Genotyping by sequencing can reveal the complex mosaic genomes in gene pools resulting from reticulate evolution: a case study in diploid and polyploid citrus
Dalel Ahmed, Aurore Comte, Franck Curk, Gilles Costantino, François Luro, Alexis Dereeper, Pierre Mournet, Yann Froelicher, Patrick Ollitrault

To cite this version:
Dalel Ahmed, Aurore Comte, Franck Curk, Gilles Costantino, François Luro, et al.. Genotyping by sequencing can reveal the complex mosaic genomes in gene pools resulting from reticulate evolution: a case study in diploid and polyploid citrus. Annals of Botany, Oxford University Press (OUP), 2019, 123 (7), pp.1231-1251. 10.1093/aob/mcz029. hal-02622898

HAL Id: hal-02622898
https://hal.inrae.fr/hal-02622898
Submitted on 26 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Genotyping by sequencing can reveal the complex mosaic genomes in gene pools resulting from reticulate evolution: a case study in diploid and polyploid citrus

Dalel Ahmed1, Aurore Comte2,3, Franck Curk4, Gilles Costantino1, François Luro1, Alexis Dereeper2,3, Pierre Mournet4,5, Yann Froelicher4,6 and Patrick Ollitrault4,6,*

1UMR AGAP, INRA, CIRAD, Montpellier SupAgro, Université de Montpellier, F-20230 San Giuliano, France, 2IRD, CIRAD, Université de Montpellier, IPME, F-34394 Montpellier, France, 3South Green Bioinformatics Platform, Bioversity, CIRAD, INRA, IRD, F-34394 Montpellier, France, 4UMR AGAP, INRA, CIRAD, Montpellier SupAgro, Université de Montpellier, F-34398 Montpellier, France, 5CIRAD, UMR AGAP, F-34398 Montpellier, France and 6CIRAD, UMR AGAP, F-20230 San Giuliano, France.

* For correspondence. E-mail patrick.ollitrault@cirad.fr

Received: 6 November 2018 Returned for revision: 17 January 2019 Editorial decision: 12 February 2019 Accepted: 18 February 2019

• Background and Aims Reticulate evolution, coupled with reproductive features limiting further interspecific recombinations, results in admixed mosaics of large genomic fragments from the ancestral taxa. Whole-genome sequencing (WGS) data are powerful tools to decipher such complex genomes but still too costly to be used for large populations. The aim of this work was to develop an approach to infer phylogenomic structures in diploid, triploid and tetraploid individuals from sequencing data in reduced genome complexity libraries. The approach was applied to the cultivated Citrus gene pool resulting from reticulate evolution involving four ancestral taxa, C. maxima, C. medica, C. micrantha and C. reticulata.

• Methods A genotyping by sequencing library was established with the restriction enzyme ApelK1 applying one base (A) selection. Diagnostic single nucleotide polymorphisms (DSNPs) for the four ancestral taxa were mined in 29 representative varieties. A generic pipeline based on a maximum likelihood analysis of the number of read data was established to infer ancestral contributions along the genome of diploid, triploid and tetraploid individuals. The pipeline was applied to 48 diploid, four triploid and one tetraploid citrus accessions.

• Key Results Among 43 598 mined SNPs, we identified a set of 15 946 DSNPs covering the whole genome with a distribution similar to that of gene sequences. The set efficiently inferred the phylogenomic karyotype of the 53 analysed accessions, providing patterns for common accessions very close to that previously established using WGS data. The complex phylogenomic karyotypes of 21 cultivated citrus, including bergamot, triploid and tetraploid limes, were revealed for the first time.

• Conclusions The pipeline, available online, efficiently inferred the phylogenomic structures of diploid, triploid and tetraploid citrus. It will be useful for any species whose reproductive behaviour resulted in an interspecific mosaic of large genomic fragments. It can also be used for the first generations of interspecific breeding schemes.

Key words: Citrus, reticulate evolution, mosaic genome, GBS, polyploids, SNPs, phylogenomic karyotype.

INTRODUCTION

Reticulate evolution is recognized as a major evolutionary process of eukaryotes and as a source of genetic diversity (Arnold, 2006). Interspecific and introgressive hybridization, recombination between genes, horizontal gene transfer and infectious heredity are the main mechanisms involved (Posada and Crandall, 2001; Linder and Rieseberg, 2004; Makarenkov and Legendre, 2004). Hybridization of genetically distinguishable populations, groups or taxa, leading to the production of viable hybrids (Barton and Hewitt, 1985; Mallet, 2005), has long been known to be involved in the emergence of plant species (Stebbins, 1950, 1959; Rieseberg, 1997; Abbott et al., 2010, 2013). Hybridization between species or subspecies has a significant weight in evolving processes including speciation, adaptation and extinction (Dowling and Secor, 1997; Barton, 2001; Yakimowski and Rieseberg, 2014). It can lead to rapid genomic changes (Baack and Rieseberg, 2007) and is an important source of genetic variability. Stebbins (1959) suggested that a high degree of genetic variability was required for major evolutionary advances; hence interspecific hybridization appears to be a predominant evolutionary force in plants. The evolutionary history of the concerned species cannot be correctly described using phylogenetic trees, but rather appears as a network (Stebbins, 1950; Grant, 1981; Arnold, 1997; Doolittle, 1999; Otto and Whitton, 2000) or a ‘Web of life’ (Arnold and Fogarty, 2009), generating phylogenetic discordance between nuclear and cytoplasmic (mitochondrial and chloroplast) genomes, and between different regions of the same nuclear genome (Pamilo and Nei, 1988; Rieseberg and Soltis, 1991; Linder and Rieseberg, 2004; Beiko and Hamilton, 2006). Reticulations lead not only to faulty phylogenetic conclusions, but also to interspecific heterozygosity of large portions of the genome when vegetative propagation involving aponexomes, bulbs, tubers, corms, suckers, etc. takes place immediately or...
a few generations after reticulation events as described in fern (Dyer et al., 2012), banana (Perrier et al., 2009, 2011) or citrus (Curk et al., 2014). Deciphering this type of complex genome needs appropriate analytical approaches and tools based on a whole-genome scan.

The emergence of NGS (next-generation sequencing) technologies has considerably changed ways of analysing plant evolution, moving from phylogenetics to phylogenomics. The analysis of whole-genome variability has become possible and has already provided new information on the history of domestication of some cereals (Mascher et al., 2016; Meyer et al., 2016; Ramos-Madrigal et al., 2016; Pankin et al., 2018) and fruit crops, including grapes (Zhou et al., 2017), apples (Duan et al., 2017) and citrus (Wu et al., 2014, 2018). However, whole-genome re-sequencing (WGS) remains costly for studying large populations. Therefore, cost-effective methods combining NGS and a reduction of the complexity of genomes have been developed, such as genotyping by sequencing (GBS) (Elshire et al., 2011), restriction-site-associated DNA sequencing (RADseq) (Miller et al., 2007; Baird et al., 2008; Davey and Blaxter, 2011; Peterson et al., 2012) and sequenced-based genotyping (SBG) (Truong et al., 2012). These methods allow sufficient coverage of the genomes and are robust means for sampling whole genomes. They enable the analysis of large segregating progenies and marker trait association studies based on linkage disequilibrium and even genomic selection (Baxter et al., 2011; Ma et al., 2012; Poland et al., 2012; Ward et al., 2013; Wang et al., 2016; Curtolo et al., 2017). The efficiency of these methods has been demonstrated not only by constructing genetic maps and conducting genetic association studies, but also by carrying out diversity analyses and revealing phylogenetically informative variation (Garcia et al., 2013; Escudero et al., 2014; Penjor et al., 2014, 2016; Hamon et al., 2017; Oueslati et al., 2017; Stetter and Schmid, 2017). More specifically, GBS has been used to perform genetic studies of numerous diploid and polyploid species, including maize (Crossa et al., 2013), wheat (Poland et al., 2012; Heslot et al., 2013), burley (Poland et al., 2012; Liu et al., 2014), rice (Huang et al., 2009; Courtois et al., 2013; Spinadel et al., 2013), ryegrass (Byrne et al., 2013), soybean (Sonah et al., 2013), chickpea (Verma et al., 2015), sugarcane (Almeida Balsalobre et al., 2017), banana (Martin et al., 2016) and citrus (Oueslati et al., 2017). However, for polyploid species, due to the generally low read depths at individual single nucleotide polymorphism (SNP) loci, genotyping has been limited to the identification of homozygous genotypes (nulliplex or quadruplex for a tetraploid) or heterozygous genotypes, joining the different classes of heterozygosity (simplex, duplex, triplex for a tetraploid) in the same genotyping class (Clevenger et al., 2015; Rocher et al., 2015; Almeida Balsalobre et al., 2017; Yang et al., 2017). For tetraploid potatoes, a technical solution has been proposed to improve the individual SNP read depths by combining GBS with enriched cultivar-specific DNA sequencing libraries using an in-solution hybridization method (SureSelect), reducing the genome to 807 target genes distributed across the genomes (Uitdeewiligen et al., 2013). New analytical methods have also been proposed to deal with the low read depths. Rather than calling genotypes, Ashraf et al. (2014) and Sverrisdóttir et al. (2017) directly used the variant allele frequencies at each data point for association studies and genomic selection from GBS data. New pipelines have also been proposed to estimate allele doses at an individual locus (McKinney et al., 2018; Bastien et al., 2018), but it remains challenging.

The Citrus genus is a good example of a gene pool resulting from reticulate evolution, where apomixes and vegetative propagation have fixed ancient reticulation events and limited further interspecific recombination, resulting in mosaics of large genome fragments from different species (Nicolosi et al., 2000; Wu et al., 2014, 2018; Curk et al., 2016). Molecular marker analyses enabled the main lines of the phylogeny of the different cultivated species of Citrus to be drawn and the identification of the various domestication events (Federici et al., 1998; Nicolosi et al., 2000; Barkley et al., 2006; Li et al., 2010; Garcia-Lor et al., 2012, 2013; Ollitrault et al., 2012a, b; Ramadugu et al., 2013; Curk et al., 2016). Four taxa [C. medica L. (citron), C. reticulata Blanco (mandarin), C. maxima (Burm.) Merr. (pummelo) and C. micrantha Wester (papeda)] have been identified as being the ancestors of most of the cultivated citrus (Nicolosi et al., 2000; Garcia-Lor et al., 2012; Ollitrault et al., 2012b; Ramadugu et al., 2013; Curk et al., 2014, 2015; Wu et al., 2018). These four ancestral taxa, which are still sexually compatible, were differentiated by foundation effects and allopatric evolution in four South-East Asian geographic regions ranging from the southern Himalayas to Indonesia. Pummelos originated in the Malay Archipelago and Indonesia. Citrons evolved in north-eastern India and in the nearby areas of Myanmar and China. Mandarins were diversified over a wide region which includes Vietnam, southern China and Japan, while C. micrantha is endemic to the Philippine islands (Wester, 1915; Tanaka, 1954; Webber et al., 1967; Scora, 1975). Secondary species [C. sinensis (L.) Osb. (sweet orange), C. aurantium L. (sour orange), C. paradisi Macf. (grapefruit), C. limon (L.) Burnm. (lemon) and C. aurantifolia (Christm.) Swing. (lime)] and modern cultivars are the result of hybridizations between the four basic taxa (Nicolosi et al., 2000; Garcia-Lor et al., 2013; Curk et al., 2016) engendering the wide genetic and phenotypic diversity observed among them. In terms of morphological characteristics (Ollitrault et al., 2003), carotenoid compositions (Fanciullino et al., 2006) and the distribution of coumarins and furanocoumarins (Dugrand-Judek et al., 2015), the structure of phenotypic variability is closely linked with the reticulate evolution of the gene pool. Therefore, in parallel with the search for the origin of cultivated forms and the optimization of genetic resources management, deciphering the phylogenomic structures of modern cultivars will open the way for association studies based on ancestral haplotypes and phylogenomic-based reconstruction breeding strategies (Rouiss et al., 2018). The accurate study of citrus interspecific mosaic genomes started with the release of the first high-quality citrus reference haploid genome by the International Citrus Genomics Consortium (ICGC; Wu et al., 2014). WGS data revealed Citrus maxima introgressions in traditional mandarin genomes (Wu et al., 2014) and the interspecific mosaic structure of sweet orange (Xu et al., 2013; Wu et al., 2014), sour orange and clementine (Wu et al., 2014). More recently, WGS data (Wu et al., 2018), including the four Citrus ancestral species and modern varieties, revealed the mosaic genome structures of the other most important horticultural groups, such as grapefruit, lemon and lime, and confirmed C. maxima introgressions in all domesticated mandarins.
A GBS approach was recently applied to analyse the interspecific admixture of diploid secondary species and modern varieties resulting from two Citrus gene pools, C. reticulata and C. maxima (Oueslati et al., 2017). To date, the phylogenomic structures of the citrus polyploid germplasm remain unpublished.

The objectives of the present work were to (1) develop a GBS approach in Citrus with a dense genotyping and a good depth, to decipher – at limited cost – the phylogenomic structures of large diploid and polyploid populations originating from a limited number of interspecific recombinations between C. reticulata, C. maxima, C. medica and C. micrantha gene pools; (2) provide a reference matrix of diagnostic SNP (DSNP) markers for the four Citrus ancestral taxa; (3) implement a generic workflow for mosaic genome analysis from GBS data of diploid and polyploid populations resulting from reticulate evolution; and (4) analyse the phylogenomic structure of modern varieties of the main citrus diploid and polyploid horticultural groups. As proof of concept, 53 citrus accessions, including several varieties already analysed using WGS (Wu et al., 2014, 2018), were sequenced in a single Illumina HiSeq 2000 line, using the restriction enzyme ApeKI and a selective PCR for GBS library preparation. Close to 16 000 DSNPs were identified and successfully used to decipher the complex genomes of the 53 accessions, using a workflow based on maximum likelihood analysis of multilocus ancestral read numbers. The GBS approach we developed combined with the reference DSNP matrix will be useful for any study of germplasm and hybrids resulting from breeding within the Citrus genus. The implemented workflow for the analysis of mosaic genomes is available online and will be useful for species with any number of identified ancestral taxa, for diploid, triploid and tetraploid accessions.

MATERIALS AND METHODS

Plant material

The study covered 53 accessions from the collection of the Inra-Cirad Citrus Biological Resource Center in San-Giuliano, Corsica, France (Luro et al., 2018). The varieties belong to the Citrus genus, and 29 of them are representative of the four ancestral taxa: 15 mandarins, six pummelos, six citrons and two papedas. They were used to identify diagnostic markers of the basic taxa. The other varieties, which are diploid, triploid and tetraploid, came from admixtures of the four ancestral taxa: two sour oranges (C. aurantium), two sweet oranges (C. sinensis), five lemons (C. limon, C. limonia Osb., C. meyeri Y. Tan. and C. jambhiri Lush.), eight limes (C. aurantifolia, C. latifolia Tan., C. excelsa Wester, C. limetoides Tan.), one ‘Alemon’ (C. macrophylla Wester), three grapefruits (C. paradisi), one bergamot (C. bergamia Risso & Poit.), one clementine (C. clementina Hort. ex Tan.) and one limonette (C. limetta Risso). In order to validate our method of deciphering the citrus interspecific mosaic structure, we included some accessions already described from WGS data by Wu et al. (2014, 2018). A summary list of the varieties analysed with their classification in two widely used taxonomic systems [the Tanaka (1954) and Swing and Reece (1967) systems] is available in Supplementary Data Table S1. Recent genetic and genomic studies demonstrated the limits of both systems resulting from reticulate evolution of the citrus gene pool and vegetative propagation of interspecific combination by apomictic seeds (Curk et al., 2016; Wu et al., 2018). Herein we refer to the Tanaka system for the secondary species (the types issued from interspecific combinations); indeed, although they cannot be considered as true species, the Tanaka classification has the advantage of distinguishing secondary taxa that have arisen from different reticulation events. Supplementary Data Table S1 also specifies whether the phylogenomic structure of each accession has already been analysed from WGS (Wu et al., 2014, 2018) or GBS (Oueslati et al., 2017) or was analysed for the first time in the present study.

GBS analysis

Library preparation and sequencing. Following the protocol of Oueslati et al. (2017), genomic DNA was isolated using the Plant DNAeasy® kit (Qiagen), according to the manufacturer’s instructions. Several in silico tests were carried out using numerous types of restriction enzymes and selective primers. The method selected consists of using the restriction enzyme ApeKI and adding a selective base (A) during the PCR step of GBS library preparation as it was found to provide a good combination of tag density and read numbers per tag. ApeKI also has the advantage of cutting DNA preferentially in gene sequences enabling better quality genotype calling (Oueslati et al., 2017). The genomic DNA concentration was adjusted to 20 ng mL⁻¹, and ApeKI GBS libraries were prepared following the protocol described by Eslhire et al. (2011). DNA of each sample (200 ng) was digested with ApeKI (New England Biolabs, Hitchin, UK). Digestion took place at 75 °C for 2 h and then at 65 °C for 20 min to inactivate the enzyme. The ligation reaction was completed in the same plate as the digestion, again using T4 DNA ligase (New England Biolabs) at 22 °C for 1 h, and the ligase was inactivated prior to pooling the samples by holding it at 65 °C for 20 min. Ligated samples were pooled and PCR-amplified in a single tube. Complexity was further reduced using PCR primers with one selective base (A) as performed by Sonah et al. (2013). Single-end sequencing was performed on a single lane of an Illumina HiSeq2000. The Illumina HiSeq 2000 sequencing raw data are available in the NCBI SRA (Sequence Read Archive), under the accession numbers SRP109295 for the 21 mandarin, pummelo, orange, grapefruit and clementine sequences already published in Oueslati et al. (2017; Supplementary Data Table S1) and PRJNA388540 for the 32 new citrus accessions. Keygene N.V. owns patents and patent applications protecting its Sequence Based Genotyping technologies.

SNP genotype calling for diploid germplasm

The Tassel 4.0 pipeline (Glaubitz et al., 2014) was used to call SNPs from the DNA sequence reads from the Illumina raw data (unfiltered fastq file). The Tassel 4.0 GBS pipeline identified good quality, unique, sequence reads with barcodes.
These sequence tags were aligned to the C. clementina 1.0 reference genome (https://phytozome.jgi.doe.gov/pz/portal.html#info?alias=Org_Cclementina) using Bowtie2 v2.2.6 (Langmead and Salzberg, 2012). For genotype calling, five reads were considered as a minimum below which they were considered as missing data (Danec et al., 2011). We finally only considered diallelic polymorphic positions with <30% of missing data for the 29 representative accessions of C. reticulata, C. maxima, C. medica and C. micrantha, and a minor allele frequency (MAF) >0.05.

**Genetic parameters**

The following parameters were used to describe the genetic diversity within and between the ancestral taxa: $H_s$, the observed heterozygosity; $H_e$, the expected proportion of heterozygous loci per individual under Hardy–Weinberg equilibrium defined as $H_e = 1 − \sum pi^2$, with $pi$ the frequency of a given allele in the sub-population concerned or in the whole population; and $F_w$, the fixation index (Wright, 1951) defined as follows:

$$F_w = 1 − \frac{H_o}{H_e}$$

They were calculated using GENETIX v. 4.03 software (Belkhir et al., 1996–2004) based on the 43 598 diallelic selected markers.

The analysis consisting of identifying the diagnostic markers of the four basic taxa was mainly based on $G_{ST}$ parameter estimations (Nei, 1973). $G_{ST}$ is the coefficient of gene differentiation which measures differentiation among sub-populations. It is equivalent to Wright’s $F_{ST}$ for two alleles and ranges from zero to one. The higher the value, the more differentiated the ancestral taxa.

The identification of diagnostic markers of the four ancestral taxa from the GBS data is schematized in the workflow in Fig. 1. Some of the accessions cited above, mostly in the mandarin group, are already known to be non-pure (Curk et al., 2014; Wu et al., 2014, 2018). They were the result of a domestication process of the real ancestors which led to interspecific introgressions. Consequently, implementing a diagnostic marker set required the identification of interspecific introgressions among the varieties considered as representatives of the ancestral taxa, and of removing these regions of the variety under consideration from the analysis. This process provided a better estimation of the allelic frequencies of the ancestral taxa and hence of the $G_{ST}$ parameter in the four basic taxa. The identification of the interspecific introgressed areas was based on the pattern of two parameters along the genome using consecutive non-sliding 20 SNP windows: (1) the average heterozygosity estimated from the matrix of SNP positions and (2) the similarity of the accession to the centroid of each of the four horticultural representative groups (the allelic frequencies of the centroid being the average frequency of the varieties of the considered group). It was expected that introgressed areas would display significant discontinuity of these patterns according to the level of differentiation between the two taxa involved. Indeed, heterozygous introgressions resulted in regions with an increase in heterozygosity and a decrease in the similarity, while homozygous introgressions resulted in a deep variation in the similarity patterns. To better visualize the pattern discontinuities, SNPs that were informative for the differentiation of one out of the four horticultural groups, representative of the ancestral taxon ($G_{ST}$ >0.5), were filtered out. Once the interspecific introgressions were removed (considered as missing data), the allelic frequencies in the four ancestral taxa and the $G_{ST}$ parameter between each ancestral taxon and the three others were estimated again. We then considered SNPs with a $G_{ST}$ value (the taxon
concerned relative to a sub-population of the three other ancestral taxa) >0.9 as diagnostic markers of a given taxon.

**Analysis of the interspecific mosaic structure of complex genomes**

The objective was to develop a generic pipeline to decipher complex genomes resulting from reticulate evolution at diploid and polyploid levels, based on the availability of a set of diagnostic markers of the ancestral taxa involved (all along the genome) and GBS data of new populations obtained with the same experimental procedure as the reference DSNP set. According to our experimental data (see below) and reports in the literature (Bastien et al., 2018; McKinney et al., 2018), it is often difficult to estimate allelic doses at a single locus accurately in heterozygous polyploids from relative allele read frequencies resulting from GBS experiments. We developed an approach based on maximum likelihood analysis applied to multilocus numbers of reads of consecutive DSNPs of the same ancestor, that can be used for diploid, triploid and tetraploid plants. This approach is described below in the concrete case of citrus with four ancestral taxa, but the tool we developed can be used with models of any number of ancestral taxa. An illustration of the process for a triploid plant is provided in Fig. 2.

The first step aims to estimate the doses of the ancestral genome fragment along the genome. For each ancestral taxon, the citrus genome was segmented in windows of \( w \) consecutive DSNPs (Fig. 2A) and the doses of the ancestral taxon considered were estimated for each window by maximum likelihood analysis (Fig. 2C). The detail for the maximum likelihood analysis for diploid, triploid and tetraploid individuals is provided in Supplementary Data Text S1.

During the preceding step, the number and position of windows varied between the ancestral taxa according to the density and positions of the DSNPs. Therefore, the next step was to integrate the information obtained for the different ancestral taxa doses along the genome.

The genome was physically sub-divided into successive fragments of \( z \) kb (by default: \( z = 100 \)) (Fig. 2D). For each ancestor and for each genomic fragment, the corresponding window of \( w \) DSNPs was identified and the ancestral dose of this window was attributed to the genomic fragment. A non-phased representation of karyotypes with two, three and four chromosomes for diploid, triploid and tetraploid plants, respectively, was then generated from the ancestral doses of each genome fragment (Fig. 2F). For a given genome fragment, if the sum of the allelic doses of the different ancestors differed from the ploidy level of the plant concerned, the phylogenomic origin of the fragment was considered as undefined. Likewise, if one of the doses of the different ancestors was undefined, the phylogenomic origin of the fragment was considered as undefined (Fig. 2E). When phased haplotypes were known for the parental genomes, we proposed manually phased karyotypes for the concerned accession, assuming the lower number of recombination events as the best model.

The tool we developed (TraceAncestor) allows the user to define the number of DSNPs per window (by default: \( w = 10 \)), the sequencing error rate (by default: \( e = 0.01 \)) and the threshold for
Fig. 2. Example of local ancestor allele dose estimation for a triploid accession. (A) Definition of non-overlapping windows of ten DSNPs for each ancestral taxon: \( \text{w} \), window of ten DSNPs. (B) Number of reads of the considered ancestor allele/number of reads of the alternative allele. (C) Estimation of allelic dosage of each ancestor per window of ten DSNPs [each pair of dose hypotheses are compared by maximum likelihood (LOD) test; if, for a pair including the more probable hypothesis, \(-3 < \text{LOD} < 3 \rightarrow \) indeterminacy]. (D) Division of the chromosome into non-overlapping windows of 100 kb; the allelic dosage of this window is deduced from that of the ten DSNPs window that include the 100 kb window. (E) If the sum of allelic dosage of the four classes of DSNPs is different from the expected ploidy (here 3) \( \rightarrow \) indeterminacy (grey). (F) Unphased karyotype automatic drawings. Blue, \( C. \text{maxima} \); yellow, \( C. \text{medica} \); green, \( C. \text{micrantha} \); red, \( C. \text{reticulata} \); grey, indeterminacy.
Genotype calling and varietal diversity

Figure 1 shows the workflow for the identification of diagnostic markers. The 53 varieties considered were part of two 55-plex libraries sequenced in two lanes of a Hiseq 2000 according to the Cornell GBS methodology (Elshire et al., 2011) using ApelKI as the restriction enzyme and a selective primer. A total of 344.8 million reads were obtained. The Tassel pipeline was used for genotype calling, and 314.2 million of these reads were validated (bar code, restriction site plus insert), and 290.7 million were mapped on the clementine reference genome (Wu et al., 2014). The average number of reads per variety was 2.2 million, ranging from 609 890 for ‘Meyer’ lemon to 5.68 million for ‘Shekwasha’ mandarin (Supplementary Data Fig. S1). A total of 2.045 million tags (unique sequence with at least five reads) were identified, of which half were only mapped once on the clementine reference genome. Genotype calling from the tags with a single hit map was undertaken considering a position with less than five reads as missing data. A total of 43 598 diallelic SNPs were selected, and filtered for sites with <30 % of missing data on the 39 representative accessions. The 35 and 84 % of the SNPs retained had, respectively, <5 % and <25 % of missing data (Supplementary Data Fig. S2A). At the individual level, 29.6 and 90.7 % of the varieties had, respectively, <5 % and <25 % of missing data (Supplementary Data Fig. S2B). ‘Meyer’ lemon had the highest rate of missing data: 35 %. The distribution of the reads number per marker (Supplementary Data Fig. S3) appeared to be globally homogeneous among the nine chromosomes, with a mean value of 1024 reads. However, a decrease in the number of reads was observed in the middle of chromosomes 2, 4, 5, 8 and 9.

The distribution of the 43 598 mined polymorphisms on the nine chromosomes is reported in Table 1. The number of diallelic SNPs varied between 3611 SNPs on chromosome 8 and 7743 SNPs on chromosome 3. Little variation was observed among the expected heterozygosity values along the nine chromosomes, with an average of 0.309, or in the observed heterozygosity values which ranged between 0.197 (chromosome 2) and 0.227 (chromosome 6), with an average of 0.213. According to the Hardy–Weinberg equilibrium, the analysed population displayed a heterozygote deficiency with the $F_w$ parameter equal to 0.282.

Based on the 43 598 diallelic SNPs, we performed a three-dimensional representation of the PCA to examine the genetic diversity of the 48 diploid citrus accessions (Fig. 3). The four main observed clusters corresponded to the four ancestral taxa (pummelos, mandarins, citrons and papedas). The first three axes represent 61.54 % of total diversity and clearly separate the four clusters of the ancestral taxa. Other clusters made of secondary species appeared between the ancestral clusters and revealed their genetic relationship. Lemons ['Lisbon' lemon (33), 'Meyer' lemon (34), 'Eureka' lemon (35), 'Rough' lemon (47) and 'Volkamer' lemon (48)], ‘Palestine’ sweet lime (38), ‘Marrakech’ limonette (39) and ‘Rangpur’ lime (46) were in an intermediate position between C. reticulata and C. medica clusters. Bergamot (30) was located close to the mandarin group but still in an intermediate position between the mandarin, pummelo and citrus groups. Grapefruits ['Duncan' (43), ‘Marsh’ (44) and ‘Star Ruby’ (45)], sour oranges ['Seville’ (31) and ‘Bouquetier de Nice’ (32)] and sweet oranges ['Valencia late' (41) and ‘Washington navel’ (42)], rather logically given their origin revealed by markers (Curt et al., 2015), previous GBS studies (Oueslati et al., 2017) and WGS analysis (Wu et al., 2014, 2018), were in an intermediate position between C. reticulata and C. maxima. ‘Nestour’ lime (36) and ‘Alemow’ (40) were located between C. medica and C. micrantha, in agreement with their origin proposed by Curt et al. (2016).

Diversity among the four ancestral taxa and search for diagnostic markers

Genetic parameters. Analyses of the diversity among the 29 representative accessions (Table 2) revealed a marked decrease in the number of reads was observed in the middle of chromosomes 2, 4, 5, 8 and 9.

### Table 1. Polymorphisms mined from GBS data on 53 citrus varieties along the nine chromosome

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>$H_o$</th>
<th>$H_e$</th>
<th>$F_w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>4180</td>
<td>0.208 ± 0.115</td>
<td>0.314 ± 0.134</td>
<td>0.308 ± 0.279</td>
</tr>
<tr>
<td>C2</td>
<td>5536</td>
<td>0.197 ± 0.106</td>
<td>0.312 ± 0.135</td>
<td>0.323 ± 0.278</td>
</tr>
<tr>
<td>C3</td>
<td>7743</td>
<td>0.211 ± 0.122</td>
<td>0.308 ± 0.137</td>
<td>0.285 ± 0.280</td>
</tr>
<tr>
<td>C4</td>
<td>4586</td>
<td>0.200 ± 0.109</td>
<td>0.307 ± 0.137</td>
<td>0.306 ± 0.265</td>
</tr>
<tr>
<td>C5</td>
<td>5565</td>
<td>0.215 ± 0.135</td>
<td>0.311 ± 0.136</td>
<td>0.280 ± 0.317</td>
</tr>
<tr>
<td>C6</td>
<td>3875</td>
<td>0.227 ± 0.130</td>
<td>0.309 ± 0.139</td>
<td>0.249 ± 0.264</td>
</tr>
<tr>
<td>C7</td>
<td>3739</td>
<td>0.213 ± 0.117</td>
<td>0.306 ± 0.137</td>
<td>0.276 ± 0.261</td>
</tr>
<tr>
<td>C8</td>
<td>3611</td>
<td>0.222 ± 0.127</td>
<td>0.309 ± 0.135</td>
<td>0.256 ± 0.282</td>
</tr>
<tr>
<td>C9</td>
<td>4763</td>
<td>0.224 ± 0.133</td>
<td>0.309 ± 0.135</td>
<td>0.255 ± 0.295</td>
</tr>
<tr>
<td>Total</td>
<td>43 598</td>
<td>0.213 ± 0.01</td>
<td>0.309 ± 0.002</td>
<td>0.282 ± 0.025</td>
</tr>
</tbody>
</table>

$n$, number of polymorphisms; $H_o$, observed heterozygosity; $H_e$, expected heterozygosity; $F_w$, Wright fixation index; C1–C9, the nine chromosomes of the reference clementine genome (Wu et al., 2014).
Ahmed et al. — GBS reveals the genome admixture of diploid and polyploid citrus


Search for ancestral taxa diagnostic markers (DSNPs). Removing the interspecific introgressed areas from the varieties representative of the four ancestral taxa was an important step to estimate effectively the allelic frequencies of the ancestral taxa and the differentiation parameter \( G_{st} \) between the four ancestral taxa at each polymorphic position. The introgressions were identified through the analysis of the discontinuity in the pattern of two parameters along the genome: the heterozygosity and the similarity between the accession and the centroids of each horticultural group, representative of the ancestral taxa.

We examined the distribution of the observed heterozygosity of the diploid accessions with 100 polymorphic positions per window. Two main modes of distribution were observed among the varieties plotted individually (Fig. 5) or in sets (Fig. 6). These two modes correspond to intraspecific and interspecific heterozygosity, with values ranging between 0 and 0.2 and 0.2 and 0.7, respectively. Three distinct types of accessions were highlighted. The first type displayed a unimodal distribution with a high value (the average value of each accession was between 0.3 and 0.4) corresponding to interspecific heterozygosity. Accessions of this type probably result from direct two-way or three-way interspecific hybridization.
Ahmed et al. — GBS reveals the genome admixture of diploid and polyploid citrus

Table 2. Diversity of the 29 accessions representative of the four ancestral taxa

<table>
<thead>
<tr>
<th>Accession</th>
<th>n</th>
<th>$H_o$</th>
<th>$H_e$</th>
<th>$F_w$</th>
<th>$G_{st}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandarins (Na = 15)</td>
<td>18 567</td>
<td>0.121 ± 0.200</td>
<td>0.110 ± 0.162</td>
<td>-0.107</td>
<td></td>
</tr>
<tr>
<td>Pummelos (Na = 6)</td>
<td>7325</td>
<td>0.086 ± 0.212</td>
<td>0.070 ± 0.154</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>Citrons (Na = 6)</td>
<td>7156</td>
<td>0.041 ± 0.163</td>
<td>0.044 ± 0.128</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>Papedas (Na = 2)</td>
<td>2285</td>
<td>0.016 ± 0.068</td>
<td>0.025 ± 0.113</td>
<td>-0.907</td>
<td></td>
</tr>
<tr>
<td>Total (Na = 29)</td>
<td>35 333</td>
<td>0.066 ± 0.04</td>
<td>0.063 ± 0.031</td>
<td>-0.1237</td>
<td>0.7831139</td>
</tr>
</tbody>
</table>

$n$, number of polymorphisms; $H_o$, observed heterozygosity; $H_e$, expected heterozygosity; $F_w$, Wright fixation index; $G_{st}$, interpopulation differentiation parameter; Na, number of accessions per taxon.

Fig. 4. Hierarchical cluster analysis of the 29 representative accessions computed from the 43 598 diallelic SNPs.

Sour oranges, all lemons, ‘Marrakech’ limonette, ‘Rangpur’, ‘Palestine’ and ‘Nestour’ limes, as well as ‘Alemow’ displayed this pattern. A higher mid-value was observed for ‘Rough’ lemon than for sour orange [explained by the greater differentiation of $C. \text{reticulata}$ (mandarins) with $C. \text{medica}$ (citrons) than $C. \text{maxima}$ (pummelos)]. Indeed, from WGS data, Wu et al. (2014, 2018) concluded that ‘Rough’ lemon and sour orange resulted from direct interspecific hybrids of $C. \text{reticulata}$ with $C. \text{medica}$ and $C. \text{maxima}$, respectively. The second type grouped the representative accessions of the basic taxa, except for the majority of mandarins. Pummelos, citrons and papedas displayed unimodal distribution, with average values of 0.09, 0.04 and 0.05, respectively. The representative mandarins belong to the third type of accessions with a bimodal distribution of heterozygosity, such as sweet orange, grapefruit, clementine and bergamot. The interspecific admixture among these accessions was highlighted. The same pattern of distribution of heterozygosity in sweet orange was reported in Wu et al.
Ahmed et al. — GBS reveals the genome admixture of diploid and polyploid citrus

Fig. 5. Violin plots of the heterozygosity distribution in the 48 diploid accessions computed from the average values in successive windows of 100 polymorphic positions along the genome. White dot, median; bar limits; upper and lower quartiles; whiskers, 1.5× interquartile range; light blue, intraspecies; light pink, interspecies.

Ahmed et al. — GBS reveals the genome admixture of diploid and polyploid citrus

Intra species Inter species

Mandarins

‘Shekwasha’ mandarin
‘de Chios’ mandarin
‘Owari Satsuma’ mandarin
‘Nan feng mi chu’ mandarin
‘San Hu Hong Chu’ mandarin
‘Se Hui Gan’ mandarin
‘Sizibat’ mandarin
‘Cleopatra’ mandarin
‘Dancy’ mandarin
‘Fuzhu’ mandarin
‘King’ mandarin
‘Ladu’ mandarin
‘Ponkan’ mandarin
‘Sunki’ mandarin
‘Willowleaf’ mandarin

Pummelos

‘Chandler’ pummelo
‘Timor’ pummelo
‘Deep red’ pummelo
‘Kao Pan’ pummelo
‘Pink’ pummelo
‘Tahtian’ pummelo

Citrons

‘Corsican’ citron
‘Buddha’s Hand’ citron
‘Etrog’ citron
‘Humpang’ citron
‘Mac Veu de Montagne’ citron
‘Poncire commun’ citron

Papedas

‘Small flower’ papeda1
‘Small flower’ papeda2
‘Nules’ clementine

Oranges/Grapefruits

‘Seville’ sour orange
‘Bouquetier de Nice’ sour orange
‘Valencia late’ sweet orange
‘Washington navel’ sweet orange
‘Duncan’ grapefruit
‘Marsh’ grapefruit
‘Star Ruby’ grapefruit

Lemons/Limes/Bergamots

‘Bergamot’
‘Marakech’ limonette
‘Lisbon’ lemon
‘Meyer’ lemon
‘Eureka’ lemon
‘Rough’ lemon
‘Volkamer’ lemon
‘Rangpur’ lime
‘Palestine’ sweet lime
‘Nestour’ lime
‘Alemow’

Heterozygosity

0 0.2 0.4 0.6

Intra species Inter species

Fig. 5. Violin plots of the heterozygosity distribution in the 48 diploid accessions computed from the average values in successive windows of 100 polymorphic positions along the genome. White dot, median; bar limits; upper and lower quartiles; whiskers, 1.5× interquartile range; light blue, intraspecies; light pink, interspecies.

(2014, 2018) from WGS data and in Oueslati et al. (2017) from GBS data. More specifically, the set of mandarins showed a first peak around 0.1, close to the peak of the set of pummelos, and a second slight peak with a mode of approx. 0.3–0.35 (Fig. 6), as observed by Oueslati et al. (2017). At the individual level, the bimodal distribution in ‘Owari Satsuma’ mandarin and
'King' mandarin was particularly clear, a result consistent with those of Wu et al. (2018). As proposed by Wu et al. (2014) and adopted by Oueslati et al. (2017), when examining the representative accessions, we considered that regions with low heterozygosity represent diploid segments which combine two haplotypes from the same species, while regions with high heterozygosity were considered to be hybrid segments combining two haplotypes from two different species. Thus, regions with heterozygosity values >0.2 were assumed to be introgressed and were removed.

The patterns of similarity between each accession and the centroid of the four horticultural groups were also examined. The regions with an increase in heterozygosity were associated with a decrease in similarity to their representative horticultural group and an increase in similarity to the horticultural group involved in the introgression. An example is given for chromosome 2 of the ‘King’ mandarin (Supplementary Data Fig. S4). A heterozygous introgression was clearly identified at the end of the chromosome. Heterozygosity increased with a decrease in similarity, starting at 25 Mb and continuing to the end of the chromosome. Similarity analysis was particularly useful to identify homozygous introgressions as described by Oueslati et al. (2017) for the ‘Ponkan’ variety. Indeed, respective similarity with the reference taxa and the introgressed taxa decreased and increased abruptly. The search for introgressions, based on the patterns of heterozygosity and similarities with centroids of the horticultural groups, was systematically performed on the nine chromosomes of the 29 representative accessions.

Allelic frequencies of the ancestral taxa and the differentiation parameter ($G_{ST}$) were then re-estimated considering the introgressed areas as missing data. All SNPs with $G_{ST}$ >0.9 for one ancestral taxon compared with all others were considered as diagnostic of the taxon concerned. A total of 15 946 DSNPs of the four ancestral taxa distributed along the nine chromosomes (Table 3; Supplementary Data Table S2) were then identified. DSNPs of C. medica represented more than one-third (37.60%) of the total number of DSNPs. The low intraspecific heterozygosity of C. medica described above explains the higher number of diagnostic SNPs in this taxon (5997), and the same is true for the C. micrantha taxon whose DSNPs represent 27.41% of the total. Citrus reticulata and C. maxima are represented by 21.9 and 13.09%, respectively, of the total number of DSNPs. The distribution of the 15 946 DSNPs along the nine chromosomes closely resembled the distribution of the whole set of polymorphisms and is closely linked with the distribution of the gene sequences (Supplementary Data Fig. S5). The selected DSNPs were used to decipher the phylogenomic mosaic structures of the 53 varieties.

**Phylogenomic structure of modern varieties**

Our main objective was to develop a pipeline for the analysis of GBS data which would make it possible to establish the phylogenomic karyotype in diploid, triploid and tetraploid germplasm. For polyploid germplasm, this requires the ability to estimate allelic doses for heterozygous genotypes. Looking at individual SNP loci for the DSNPs of C. medica in the triploid ‘Persian’ lime (Supplementary Data Fig. S6), the frequency of C. medica allele reads per locus did not display a clear bimodal distribution for heterozygous loci (Supplementary Data Fig. S6A) and, consequently, estimated allelic doses are subject to high uncertainty. When working with all reads of ten consecutive loci, the bimodal distribution of the C. medica allele frequency was much clearer (Supplementary Data Fig. S6B), enabling efficient estimation of the dose of C. medica (1/3 and 2/3) in the genomic fragment corresponding to the ten markers considered. For the analysis of diploid and triploid Citrus germplasm, we kept ten DSNPs per window as default to estimate the doses for each ancestral taxon.

Using the TraceAncestor tool that we developed, we inferred the unphased phylogenomic karyotypes of the 53 accessions (Fig. 7). The average phylogenomic contributions of C. reticulata, C. maxima, C. medica and C. micrantha to the modern varieties are presented in Supplementary Data Text S2.

**Validation of the karyotypes inferred from GBS data**

We compared karyotypes obtained from GBS data with those proposed by Wu et al. (2014, 2018) from WGS data.

---

**Fig. 6.** Distribution of the heterozygosity in mandarins, pummelos, citrums, papedas, all the diploid varieties, the ‘Seville’ sour orange and the ‘Rough’ lemon computed from the average values in successive windows of 100 polymorphic positions along the genome.
Ahmed et al. — GBS reveals the genome admixture of diploid and polyploid citrus

Table 3. Distribution of the 15,946 diagnostic SNPs (DSNPs) per taxon and per chromosome along the nine chromosomes

<table>
<thead>
<tr>
<th>C. reticulata</th>
<th>C. maxima</th>
<th>C. medica</th>
<th>C. micrantha</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>404</td>
<td>274</td>
<td>604</td>
<td>430</td>
</tr>
<tr>
<td>C2</td>
<td>429</td>
<td>257</td>
<td>826</td>
<td>555</td>
</tr>
<tr>
<td>C3</td>
<td>593</td>
<td>328</td>
<td>1089</td>
<td>817</td>
</tr>
<tr>
<td>C4</td>
<td>388</td>
<td>245</td>
<td>630</td>
<td>503</td>
</tr>
<tr>
<td>C5</td>
<td>423</td>
<td>264</td>
<td>719</td>
<td>490</td>
</tr>
<tr>
<td>C6</td>
<td>321</td>
<td>228</td>
<td>564</td>
<td>428</td>
</tr>
<tr>
<td>C7</td>
<td>318</td>
<td>179</td>
<td>494</td>
<td>343</td>
</tr>
<tr>
<td>C8</td>
<td>261</td>
<td>130</td>
<td>480</td>
<td>364</td>
</tr>
<tr>
<td>C9</td>
<td>354</td>
<td>182</td>
<td>591</td>
<td>441</td>
</tr>
<tr>
<td>Total</td>
<td>3491</td>
<td>2087</td>
<td>5997</td>
<td>4371</td>
</tr>
<tr>
<td>%</td>
<td>21.9</td>
<td>13.09</td>
<td>37.6</td>
<td>27.41</td>
</tr>
</tbody>
</table>

C1–C9, the nine chromosomes of the reference clementine genome (Wu et al., 2014); %, percentage of DSNPs for the taxon.

for four citrons (‘Buddha’s Hand’, ‘Corsican’, ‘Humpang’ and ‘Mac Veu de Montagne’), C. micrantha, seven mandarins (‘Ponkan’, ‘Owari Satsuma’, ‘King’, ‘Dancy’, ‘Sunki’, ‘Cleopatra’ and ‘Willowleaf’), ‘Chandler’ pummelo, ‘Washington Navel’ sweet orange, ‘Seville’ sour orange, ‘Nules’ clementine, ‘Marsh’ grapefruit, ‘Rough’ lemon, ‘Rangpur’ lime and ‘Eureka’ lemon. For example, Supplementary Data Fig. S7A shows the phylogenomic karyotypes of the ‘Washington Navel’ sweet orange and the ‘Owari Satsuma’ mandarin inferred from our GBS data and from WGS data (Wu et al., 2014). As concluded by Wu et al. (2018), the four citrons common to both studies and the two ‘small flower’ papeda were fully homozygous with C. medica and C. micrantha, respectively. Regarding ‘Chandler’ pummelo, only a small genomic area considered by Wu et al. (2014, 2018) to be introgressed in heterozygosity by C. reticulata on chromosome 2 (C2) coincided with an undetermined area in our karyotype generated from GBS data (Fig. 7A; Wu et al., 2018). For the rest of the genome, like Wu et al. (2014, 2018), we concluded homozygosity for C. maxima. For the representative mandarins, the karyotypes inferred from GBS data completely matched those inferred from WGS (Wu et al., 2014, 2018) except for two small genomic regions. A small C. reticulata homozygous fragment in the C6 of ‘Owari Satsuma’ mandarin and a small heterozygous introgression of C. maxima at the beginning of the C2 of ‘Willow leaf’ mandarin were not detected by the GBS analysis. Focusing on the areas determined in our GBS analysis, we detected no differences between our results for sweet orange, sour orange, clementine, grapefruit, lime and lemons common to both analyses (Fig. 7B, C) and those obtained by Wu et al. (2018). Moreover, we checked the repeatability of the analysis through three experimental replicates (three independent samples during preparation of the GBS library) of ‘Nules’ clementine. The determined areas of the three replicates displayed exactly the same pattern (Supplementary Data Fig. S7B). Overall, phylogenomic karyotypes were successfully inferred from GBS data but with more undetermined regions than those inferred from WGS data. Given these positive results, we considered that our GBS workflow was validated and the karyotypes inferred for all the remaining varieties as a good approximation of the phylogenomic structure.

New karyotypes of diploid varieties

The analysis of the additional varieties representative of the four ancestral taxa revealed introgressions of C. maxima fragments in all mandarins except ‘Shekewasha’ mandarin. It varied between 1.39 % for ‘Se Hui Gan’ mandarin to 4.41 % in ‘San Hu Hong Chu’ mandarin, with variable introgression positions in C2, C3, C4, C6, C8 and C9. ‘Shekewasha’ mandarin displayed a small introgression of C. micrantha in C3. In the case of pumellos, GBS data identified a small introgressed area by C. medica in the C7 of ‘Timor’ pummelo, while ‘Pink’, ‘Tahitian’, ‘Kao Pan’ and ‘Deep red’ pummelos appeared fully homozygous for C. maxima (Fig. 7A). In the same way, the two C. medica not analysed in the study of Wu et al. (2018) (‘Etrog’ and ‘Poncire commun’ citrons) appeared fully homozygous for C. medica.

For the secondary species, ‘Bouquetier de Nice’ sour orange displayed the same karyotype as ‘Seville’ sour orange with full C. reticulata/C. maxima heterozygosity. Examining the determined areas, ‘Valencia late’ sweet orange was found to be identical to ‘Washington navel’, displaying C. reticulata/C. maxima heterozygosity or C. reticulata homozygosity all along the genome except on two fragments on C2 and C8, which appeared in C. maxima homozygosity. In the same way, ‘Duncan’ and ‘Star Ruby’ grapefruits were found to be identical to ‘Marsh’ (Fig. 7B). ‘Volkamer’ lemon appeared to be fully heterozygous for C. reticulata/C. medica along the nine chromosomes, as previously observed for ‘Rangpur’ lime and ‘Rough’ lemon (Wu et al., 2018; this study). Karyotypes of ‘Palestine’ sweet lime, ‘Marrakech’ limonette, and ‘Meyer’ and ‘Lisbon’ lemons displayed interspecific heterozygous fragments of C. medica/C. reticulata and C. medica/C. maxima (Fig. 7B) as previously described for ‘Eureka’ lemon (Wu et al., 2018; our present results from GBS). Moreover ‘Lisbon’ and ‘Eureka’ lemons were strictly identical in their determined areas. Bergamot displayed a much more complex admixture of C. maxima, C. reticulata and C. medica genomes. Indeed, in addition to the C. medica/C. reticulata and C. medica/C. maxima heterozygosity regions, we found fragments in C. reticulata/C. maxima heterozygosity, C. reticulata homozygosity (C7) and C. maxima homozygosity (C3, C4, C6 and C7). Referring to the hypothesis that the bergamot comes from a hybridization between a sour orange and a lemon (Gallesio, 1811; Curk et al., 2016), we examined the ancestor allelic dosage of the 100 kb windows of
Ahmed et al. — GBS reveals the genome admixture of diploid and polyploid citrus

this variety and its assumed parents. A total of 99.12% of them completely fit with the hypothesis, each parental gamete bringing the ancestor allelic doses observed in the bergamot. The remaining 0.88% corresponds to C. reticulata/C. maxima heterozygosity regions located in the C1 and C6 undetermined in lemon. Considering this origin hypothesis and the haplotype structure of the parental genomes, we have been able to draw the bergamot phased karyotype (Fig. 8; Supplementary Data Fig. S8).

Karyotypes of polyploid varieties

The phylogenomic structures of ‘Tanepao’, ‘Coppenrhad’, ‘Tahiti’ and ‘Persian’ triploid limes (Fig. 7C) and ‘Giant Key’ tetraploid lime were also inferred with the ‘TraceAncestor’ tool (Fig. 7D). ‘Tahiti’ and ‘Persian’ limes involving the contribution of the four basic taxa and, excluding undetermined areas, noticeably had the same phylogenomic karyotype. The quasi-systematic single dose of C. micrantha, the frequent double dose of C. medica and the occurrence of a double dose of C. micrantha (C3 and C5) and a triple dose of C. medica on small fragments, while C. reticulata and C. maxima were found only in single doses, fit the hypothesis that these limes derive from the union of a diploid ovule of ‘Mexican’ lime (C. aurantiifolia = C. micrantha × C. medica) and haploid pollen of lemon [C. limon = (C. maxima × C. reticulata) × C. medica] as proposed by Curk et al. (2016) and Rouiss et al. (2018). Therefore, following this hypothesis, we propose a phased karyotype identifying the haploid and diploid gametes from which this triploid lime originated (Fig. 8). For all the chromosomes, except C3 and C5, we observed a total restitution of the ‘Mexican’ lime-like parent by the diploid gamete. The representation of chromosomes 3 and 5 is just one of the different possibilities of C. medica and C. micrantha fragment phases in the diploid gamete. The interspecific recombination points in the diploid C. aurantiifolia and haploid C. limon gametes were clearly identified (Fig. 7C).
For determined areas, ‘Coppenrhad’ and ‘Tanepao’ limes displayed an identical pattern involving only C. *medica* and C. *micrantha* with single doses of C. *micrantha* and double doses of C. *medica* all along the nine chromosomes.

For the tetraploid ‘Giant key’ lime, the phylogenomic analysis with ten DSNPs per window produced many undetermined regions (60.58 %), due to a relatively low coverage (Supplementary Data Fig. S1) and the higher difficulty to
distinguish 1/3, 2/2 and 3/1 doses for heterozygous genotypes. Therefore, we tested the inference with 20 and 30 DSNPs (Supplementary Data Fig. S9). The karyotype we obtained with 30 DSNP windows allowed a better estimation of the allelic doses and was able to reduce the undetermined regions to only 20%. It showed full C. medicala/C. microantha heterozygosity along the genome.

**DISCUSSION**

The DSNP-based approach is powerful to decipher the admixture genomic structure in Citrus

Recent studies based on NGS (WGS and GBS) analysed the admixture of modern citrus varieties. They were based on the identification of diagnostic polymorphism (mainly SNPs) of the ancestral taxa considered. Wu et al. (2014) were the first to develop the DSNP approach to decipher the genomic structures of modern varieties originating from two ancestral taxa, C. reticulata and C. maxima, from WGS data. They used a small panel of mandarins (three varieties) and pummelos (two varieties), as representative of C. reticulata and C. maxima, to identify SNPs that distinguish these two ancestral taxa. The patterns of heterozygosity and similarity to the other mandarins and pummelos were used to identify introgressed areas in the different varieties in the two panels. The study revealed unexpected C. maxima introgressions in ‘Ponkan’ and ‘Willowleaf’ mandarins which were previously believed to be pure representatives of the C. reticulata taxon. The very large set of identified DSNPs was highly efficient to decipher the phylogenomic structures of clementine, sweet and sour oranges and ‘Afourer’ mandarin (W Murcott). Oueslati et al. (2017) showed that a similar approach can be used with GBS data using the ApeKI restriction enzyme. They expanded the phylogenomic analysis to 55 citrus varieties composed of representatives of C. maxima and C. reticulata taxa and hybrids assumed to derive from the admixture of these two taxa (mandarins, tangors, tangelos, orangelos and grapefruits). From a larger panel of representative mandarins (11 varieties) and pummelos (six varieties), these authors identified a set of 11 133 diagnostic polymorphisms, mostly SNPs (89%), with a very similar pattern of distribution along the genome to those of gene sequences. This allowed them to infer the phylogenomic karyotypes of all the accessions by analysing the relative proportion of diagnostic markers homozygous in successive windows of 20 diagnostic markers.

Curk et al. (2015) were the first to publish sets of DSNPs for the four Citrus ancestral taxa. They identified 273 DSNPs from 454 amplicon sequencing data of 57 gene fragments dispersed on the nine chromosomes. They then developed allele competitive PCR markers (using KASPPar technology) for 105 of these DSNPs and successfully analysed the interspecific origin of >200 Citrus accessions (Curk et al., 2015, 2016) and revealed systematic introgression of C. maxima in edible mandarins. However, the low number of DSNPs used in these studies did not make it possible to infer the phylogenomic karyotypes of the analysed varieties.

Wu et al. (2018) mined DSNPs which differentiate three of the four basic taxa (C. maxima, C. medica and C. reticulata) using only two pure Chinese mandarins, two citrons and three pummelos. They identified a total of 588 583 DSNPs (169 963 for C. reticulata, 116 803 for C. maxima and 301 817 for C. medica) and used them to decipher the phylogenomic karyotype of 47 Citrus varieties.

Whether the studies dealt with WGS (Wu et al., 2014, 2018), GBS (Oueslati et al., 2017) or DSNP markers (Curk et al., 2015, 2016), the analyses have always identified C. maxima introgressions in most cultivated mandarins. If the corresponding sequences are taken into account when estimating the allelic frequencies of the ancestral taxa, this introduces a bias in the estimation of the diversity parameters (allelic frequencies of the ancestral taxa and G_{ST}) and hence in the detection of diagnostic polymorphisms of the four ancestral taxa. This is why Wu et al. (2018) drastically limited their representative panel to only two pure genetically close mandarins. However, such a small panel could result in considering specific SNPs of the considered accessions as diagnostic of C. reticulata, whereas in fact polymorphisms existed within the species. Therefore, for our analysis, we preferred to keep the panel of representatives of the ancestral taxa as large as possible and used 15 mandarins, six pummelos, six citrons and two ‘small flowered’ papada as representative of C. reticulata, C. maxima, C. medica and C. microantha, respectively. Therefore, like Oueslati et al. (2017), we first identified introgression areas along the genome of the 29 representative accessions of the basic taxon according to the pattern along the genome of heterozygosity and to similarity with centroids of mandarins, pummelos, citrons and papadas. After removing the identified introgression regions, we computed the differentiation parameters again and filtered for polymorphic positions with G_{ST}>0.9. We selected 15 946 DSNPs and developed a pipeline to infer the phylogenomic structures of the 53 citrus accessions. Taking into account the difficulty involved in correctly estimating the allelic doses in triploid and tetraploid accessions at individual SNP loci (McKinney et al., 2018; Bastien et al., 2018; our data) and according to our choice of using the same analytical approach for diploids, triploids and tetraploids, we based our pipeline on the relative number of reads of each ancestor in windows of ten DSNPs of the considered taxon (while Wu et al., 2014, 2018 and Oueslati et al., 2017 performed their analysis in diploid accessions from genotyping data at individual loci) and on maximum likelihood analysis.

For diploid accessions common to both studies, our GBS data produced highly similar results to those obtained from WGS data (Wu et al., 2014, 2018), apart from the undetermined genomic areas, which were more frequent for GBS data, due to a lower density of DSNPs.

Therefore, GBS combined with our analytic pipeline proves to be a powerful approach to correctly analyse the phylogenomic admixture of diploid, triploid and tetraploid citrus varieties along the genome at significantly lower cost than the WGS approach. The panel of DSNPs can be used as reference for further GBS analyses using the same protocol (ApeKI; selection base A) to decipher the phylogenomic karyotypes in large citrus populations (germplasm or recombining populations). It opens the way for genetic association studies, quantitative trait locus (QTL) analyses and genomic selection based on phylogenomics.

We developed a generic pipeline to decipher admixture in diploid, triploid and tetraploid genomes from an unlimited number of ancestors, allowing the user to define the number of DSNPs per window for the analysis of the dose contributed by
each ancestor, the error rate considered for homozygous genotypes, the threshold for the LOD test of the maximum likelihood and the size of the window used to integrate information on the doses from the different ancestors to generate the phylogenetic karyotypes. This pipeline is available at http://galaxy.southgreen.fr/galaxy/ and should be useful for any species whose reproductive behaviour (vegetative propagation, preferential chromosome pairing associated with preferential disomic segregation) limited the number of interspecific recombinations after reticulation events and resulted in interspecific mosaics of large genomic fragments. It can also be used for the first generations of interspecific breeding schemes to identify interspecific recombination points. The selection of the ApeKI enzyme results in a marker density closely linked with gene sequence density and, consequently, in high coverage of the high recombining areas of the genome and low coverage of centromeric and paracentromeric areas with very low recombination rates (Aleza et al., 2015). This is a major advantage to trace interspecific recombination from GBS data efficiently. The main limitation of the approach is that it is based on the assumption of conserved physical genomic structure among the considered ancestors. In citrus, the overall high level of synthy and conserved collinearity of markers observed for the genetic maps of clementine, sweet orange and pummelo (Ollitrault et al., 2012a), sour orange, pummelo, Poncirus trifoliata and ‘Fortune’ mandarin (Bernet et al., 2010), and sweet orange and Poncirus (Chen et al., 2008) justifies the use of the clementine reference genome as the genomic template to establish the phylogenomic karyotypes from either WGS data (Wu et al., 2014, 2018) or GBS data (Oueslati et al., 2017; this study). For plants with known large structural variations, a specific approach will be needed to describe the phylogenomic structures correctly in the genomic areas concerned.

The phylogenetic structures of 48 diploid varieties were deciphered; 16 for the first time

The representative accessions of the four basic taxa. We analysed 15 mandarins assumed to be good representatives of C. reticulata species. Twelve of them displayed C. maxima introgressions and one, ‘Shekwasha’ mandarin, has a small introgression of C. micrantha. No C. maxima introgressions were detected in ‘Shekwasha’, ‘Cleopatra’ and ‘Sunki’ mandarins. Limited introgressions were identified in ‘Szibat’ mandarin (1.49%), ‘Ladu’ mandarin (1.72%), ‘Nan Feng Mi Chu’ mandarin (1.74%) and ‘Se Hui Gan’ mandarin (1.39%). ‘Satsuma’ and ‘King’ mandarins were distinguished from all the other introgressed mandarins by their higher rate of C. maxima introgressions (22.6 and 19.5%, respectively). Our results for newly studied varieties confirm that most edible mandarins are introgressed by C. maxima fragments as previously detected from WGS (Wu et al., 2014, 2018), 454 amplicon sequencing data (Curk et al., 2015) and GBS data (Oueslati et al., 2017). Wu et al. (2018) showed that some Chinese mandarins were not introgressed, and proposed three types of mandarins. The first type corresponds to unintrogressed genomes; type II includes mandarins with limited early introgression of the same two C. maxima haplotypes; and type III comprises mandarins derived from type II after more recent additional C. maxima introgression, probably resulting from hybridization with sweet orange. Based on our GBS analysis, ‘Szibat’, ‘Ladu’, ‘Nan Feng Mi Chu’ and ‘Se Hui Gan’ mandarins should be included in type II mandarins.

Despite the small C. reticulata introgressions in two pummmelos (Wu et al., 2014, 2018; Oueslati et al., 2017) and the C. medica introgression in ‘Timor’ pummelo, our analysis confirms that modern pummmelos can be considered as good representatives of the C. maxima species, as previously argued by several authors (Wu et al., 2014, 2018; Curk et al., 2015; Oueslati et al., 2017).

In our study, neither citrons nor ‘small flowered’ papedas displayed introgression areas. These results are in agreement with the conclusions drawn by Curk et al. (2015) and Wu et al. (2018). The analysed citrons and papedas therefore appear to be good representatives of the C. medica and C. micrantha species, respectively. Our results reveal the high level of homozygosity of citron accessions, including genomic areas with no revealed heterozygosity. Molecular marker studies (Barkley et al., 2006; Garcia-Lor et al., 2012; Luro et al., 2012; Curk et al., 2016) and WGS data (Wu et al., 2018) previously provided evidence for the low polymorphism of citrons and their high level of homozygosity. This can be linked with the cleistogamy of citron resulting in inbreeding and complete homozygosity of certain genome areas.

Secondary diploid species. The phylogenomic structures of accessions resulting from interspecific C. reticulata/C. maxima admixture are in full agreement with previous results and with hypotheses on their origins (Nicolosi et al., 2000; Ollitrault et al., 2012b; Curk et al., 2014; Wu et al., 2014, 2018; Oueslati et al., 2017). Thus, grapefruits, which are hybrids between C. maxima and sweet orange, display genome fragments in C. reticulata/C. maxima heterozygosity and C. maxima homozygosity. We found identical GBS-derived phylogenomic karyotypes for the three grapefruits analysed (‘Marsh’, ‘Duncan’ and ‘Star Ruby’) and that of ‘Marsh’ inferred from WGS (Wu et al., 2014, 2018). This confirms that these different cultivars derived from a single hybrid ancestor with no further sexual recombination. Citrus maxima and C. reticulata contributed equally to sour orange structure, and our results reveal an identical phylogenomic karyotype for ‘Bouquetier de Nice’ and ‘Seville’ sour oranges. The two sweet orange cultivars analysed displayed the same karyotypes with C. reticulata homozygosity fragments as well as C. maxima homozygosity and C. reticulata/C. maxima heterozygosity, in complete agreement with the study of ‘Washington Navel’ sweet orange by Wu et al. (2014). These results are evidence for the absence of sexual recombination during the diversification of these sweet oranges, whose polymorphisms are hypothesized to result from sporadic mutations, inheritable epigenetic changes and movements of transposable elements, as demonstrated for the anthocyanin content of blood oranges (Butelli et al., 2012).

The karyotype analysis of acidic citrus (limes and lemons) of Wu et al. (2018) was limited to ‘Rangpur’ and ‘Mexican’ limes and ‘Eureka’ and ‘Rough’ lemons. We expanded the analysis to ‘Alemow’, ‘Nestour’ lime, ‘Lisbon’, ‘Meyer’ and ‘Volkamer’ lemons, ‘Marrakech’ limonette and ‘Palestine’ sweet lime. ‘Rangpur’ lime, ‘Rough’ and ‘Volkamer’
lemons displayed the same pattern, with equal contributions of *C. reticulata* and *C. medica* along the nine chromosomes. These results support the hypothesis that they both derive from direct *C. reticulata × C. medica* hybridization as proposed by Curk et al. (2016) and Wu et al. (2018) for ‘Rangpur’ lime and ‘Rough’ lemon. In both previous studies, the contribution of *C. medica* as male parent was proved by chloroplast phylogeny. Our results also agree with the cytogenetic studies of Carvalho et al. (2005) in which ‘Alemow’ and ‘Nestour’ lime displayed the same pattern with *C. micrantha/C. medica* heterozygosity over all nine chromosomes, and confirm the hypothesis proposed by Curk et al. (2016), i.e. that these two acidic citrus resulted from direct hybridization between *C. micrantha* and *C. medica*. Using simple sequence repeat (SSR) markers in addition to DSNPs and cytoplasmic markers, Curk et al. (2016) also demonstrated that these two varieties resulted from independent reticulation events and that citron was the male parent. The phylogenomic karyotypes we obtained for ‘Eureka’ and ‘Limon’ lemons were identical and in full agreement with that proposed for ‘Eureka’ lemon by Wu et al. (2018). Probably, ‘Meyer’ lemon, ‘Palestine’ sweet lime and ‘Marrakech’ limonette involve the same three species *C. maxima*, *C. reticulata* and *C. medica*. Considering that *C. medica* is present as a single dose all over their genomes, we propose that they all result from hybridization between *C. maxima/C. reticulata* admixed genotypes and a *C. medica*. According to previous maternal phylogeny studies (Nicolosi et al., 2000; Luro et al., 2012; Carbonell-Caballero et al., 2015; Curk et al., 2016), *C. medica* is assumed to be the male parent in all cases. Previous molecular marker analyses of ‘Limon’ and ‘Eureka’ type yellow lemons (Nicolosi et al., 2000; Curk et al., 2016) suggested that they resulted from a single direct hybridization event between sour orange and citron. The same conclusion was drawn recently for ‘Eureka’ lemon based on WGS data (Wu et al., 2018). According to a maternal phylogenomic study (Curk et al., 2016) and nuclear data (Curk et al., 2016; this study), the ‘Marrakech’ limonette is hypothesized to have the same phylogenetic origin but to derive from an independent interspecific hybridization event. Maternal phylogeny studies revealed that ‘Meyer’ lemon and ‘Palestine’ sweet lime have the same cytoplasmic profile as sweet oranges and pummelos (Curk et al., 2016). However, their exact maternal parent remains to be determined.

The phylogenomic structure of bergamot also displays the admixture of the same three ancestral taxa, but the karyotype appears to be much more complex than that of the lemons, sweet lime and limonette discussed above. Many researchers have attempted to identify the origin of bergamot. In 1811, Gallesio proposed that it derives from a sour orange × lemon parentage. Several other origins have also been proposed, as reviewed and tested by Curk et al. (2016) in a nuclear and cytoplasmic marker study. Their results supported that proposed by Gallesio (1811). Our comparison of the karyotypes of bergamot and the karyotypes of sour orange and yellow lemons (‘Eureka’ and ‘Limon’) totally fits with the hypothesis that bergamot results from hybridization between a sour orange and a yellow lemon. It was therefore possible to draw a phased karyotype of the bergamot distinguishing the gamete originating from lemon and that originating from sour orange. Considering their modern distribution, it is probable that bergamot and ‘Marrakech’ limonette resulted from hybridization that occurred in the Mediterranean Basin, where the presence of citrons dates from the second century BC and the introduction of sour orange dates to the Arab era in the seventh century (Webber et al., 1967; Swingle and Reece, 1967; Nicolosi et al., 2005). This confirms the importance of this region as a secondary area of citrus diversification.

The phylogenomic karyotype of triploid and tetraploid limes were deciphered

Leaving aside the undetermined regions, our phylogenomic inference resulted in identical structures for ‘Tahiti’ and ‘Persian’ limes, with a contribution of the four ancestral taxa. As reported in Curk et al. (2016), our results also revealed single or double doses of *C. medica* and *C. micrantha*, while *C. maxima* and *C. reticulata* contributed no dose or a single dose along the genome. Curk et al. (2016) proposed that this type of lime resulted from the fusion of a haploid lemon ovule and a diploid pollen of a diploid ‘Mexican’-like lime. Our analysis of the four ancestor does all along the genome perfectly fits this hypothesis at the nuclear level. The diploid gamete of ‘Mexican’ lime type restituted 84.65 % of the parental interspecific heterozygosity and displayed only 2.47 and 0.80 % of *C. micrantha* and *C. medica* homozygote fragments, respectively. This high level of heterozygosity restitution and the heterozygosity for the centromeric areas of the nine chromosomes preclude the hypothesis of an unreduced gamete from a diploid ‘Mexican’ lime resulting from second division restitution (SDR) of the meiosis. They suggest that the diploid gamete comes from a doubled diploid parent with a preferential disomic segregation, or from first division restitution (FDR) of a diploid parent. Indeed, SDR 2n gametes contain sister chromatids and are homozygous from the centromere until the first crossing-over, while, under FDR, 2n gametes contain two non-sister chromatids allowing the entire conservation of parental heterozygosity from the centromere until the first crossing-over (Park et al., 2007; Ollitrault et al., 2008; Peloquin et al., 2008; Cuencea et al., 2011; Storme and Geelen, 2013); as a consequence, FDR gametes transmit 70–80 % of the parental heterozygosity, whereas this is only about 30–40 % for SDR (Barone et al., 1995; Douchez and Mass, 1998; Dewitte et al., 2012; Aleza et al., 2016).

Molecular marker inheritance proved that doubled diploid ‘Mexican’ lime had preferential disomic inheritance with 90.2 % of heterozygosity restitution on average (Rouiss et al., 2018). Therefore, the phylogenomic karyotype of ‘Tahiti’ lime fits well with the interploid (diploid citron × tetraploid lime) origin hypothesis proposed by Rouiss et al. (2018). However, the unreduced FDR gamete hypothesis cannot be totally ruled out. Indeed, the FDR mechanism has been recently described at the origin of 2n pollen in citrus (Rouiss et al., 2017), and it can also result in a very high level of heterozygosity restitution. ‘Tanepao’ and ‘Coppenrhad’ limes presented identical patterns with single doses of *C. micrantha* and double doses of *C. medica* all along the nine chromosomes. Rouiss et al. (2018) observed that the preferential disomic inheritance of the doubled diploid ‘Mexican’ lime resulted in the production of 7 %
of gametes with full interspecific heterozygosity. Therefore, an interpollen backcross hybridization of a doubled diploid ‘Mexican’ lime type with a diploid citron may be at the origin of these limes, as proposed by Curk et al. (2016) and Rouiss et al. (2018). However, FDR coupled with asynapsis of ‘Mexican’ lime, which is dependent on low temperatures (Iwamasa and Iwasaki, 1963), could also produce fully heterozygous diploid gametes from a diploid ‘Mexican’ lime parent. Therefore, fertilization of an FDR ovule of ‘Mexican’ lime type by a haploid pollen of citron cannot be eliminated. The tetraploid ‘Giant key’ lime displayed a full heterozygous pattern with a double dose of C. medica and a double dose of C. micrantha along its genome. In a molecular marker study, Curk et al. (2016) obtained identical patterns for ‘Giant key’ and ‘Mexican’ limes. They suggested that ‘Giant key’ lime emerged from the natural duplication of chromosomes of a ‘Mexican’ lime type which derives from a C. micrantha × C. medica natural hybridization. Our results agree with these conclusions. To limit the undetermined area for ‘Giant Key’ lime, we had to perform the likelihood analysis in windows of 30 DSNPs. This was required by the low coverage of this accession and also because more reads are necessary to conclude significantly between hypotheses of a 1/3, 2/2 and 3/1 ratio for heterozygous loci in tetraploids than a single homozygous/heterozygous distinction in diploid or 1/2 vs. 2/1 discrimination in triploids.

Conclusion

Genotyping by sequencing, using the ApelKI restriction enzyme, to focus on gene areas, and a selective base (A), to improve the depth of the analysis, was successfully applied to diploid, triploid and tetraploid citrus. The analysis of 29 representative accessions of the four citrus ancestral taxa allowed us to identify 15 946 DSNPs among 43 598 mined SNPs. The generic pipeline developed to infer phylogenomic karyotypes is based on the relative number of reads of ancestral and alternative alleles at DSNP loci. For each ancestral taxon, maximum likelihood tests were performed to infer doses of ancestral taxa in successive windows of ten DSNPs of the taxa considered. This approach provided results which closely resembled previously published results from WGS data. It revealed the phylogenomic structure for new diploid species and cultivars including direct interspecific hybrids such as C. limonia, ‘Volkamer’ lemon (C. reticulata × C. medica), C. macrophylla ‘Alemow’ and C. excelsa ‘Nestour’ lime (C. micrantha × C. medica), but also more complex structures involving three ancestors such as C. limetta ‘Marrakech’ limonette [(C. maxima × C. reticulata) × C. medica; sour orange × citron], C. limettiodes Tan. ‘Palestine’ Sweet lime and C. meyeri [(C. maxima × C. reticulata) × C. medica] or C. bergamia bergamot [(C. maxima × C. reticulata) × C. medica] (C. maxima × C. reticulata); lemon × sour orange]. The phylogenomic karyotypes of triploid limes were also revealed, confirming the highly complex structure of ‘Tahiti’ and ‘Persian’ limes involving the four ancestral taxa [(C. maxima × C. reticulata) × C. medica] (C. micrantha × C. medica); lemon haploid ovule × ‘Mexican’ lime-like diploid pollen], and are in agreement with the probable origin of ‘Tanepao’ and ‘Coppenrhad’ limes from the interpollen backcross [(C. micrantha × C. medica) × C. medica; ‘Mexican’ lime-like diploid ovule × citron haploid pollen]. The GBS approach and analytical pipeline combined with the reference DSNP matrix will be useful for any study of germplasm and hybrids resulting from breeding within the Citrus genus. The workflow implemented for mosaic genome analysis is available online and can also be used for other species with unlimited numbers of identified ancestral taxa, for diploid, triploid and tetraploid accessions. Considering the density of DSNPs along the genome revealed by GBS, it will probably be particularly useful for any species whose reproductive behaviour has limited the number of interspecific recombinations after reticulation events and resulted in interspecific mosaics of large genomic fragments. It can also be used to localize interspecific recombinating points in the first generations of interspecific breeding schemes.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/aob and consist of the following. Text S1: maximum likelihood test for diploid, triploid and tetraploid individuals. Text S2: average phylogenomic contribution of the four ancestral taxa to the modern varieties. Figure S1: number of reads per accession. Figure S2: distribution of missing data among markers and individuals. Figure S3: distribution of the number of reads per marker along the nine chromosomes. Figure S4: identification of the interspecific introgressions in representative accessions of the ancestral taxa: example of the chromosome 2 of ‘King’ mandarin. Figure S5: distribution along the nine chromosomes of DSNPs, the whole set of diallelic SNPs and gene sequences. Figure S6: estimation of C. medica allele doses in triploid ‘Persian’ lime. Figure S7: validation of the GBS approach. Figure S8: phased karyotype of the bergamot based on the gametates of the lemon and the sour orange. Figure S9: the inferred karyotypes of the ‘Giant key’ lime. Table S1: plant material. Table S2: list of the 15 946 diagnostic markers, specifying their genomic position, their reference and alternative alleles, the ancestral taxon they are diagnostic for and the gene name where they are located if available.

FUNDING

This work was supported by the FEDER Guadeloupe ‘CAVALBIO’ project, by the FEDER Corsica ‘InnovAgromus’ project and the Agropolis Fondation ‘GenomeHarvest project’ (ID 1504-006) through the ‘Investissements d’avenir’ program (Labex Agro:ANR-10-LABX-0001-01).

LITERATURE CITED


Ahmed et al. — GBS reveals the genome admixture of diploid and polyploid citrus


