HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2015 January 01.

Published in final edited form as:

Nat Biotechnol. 2014 July ; 32(7): 656–662. doi:10.1038/nbt.2906.

Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication

A full list of authors and affiliations appears at the end of the article.

Abstract

The domestication of citrus, is poorly understood. Cultivated types are selections from, or hybrids of, wild progenitor species, whose identities and contributions remain controversial. By comparative analysis of a collection of citrus genomes, including a high quality haploid reference, we show that cultivated types were derived from two progenitor species. Though cultivated pummelos represent selections from a single progenitor species, *C. maxima*, cultivated mandarins are introgressions of *C. maxima* into the ancestral mandarin species, *C. reticulata*. The most widely cultivated citrus, sweet orange, is the offspring of previously admixed individuals, but sour orange is an F1 hybrid of pure *C. maxima* and *C. reticulata* parents, implying that wild mandarins were part of the early breeding germplasm. A wild “mandarin” from China exhibited substantial

*Correspondence to: D.S.R. (dsrokhsar@gmail.com); F.G.G. (fgmitter@ufl.edu).
‡Current address: Life Technologies Corp., Grand Island, NY 14072, USA.
¶Current address: USDA, ARS, Southeastern Fruit and Tree Nut Research Laboratory, Byron, GA, USA
†These authors contributed equally.

Author Contributions

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

Competing Financial Interests Statement

The authors declare no competing financial interests.

Accessions

The reference haploid Clementine assembly and annotation is deposited inNCBI’s genome database (http://www.ncbi.nlm.nih.gov/genome) under the accession AMZM00000000. Sanger WGS for Clementine is deposited in the NCBI trace archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi) under SPECIES_CODE=’CITRUS CLEMENTINA’. The sweet orange assembly and annotation is deposited in NCBI’s genome database under Accession XXXXX000000. WGS sequence for the sweet orange is deposited in NCBI’s Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/) under BioProject PRJNA225968 (454 data) and in the NCBI trace archive under SPECIES_CODE=’CITRUS SINESIS’ AND CENTER_NAME=’JGI’. Citrus resequencing data are in the Sequence Read Archive at NCBI under the following accessions:SRX372786 (sour orange), SRX372703 (sweet orange), SRX372702 (Low acid pummelo), SRX372688 (Chandler pummelo), SRX372685 (Willowleaf mandarin), SRX372687 (W. Murcott mandarin), SRX372665 (Ponkan mandarin), SRX371962 (Clementine mandarin).

divergence from *C. reticulata*, suggesting the possibility of other unrecognized wild citrus species. Understanding citrus phylogeny through genome analysis clarifies taxonomic relationships and enables sequence-directed genetic improvement.

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

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

Two wild species are believed to have contributed to domesticated pummelos, mandarins and oranges (Supplementary Note 1). Based on morphology and genetic markers, "pummelos" have generally been identified with the wild species *C. maxima* (Burm.) Merrill that is indigenous to Southeast Asia. Although "mandarins" are similarly widely identified with the species *C. reticulata* Blanco[4–6], wild populations of *C. reticulata* have not been definitively described. Various authors have taken different approaches to classifying mandarins, and several naming conventions have been developed[7, 8]. Here we emphasize that the term "mandarin" is a commercial or popular designation referring to citrus with small, easy-peeling, sweet fruit, and not necessarily a taxonomic one. We use the qualifier "traditional" to refer to mandarins without previously suspected admixture from other ancestral species, to distinguish them from mandarin types that are known or believed to be recent hybrids. For clarity we use "×" in the systematic name of such known hybrids (see *e.g.*, Ref.[9]). Recognizing that genome sequencing and diversity analysis has provided insights into the domestication history of several other fruit crops[10, 11], cereals[12, 13] and other crops (reviewed in Ref.[14]), we sequenced and analyzed the genomes of a diverse collection of cultivated pummelos, mandarins and oranges to test the pummelo-mandarin species hypothesis and to uncover the origins of several important citrus cultivars.

Results

A high quality reference Clementine genome

To provide a genomic platform for analyzing *Citrus*, we generated a high quality reference genome from ~7× Sanger dideoxy whole genome shotgun coverage of a haploid derivative of Clementine "mandarin" (*C.* × *clementina* cv. Clemenules)[15] (Supplementary Note 2). The use of haploid material (derived from a single ovule after induced gynogenesis[15, 16]) removes complications that arise when assembling outbred diploid genomes. The resulting 301.4 Mbp reference sequence is nearly complete, with superior assembly contiguity (contig L50 = 119 kbp) and scaffolding (scaffold L50 before pseudochromosome construction = 6.8 Mbp) compared to a recently published sweet orange draft sequence[17] (Supplementary Note 2). The long scaffolds allowed us to construct pseudochromosomes by assigning 96% of the assembly to a location on the nine citrus chromosomes using the latest citrus genetic map[18], compared with only 79% in the sweet orange draft[17](Supplementary Note 2). From sequence data we also inferred the phase of the two diploid Clementine haplotypes, identifying ten crossovers from the meiosis that produced the haploid Clementine (Supplementary Fig. 1), and annotated nominal centromeres as large regions of low recombination (Supplementary Figs. 2–11). Independently we also sequenced and assembled a draft genome of the (diploid) sweet orange variety 'Ridge Pineapple' by combining deep 454 sequence with light Sanger sampling (Supplementary Note 3) and inferred chromosome phasing using the recently reported rough draft genome of a sweet-orange-derived dihaploid[17].

The citrus genome retains substantial segmental synteny (that is, local co-linearity) with other eudicots, although it has experienced extensive large-scale rearrangement on the chromosome scale (Supplementary Note 4). Based on analysis of synteny we propose a specific model for the origin of the citrus genome from the paleo-hexaploid eudicot ancestor[19] through a series of chromosome fissions and fusions (Supplementary Figs. 12,13). Despite the compactness of the citrus genome, 45% is repetitive, with long-terminal repeat retrotransposons and numerous uncharacterized elements, each making up nearly half of the repetitive content; the remainder comprises DNA transposons and LINEs (Supplementary Note 5). We identified ~25,000 protein-coding gene loci in both Clementine and sweet orange by computational methods combined with extensive long-read 454 and Sanger expressed sequence tags (Supplementary Note 5).

Investigation of citrus ancestry

To investigate the origin of cultivated varieties, we sequenced the genomes of four mandarins (including Clementine), two pummelos and one sour orange, as well as the sweet orange genome reported above (Table 1, Supplementary Table 1, Supplementary Notes 1,6). (Cultivars derived from *C. medica* (the third purported wild species), *i.e.*, citrons, limes and lemons, were not part of this study.) Two distinct types of chloroplast genomes (cpDNA) were readily identified, with mandarins all having one type (which we define as "M" for mandarin or *C. reticulata*) and pummelos and oranges sharing another type (defined as "P" for pummelo or *C. maxima*), with limited variation within each cpDNA type (Supplementary Note 6), consistent with prior studies of mitochondrial markers[20]. Citrus

nuclear genomes tell a more complex story (Supplementary Notes 7, 8). We find that while the sequenced pummelos are evidently genotypes from the sexual *C. maxima* species with minimal introgression of other species, all the mandarin-type citrus we sequenced show substantial admixture with pummelo and therefore cannot simply be selections from an ancestral *C. reticulata* population (Fig. 2,3). The sweet and sour oranges are also hybrids of varying complexity, with pummelo-type chloroplast genomes in both cases.

Ancestry of pummelos

The two diploid pummelos that we sequenced contain three distinct haplotypes, since Low acid (Siamese Sweet) pummelo is the known female parent of Chandler pummelo[21], so that the two pummelos share one haplotype at each locus (Supplementary Note 9). Within the two sequenced pummelos and between their non-shared alleles (derived from the other parent of Chandler, *i.e.*, Siamese Pink pummelo) modest levels of heterozygosity were observed, with a genome-wide nucleotide heterozygosity of 5.7 heterozygous (het) sites/kb (Fig. 2a). The presence of a second low-heterozygosity peak (~1 het site/kb) in the distribution can be explained by a strong ancient bottleneck in the *C. maxima* population ~100–300 kya (Supplementary Note 10). Our reanalysis of three Chinese pummelos previously reported[17] (including the Wusuan pummelo that we identify as from the same somatic lineage as Siamese Sweet pummelo), shows that both Thai and Chinese pummelos are derived from the same wild population (Supplementary Note 11). Only a single short 1.5 Mb segment on chromosome 2 of Chandler shows unusually high heterozygosity that could reflect interspecific introgression. These observations are consistent with pummelo domestication by selection from a wild sexual *C. maxima* population.

Ancestry of mandarins

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

Thus, our surprising conclusion is that “traditional” mandarin types like Ponkan and Willowleaf, are in fact interspecific introgressions of *C. maxima* (pummelo) into *C. reticulata* (wild mandarin). Furthermore, although these traditional mandarins were previously thought to be unrelated, we detect extensive haplotype sharing between them (Supplemental Note 10). Because microsatellite-based population structure analyses of a wide range of citrus genotypes shows mandarins as a defined cluster of genotypes[22], such admixture is likely widespread among mandarin types. Indeed, reanalysis of a recently sequenced Chinese mandarin[17] in the light of our discovery of interspecific introgression in multiple mandarin types, shows that the traditional Chinese Huanglingmiao mandarin (incorrectly treated previously[16] as a pure *C. reticulata*) also exhibits unsuspected admixture between *C. reticulata* and *C. maxima* (Supplementary Note 11).

Although none of our cultivated mandarin genotypes represent pure *C. reticulata*, we can nevertheless extract wild mandarin alleles from our data by comparing the (admixed) cultivated mandarins with each other and the two pure pummelos. By such genome-wide comparisons we identified 1,537,264 putative fixed single nucleotide differences between *C. reticulata* and *C. maxima* (Supplementary File 1, Supplementary Note 7). These diagnostic variants can in turn be used to partition the mandarin, pummelo and orange genomes into segments according to their species ancestry (Fig. 3). The characterization of *C. reticulata* genomic segments from modern mandarins is analogous to the extraction of African haplotypes from Mexican Americans[23][SEP1]and native American haplotypes from extant ethnic human populations that are admixtures with American, African and European roots[24].

We can estimate the parameters of a simple population genetic model for the divergence of *C. reticulata* and *C. maxima* from an ancestral south Asian citrus founder population, using a coalescent framework and our collection of fixed interspecific differences and intraspecific variation (Supplementary Note 9). This analysis is consistent with effective population sizes of several hundred thousand trees for *C. maxima* and somewhat fewer for *C. reticulata*, with larger effective population size for pummelos in keeping with their higher heterozygosity. Note that the likely occurrence of apomixis in wild mandarin populations, a trait that seems to be absent in *C. maxima*, may contribute to reducing the effective *C. reticulata* population size relative to the census size. If we assume a per site mutation rate of $\mu \sim 1 - 2 \times 10^{-9}$/yr (comparable to that observed in poplar trees[25]) then we can estimate that *C. reticulata* and *C. maxima* diverged ~1.6–3.2 Mya, consistent with the divergence between *Citrus* and the related genus *Poncirus*, which is estimated at 4–9.6 Mya[26]. As noted, the excess of low heterozygosity segments in pummelo is consistent with a substantial population bottleneck several hundred thousand years ago and prior to the separation of Thai and Chinese pummelo lineages (Supplementary Notes 9, 11).

Some specific citrus genotypes are generally recognized as “hybrid” varieties. For example, Clementine mandarin (also known as Algerian tangerine) is believed to be a chance seedling from a Mediterranean mandarin (*e.g.*, Willowleaf) selected just over a century ago in Algeria[27]. Although various male parents have been proposed, serological and molecular studies demonstrated that the Clementine was likely a mandarin × sweet orange hybrid[6, 18, 28]. We confirm this hypothesis at the sequence level by definitively identifying a Willowleaf and sweet orange allele at each Clementine locus; demarcating the

recombination breakpoints in the meiosis that produced the haploid Clementine sequence; and determining the Willowleaf and sweet orange haplotypes that contributed to diploid Clementine (Supplementary Note 10, Supplementary Fig. 14,15). Similarly, the W. Murcott mandarin is believed to be a chance zygotic seedling of Murcott tangor, itself a presumed F1 hybrid of sweet orange and an unknown mandarin. Our sequence analysis is consistent with the suspected grandparent/grandchild relationship between sweet orange and W. Murcott (Supplementary Note 10). Although the other parent and grandparent of W. Murcott are not known (but see[29]), a search for these ancestors will be enabled by the other observed alleles.

Ancestry of oranges

Sweet orange (*C.* × *sinensis* L. Osbeck) is the citrus type most widely cultivated for fruit and juice and is widely believed to be an interspecific hybrid, but its origin is unknown[4, 6]. Different sweet orange cultivars share the same genomic organization with little sequence variation, having arisen by mutation from the original sweet orange domesticate (see, e.g. Ref.[30]). Using our genome-wide catalog of fixed *C. reticulata* vs. *C. maxima* alleles, we can represent the sweet orange genome as segments of these two parental species or hybrid segments thereof (Supplementary Note 10; Fig. 2d), with clear boundaries between different segments types (Fig. 3a). A recently proposed "(P×M)×M" backcross scheme for the derivation of sweet orange from mandarin and pummelo[17], however, is easily ruled out by the presence of clear "P/P" (*i.e.*, *C. maxima*/*C. maxima*) segments in sweet orange, which requires both parents to have some pummelo ancestry. (The P/P segment on chromosome 2 has been confirmed by directed resequencing of three genes in this region[31].)

Unexpectedly, in our analysis we found that sweet orange shares alleles with Ponkan mandarin across nearly three-quarters of the genome, and many of the same segments are also shared with Willowleaf and Huanglingmiao (Supplementary Note 10; Supplementary Fig. 16). This leads to the surprising conclusion that these three traditional mandarins, previously considered independent selections, in fact show substantial kinship with each other and an ancestor of sweet orange, suggesting much more limited genetic diversity among the traditional mandarins than previously recognized (Supplementary Note 10). The nature of the other parent of sweet orange is more difficult to infer, but the distribution of heterozygous segments in sweet orange (Supplementary Fig. 17) and its pummelo-type chloroplast genome are more readily accounted for if the female parent was itself a pummelo with substantial introgression of wild mandarin (Supplementary Note 9).

Finally, Seville or sour orange (also known as *C.* × *aurantium*), which has historically been an important rootstock for citrus and, more familiarly, is used in marmalade and other products, is another traditional cultivar type that is widely regarded as a pummelo-mandarin hybrid. Our genomic analysis shows that sour orange is indeed the direct result of a simple interspecific F1 cross between a pummelo (*C. maxima*) seed parent and a wild mandarin (*C. reticulata*) pollen parent (Supplementary Note 10). Surprisingly in light of our discovery of widespread pummelo admixture among traditional mandarins, no such admixture is found in the *C. reticulata* parent of sour orange, but the specific parental genotypes remain unknown. Sour orange may have arisen as a natural hybrid of two wild *Citrus* species, and persisted by

virtue of its reproduction through apomixis, followed by deliberate human cultivation and distribution. We found no detectable recent relationship between sweet and sour orange.

Chinese Mangshan represents a distinct species, *C. mangshanensis*

Among cultivars traditionally classified as "mandarins", however, we found another surprise. Our analysis of the genome of a presumed "wild mandarin" from Mangshan, China[17] (CMS) shows (i) a chloroplast genome that is distinct from both *C. reticulata* and *C. maxima* (Fig. 4a); (ii) limited heterozygosity (Fig. 4b), again uniformly distributed across the genome, and no segments of pummelo or mandarin ancestry, indicating no admixture; (iii) ~2% homozygous differences from both *C. reticulata* and *C. maxima* uniformly across the genome, a rate comparable to the divergence between *C. maxima* and *C. reticulata* (Fig. 4b). At the level of nucleotide diversity, CMS is as diverged from C. *maxima* and *C. reticulata* as they are from each other (Fig. 4b) and is clearly separated from pummelos, oranges and mandarins by principal coordinate analysis (Fig. 4c, Supplementary Note 11). By all these measures, we find that Mangshan "mandarin"is unrelated to the other cultivated mandarins discussed above (including Huanglingmiao mandarin). We therefore propose that despite its morphology Mangshan "mandarin" represents a distinct species from *C. reticulata*, supporting the nomenclature *C. mangshanensis*[32].

Discussion

Our genomic analyses clarify some of the murky early history of citrus domestication. The nuclear and chloroplast genomes of cultivated pummelos are consistent with the identification of pummelos as a single *Citrus* species, *C. maxima*. In contrast, the nuclear genomes of sequenced "mandarin" type cultivars all contain substantial admixture of *C. maxima*, despite the similarity of mandarin chloroplast sequences. Our results thus show that the various conventional *Citrus* taxonomies that associate mandarin citrus types with the ancestral *Citrus* species *C. reticulata* are too simplistic. It is particularly surprising that even the traditional mandarin types with no prior suspicion of relatedness or admixture such as Ponkan, Willowleaf and Huanglingmiao mandarin show substantial haplotype sharing and all include introgressed pummelo segments. A supposed "wild mandarin" from Mangshan, China, turns out to represent a distinct taxon only distantly related to *C. reticulata*, based on analysis of its nuclear and chloroplast genomes. (In a previous analysis of sweet orange ancestry[17], Mangshan "mandarin" Clementine and Huanglingmiao were used to represent *C. reticulata*. Our discovery of substantial pummelo admixture in Clementine and Huanglingmiao, and the distinctness of Mangshan "mandarin" from *C. reticulata*, further invalidates their conclusions.)

Remarkably, even in the absence of a pure type specimen for *C. reticulata*, we can characterize the genome of this wild mandarin progenitor species from genome-wide comparative analysis of admixed descendants[23]. Our collection of 1,537,264 SNPs (Supplementary File 1) that differentiate *C. reticulata* from *C. maxima* can be used to guide the search for pure *C. reticulata* mandarin types (or recognize other cryptic species) among the hundreds of known cultivars and other germplasm accessions. Small-fruited mandarins that are less desirable for fresh consumption based on appearance, flavor, texture and aroma

may be considered likely candidates. With the discovery that *C. mangshanensis* is a distinct group, the possibility of additional undescribed wild *Citrus* species must also be considered.

The prevalence of interspecific admixture in cultivated citrus suggests that either early in domestication or in a natural hybrid zone prior to domestication, *C. reticulata* and *C. maxima* interbreeding occurred. Given the typical size of the hybrid blocks, only a few generations of introgression occurred prior to the selection of attractive cultivars, which were then propagated asexually by apomictic or vegetative means, perhaps in southern China[33]. Our analysis of sweet orange and sour orange shows that these ancient and widely cultivated genotypes are pummelo-mandarin admixtures that are unrelated to each other, despite some degree of phenotypic similarity[34]. The discovery that sour orange is a simple F1 hybrid of *C. maxima* and *C. reticulata* implies that pure *C. reticulata* individuals were part of the breeding germplasm at the origin of sour orange. Remarkably, we found that extant Ponkan, Willowleaf and Huanglingmiao mandarins are related to each other and to the male parent of sweet orange. Although the female parent of sweet orange remains unknown, it cannot have been a pure pummelo (though it had pummelo cytoplasm, based on cpDNA and mtDNA[20]). Its identity is constrained by the high proportion of hybrid P/M segments in sweet orange, which can be naturally explained if the female parent of sweet orange were (P×M)×P.

Like many other agricultural enterprises, the global citrus industry relies substantially on large-scale monoculture which makes it particularly challenging to meet consumer demand for greater product diversity while trying to incorporate tolerance and/or resistance to biotic and potentially catastrophic abiotic stresses[35]. Advances in citrus genomics[36, 37] should soon allow the identification of the somatic mutations that, with their ancient genetic backgrounds, underlie the diversity of citrus color, flavor and aroma in modern cultivars. Our analysis of the relationships between cultivated citrus and the ancestral species from which they were derived emphasizes the limited ancestral germplasm that contributed to the commercially important cultivar types like sweet orange, and highlights the opportunities for the creation of new combinations of the ancestral citrus types with novel fruit quality traits or even the re-creation of sweet orange with improved disease resistance *via* sexual hybridization, beyond the current approaches based on somatic mutations and genetic engineering.

Online Methods

Haploid *C.* × *clementina* ‘Clemenules’ sequencing and assembly

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

C. × *sinensis* genome sequencing and assembly

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

Annotation of repeats and genes in citrus genome assemblies

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

Evolutionary comparisons with other plant genomes

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

Analysis of resequencing datasets

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

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

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

Population genetic analysis and simulations

Population genetic analysis of the two citrus species and demographic inference were based on coalescent simulations conducted using MaCS (Supplementary Note 10).

Analysis of relatedness in citrus

Parentage and relatedness analysis for Clementine and other citrus genomes made use of homozygous SNPs in each diploid genome relative to the haploid Clementine reference as well as to the inferred second haplotype of Clementine (Supplementary Notes 9 and 11). In the same way, the haploid sweet orange assembly was used for identifying shared haplotypes with sweet orange (Supplementary Note 9). A modified identical-by-state (IBS) method was used for haplotype sharing analysis among mandarins and other citrus pairs (Supplementary Note 9).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

G. Albert Wu[1,†], Simon Prochnik[1,†], Jerry Jenkins[2], Jerome Salse[3], Uffe Hellsten[1], Florent Murat[3], Xavier Perrier[4], Manuel Ruiz[4], Simone Scalabrin[5], Javier Terol[8], Marco Aurélio Takita[6], Karine Labadie[7], Julie Poulain[7], Arnaud Couloux[7], Kamel Jabbari[7], Federica Cattonaro[5], Cristian Del Fabbro[5], Sara Pinosio[5], Andrea Zuccolo[5,26], Jarrod Chapman[1], Jane Grimwood[2], Francisco R. Tadeo[8], Leandro H. Estornell[8], Juan V. Muñoz-Sanz[8], Victoria Ibanez[8], Amparo Herrero-Ortega[8], Pablo Aleza[9], Julián Pérez-Pérez[10,27], Daniel Ramón[10], Dominique Brunel[7,11], François Luro[12], Chunxian Chen[13,¶], William G. Farmerie[14], Brian Desany[15], Chinnappa Kodira[15], Mohammed Mohiuddin[15], Tim Harkins[15,‡], Karin Fredrikson[15], Paul Burns[16,17], Alexandre Lomsadze[16,17], Mark Borodovsky[16,17,18], Giuseppe Reforgiato[19], Juliana Freitas-Astúa[6,20], Francis Quetier[7,21], Luis Navarro[9], Mikeal Roose[22], Patrick Wincker[7,21,23], Jeremy Schmutz[2], Michele Morgante[5,24], Marcos Antonio Machado[6], Manuel Talon[8], Olivier Jaillon[7,21,23], Patrick Ollitrault[4], Frederick Gmitter[13,*], and Daniel Rokhsar[1,25,*]

Affiliations

[1]US-Department of Energy Joint Genome Institute, Walnut Creek, CA, USA

[2]HudsonAlpha Biotechnology Institute, Huntsville, AL, USA

[3]INRA/UBP UMR 1095 GDEC, Clermont Ferrand, France

[4]CIRAD, UMR AGAP, Montpellier, France

[5]Istituto di Genomica Applicata, Udine, Italy

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

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

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

[9]Centro de Protección Vegetal y Biotecnología-IVIA, Moncada, Valencia, Spain

[10]Lifesequencing SL, Valencia, Spain

[11]INRA, US EPGV_1279, Evry, France

[12]INRA GEQA, San Giuliano, France

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

[14]Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL, USA

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

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

[17]School of Computational Science & Engineering, Georgia Institute of Technology, Atlanta, GA, USA

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

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

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

[21]Département de Biologie, Université d'Evry, Evry, France

[22]Department of Botany and Plant Sciences, University of California, Riverside, CA, USA

[23]Centre National de Recherche Scientifique (CNRS), Evry, France

[24]Department of Agriculture and Environmental Sciences, University of Udine, Udine, Italy

[25]Division of Genetics, Genomics, and Development, University of California, Berkeley, CA, USA

[26]Institute of Life Sciences, Scuola Superiore Sant'Anna, Pisa, Italy

[27]Secugen SL, Madrid, Spain

Acknowledgments

National Science and Technology Institute of Genomics for Citrus Breeding (FAPESP 2008/57909-2 and CNPq573848/08-4) and Brazilian Agricultural Research Corporation (Embrapa) (MAT, JF-A, MAMach) and Embrapa-Monsanto Agreement (JF-A); Agence nationale de la recherche (ANR) grant CITRUSSEQ PCS-08-GENO (OJ, XP, MR, PO, FL, KJ) and program ANR Blanc-PAGE, ref: ANR-2011-BSV6-00801 (JS, FM);US National Institute of Health grant HG00783 (MB, PB, AL); Grant PrometeoII/2013/008 from the Generalitat Valenciana, Spain and grant (AGL2011-26490) from the Ministry of Economy and Innovation-Fondo Europeo de Desarrollo Regional (FEDER), Spain (PA, LN); Conselleria de Agricultura, Pesca, Alimentación y Água from the Generalitat Valenciana (JPP, DR); Ministerio de Economia e Innovación grants PSE-060000-2009-8 and IPT-010000-2010-43 and Citruseq-Citrusgenn consortium companies (Anecoop S. Coop., Eurosemillas S.A., Fundación Ruralcaja Valencia, GCM Variedades Vegetales A.I.E, Investigación Citrícola Castellón, S.A and Source Citrus Genesis – SNFL) (JT, FRT, LHE, JVM-S, VI, AH-O, MT); Florida Citrus Production Research Advisory Council (FCPRAC), the Florida Department of Agriculture and Consumer Services (Grant # 013646), the Florida Department of Citrus (FDOC), and the Citrus Research and Development Foundation, Inc. (Grant #71), on behalf of the Florida citrus growers (FG, CC, WGF); Project Citrustart from Ministero delle Politiche Agricole Alimentari e Forestali; Project IT-Citrus Genomics PON_01623 from MIUR, Programma Operativo Nazionale "Ricerca e Competitività" 2007–2013 (MMorg, SS, FC, CDF, SPin, AZ). Pineapple Ridge sweet orange sequencing by 454 Life Sciences, a Roche company. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

1. Bové J. HUANGLONGBING: A DESTRUCTIVE, NEWLY-EMERGING, CENTURY-OLD DISEASE OF CITRUS. J. Plant Path. 2006; 88:7–37.
2. Reuther, W.; Webber, HJ.; Batchelor, LD., editors. The Citrus Industry. Edn. 1. Vol. Vol. 1. Berkeley: University of California, Division of Agricultural Sciences; 1967.
3. Spiegel-Roy, P.; Goldschmidt, EE. Biology of citrus. Cambridge; New York: Cambridge University Press; 1996.
4. Scora RW. On the history and origin of Citrus. Bull. Torrey Botanical Club. 1975; 102:369–375.

5. Barrett HC, Rhodes AM. A numerical taxonomic study of affinity relationships in cultivated citrus and its close relatives. Syst. Biol. 1976; 1:105–136.
6. Nicolosi E, et al. Citrus phylogeny and genetic origin of important species as investigated by molecular markers. Theor. Appl. Genet. 2000; 100:1155–1166.
7. Swingle, WT.; Reece, HC. The Citrus Industry. Edn. 2nd edition. Reuther, W.; Webber, HJ.; Batchelor, LD., editors. Vol. Vol. 1. Berkeley: University of California Press; 1967. p. 190-430.
8. Tanaka T. Fundamental discussion of Citrus classification. Studia Citrologica. 1977; 14:1–6.
9. Moore GA. Oranges and lemons: clues to the taxonomy of Citrus from molecular markers. Trends Genet. 2001; 17:536–540. [PubMed: 11525837]
10. Cornille A, et al. New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. PLoS Genet. 2012; 8:e1002703. [PubMed: 22589740]
11. Myles S, et al. Genetic structure and domestication history of the grape. Proc. Natl. Acad. Sci. 2011; 108:3530–3535. [PubMed: 21245334]
12. Huang X, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature. 2012; 490:497–501. [PubMed: 23034647]
13. Hufford MB, et al. Comparative population genomics of maize domestication and improvement. Nat. Genet. 2012; 44:808–811. [PubMed: 22660546]
14. Morrell PL, Buckler ES, Ross-Ibarra J. Crop genomics: advances and applications. Nat. Rev. Genet. 2011; 13:85–96. [PubMed: 22207165]
15. Germana MA, et al. Cytological and molecular characterization of three gametoclones of Citrus clementina. BMC Plant Biol. 2013; 13:129. [PubMed: 24020638]
16. Aleza P, et al. Recovery and characterization of a Citrus clementina Hort. ex Tan. 'Clemenules' haploid plant selected to establish the reference whole Citrus genome sequence. BMC Plant Biol. 2009; 9:110. [PubMed: 19698121]
17. Xu Q, et al. The draft genome of sweet orange (Citrus sinensis). Nat. Genet. 2013; 45:59–66. [PubMed: 23179022]
18. Ollitrault P, et al. A reference genetic map of C. clementina hort. ex Tan.; citrus evolution inferences from comparative mapping. BMC Genomics. 2012; 13:593. [PubMed: 23126659]
19. Salse J. In silico archeogenomics unveils modern plant genome organisation, regulation and evolution. Curr. Opin. Plant Biol. 2012; 15:122–130. [PubMed: 22280839]
20. Froelicher Y, et al. New universal mitochondrial PCR markers reveal new information on maternal citrus phylogeny. Tree Genetics & Genomes. 2011; 7:49–61.
21. Cameron, JWaS. R K Chandler – an early-ripening hybrid pummelo derived from a low-acid parent. Hilgardia. 1961; 30:359–364.
22. Barkley NA, Roose ML, Krueger RR, Federici CT. Assessing genetic diversity and population structure in a citrus germplasm collection utilizing simple sequence repeat markers (SSRs). Theor. Appl. Genet. 2006; 112:1519–1531. [PubMed: 16699791]
23. Johnson NA, et al. Ancestral components of admixed genomes in a Mexican cohort. PLoS Genet. 2011; 7:e1002410. [PubMed: 22194699]
24. Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. Nature. 2011; 475:163–165. [PubMed: 21753830]
25. Tuskan GA, et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science. 2006; 313:1596–1604. [PubMed: 16973872]
26. Pfeil BE, Crisp MD. The age and biogeography of Citrus and the orange subfamily (Rutaceae: Aurantioideae) in Australasia and New Caledonia. Am. J. Bot. 2008; 95:1621–1631. [PubMed: 21628168]
27. Trabut JL. L'hybridation des Citrus: une nouvelle tangérine "la Clémentine". Revue Horticole. 1902; 10:232–234.
28. Samaan LG. Studies on the origin of Clementine tangerine (Citrus reticulata Blanco). Euphytica. 1982; 31:167–173.
29. Luro, F., et al. Eleventh International Citrus Congress. Wuhan, China: 2008.

30. Novelli VM, Cristofani M, Souza AA, Machado MA. Development and characterization of polymorphic microsatellite markers for the sweet orange (Citrus sinensis L. Osbeck). Genetics and Molecular Biology. 2006; 29:90–96.
31. Garcia-Lor A, et al. A nuclear phylogenetic analysis: SNPs, indels and SSRs deliver new insights into the relationships in the 'true citrus fruit trees' group (Citrinae, Rutaceae) and the origin of cultivated species. Ann. Bot. 2013; 111:1–19. [PubMed: 23104641]
32. Liu GF, He SW, Li WB. Two new species of citrus in China. Acta Botanica Yunnanica. 1990; 12:287–289.
33. Gmitter FG, Hu X. The possible role of Yunnan, China, in the origin of contemporary citrus species (Rutaceae). Economic Botany. 1990; 44:267–277.
34. Morton, JF. Fruits of Warm Climates. Miami, Florida, USA: Florida Flair Books; 1987.
35. Gottwald TR. Current epidemiological understanding of citrus Huanglongbing. Annu. Rev. Phytopathol. 2010; 48:119–139. [PubMed: 20415578]
36. Talon M, Gmitter FG Jr. Citrus genomics. Int. J. Plant Genomics. 2008; 2008:1–17.
37. Gmitter FG, et al. Citrus genomics. Tree Genetics & Genomes. 2012; 8:611–626.
38. Bausher MG, Singh ND, Lee SB, Jansen RK, 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. 2006; 6:21. [PubMed: 17010212]

Figure 1.

A selection of mandarin, pummelo and orange fruits, including cultivars sequenced in this study. Pummelos (numbered 1, 2 in outline, on left) are large trees that produce very large fruit, with white, pink or red flesh color (2) and yellow or pink rinds. Most cultivars have large leaves having petioles with prominent wings. Apomictic reproduction is absent and most selections are self-incompatible. Mandarins (3–7) are smaller trees bearing smaller fruit, with orange flesh (9, 11) and rind color. Mandarins have both apomictic and zygotic reproduction and some are self-compatible. Oranges (8, 10) are generally intermediate in

tree and fruit size, flesh (10) and rind color is commonly orange, and apomictic reproduction is always present. (The sour orange shown (12) is immature.)

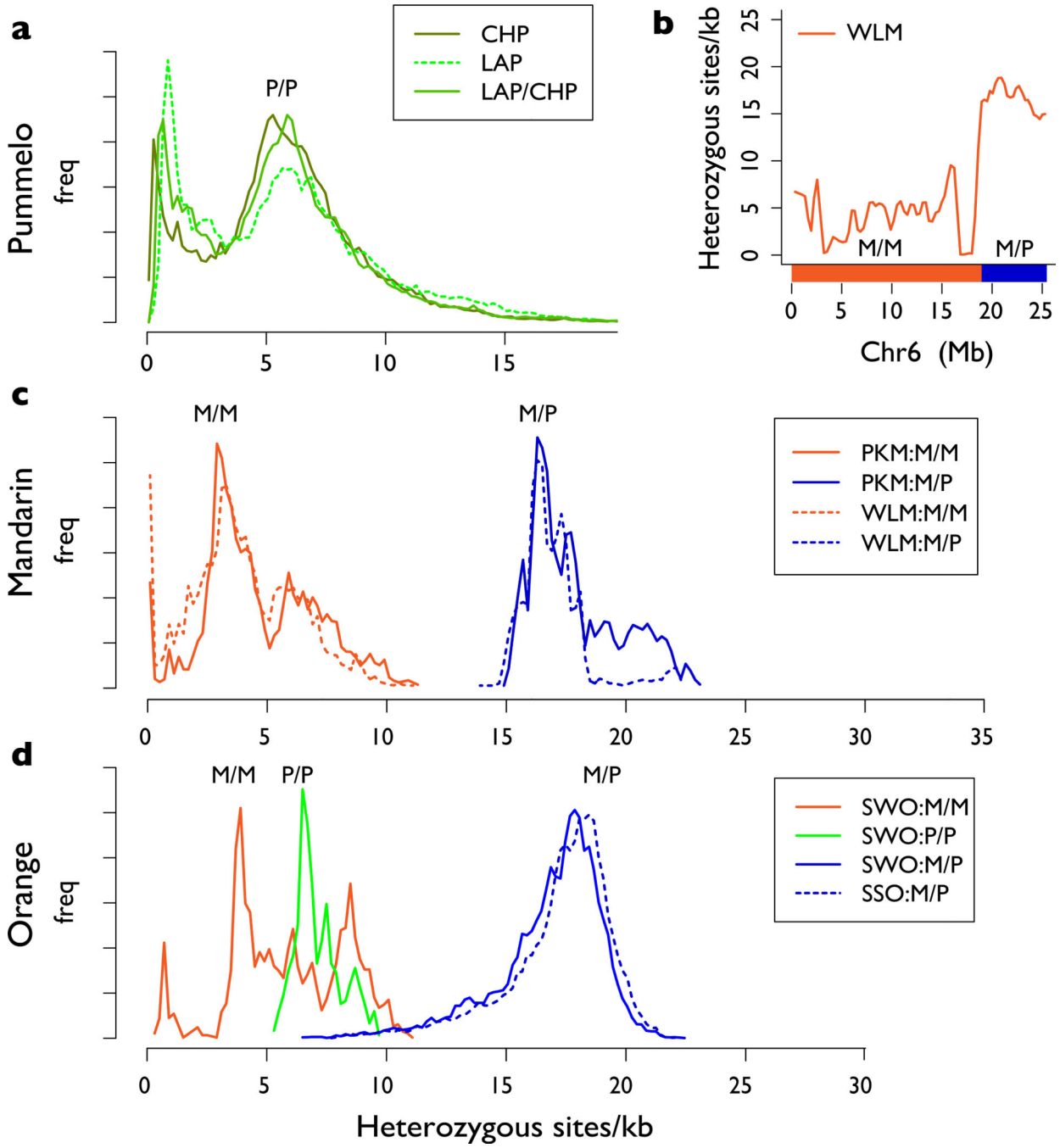

Figure 2.

Nucleotide diversity distribution in citrus.(**a**) Nucleotide heterozygosity distribution computed in overlapping 100kb windows (with 5 kb step size) across the Low acid (LAP) and Chandler (CHP) pummelo genomes and between the non-shared haplotypes of this parent-child pair (LAP/CHP) is shown. The peak at ~6 heterozygous sites/kb in all three pairwise comparisons represents the characteristic nucleotide diversity of the species *C. maxima*; the peak near ~1 heterozygous site/kb reflects a bottleneck in the ancestral *C. maxima* population after divergence from *C. reticulata* (Supplementary Note 10). **(b)**

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

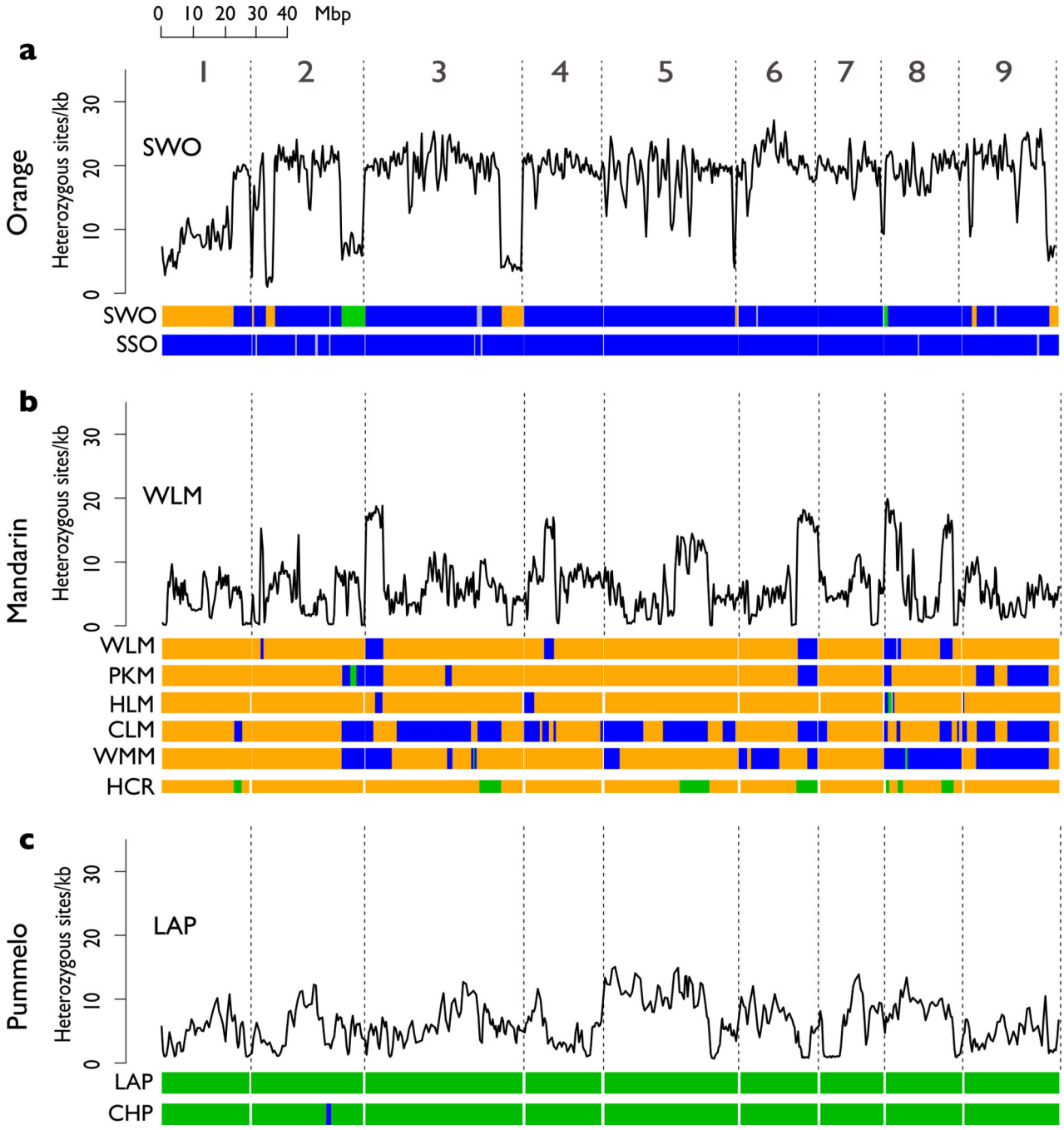

Figure 3.

Admixture patterns and nucleotide diversity in cultivated citrus. For each of the three groups of sequenced citrus, variation in nucleotide diversity (averaged over 500kb windows with step size 250kb) is shown across the genome for one representative cultivar above genotype maps (horizontal bars: green = *C. maxima/C. maxima*; blue = *C. maxima/C. reticulata*; orange= *C. reticulata/C. reticulata*; grey=unknown; the 9 chromosomes are numbered at the top). **(a)** Sweet orange (SWO) nucleotide diversity with genotype maps for SWO and sour orange (SSO). Note the *C. maxima/C. maxima* genotype (green segments present on

chromosomes 2 and 8) in SWO. **(b)** Willowleaf mandarin (WLM) nucleotide diversity and genotype maps for three traditional mandarins (Ponkan mandarin (PKM), WLM, Huanglingmiao (HLM)) and three recent mandarin types (Clementine (CLM), W. Murcott mandarin (WMM), haploid Clementine reference (HCR)). For the haploid Clementine reference sequence (HCR), red and green segments indicate *C.reticulata* and *C. maxima* haplotypes, respectively. All five mandarin types show pummelo introgressions (blue or green segments). **(c)** Low acid pummelo (LAP) nucleotide diversity and genotype maps for two pummelos (LAP, Chandler pummelo (CHP)).

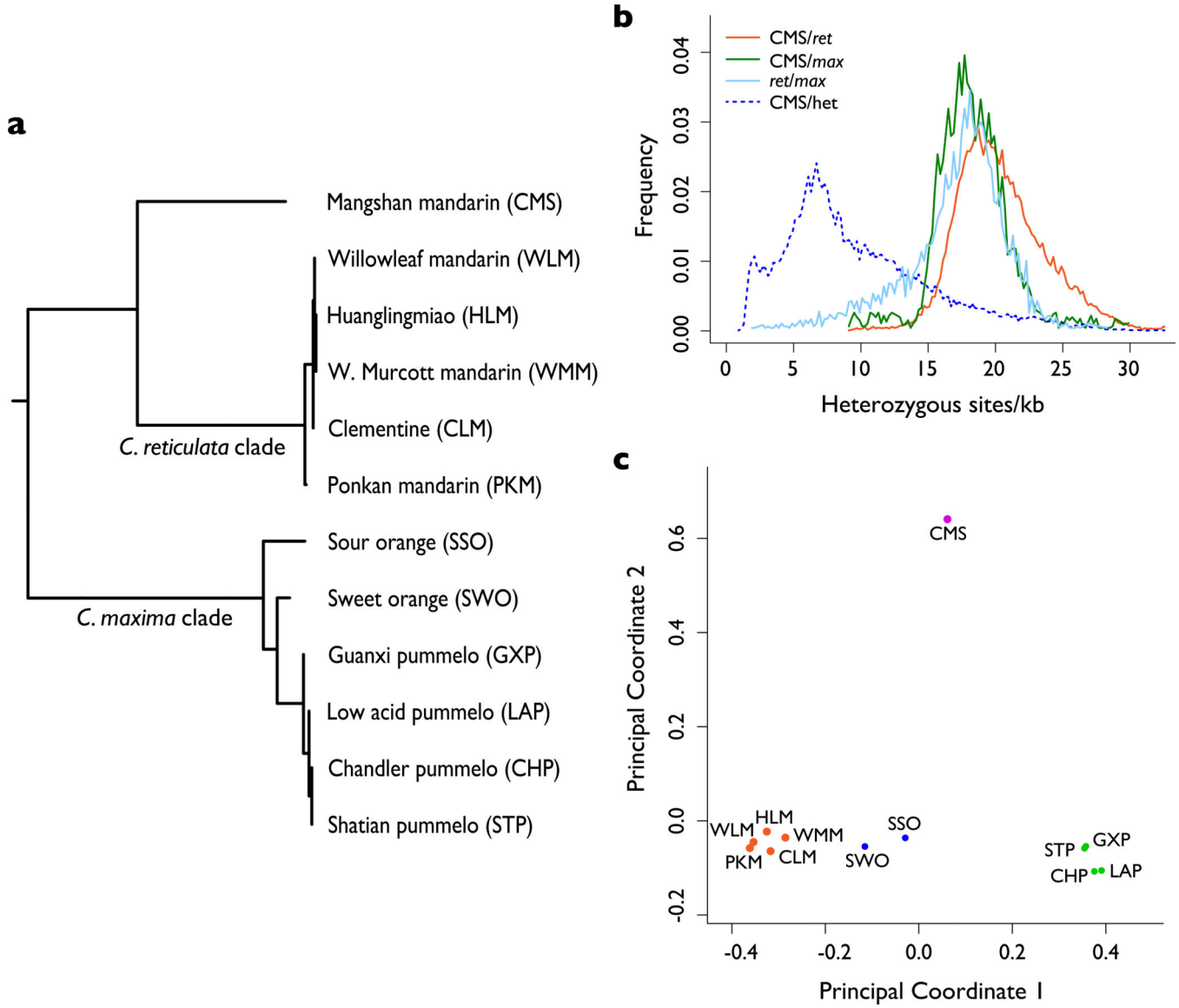

Figure 4. Mangshan mandarin is a species distinct from *C. maxima* and *C. reticulata*

(**a**) Midpoint-rooted neighbor-joining phylogenetic tree of citrus chloroplast genomes. (**b**) The frequency distributions of the pairwise sequence divergences (across 100 kb windows) between Mangshan mandarin (CMS) and *C. maxima* (green), CMS and *C. reticulata* (orange), *C. reticulata* and *C. maxima* (light blue), as well as the distinctly lower CMS intrinsic nucleotide diversity (dashed blue). (c) The first two coordinates of principal coordinate analysis of the citrus nuclear genomes, based on pairwise distances and the metric multidimensional scaling. The *C. maxima* - *C. reticulata* axis (Principle coordinate 1, 47.5% variance) separates pummelos (green) from mandarins (orange), with oranges (blue) lying in between; Principle coordinate 2 (19.6% of variance) separates CMS (purple) from the others.

Table 1

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

Three letter abbreviations as used throughout this work and common systematic designation are shown. Sequence depth reported as count of aligned reads to reference, after removal of duplicate reads. Chloroplast genome type inferred from shotgun reads aligning to the sweet orange chloroplast genome[38], with M indicating mandarin type and P indicating pummelo type. Diploid nuclear genotype proportions refer to fraction of genome in megabases using the HCR physical map (proportions of unknown genotype are not shown but can be inferred by subtracting the three genotype proportions from 100%). The last two columns show proportions of *C. maxima* and *C. reticulata* haplotypes, and are derived from the three genotype proportions. max. = *C. maxima*; ret. = *C.reticulata*.

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

[*] Ponkan mandarin is widely assumed to represent *C. reticulata*, but as shown here it has substantial admixture from *C. maxima*.