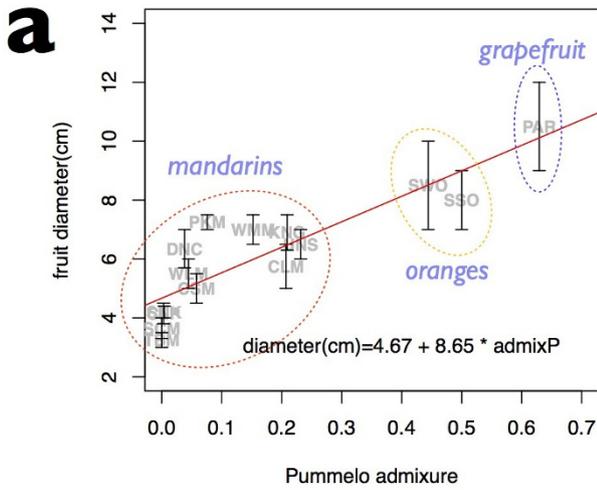
Extended Data Figure 3 | Pattern of pummelo introgression in mandarins. a, Distinct admixed pummelo haplotypes among mandarins, oranges and grapefruit are shown in different colours; the *C. reticulata* haplotypes are masked. The admixture pattern separates the mandarins into three groups, with type-1 representing pure mandarins. Type-2 mandarins contain a small amount of pummelo admixture derived from two *C. maxima* haplotypes: P1 (light blue colour) and P2 (dark blue), suggesting as few as one common pummelo ancestor in the distant past. Type-3 mandarins are characterized by both marked pummelo admixture and additional pummelo haplotypes besides P1 and P2. **b,** Haplotype trees for two chromosome segments where pummelo haplotypes of

type-2 mandarins are in green. Left, haplotype tree for chr3:3.2–5.2 Mb. Sweet orange, sour orange, and twelve of the sequenced mandarins are interspecific hybrids, and their phased *C. maxima* and *C. reticulata* haplotypes are denoted by prepending, respectively, ‘p’ and ‘m’ to the corresponding accession codes. The nine type-2 mandarins share the same pummelo haplotype (P1). Right, the haplotype tree for chr2:31.4–33.4 Mb. Two pummelo haplotypes (P1, P2) are shared among seven type-2 mandarins, with Ponkan mandarin containing both P1 and P2. Sweet orange also carries two pummelo haplotypes at this locus, denoted by pC.SWO (shared with Clementine) and pA.SWO (alternate haplotype).

a**b**

Extended Data Figure 4 | Haplotype sharing in mandarins. **a**, Runs of homozygosity in mandarins. Heterozygosity is plotted in non-overlapping windows of 200 kb along the nine chromosomes of 7 mandarin accessions with the highest degree of inbreeding. Runs of homozygosity correspond to regions with zero heterozygosity as a result of haplotype sharing between the parents. **b**, Haplotype sharing between sweet orange and mandarins. The two haplotypes of sweet orange are denoted by hapC (transmitted to Clementine) and hapA (alternate), respectively. The hapC haplotype is coloured in red (denoted by M_C) or dark green (denoted

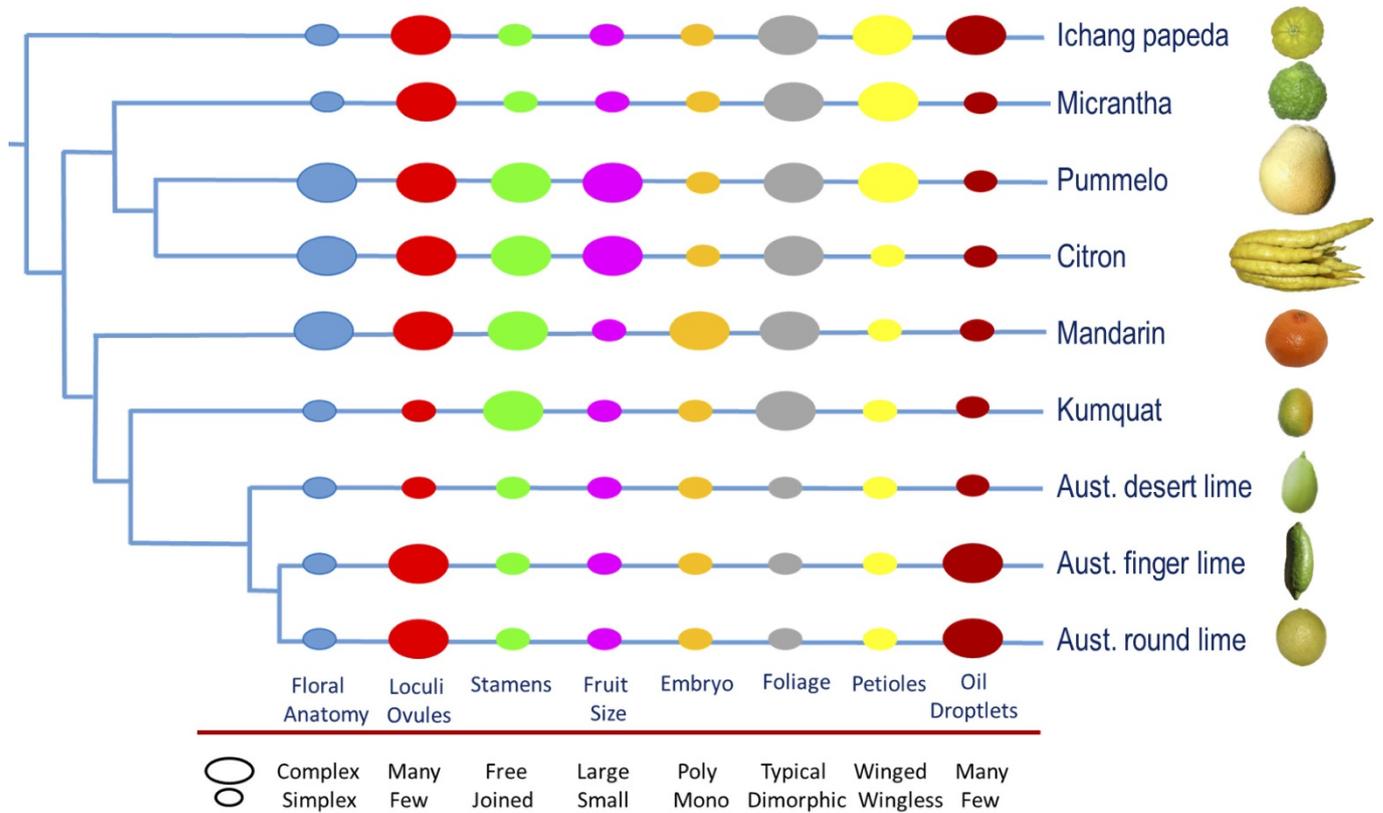
by P_C) if its genetic ancestry is *C. reticulata* or *C. maxima*, respectively. Similarly, hapA can take the form of *C. reticulata* (M_A in orange colour) or *C. maxima* (P_A in light green) depending on its genetic ancestry. Shared sweet orange haplotypes in mandarins are coloured accordingly, except when both haplotypes of sweet orange are shared (IBD2) either as two *C. reticulata* haplotypes (M_{AC} , dark red) or as interspecific hybrid ($P + M$, blue). Regions coloured in grey denote the absence of haplotype sharing between sweet orange and mandarin.



Extended Data Figure 5 | See next page for caption.

Extended Data Figure 5 | Fruit size and acidity correlated with pummelo introgression. **a**, Fruit size strongly correlated with pummelo admixture. The diameters of mandarins, oranges and grapefruit are plotted against the corresponding pummelo admixture proportions. A simple linear regression is shown in red. The strong correlation (Pearson correlation coefficient $r = 0.88$) between fruit size and pummelo admixture is apparent, especially among the three taxonomic groups of mandarins, oranges and grapefruit. The outliers (Ponkan mandarin and four acidic mandarins) suggest that certain genomic loci could be more important than others in fruit size determination. Accessions without size information are not included. Data are mean \pm s.d. from a set of 25 measurements for each of the 15 accessions. **b**, Fruit size correlation with pummelo allelic proportion with the addition of two pummelos. A polynomial regression provides a better fit than simple linear regression (adjusted $R^2 = 0.92$). Data are mean \pm s.d. from a set of 25 measurements for each of the 17 accessions. **c**, Genome scan of

significant loci associated with citrus acidity. Manhattan plot of a case-control analysis of a genome-wide association study (GWAS) of $n = 37$ citrus accessions with known acidity profile. The horizontal dashed line denotes the conservative Bonferoni-corrected P value of 7.9×10^{-8} for genome-wide significance ($\alpha = 0.05$). **d**, Manual inspection of candidate regions identified by GWAS ($n = 37$ accessions) demonstrates that in addition to the locus at chr1:23512067, single-nucleotide polymorphisms located in chromosome 8 are discriminatory for acidity. Shown are 40-kb zoom windows containing focal single-nucleotide polymorphisms (red vertical lines and green arrows) and gene models depicted by blue boxes in these two regions. *TOE1*, target of EGR1 protein 1 (Ciclev10007611; <https://phytozome.jgi.doe.gov/pz/portal.html>); NAD⁺-IDH, NAD⁺-dependent isocitrate dehydrogenase (*IDH*, Ciclev10028714); GH10, glycosyl hydrolase family 10 protein (Ciclev10028121) and MTOR, serine/threonine protein kinase (Ciclev10027661).



Extended Data Figure 6 | Species characteristics of citrus. Reproductive and vegetative characteristics among several species of the genus *Citrus* and related genera according to refs 1 and 55. The tree topology represented is that of the chronogram shown in Fig. 1c and citrus fruit

images are not drawn to scale. Most mandarins in our collection are polyembryonic, though a few are monoembryonic, including Clementine. Other exceptions to the generalized description concerning embryo numbers, in kumquat and other citrus species, can also be found.

Extended Data Table 1 | Citrus interspecies divergence and intraspecies diversity

ADR	AFR	ARL	BUD	LAP	SCM	MIC	FOR	ICH	CMS	PON	SVR	
4.291	13.85	15.19	21.64	18.64	18.11	20.23	18.61	17.87	19.34	24.4	36.76	Aus. desert lime
	6.246	11.37	18.92	15.76	16.96	17.67	15.97	15.31	16.36	21.84	34.4	Aus. finger lime
		5.483	21.98	17.85	20.14	20.59	18.82	18.22	18.02	24.97	37.12	Aus. round lime
			1.092	15.66	19.88	18.84	18.9	17.46	17.68	24.27	36.24	Citron
				3.786	15.44	14.71	14.6	12.77	14.45	19.85	32.46	Pummelo
					3.678	18.51	15.86	14.66	14.87	22.33	34.17	Mandarin
						3.893	17.28	16.02	16.57	23.24	35.56	Micrantha
							5.836	14.32	14.49	21.39	34.01	<i>Fortunella</i>
								4.758	13.16	19.79	32.77	Ichang papeda
									3.082	18.8	32.55	<i>C. mangshanensis</i>
										3.16	36.31	<i>Poncirus</i>
											6.479	<i>Severinia</i>

Average pairwise sequence divergences between ten citrus species and two out groups (*Poncirus* and *Severinia*) are listed, in unit of 10^{-3} . Each citrus species is represented by one diploid genotype free from interspecific admixture. Intraspecies variation is measured by the nucleotide diversity (that is, mean sequence divergence between the two haploid sequences of a diploid), and is represented along the diagonal in units of 10^{-3} . The species names and the codes of the representative accessions are given in the last column and first row, respectively. Note the wide separation between interspecies divergence and intraspecies variation. See Supplementary Table 2 for details and definitions.

Extended Data Table 2 | Admixture proportions of 50 citrus accessions derived from five progenitor species

	Code	PU/PU	PU/MA	MA/MA	UNK
Type-1 mandarins	TBM,SCM,M01,M02,M04	0	0	1	0
Type-2 mandarins	CLP	0.000	0.019	0.977	0.004
	SNK	0.000	0.025	0.975	0.000
	M17	0.000	0.045	0.949	0.006
	M12	0.000	0.047	0.944	0.009
	M03	0.000	0.055	0.937	0.007
	HLM	0.008	0.043	0.928	0.021
	M10	0.000	0.063	0.934	0.003
	M15	0.000	0.080	0.915	0.005
	CSM	0.000	0.081	0.901	0.017
	M16	0.000	0.082	0.911	0.007
	M11	0.000	0.088	0.894	0.018
	DNC	0.000	0.114	0.882	0.004
	WLM	0.000	0.118	0.870	0.012
	M14	0.000	0.120	0.868	0.013
	PKM	0.010	0.124	0.863	0.003
M08	0.015	0.173	0.804	0.008	
Type-3 mandarins	CLM	0.000	0.237	0.737	0.027
	WMM	0.000	0.252	0.732	0.016
	M19	0.031	0.214	0.747	0.008
	M21	0.015	0.281	0.672	0.032
	UNS	0.018	0.411	0.563	0.008
	KNG	0.000	0.471	0.519	0.009
	M20	0.099	0.565	0.327	0.009
Sweet orange	SWO	0.044	0.743	0.200	0.013
Ambersweet or.	SO5	0.104	0.543	0.318	0.035
Sour oranges	SSO,BO2	0.000	1.000	0.000	0.000
	BO3	0.046	0.942	0.000	0.012
Grapefruit	PAR	0.328	0.670	0.000	0.002
Cocktail grapef.	GF0	0.067	0.933	0.000	0.000
Pummelos	CHP	0.990	0.005	0.000	0.005
	LAP,GXP,STP	1	0	0	0
	Code	CI/CI	CI/PU	CI/MA	UNK
Rangpur lime	LMA	0	0	1	0
Rough lemon	RRL	0	0	1	0
Lemon	LIM	0	0.364	0.620	0.016
Citrons	BUD,COR,HUM,VEU	1	0	0	0
	Code	CI/CI	CI/MC	MC/MC	UNK
Micrantha	MIC	0	0	1	0
Mexican lime	MXL	0	1	0	0
	Code	MA/MA	MA/FO	FO/FO	UNK
Kumquat	FOR	0	0	1	0
Calamondin	CAL	0	1	0	0

These five species are *C. medica* (CI), *C. maxima* (PU), *C. reticulata* (MA), *C. micrantha* (MC), *Fortunella* (FO). Estimates based on genetic map lengths. UNK, unknown.

Extended Data Table 3 | Alleles of candidate single-nucleotide polymorphisms associated with citrus palatability

Code	1:415175	1:23512067	1:23679916	1:24219222	2:15484525	2:15702160	5:35094706	5:35098538	8:325527	8:631678	8:927020	8:1149577	8:1149586	8:1174414	8:1413967	8:1651338	8:1655701	8:1722788	8:2058824	8:2060290	8:2063416	8:2137063	8:2174360	9:30789594
NON-ACIDIC VARIETIES																								
CLM	C	C	C	A	G	G	G	C	A/G	T/C	C/T	G/T	C/T	T/A	A	G	G	T	C	T	T	G	C	T
KNG	C	C/T	C/A	A/G	G/T	G/T	G	C	A/G	T/C	C/T	G/T	C/T	T/A	A/G	G/A	G/T	T/C	C/T	T/C	T/C	G/C	C/T	T
PKM	C	C	C	A	G/T	G/T	G	C	A/G	T/C	C/T	G/T	C/T	T/A	A/G	G/A	G/T	T/C	C/T	T/C	T/C	G/C	C/T	T
DNC	C	C	C	A	G/T	G/T	G	C	A/G	T/C	C/T	G/T	C/T	T/A	A/G	G/A	G/T	T/C	C/T	T/C	T/C	G/C	C/T	T
CSM	C	C	C	A	G/T	G/T	G	C	G	C	C/T	G/T	C/T	T/A	A/G	G/A	G/T	T/C	C/T	T/C	T/C	G/C	C/T	T
WMM	C	C	C	A	G/T	G/T	G	C	A/G	T/C	C/T	G/T	C/T	T/A	A/G	G/A	G/T	T/C	C/T	T/C	T/C	G/C	C/T	T
HLM	C	C	C	A	G/T	G/T	G	C	A/G	T/C	C/T	G	C	T	G	A	T	C	T	C	C	C	T	T
UNS	C	C	C	A	G/T	G/T	G/A	C/T	A/G	T/C	C/T	G	C	T	G	A	T	C	T	C	C	C	T	T
WLM	C	C	C	A	G/T	G	G/A	C/T	A/G	T/C	C/T	G/T	C/T	T/A	A/G	G/A	G/T	T/C	C/T	T/C	T/C	G/C	C/T	T
SWO	C	C	C	A	G/T	G/T	G/A	C/T	A	T	C	G	C	T	A/G	G/A	G/T	T/C	C/T	T/C	T/C	G/C	C/T	T/C
CHP	T	C	C	A	T	T	A	T	A/G	T	C	G/T	C/T	T	G	A	T	C	C/T	T/C	C	C	T	C
LAP	T	C	C	A	T	T	A	T	A/G	T	C	G/T	C/T	T	G	A	T	C	T	C	C	C	T	C
GXP	T	C	C	A	T	T	A	T	A	T	C	G/T	C/T	T	G	A	T	C	T	C	C	C	T	C
STP	T	C	C	A	T	T	A	T	A	T	C	G	C	T	G	A	T	C	T	C	T/C	G/C	T	C
PAR	C/T	C	C	A	T	T	A	T	A/G	T	C	G	C	T	A/G	G/A	G/T	T/C	C/T	T/C	T/C	G/C	C/T	C
ACIDIC VARIETIES																								
SNK	T	C/T	C	A/G	T	T	G/A	C/T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
TBM	T	T	C/A	A/G	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
CLP	T	C/T	C/A	A/G	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
SCM	C/T	C/T	C/A	A/G	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
SSO	T	C/T	C/A	A/G	T	T	A	T	G	C	C/T	T	T	T/A	A/G	G/A	G/T	T	C	T	T	G	C	C
RRL	C/T	T	C/A	A/G	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	T/C
LMA	T	C/T	C/A	A/G	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
LIM	T	T	C/A	A/G	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
CAL	T	T	C/A	A/G	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
FOR	T	T	C	A	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
MIC	T	T	C	A	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
MXL	T	T	C	A	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
COR	T	T	C	A	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
VEU	T	T	C	A	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
BJD	T	T	C	A	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
HUM	T	T	C	A	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
ICH	T	T	C	A	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
AFL	T	T	C	A	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
ARL	T	T	C	A	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
AFR	T	T	C	A	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
ARR	T	T	C	A	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C
ADR	T	T	C	A	T	T	A	T	G	C	T	T	T	A	A	G	G	T	C	T	T	G	C	C

The association study is based on a case-control GWAS analysis of 37 accessions with known palatability. Grey, ancestral alleles detected in *Severinia* and *Poncirus*; red, derived alleles; yellow, heterozygous single-nucleotide polymorphisms.

Extended Data Table 4 | The *IDH* gene variants

Species	Cultivar	Code	SNP position		
			Exon 1 8:324328	Exon 4 8:326594	Exon 4 8:326608
NON-ACIDIC VARIETIES					
<i>Citrus maxima</i>	Chandler pummelo	CHP	T/T	G/G	C/C
<i>Citrus maxima</i>	Low-acid pummelo/	LAP	T/T	G/G	C/C
<i>Citrus maxima</i>	Guanxi pummelo	GXP	T/T	G/G	C/C
<i>Citrus maxima</i>	Shatian pummelo	STP	T/T	G/G	C/C
<i>Citrus sinensis</i>	Sweet orange	SWO	T/T	G/G	C/C
<i>Citrus paradisi</i>	Marsh grapefruit	PAR	T/T	G/G	C/C
<i>Citrus reticulata</i>	Ponkan mandarin	PKM	T/T	G/T	C/G
<i>Citrus clementina</i>	Clementine mandarin	CLM	T/T	G/T	C/G
<i>Citrus tangerina</i>	Dancy mandarin	DNC	T/T	G/T	C/G
<i>Citrus deliciosa</i>	Willowleaf mandarin	WLM	T/T	G/T	C/G
<i>Citrus unshiu</i>	Satsuma mandarin	UNS	T/T	G/A	C/G
<i>Citrus reticulata</i>	Huanglingmiao mandarin	HLM	T/T	G/A	C/G
<i>Citrus nobilis</i>	King mandarin	KNG	T/G	G/A	C/G
<i>Citrus reticulata</i>	Changsha mandarin	CSM	T/T	A/T	G/G
<i>Citrus reticulata</i>	W. Murcott mandarin	WMM	T/G	G/A	C/G
ACIDIC VARIETIES					
<i>Citrus aurantium</i>	Sour orange	SSO	T/G	G/A	C/G
<i>Citrus sunki</i>	Sunki mandarin	SNK	T/T	A/T	G/G
<i>Citrus reshni</i>	Cleopatra mandarin	CLP	T/G	A/A	G/G
<i>Citrus ichangensis</i>	Ichang papeda	ICH	T/G	A/A	G/G
<i>Citrus limon</i>	Eureka lemon	LIM	T/T	A/A	G/G
<i>Citrus micrantha</i>	Micrantha	MIC	A/A	A/A	G/G
<i>Citrus aurantifolia</i>	Mexican lime	MXL	A/A	A/A	G/G
<i>Fortunella margarita</i>	Kumquat. Nagami	FOR	T/A	A/A	G/G
<i>Citrus limonia</i>	Rangpur lime	LMA	T/A	A/A	G/G
<i>Citrus madurensis</i>	Calamondin	CAL	T/A	A/A	G/G
<i>Citrus medica</i>	Mac Veu Montain Citron	VEU	T/T	A/A	G/G
<i>Citrus medica</i>	Corsican citron	COR	T/T	A/A	G/G
<i>Citrus medica</i>	Buddha's hand citron	BUD	T/T	A/A	G/G
<i>Citrus medica</i>	Humpang citron	HUM	T/T	A/A	G/G
<i>Microcitrus australasica</i>	Australian finger lime (BC2)	AFL	G/G	A/A	G/G
<i>Microcitrus australis</i>	Australian round lime (Pure)	ARL	G/G	A/A	G/G
<i>Microcitrus australasica</i>	Australian finger lime (Pure)	AFR	G/G	A/A	G/G
<i>Microcitrus australis</i>	Australian round lime (Pure)	ARR	G/G	A/A	G/G
<i>Eremocitrus glauca</i>	Australian desert line	ADR	G/G	A/A	G/G
<i>Citrus jambhiri</i>	Red rough lemon	RRL	T/T	A/T	G/G
<i>Citrus reticulata</i>	Sun Chu Sha Kat mandarin	SCM	T/T	T/T	G/G
<i>Citrus tachibana</i>	Tachibana mandarin	TBM	G/G	A/A	G/G

Alleles of non-synonymous single-nucleotide polymorphisms of the NAD⁺-dependent isocitrate dehydrogenase (*IDH*) gene (Ciclev10028714) in 37 citrus accessions with known palatability.

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

▶ Experimental design

1. Sample size

Describe how sample size was determined.

Thirty-seven citrus accessions with known acidity profile were used in the GWAS analysis for preliminary and tentative inference (Supplementary Note 10.2).

2. Data exclusions

Describe any data exclusions.

Recently published citrus accessions (Wang et al 2017 Nat. Gen.) don't have information on fruit size and acidity profile, and are excluded in association analysis. Variant calls failing the allele balance and other filters are excluded from consideration (Supplementary Note 3.1). For nuclear genome phylogenetic reconstruction, we used SNVs from the non-repetitive, non-genic and non-pericentromeric regions of the genome (Supplementary Note 8).

3. Replication

Describe whether the experimental findings were reliably reproduced.

N/A

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

N/A

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

N/A

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- A statement indicating how many times each experiment was replicated
- The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
- A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted
- A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- Clearly defined error bars

See the web collection on [statistics for biologists](#) for further resources and guidance.

► Software

Policy information about [availability of computer code](#)

7. Software

Describe the software used to analyze the data in this study.

For single nucleotide variant calls: BWA, PICARD, GATK. For multidimensional scaling: R function cmdscale. For phylogenomic inference: PhyML, MrBayes, R package APE. For GWAS: Gemma.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). [Nature Methods guidance for providing algorithms and software for publication](#) provides further information on this topic.

► Materials and reagents

Policy information about [availability of materials](#)

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials were used in this study.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

N/A

b. Describe the method of cell line authentication used.

N/A

c. Report whether the cell lines were tested for mycoplasma contamination.

N/A

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

N/A

► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

N/A

Policy information about [studies involving human research participants](#)

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

N/A