Mol Genet Genomics (2012) 287:77–94
DOI 10.1007/s00438-011-0658-4

ORIGINAL PAPER

Comparative use of InDel and SSR markers in deciphering the interspecific structure of cultivated citrus genetic diversity: a perspective for genetic association studies

Andrés García-Lor · François Luro · Luis Navarro · Patrick Ollitrault

Received: 12 July 2011 / Accepted: 3 November 2011 / Published online: 11 December 2011

Abstract Genetic stratification associated with domestication history is a key parameter for estimating the pertinence of genetic association study within a gene pool. Previous molecular and phenotypic studies have shown that most of the diversity of cultivated citrus results from recombination between three main species: *C. medica* (citron), *C. reticulata* (mandarin) and *C. maxima* (pummelo). However, the precise contribution of each of these basic species to the genomes of secondary cultivated species, such as *C. sinensis* (sweet orange), *C. limon* (lemon), *C. aurantium* (sour orange), *C. paradisi* (grapefruit) and recent hybrids is unknown. Our study focused on: (1) the development of insertion–deletion (InDel) markers and their comparison with SSR markers for use in genetic diversity and phylogenetic studies; (2) the analysis of the contributions of basic taxa to the genomes of secondary species and modern cultivars and (3) the description of the organisation of the *Citrus* gene pool, to evaluate how genetic association studies should be done at the cultivated *Citrus* gene pool level. InDel markers appear to be better phylogenetic markers for tracing the contributions of the three ancestral species, whereas SSR markers are more useful for intraspecific diversity analysis. Most of the genetic organisation of the *Citrus* gene pool is related to the differentiation between *C. reticulata*, *C. maxima* and *C. medica*. High and generalised LD was observed, probably due to the initial differentiation between the basic species and a limited number of interspecific recombinations. This structure precludes association genetic studies at the genus level without developing additional recombinant populations from interspecific hybrids. Association genetic studies should also be affordable at intraspecific level in a less structured pool such as *C. reticulata*.

Keywords *Citrus* · Genetic diversity · Linkage disequilibrium · Association genetics · InDels · SSRs

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-011-0658-4) contains supplementary material, which is available to authorized users.

A. García-Lor · L. Navarro (✉) · P. Ollitrault
Centro de Protección Vegetal y Biotecnología,
Instituto Valenciano de Investigaciones Agrarias (IVIA),
46113 Moncada, Valencia, Spain
e-mail: lnavarro@ivia.es

F. Luro
Unité de Recherche GEQA, Institut National de la Recherche Agronomique (INRA), 20230 San Giuliano, France

P. Ollitrault (✉)
Department BIOS, TGU AGAP, International Center for of Agricultural Research for Development (CIRAD),
Avenue Agropolis, TA A-75/02,
34398 Montpellier Cedex 5, France
e-mail: patrick.ollitrault@cirad.fr

Introduction

Genetic association studies based on linkage disequilibrium (LD) are similar to quantitative trait locus (QTL) mapping. However, whereas QTL mapping considers only variations between two crossed individuals, LD mapping exploits the phenotypic and genetic variation present across a natural population. This method has been successfully applied in studies of cultivated plants (Thornsberry et al. 2001; Casa et al. 2008; Zhu et al. 2008). However, the presence of population stratification and an unequal distribution of alleles within these groups can result in spurious associations (Abdurakhmonov and Abdukarimov 2008).

Breeding systems and domestication history are determinant factors of the LD structure in cultivated species germplasm. The extent of LD is generally higher for species with selfing mating system (Arabidopsis, Nordborg et al. 2002; rice, Garris et al. 2003; and sorghum, Deu and Glaszmann 2004) than for outcrossing organisms (maize, Remington et al. 2001; populus, Ingvarsson 2005; and Norway spruce, Rafalski and Morgante 2004). To our knowledge, no data are available for LD in agamic complexes.

Citrus is one of the most important fruit crops in the world, and its diversity (Krueger and Navarro 2007) and origin (Webber et al. 1967; Calabrese 1992) have been widely studied. The taxonomy of citrus remain controversial, due to the conjunction of broad morphological diversity, total sexual interspecific compatibility within the genus and partial apomixis of many cultivars. Fixing complex genetic structures through seedling propagation via apomixis has led some taxonomists to consider clonal families of interspecific origin as new species (Scora 1975). Two major systems are widely used to classify *Citrus* species: the Swingle and Reece (1967) classification that considers 16 species and Tanaka's (1961) one that identifies 156 species. More recently, Mabberley (1997) proposed a new classification of edible citrus recognising three species and four hybrid groups. In this paper, we will use the Swingle and Reece (1967) classification system. Indeed, this taxonomic system is widely used in the citrus scientist community and, as mentioned below, mostly agrees with molecular data.

Despite the difficulties involved in establishing a consensual classification of edible citrus, most authors now agree on the origins of most cultivated forms. Early studies by Scora (1975) and Barrett and Rhodes (1976) based on biochemical and morphological polymorphisms, respectively, suggested that most of the cultivated citrus originated from three main species (*C. medica* L., citrons; *C. reticulata* Blanco, mandarins; and *C. maxima* L. Osbeck, pummelos). More recent studies involving the diversity of morphological characteristics (Ollitrault et al. 2003) and secondary metabolites (Fanciullino et al. 2006) confirmed that the majority of the phenotypic diversity of edible citrus results from the differentiation between these three basic taxa. Isoenzymes (Herrero et al. 1996; Ollitrault et al. 2003), RFLP (Federici et al. 1998), RAPD, SCAR (Nicolosi et al. 2000), AFLP (Liang et al. 2007) and SSR (Luro et al. 2001; Barkley et al. 2006) molecular markers generally support the following conclusions for the origin of the other cultivated *Citrus* species (Nicolosi 2007): (1) *C. sinensis* (L.) Osb. (sweet oranges) and *C. aurantium* L. (sour oranges) are related with *C. reticulata* but display introgressed traits and markers of *C. maxima*. The closer relation with *C. reticulata* suggests that they are not direct hybrids but are probably backcrossed hybrids of first or second generation crosses with the *C. reticulata* gene pool. Analysis of chloroplastic (Green et al. 1986; Nicolosi et al. 2000) and mitochondrial genomes (Froelicher et al. 2011) indicate a *C. maxima* maternal phylogeny. (2) *C. paradisi* Macf. (grapefruits) is close to *C. maxima*, and could result from hybridization between *C. maxima* and *C. sinensis* (Barrett and Rhodes 1976; Scora et al. 1982; De Moraes et al. 2007). (3) *C. medica* is clearly a progenitor of *C. aurantifolia* (Christm.) Swing (limes) and *C. limon* Osb. (lemons). Chloroplast and nuclear data analysis indicate that the genetic pools of *C. reticulata* and *C. maxima* also contributed to the genesis of *C. limon*. Nicolosi et al. (2000) proposed that this species resulted from direct hybridisation between *C. aurantium* and *C. medica*. This assumption is supported by Gulsen and Roose (2001) and Fanciullino et al. (2007). The origin of *C. aurantifolia* is more controversial. However, molecular data (Federici et al. 1998; Nicolosi et al. 2000) support the hypothesis of Torres et al. (1978) that the Mexican lime is a hybrid between *C. medica* and a Papeda species. Nicolosi et al. (2000) proposed that *C. micrantha* might be the parental Papeda. These previous molecular studies have provided a better understanding of citrus maternal phylogeny, hybrid origin and parentage determination of many species. However, little is known about the precise contribution of the basic edible species to the nuclear genome constitution of secondary cultivated species (*C. sinensis*, *C. limon*, *C. aurantium*, *C. paradisi* and *C. aurantifolia*) and recent hybrids from twentieth century breeding programs. Furthermore, the impact of this domestication history on global genetic organisation and the extent of linkage disequilibrium (LD) on the *Citrus* gene pool have not been studied. The distance over which LD persists is a fundamental parameter to determine how association studies may be conducted on a gene pool. Regarding the important phenotypic differentiation between the basic taxa and the interspecific origin of most cultivated citrus, a better knowledge of the contribution of the nuclear genome of the basic taxa to the secondary species and modern cultivated citrus, as well as the analysis of the LD extent, appear as prerequisites to undergo association studies in the *Citrus* gene pool.

Among the codominant markers used for citrus genetic studies, simple sequence repeats (SSRs) (Luro et al. 2001, 2008; Gulsen and Roose 2001; Barkley et al. 2006; Ollitrault et al. 2010) are regarded as powerful tools because they are highly polymorphic, codominant, generally locus-specific and randomly dispersed throughout the plant genome. Thus, the use of mapped SSR markers should be particularly useful to analyse the extent of LD. However, Barkley et al. (2009) showed that homoplasy may limit the usefulness of SSR markers in identifying the phylogenetic origin of DNA fragments in citrus. Insertion or deletion (InDel) markers generally have low frequency of

homoplasy. Indeed, there is a sufficiently low probability of two InDel mutations of exactly the same length occurring at the same genomic position, that shared InDels can confidently be related to identity-by-descent. In general, InDels arise from the insertion of retroposons or other mobile elements, slippage in simple sequence replication or unequal crossover events (Britten et al. 2003). At the technical level, InDels can be genotyped with simple procedures based on size separation after targeted PCR (Vasemägi et al. 2010). InDels have been used successfully for genetic studies in wheat (Raman et al. 2006), rice (Hayashi et al. 2006) and natural populations (Väli et al. 2008).

Our study focused on three basic species (*C. medica*, *C. reticulata* and *C. maxima*), the secondary species that they generated (*C. sinensis*, *C. aurantium*, *C. paradisi* and *C. lemon*) and some known or putative interspecific hybrids. Twelve InDel markers were developed from gene sequencing, and their polymorphism organisation was compared with 50 SSR markers. Next, the complete set of markers was used to answer the following three questions: (1) what is the intraspecific diversity of InDel markers and are they more useful than SSRs as tag of DNA fragments in studies of phylogenetic origin? (2) What is the contribution of the three basic edible taxa to the genomes of secondary species and modern cultivars? (3) Are the genetic organisation of the *Citrus* gene pool and the extent of linkage disequilibrium adapted for association genetics? Furthermore, we propose a subset of markers (core markers) for quick and inexpensive systematic germplasm genotyping that maintains most of the organisation and intraspecific polymorphism information.

Materials and methods

Interspecific InDel polymorphism research

Plant material and DNA extraction

With the objective to identify InDel polymorphism differences between the basic citrus taxa, we selected two cultivars of *C. medica* (Corsican and Buddha's hand citrons), two cultivars of *C. reticulata* (Cleopatra and Willow Leaf mandarins) and two cultivars of *C. maxima* (Chandler and Pink pummelos). High molecular weight genomic DNA was extracted from leaf samples using the DNeasy Plant Mini Kit (Qiagen S.A.; Madrid, Spain) according to the manufacturer's instructions.

Gene sequence amplification and sequencing

Primers were designed from EST sequences corresponding to 16 genes available in public databases. Thirteen genes [Chalcone isomerase (CHI), Chalcone synthase (CHS), Flavonol synthase (FLS), Malic enzyme (EMA), Malate dehydrogenase (MDH), Vacuolar citrate/H+ symporter (TRPA), Phosphoenolpyruvate carboxylase (PEPC), Phosphofructokinase (PKF), Lycopene β-cyclase (LCY2), β-Carotene hydroxylase (Hy-b), Phytoene synthase (PSY), 1-deoxyxylulose 5-phosphate synthase (DXS) and Lycopene β-cyclase (LCYB)] are involved in primary and secondary metabolite biosynthesis pathways that determine the quality of citrus fruit (sugars, acids, flavonoids and carotenoids). In addition, three candidate genes for salt tolerance [CAX1 (cation/H+ membrane antiporter), AtGRC (raffinose synthase) and AVP (vacuolar H+ pyrophosphatase)] were used. Primers (Table 1) were designed to amplify fragments with a length between 166 and 1,201 bp. The PCR mixture consisted of 1 ng/µl template DNA, 0.2 mM dNTPs, 0.2 µM forward primer, 0.2 µM reverse primer, 10× PCR buffer (Fermentas), 1.5 mM $MgCl_2$ and 0.027 U/µl *Taq* DNA polymerase (Fermentas), in a final volume of 15 µl. PCR reactions were carried out with the following program: 5 min at 94°C; 40 cycles of 30 s at 94°C, 30 s at 50–58°C and 2 min at 72°C with a final extension of 4 min at 72°C.

Amplicons of the six selected genotypes were sequenced by the Sanger method from the 5′ end using dideoxynucleotides labelled by fluorescence (Big Dye Terminator Cycle Sequencing Kit v3.1). The sequencing reaction was carried out in a thermal cycler (ABI GeneAmp PCR System 9700), and the resolution and analysis of the labelled products were performed in a capillary sequencer (ABI 3100).

InDel identification and design of new primers for diversity studies

BioEdit (Hall 1999) was used to align sequences from which InDel polymorphisms were identified. For genes with InDel polymorphisms, new primer pairs in conserved regions flanking the InDel polymorphism were designed using Primer3 software (http://biotools.umassmed.edu/bioapps/primer3) (Table 2) to amplify fragments smaller than 350 bp that were subsequently analysed in a capillary fragment analyser (see below).

Diversity analysis

Plant material

Ninety genotypes from the citrus germplasm bank of IVIA (Spain) and INRA/CIRAD (France) were used for the diversity study with SSR and InDel markers (Online Resource 1). According to the Swingle and Reece classification system (1967), 45 genotypes belong to the three

Table 1 Primers of candidate genes

Process involved	Gene	Primers	AT	High-quality sequence (bp)	EST size (bp)	Genomic size (bp)	Genebank accessions
Flavonoids biosynthesis	Chalcone isomerase	F:TTGTTCTGATGGCCTAATGG	55	647	721	721	aCL6103Contig1
		R:AAAGGCTGTCACCGATGAAT					
	Chalcone synthase	F:GATGTTGGCCGAGTAATGCT	55	565	659	659	aCL6909Contig1
		R:ATGCCAGGTCCAAAAGCTAA					
	Flavonol synthase	F:GGAGGTGGAGAGGGTCCAAG	55	710	763	763	AB011796
		R:GGGCCACCACTCCAAGAGC					
Acids biosynthesis	Malic enzyme	F:ACATGACGACATGCTTCTGG	55	420	166	420	CB417399
		R:CGTAGCCACGCCTAGTTCAT					
	Malate dehydrogenase	F:ATGGCCGCTACATCAGCTAC	55	705	1,209	1,250	DQ901430
		R:TGCAACCCCCTTTTCAATAC					
	Vacuolar citrate/H+ symporter	F:GGCGCCACTCCTACCTTCCC	58	715	987	1,300	EF028327
		R:CGGTCATTGAAGAGTGCTCCCC					
Sugars biosynthesis	Phosphoenolpyruvate carboxylase	F:AGCCAATGGGATTTCTGACA	55	669	1,201	2,000	EF058158
		R:GCCAAGCCACACAGGTAAAT					
	Phosphofructokinase	F:CGCCGACCTCAGTCCCGTC	58	630	807	1,650	AF095520
		R:GCTGCACGCCCCATAAGCCG					
Carotenes biosynthesis	Lycopene β-cyclase 2	F:GCATGGCAACTCTTCTTAGCCCG	55	725	850	850	FJ516403
		R:AGCTCGCAAGTAAGGCTCATTCCC					
	β-Carotene hydroxylase	F:AGCCCTTCTGTCTCCTCACA	55	675	787	1,600	AF315289
		R:CCGTGGAATTTATCCGAGTG					AF296158
	Phytoene synthase	F:GCTCGTTGATGGGCCTAATGC	58	560	727	2,100	AB037975
		R:CGGGCGTAAGAGGGATTTTGC					AF220218
							AF152892
	1-Deoxyxylulose 5-phosphate synthase	F:GGCGAGGAAGCGACGAAGATGG	58	590	935	1,500	aCL303Contig1
		R:GGATCAGAACTGGCCCTGGCG					
	Lycopene β-cyclase	F:GAATTCTTGCCCCAAGTTCA	55	710	1,206	1,500	AY166796
		R:TATGGGCCACAAATCTTTCC					AF152246
							AY644699
Salt stress tolerance	Cation/H+ membrane antiporter	F:GTTGCTGATGCTACAGATG	50	840	805	1,800	aCL1735Contig1
		R:CCTCTCTCTCTTCTTTACCG					
	Raffinose synthase	F:CATGCGGAAAAGATGTACC	52	740	804	1,800	aCL3302Contig1
		R:CAGCAAGGCTGTCCATAAC					
	Vacuolar H+ pyrophosphatase	F:GCATATGCTCCCATCAGTG	53	800	831	1,650	aCL5319Contig1
		R:CAGGCTCCTGTCTGTTTGAG					

High-quality sequence resulted from cleaning the alignments. aCLxxxxContig1, sequences were obtained from the Citrus Functional Genomics Project (CFGP), http://bioinfo.ibmcp.upv.es/genomics/cfgpDB/; the rest of the sequences were obtained from the National Center for Biotechnology Information (NCBI)

AT annealing temperature

ancestral species (29 *C. reticulata*, 10 *C. maxima* and 6 *C. medica*) and 11 genotypes represented the secondary species (2 *C. aurantium*, 4 *C. sinensis*, 2 *C. paradisi* and 3 *C. limon*). Seventeen accessions are supposed of interspecific origin from their morphology or previous molecular data (46–50, 53–55, 65–66, 81, 84–89) even some of them were classified by Swingle and Reece (1967) as pure species. The last 17 accessions are hybrids from twentieth century breeding projects (67–80, 82, 83, 90).

Genotyping

Sixty-seven SSR markers were tested on the citrus population selected for our study. Fifty markers presented

Table 2 Characteristics of InDel markers

Marker name	Gene	Primers	AT	Fragment size (bp)
IDCHI	Chalcone isomerase	F:TTTCCTCTTGCTTTACGTGT	55	146–196
		R:GTCACAGGTAACGGATTTTC		
IDEMA	Malic enzyme	F:CTCTTTCTGCTTCCTGACATC	55	263–277
		R:GCCGGTGAATAAAACACAAC		
IDTRPA	Vacuolar citrate/H+ symporter	F:CCCTCGTTCTTGGTAGCTTT	55	306–309
		R:TTATGCATCCACATGCTCAC		
IDLCY2	Lycopene β-cyclase	F:CGCAAATAATTGATTCAACA	50	220–226
		R:GATGATCACGTCATATCGAA		
IDHYB1	β-Carotene hydroxylase	F:AAAAACAAAGCACCCAGAT	53	192–213
		R:GCCACCAGAACCTGTAATAA		
IDHYB2	β-Carotene hydroxylase	F:TTTGGCACATTTGCTCTCTCT	55	305–307
		R:AAAGAAGCATGCCACAGAGC		
IDPSY	Phytoene synthase	F:CCTGTCGACATTCAGGTTAG	55	246–249
		R:CTCATCACATCTTCGGTCTC		
IDPEPC1	Phosphoenolpyruvate carboxylase	F:TTTTGAACAATCGGCTAATGG	55	231–259
		R:TTGCTGGAAGAGAGACTCCAA		
IDPEPC2	Phosphoenolpyruvate carboxylase	F:TTGGAGTCTCTCTTCCAGCAA	55	128–153
		R:GTGAGAGCCACAATGCAAAA		
IDCAX	Cation/H+ membrane antiporter	F:TAAGCTGCATTTAACCCTTT	55	237–243
		R:GCAATTGGGAGATAGTCAAT		
IDAtGRC	Raffinose synthase	F:GGCAATGAAAACAATGAGAT	55	208–225
		R:TTTCAAGATTGTTGGTCCTC		
IDAPV	Vacuolar H+ pyrophosphatase	F:CAGCTATTGGAAAGGTTTGT	55	156–163
		R:GGAGACAGGCATAAAACATC		

AT annealing temperature

proper and clear results (Online Resource 2; Kijas et al. 1997; Froelicher et al. 2008; Luro et al. 2008; Aleza et al. 2011; Cuenca et al. 2011; Kamiri et al. 2011) and were used for the diversity study. Forty-seven of them were included in the Clementine genetic map (Ollitrault et al. 2011) and were well distributed between and within all linkage groups. In addition, 12 InDel markers were analysed. One of them (TRPA) is located in the Clementine genetic map (linkage group 2).

Amplification by polymerase chain reaction (PCR) was performed using wellRED forward oligonucleotides (Sigma-Aldrich; Saint-Louis, USA) for analysis with a capillary genetic fragment analyser (CEQ/GeXP Genetic Analysis Systems; Beckman Coulter; Fullerton, USA). PCR was performed in a final volume of 15 µl. Each PCR reaction consisted of 1 ng/µl template DNA, 0.2 mM dNTPs, 0.2 µM wellRED dye-labelled forward primer, 0.2 µM of non-dye-labelled reverse primer, 10× PCR buffer (Fermentas), 1.5 mM $MgCl_2$ and 0.027 U/µl *Taq* DNA polymerase (Fermentas). PCR reactions were carried out with the following program: 5 min at 94°C; 40 cycles of 30 s at 94°C, 30 s at 55 or 50°C (depending on the primer) and 1 min at 72°C with a final extension of 4 min at 72°C.

Denaturation and capillary electrophoresis were carried out on a Capillary Gel Electrophoresis CEQ™ 8000 Genetic Analysis System using linear polyacrylamide according to the manufacturer's instructions (Beckman Coulter Inc.). Genetic analysis system software (GenomeLab™ GeXP version 10.0) was used for data collection and analysis. Alleles were sized based on a DNA size standard (400 bp).

Data analysis

Neighbour-joining (NJ) analysis

Population diversity organisation was analysed with DARwin software (Perrier and Jacquemoud-Collet 2006). For each primer, bands were scored as allelic data to calculate the genetic dissimilarity matrix using the simple matching dissimilarity index (d_{i-j}) between pairs of accessions (units):

$$d_{i-j} = 1 - 1/L \sum_{l=1}^{L} m_l/2$$

where d_{i-j} is the dissimilarity between units i and j, L is the number of loci and m_l is the number of matching alleles for locus l. From the dissimilarity matrix obtained, a weighted NJ tree (Saitou and Nei 1987) was computed using the Dissimilarity Analysis and Representation for Windows (DARwin5) software version 5.0.159, and the robustness of branches was tested using 10,000 bootstraps.

To establish the genetic structure with the core set of markers, NJ under topological constraints was used. It is a modified version that forces the a priori known topology of a subset of samples and positions additional subsets on the previous organisation. Secondary species and modern cultivars were positioned under the constraint of a tree based on basic taxa.

Severinia buxifolia (Poir.) Ten, a species related to citrus, was used to root NJ trees.

Principal coordinates analysis (PCoA)

It was performed using the software GENEALEX6 (Peakall and Smouse 2006). The data from molecular markers was used to obtain the pairwise genetic distance matrix, which was standardised and used for PCoA analysis.

Population structure

It was inferred with the Structure version 2.3.3 program (http://cbsuapps.tc.cornell.edu/structure), which implements a model-based clustering method using genotype data (Pritchard et al. 2000; Falush et al. 2003). According to the general agreement on the origin of cultivated species (Scora (1975); Barrett and Rhodes (1976)), we considered an initial structure between three populations ($K = 3$): mandarin (29 samples), pummelo (10 samples) and citron (6 samples), assuming that the analysed genotypes are derived from these three ancestral taxa. The relative proportion of these ancestral populations in the secondary species and hybrids was assigned based on this assumption of an admixture model. Correlated allele frequencies were determined from the estimates of the three ancestral populations defined in this work. Ten runs of structure were performed with 500,000 steps of burning followed by 1,000,000 Monte Carlo Markov chain (MCMC) repetitions.

F_{stat} parameters

F_{is}, F_{it} and F_{st} were calculated with the software program GENETIX v. 4.03 based on the parameters of Wright (1969) and Weir and Cockerham (1984).

Linkage disequilibrium

For multi-allelic loci, LD between two loci is commonly measured by the D' estimate (Gupta et al. 2005). D' values for each pair of markers were estimated on the whole data set using the software program PowerMarker v. 3.25 (Liu and Muse 2005). D' values vary from 0 (total random association between alleles of the two considered loci) to 1 (total LD). The p value for obtaining the significance of D' was estimated by the exact test.

Selection of a subset of markers for quick genotyping

The methodology described by Jombart et al. (2010) was employed to obtain a small number of markers (core set) with good interspecific and intraspecific differentiation for quick and accurate genotyping. The procedure is based on a discriminant analysis of principal components (DAPC). Data from molecular markers are transformed with a PCoA, and the matrix obtained is employed to perform a discriminant analysis (DA). These results are used to calculate the allele contribution to the main axes, and the alleles with the highest contribution are selected. Expected heterozygosity was used as an extra parameter to select primers that allow good intraspecific differentiation.

Results

Interspecific InDel polymorphism research and InDel marker development

For the 16 genes a total of 10,701 bp by genotypes were successfully sequenced and aligned (Table 1), allowing the identification of 12 InDel polymorphic loci in 10 genes. Specific InDel polymorphisms were encountered in four loci in *C. medica* and another four loci in *C. maxima*, whereas the other InDel polymorphisms were detected in different groups.

New primers were designed to analyse the InDel diversity of these 12 loci (Table 2). In this diversity study, four loci (IDCHI, IDEMA, IDHYB1 and IDLCY2) had novel alleles not present in the six genotypes initially sequenced. Amplicons of genotypes with these new alleles were sequenced, as described previously, to analyse the origin of this pluri-allelism (Online Resource 3). At locus IDCHI, a new polymorphism was found in heterozygosis in *C. sunki*, another one was found in IDEMA (genotype *C. sunki* and others in heterozygosis), one at IDHYB1 in Cleopatra mandarin and other genotypes in heterozygosis and the last polymorphism was found in homozygosis at locus IDLCY2 in *C. sunki* and other genotypes in

heterozygosis. InDel allele sequences of the ten analysed genes are given in Online Resource 3. For multi-allelic loci, the variation of amplicon size is due to variation in size of the same InDel (IDCHI, IDHYB1 and IDLCY2) or several InDels between the two primer sites (IDCHI, IDHYB2 and IDCAX). Three loci (IDPSY, IDPEPC2 and IDAVP) displayed intra-taxon polymorphisms only in *C. medica*, and the other three loci (IDHYB2, IDPEPC1 and IDATGRC) displayed intra-taxon polymorphisms only in *C. maxima*. Polymorphisms in loci IDTRPA, IDLCY2 and IDHYB1 may be due to copy number variations of SSRs.

InDel analysis

A total of 32 alleles were detected from the InDel markers. The average number of alleles per locus was 2.67. Genetic diversity statistics were calculated for each InDel marker in the entire population and for different citrus groups, including *C. reticulata*, *C. medica* and *C. maxima* (Online Resource 4) The allele number varied between 2 (for 7 loci) and 5 for IDCAX. IDCAX displayed the highest diversity ($H_e = 0.69$) related to different alleles in the three ancestral taxa. IDAVP ($H_e = 0.12$) was the least informative marker, as it differentiated only varieties from the citron subpopulation. The best markers for genotype differentiation within mandarins, pummelo and citron were IDCAX, IDPEPC1 and IDCHI, respectively. F_{stats} parameters (Wright 1969; Weir and Cockerham 1984) were estimated to analyse the differentiation between the three ancestral taxa (*C. maxima*, *C. medica* and *C. reticulata*). F_{is} values varied from −0.474 for IDAVP to 0.125 for IDCHI. For four loci, it was not possible to calculate the F_{is} parameter because the loci were monomorphic in each of the ancestral taxa. With the exception of IDAVP, the F_{is} value confirms a situation close to the Hardy–Weinberg equilibrium within each species. In contrast, F_{it} values with a high average (0.730) showed that, in the whole population (of the subset of the 3 ancestral taxa), the inbreeding coefficient is higher than within taxa for almost all of the markers, indicating an important organisation between taxa. Only IDTRPA had a low value (−0.149) with two alleles shared by *C. maxima* and *C. reticulata*. The high F_{st} average value (0.766) and the F_{st} value of each locus (excluding IDTRPA) confirms that the inter-taxa differentiation contributes much more to the global inbreeding than does the intra-taxa component. Thus, a large portion of the total variation is explained by the differentiation between populations.

Average data over all InDel loci are given in Table 3. The average F_W value (0.433) shows a high deficit of observed heterozygous individuals in the population. Indeed, the whole population had an observed heterozygosity of 0.18, which is 38% lower than the expected heterozygosity (0.29), suggesting an organisation in differentiated sub-gene pools with limited gene flows. Individually, the different taxa had an observed heterozygosity similar to the expected. *C. reticulata* was the most polymorphic ($H_e = 0.13$) and heterozygous ($H_o = 0.14$) ancestral taxon, and *C. maxima* was the least polymorphic and heterozygous ($H_o = H_e = 0.07$) ancestral taxon.

SSR analysis

The same genetic diversity parameters were calculated for each individual SSR marker, the entire population and for the different specified citrus groups (Online Resource 5). A total of 405 alleles were detected with the SSR markers. The average number of alleles and H_e per locus was 8.1 and 0.71, respectively. The allele number varied between 3 (for loci MEST107, CAC15 and CAC23) and 14 (MEST56). TAA41 was the most informative marker with a H_e of 0.86, and CAC15 was the least informative marker ($H_e = 0.39$). Most of the markers (48 out of 50) showed H_e values higher than 0.5. When analysing the organisation among the three basic taxa, F_{is} values varied from −0.114 for CAC23 to 0.594 for mCrCIR05A04. The overall F_{is} value was close to zero (0.030), confirming that few deviations from the Hardy–Weinberg equilibrium occurred within each basic taxon. In contrast, high F_{it} and F_{st} values for almost all markers (averages of 0.454 and 0.434, respectively) are evidence of high differentiation between the three basic taxa.

Average data over all InDel loci are given in Table 3. The population displayed a deficit of average observed heterozygosity ($H_o = 0.59$) compared with the expected value under Hardy–Weinberg equilibrium ($H_e = 0.71$). This finding is confirmed by the average F_W value (0.175). Each of the three basic taxa had an observed heterozygosity close to the expected value. *C. reticulata* was the most diverse ($H_e = 0.56$) and heterozygous ($H_o = 0.56$) ancestral taxa, but citron was the lowest ($H_e = 0.28$ and $H_o = 0.17$).

Comparative diversity structure displayed by InDels and SSRs

The genetic parameters for InDel and SSR markers, respectively, were as follows: allele number per locus ranged from 2 to 5 and from 3 to 14, observed heterozygosity average was 18 and 59% and the percentage of varieties differentiated among the whole population was 57.78% (52 out of 90) and 91.11% (82 out of 90). The distribution of H_e and F_{st} between the three basic taxa (Fig. 1) confirmed that InDel markers are less polymorphic than are SSR markers (lower H_e values) but allow a better

Table 3 Statistical summary of the diversity of InDel and SSR markers

Marker type	All citrus accessions				*C. reticulata*			*C. maxima*			*C. medica*			3 basic taxa		
	N	H_o	H_e	F_W	*N*	H_o	H_e	*N*	H_o	H_e	*N*	H_o	H_e	F_{is}	F_{it}	F_{st}
InDel	2.67	0.18	0.29	0.433	1.58	0.14	0.13	1.25	0.07	0.07	1.25	0.09	0.09	−0.148	0.730	0.766
SSR	8.10	0.59	0.71	0.175	5.02	0.56	0.56	3.36	0.50	0.52	1.94	0.17	0.28	0.030	0.454	0.434

Mean values are represented in the table

N allele number, H_o heterozygosity observed, H_e heterozygosity expected, F_w Wright fixation index over the whole population, F_{is}, F_{it} and F_{st} Weir and Cockerham Index over the subset of *C. maxima*, *C. medica* and *C. reticulata* accessions

differentiation between ancestral species (higher F_{st} values). Statistics for the three ancestral groups were calculated for both types of primers (Table 3). Expected and observed heterozygosity were similar for both types of markers but were lower for InDels than SSRs within each taxon. With SSR markers, all accessions of *C. medica* and *C. maxima* were fully differentiated, whereas 96.7% of intervarietal differentiation was obtained within *C. reticulata*. The InDel intervarietal differentiations were 100, 40 and 53.3% within *C. medica*, *C. maxima* and *C. reticulata*, respectively. Twelve out of 50 SSR and 7 out of 12 InDel markers displayed significant deficits of heterozygous genotypes in the whole sample set (Online resources 4 and 5).

The F_{st} value was estimated for each pair of basic taxa, and it was systematically higher with InDel than SSR markers. The least differentiated species were *C. reticulata* and *C. maxima* (F_{st} of 0.373 and 0.422 for SSR and InDel, respectively), followed by *C. reticulata*/*C. medica* (0.427 and 0.758) and *C. maxima*/*C. medica* (0.484 and 0.844). All of these data support the conclusion that InDel markers yield higher inter-taxa discrimination compared with SSR markers.

Both NJ, Fig. 2 and principal coordinates analysis PCoA, Fig. 3 analyses revealed a clear differentiation between the three ancestral citrus taxa for both kinds of markers.

NJ trees (Fig. 2) clearly separated *C. medica* and *C. maxima* from *C. reticulata*. For InDel markers (Fig. 2a), *C. medica* was the best defined group and showed good bootstrap support in all branches of its cluster, and all of the samples were differentiated. The *C. maxima* group formed a well-defined clade, but only four profiles were differentiated among ten accessions. The intraspecific diversity of *C. reticulata* was not well resolved (low bootstrap support), perhaps due to the high number of hybrids (within mandarin) in the sample set. Fourteen genotypes were differentiated among the 29 mandarins.

SSRs allowed a complete intercultivar differentiation for *C. maxima* and *C. medica*, whereas only two *C. reticulata* cultivars (East India SG and Vohangisany Ambodiampoly) were not differentiated (Fig. 2b).

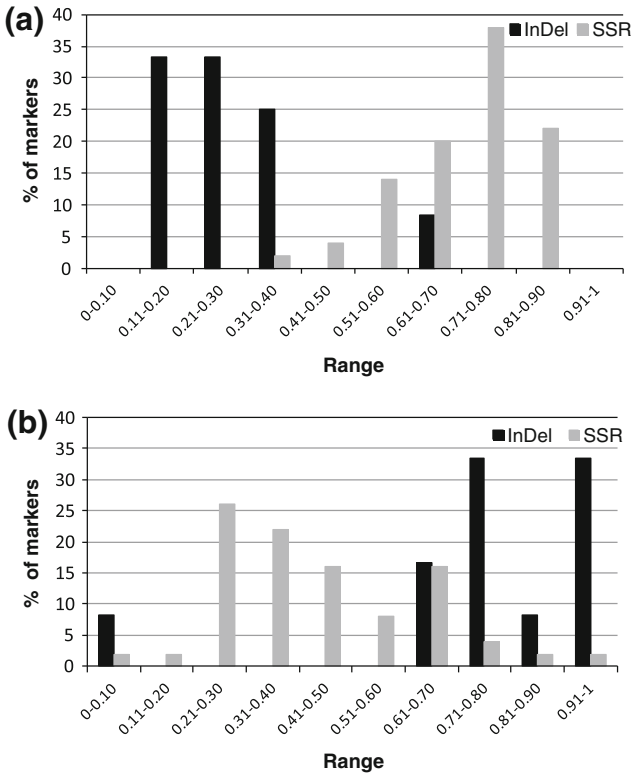

Fig. 1 Comparison between InDel and SSR markers of the expected heterozygosity (H_e) and the genetic differentiation index (F_{st}) between ancestral taxa. **a** Expected heterozygosity, **b** genetic differentiation index

NJ analysis confirmed higher intraspecific diversity with SSRs than with InDel markers. The lower differentiation obtained with InDels may be partly due to the lower number of these markers. However, it is also clearly explained by their lower allelic diversity, which is observed mostly at the interspecific level. Clustering was stronger with InDel than with SSR markers, but SSRs allowed a better intra-cluster differentiation between accessions.

PCoA (Fig. 3) is more adapted than tree representation in describing the organisation of genetic diversity when hybrids between differentiated groups are frequent in the sample. In our study, PCoA allowed us to have a better idea

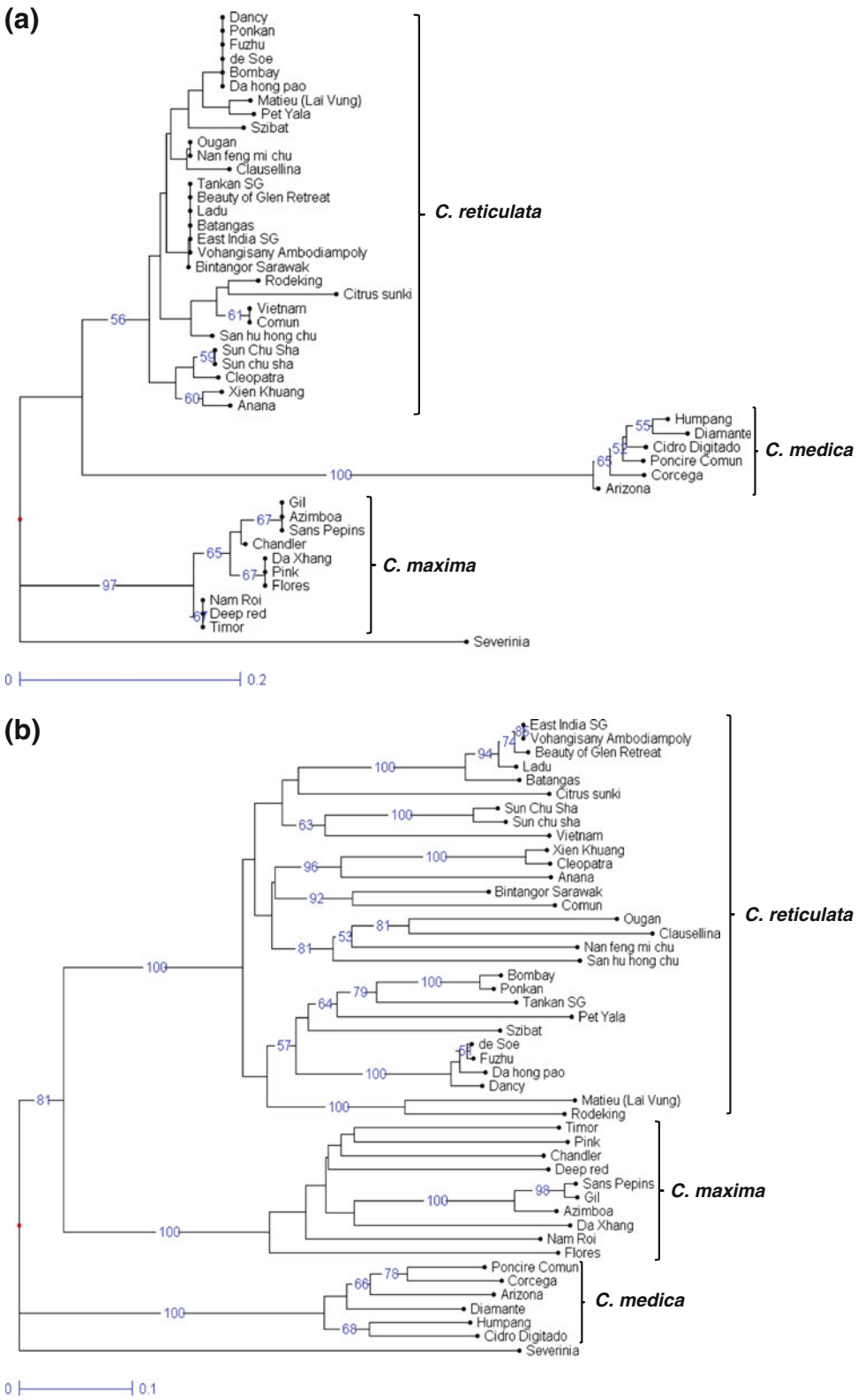

Fig. 2 NJ bootstrap consensus trees of 45 accessions of citrus (3 ancestor groups) including one outgroup, *Severinia buxifolia*. *Numbers* are bootstrap values over 50 based on 10,000 resampling. **a** InDel markers data, **b** SSR markers data

of the relative contribution of the three basic taxa to the genome constitutions of secondary species and modern hybrids. Almost all of the existing variability (92.10%) is represented in the first two axes for InDels (Fig. 3a), but only 75.89% variability is represented for SSRs (Fig. 3b). This result confirms that higher interspecific organisation is

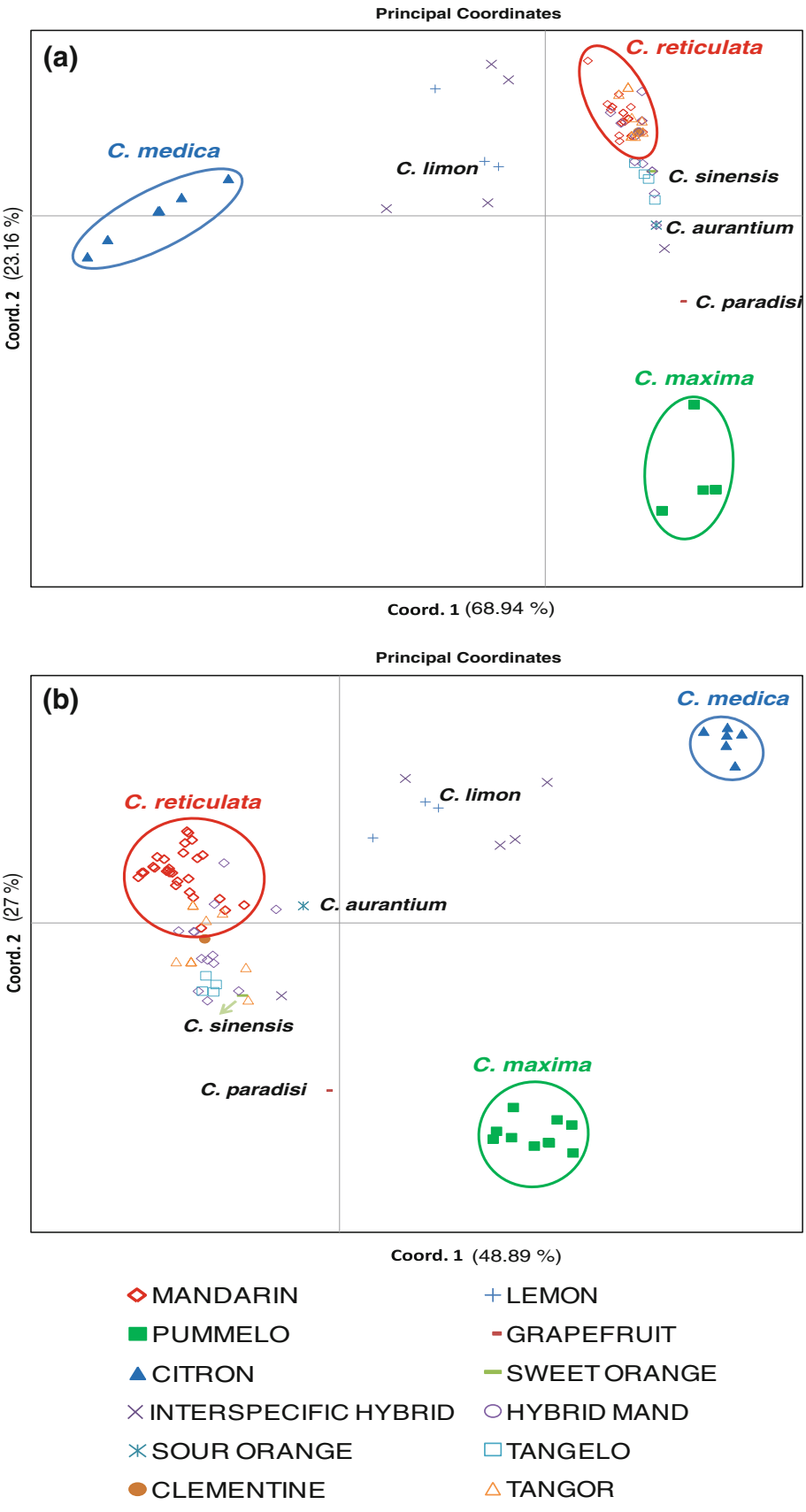

Fig. 3 Organization of cultivated *Citrus* genetic diversity; principal coordinates analysis. **a** InDel markers data, **b** SSR markers data. Mandarin (samples 1–29), pummelo (samples 30–39), citron (samples 40–45), interspecific hybrids (samples 46–50), sour orange (samples 51–52), clementine (samples 53–54), lemon (samples 56–58), grapefruit (samples 59–60), sweet orange (samples 61–64), hybrid mandarins (samples 67–76), tangelo (samples 77–80) and tangor (samples 81–90). (Sample number assignment can be found in Online Resource 1)

determined using InDel markers. For these markers, the *C. medica* group (and its hybrids with citron as one parent) was strongly differentiated from *C. reticulata* (and its hybrids) and *C. maxima* by axis 1, whereas the *C. maxima* group was differentiated from the other species by axis 2. *C. paradisi* varieties and Bali hybrid, mandarin Suntara and *C. aurantium* (the last two had exactly the same position), in this order, were closer to *C. maxima* with InDel markers than with SSR markers. Tangors (mandarin × sweet orange) were closer to the *C. reticulata* cluster and Tangelos (mandarin × grapefruit) were closer to *C. maxima*, as expected from their origin. Clementines were close to *C. reticulata* accessions and some hybrids that have clementines as a parent.

For SSRs, *C. medica* was differentiated from *C. maxima* by axis 1, and the *C. reticulata* group was differentiated from *C. medica* by axis 2. *C. reticulata* accessions were more dispersed around the axis based on SSR markers than with InDel markers. As *C. sinensis*, *C. aurantium* appeared much more related to *C. reticulata* than to *C. maxima*, *C. limon* was clearly positioned between the *C. medica* gene pool and *C. aurantium*. Some hybrids derived from *C. medica* (Poncil, Rhobs el Arsa, Kadu Mul and Damas) were positioned in a similar place, suggesting that these hybrids share similar origins as *C. limon*. Tangor was the most dispersed group, Murcott and Umatilla were the closest varieties to *C. reticulata* and Ortanique was the closest to *C. maxima*. Tangelos were similarly distanced between them. Clementines were close to the *C. reticulata* gene pool, whereas *C. paradisi* was the secondary species closest to *C. maxima*.

Contribution of the ancestral taxa to secondary species and modern hybrids; analysis with structure software

PCoA analysis provided some information on the relative contribution of the three basic taxa to the genome constitution of the secondary ones, confirming the status of *C. medica*, *C. reticulata* and *C. maxima* as parental gene pools of the other species and modern hybrids in this study. Assuming an admixture model between the three ancestral species, the relative proportion of ancestral taxa genomes in the secondary species and recent hybrids was inferred using the Structure version 2.3.3 software (Fig. 4) with the complete set of data (SSRs + InDels).

C. limon and hybrids with *C. medica* as parents (Poncil, Rhobs el Arsa, Kadu Mul and Damas) have the greatest average contribution from *C. medica* (46%). Contributions of *C. medica* lower than 2.5%, which was observed for *C. sinensis*, *C. aurantium*, *C. paradisi*, Bali pummelo, Clementine and Temple, can probably be considered artefacts and related to the relatively low number of representative genotypes of the basic taxa and probable lack of intra-taxa diversity. *C. paradisi* is the secondary species with the highest contribution from *C. maxima* (60%), followed by *C. aurantium* (30%), *C. sinensis* (25%), tangelo group (20%), tangor group (10%) and clementines (7%). *C. aurantium* varieties displayed seven rare alleles, five of which were shared with Suntara mandarin (two of them were also shared with *C. limon*), one was shared with *C. limon* and another one was only present in *C. aurantium*.

The contributions of the ancestral groups to the secondary species obtained with the Structure software was compared with direct estimations performed with the specific allele from the SSR and InDel markers derived from the mandarin, pummelo and citron groups (Table 4). No significant difference was found between the two methods of evaluation. It is interesting to note that no specific allele from *C. medica* was observed in *C. sinensis*, *C. paradisi*, Bali pummelo, Clementine and Temple, which confirms that the low values estimated for the same genotypes with Structure were not significant.

Linkage disequilibrium

Based on the data obtained with the 50 SSR markers distributed along the genome, the extent of genome-wide LD was estimated by D' for the whole population. InDel markers were not selected for this analysis because they were not mapped. D' values ranged from 0.11 to 0.9 for interchromosome pairs of loci and from 0.21 to 0.94 for intrachromosome pairs (Fig. 5). The average D' estimates for marker pairs within and between chromosomes were 0.56 and 0.51, respectively. For interchromosome and intrachromosome marker pairs, 65.69 and 53.68% of the D' values were over 0.5, respectively. The percentage of significant p values was very high for marker pairs within and between chromosomes: 99.27/99.26% (<5%) and 97.08/97.89% (<1%), respectively. When analysing the relation between LD and genetic distances between markers (Fig. 6), it appears that there is a high LD even between distant markers with a limited LD decay with increasing distances. The distribution of the interchromosome D' is highly similar. The mean value of D' was 0.5161 for the whole population and all marker pairs.

Selection of a subset of markers for quick genotyping

Identifying a subset of markers that can differentiate new accessions and study their origin could be useful for quick and inexpensive genotyping. In this study, the parameters used to select the subset of markers were high locus contribution to F1 and F2 coordinates of the PCoA analysis (interspecific organisation), high expected heterozygosity (global diversity displayed by the marker) and limited LD between the selected markers to avoid excessive redundant

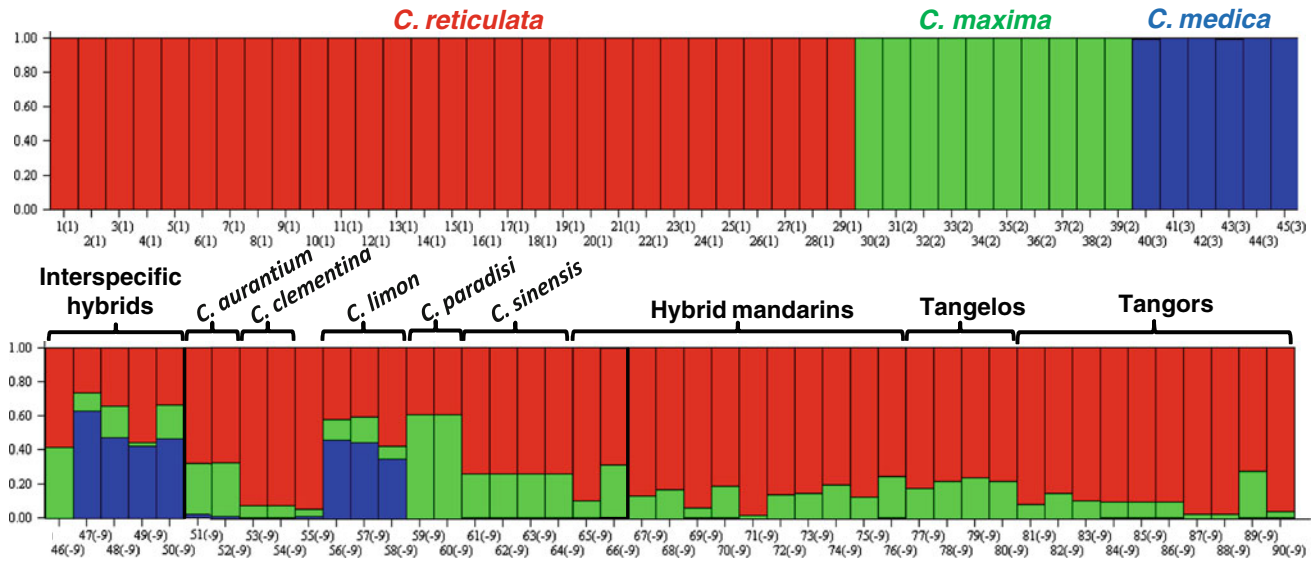

Fig. 4 Relative contribution of basic taxa to secondary species and modern cultivars; structure analysis with $K = 3$ as initial hypothesis, considering SSR and InDel data. In *parenthesis* are indicated the reference population assignment for the admixture model *1 C. reticulata* population, *2 C. maxima* population, *3 C. medica* population, *−9* population with unknown contribution from ancestors (sample number assignment can be found in Online Resource 1)

Table 4 Contribution of the ancestral taxa to secondary species: comparison between direct estimation from interspecific discriminant allele and the estimation from Structure software (admixture model)

Latin name	Common name	SSR + InDel allele specific from			Total informative alleles	Direct estimation from discriminant alleles			Structure data			
		Re	Ma	Me	SSR + InDel	Re (%)	Ma (%)	Me (%)	Re (%)	Ma (%)	Me (%)	χ^2
C. aurantium	Sevillano	32	16	1	49	65.31	32.65	2.04	67.2	30.6	2.2	0.10
C. clementina	Clemenules	49	1	0	50	98	2	0	92	7.1	0.9	2.48
C. limon	Eureka Frost	21	4	22	47	44.68	8.51	46.8	41.6	12.1	46.3	0.61
C. limon	Lisbon Limoneira	20	6	22	48	41.67	12.50	45.8	40.3	14.7	45	0.19
C. paradisi	Marsh	21	22	0	43	48.84	51.16	0	38.6	60.9	0.5	2.05
C. sinensis	Valencia late delta	37	5	0	42	88.10	11.90	0	73.3	25.6	1.1	4.79
× *C. maxima*	Bali	28	18	0	46	60.87	39.13	0	58.4	41.1	0.6	0.37
× *C. medica*	Poncil	14	5	26	45	31.11	11.11	57.8	26.3	10.8	62.9	0.59
× *C. medica*	Rhobs el Arsa	15	9	20	44	34.09	20.45	45.5	33.6	18.7	47.7	0.12
× *C. medica*	Kadu Mul	31	0	23	54	57.41	0	42.6	54.9	2.7	42.3	1.52
× *C. medica*	Damas	11	8	24	43	25.58	18.60	55.8	33.6	18.7	47.7	1.42
× *C. reticulata*	*Citrus daoxianensis*	50	1	0	51	98.04	1.96	0	94.1	4.1	1.8	1.57

Re C. reticulata, *Ma C. maxima*, *Me C. medica*, χ^2 homogeneity test on distribution with the two methods ($\alpha = 0.05$; $\chi^2 < 5.99$)

information between markers (Online Resource 6). A total of nine markers were selected: mCrCI02D04b and MEST431 were selected for their high contribution to the F1 component (which distinguished *C. reticulata* from the other two ancestors), IDCHI and IDCAX have a high contribution to F2 (axis which differentiates between *C. medica* and the other ancestors), mCrCI07F11 and mCrCI07D06 contributed in both axes (it is helpful to distinguish individuals that are intermediate) and MEST488, TAA41 and mCrCI02G12 were selected for their high expected heterozygosity. Six out of the nine linkage groups were represented by the selected marker subset. With these nine markers, the three ancestors groups were clearly differentiated (Online Resource 7). Samples in the *C. medica* group were fully separated, whereas in *C. maxima*, only 'Gil' and 'Sans Pepins' cultivars could not be differentiated. *C. reticulata* within diversity was slightly less resolved than with the whole marker set (6 mandarins were

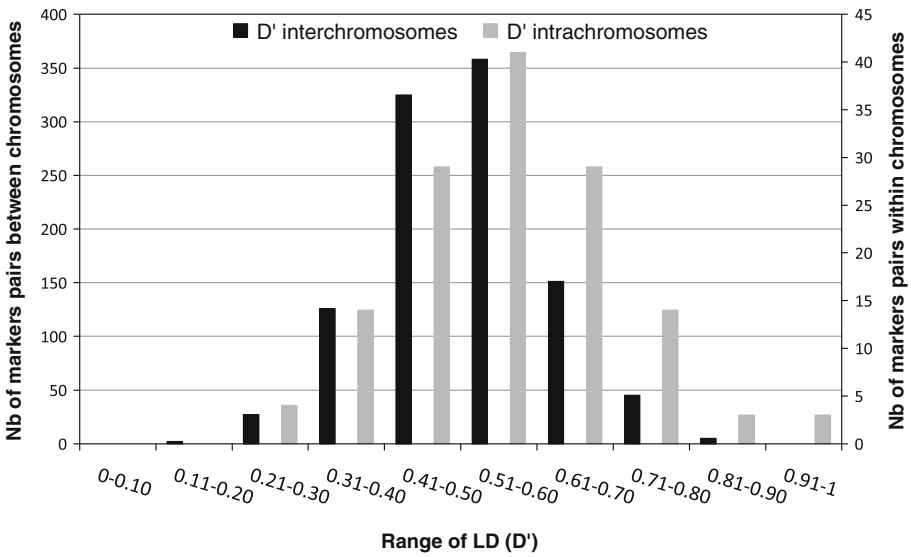

Fig. 5 Linkage disequilibrium for marker pairs within a same linkage group (*grey*) and between markers located in different chromosomes (*black*)

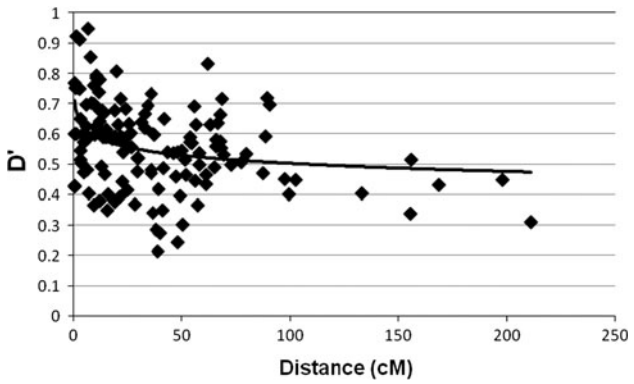

Fig. 6 Relation between LD in the population for all markers pairs within chromosomes and genetic distances (Clementine genetic map; Ollitrault et al. 2011)

not distinguished). The average observed and expected heterozygosity values were 56 and 64%, respectively, and the F_W was 0.163.

Discussion

Citrus InDel markers are less polymorphic but display higher interspecies differentiation than do SSR markers

InDels are generally considered to be interesting polymorphisms for genetic studies. However, despite increasing molecular resources in citrus, such as EST sequence information (Forment et al. 2005; Terol et al. 2007), HarvEST software Version 1.32 of "HarvEST:Citrus" (http://www.harvest-web.org) and genomic sequence information (Terol et al. 2008), no specific study has been conducted prior to the present work to analyse the value of nuclear InDels as genetic markers in *Citrus*. We searched for InDel polymorphisms in the three basic taxa (*C. reticulata*, *C. maxima* and *C. medica*) by sequencing PCR products obtained from 13 genes. Primers were designed to amplify 150–350 bp fragments flanking the 12 identified InDels, and amplicon size variation was studied by capillary electrophoresis on a sample of 90 genotypes of the *Citrus* genus.

The frequency of InDels per kb in citrus was 0.71 and 5.22 in exon and intron sequences, respectively. More sequence polymorphisms were found in non-coding regions than in coding regions. Similar results have been observed in other species. In Brassica, 0.45 and 7.42 InDel/kb were found in exons and introns, respectively (Park et al. 2010). In melon, InDels occurred less frequently in introns (approximately 0.60/kb) and no InDel was found inside coding regions (Morales et al. 2004). In maize, 0.43 and 11.76 InDels/kb were found in coding and non-coding regions, respectively (Ching et al. 2002).

The mean number of alleles per locus was 2.83 with a maximum of five alleles at the IDCAX locus. Seven of the 12 markers were diallelic. Retroposon movements, such as *Alu* or the L1 element, are known to generate such diallelic InDels (Watkins et al. 2001). In our study, pluri-allelism was caused by differences in InDel size or the presence of several InDels in the amplified fragments. InDels with a size that is not a multiple of 3 are uncommon in exons but relatively common in introns (Mills et al. 2006; The Arabidopsis Genome Initiative 2000).

Almost 60% of the whole set of samples were differentiated with the 12 InDel markers. A better differentiation may be obtained with more InDels; however, the low mean number of alleles per locus may be a limitation compared with techniques using multi-allelic markers, such as SSRs. Indeed, we found a mean value of 8.1 alleles per locus for

SSRs. With higher allelic diversity and intra-taxon diversity, SSRs are more informative than InDels at the intra-specific level. The number of repeats in microsatellites evolves at a high rate (Weber and Wrong 1993; Jarne and Lagoda 1996), which can vary depending on the number of repeats or base composition (Bachtrog et al. 2000). Thus, there are generally good markers for intra-population diversity analysis, as we observed at the intra-taxon level. However, due to this important rate of variation, homoplasy should be relatively frequent, as demonstrated in *Citrus* (Barkley et al. 2009), and should limit the value of SSRs as phylogenetic markers. Our results confirmed this hypothesis, as we observed that InDel markers displayed a much higher differentiation between the three basic taxa than SSRs, with F_{st} value averages of 0.77 and 0.43, respectively. The structure of the whole sample diversity was higher for InDels with a fixation index value (F_w; Wright 1978) of 0.433 and 0.175 for SSRs. Interestingly, the three InDel markers (IDTRPA, IDLCY2 and IDHYB1) that may result from variation in copy number of SSRs showed lower F_{st} value than the average. Therefore, these three markers provide less inter-taxa differentiation than the other InDels. The PCA also confirmed a higher level of structure of the diversity displayed by InDels markers than by SSRs with 92.2 and 75% of the whole diversity, respectively, represented by the first two axes.

Thus, we can conclude that, in the *Citrus* genus, InDel markers are less polymorphic than SSRs but display a higher organisation of genetic diversity at the interspecific level. From the 50 SSRs and 12 InDels we have selected a core set of 9 markers (2 InDels and 7 SSRs) that keep the interspecific structure, as well as a significant part of the intraspecific polymorphism information. These markers should be useful for the rapid and inexpensive assignment of a new germplasm variety to its genetic group or identification of its potential hybrid origin.

InDels play a major role in sequence divergence between closely related DNA sequences in animals, plants, insects and bacteria. InDels are responsible for many more unmatched nucleotides than are base substitutions, and human genetic data suggests that InDels are a major source of gene defects (Britten et al. 2003). InDels in coding regions probably have functional roles and are considered to be a significant source of evolutionary change in eucaryotic and bacterial evolution (Britten et al. 2003). InDels in genes with functional diversity between alleles should be highly useful for marker-assisted selection (Raman et al. 2006) or QTL mapping (Vasemägi et al. 2010). Using the increasing amounts of sequence information acquired by new technologies (454-Roche, SOLiD system-Applied biosystems or Solexa-Illumina), the development of PCR-based InDel markers will become an important source of genetic markers that are easy and inexpensive to use in phylogenetic and genetic association studies in *Citrus*.

The genetic constitution of secondary species and modern hybrids

In agreement with previous molecular studies (Barkley et al. 2006; Luro et al. 2008), no intercultivar polymorphism was found at intraspecific level for *C. sinensis*, *C. aurantium* and *C. paradisi*, whereas these species are highly heterozygous (H_o values of 0.47, 0.50 and 0.44, respectively). This finding confirms that most of the intervarietal polymorphisms within these secondary species arise from punctual mutation or movement of transposable elements (Bretó et al. 2001). These types of mutations are unlikely to be detected with SSR or InDel markers. The three lemon cultivars were differentiated. However, lemons cv. 'Lisbon' and cv. 'Eureka' only differed for five markers.

PCA using SSR or InDel markers confirmed that the differentiation between *C. reticulata*, *C. maxima* and *C. medica* gene pools was the structuring factor of the analysed edible citrus germplasm. Secondary species and modern tangor and tangelo cultivars (which display higher heterozygosity than *C. reticulata*, *C. maxima* and *C. medica*) take intermediary positions between the three basic taxa, confirming their hybrid status. Structure analysis with an admixture model considering *C. reticulata*, *C. maxima* and *C. medica* at the origin of all analysed germplasm allowed us to estimate the contribution of these taxa to the genomes of secondary species, modern cultivars and some genotypes of unclear origin.

Two accessions initially considered as representative of *C. maxima* and *C. medica* (Bali pummelo and Poncil citron, respectively) were discarded by structure analysis from the ancestor species and positioned as hybrids. Bali seemed to be a hybrid between *C. reticulata* and *C. maxima* (genome contributions of 57 and 43%, respectively) and Poncil seemed to be a tri-hybrid from *C. medica* (63%), *C. reticulata* (26%) and *C. maxima* (11%).

As proposed by Roose et al. (2009), we found that sweet orange (*C. sinensis*) exhibits close to 75% *C. reticulata* and 25% *C. maxima* contribution and thus should be the result of a backcross 1 (BC1) [(*C. maxima* × *C. reticulata*) × *C. reticulata*]. These contributions differ from the ones estimated by Nicolosi et al. (2000) where *C. sinensis* shared half of its markers with *C. reticulata* and the other half with *C. maxima* and in Barkley et al. (2006) where only 6–8% of its genome arose from *C. maxima*.

It is believed that grapefruit (*C. paradisi*) arose from a cross between pummelo and sweet orange in the West Indies where they were introduced after Christopher Columbus discovered the new world (Barrett and Rhodes

1976; Nicolosi et al. 2000). Grapefruit displays a contribution of 61% from *C. maxima* and 39% from *C. reticulata*, which are values that are close to the theoretical average values (62.5 and 37.5%, respectively) expected for a *C. maxima* × [(*C. maxima* × *C. reticulata*) × *C. reticulata*] hybrid.

Sour orange (*C. aurantium*) is thought to be derived from hybridisation between *C. maxima* and *C. reticulata* gene pools (Nicolosi et al. 2000; Barkley et al. 2006; Uzun et al. 2009). Our analysis with Structure suggests that it showed a greater contribution from *C. reticulata* (68%) than did *C. maxima* (30%) and a bit of *C. medica* (2%). Seven rare alleles were found in *C. aurantium* that were not present in the analysed germplasm of the three main ancestors. However, five of them were found in the accession Suntara mandarin. Furthermore, Suntara and *C. aurantium* share the same alleles at most loci. Thus, there is a high probability that *C. aurantium* and Suntara mandarin share parentage, but we do not have sufficient evidences to conclude whether Suntara is a parent or a hybrid from *C. aurantium*. The small contribution of *C. medica* (2%) can probably be considered an artefact by estimation with Structure software, due to an underrepresentation of *C. maxima* and *C. reticulata* diversity. It is likely that *C. aurantium* is a BC1 (*C. maxima* × (*C. maxima* × *C. reticulata*)).

In agreement with its putative *C. aurantium* × *C. medica* origin (Nicolosi et al. 2000; Gulsen and Roose 2001), we found that lemons (*C. limon*) cv. 'Eureka' and 'Lisbon' had a complex tri-hybrid structure from *C. reticulata* (41%), *C. medica* (45%) and *C. maxima* (13%). The argument that *C. aurantium* is one parent is reinforced by the fact that these two lemons shares three rare alleles with *C. aurantium*.

Mandarin-like varieties are an increasing component of the citrus fresh fruit market and include *C. reticulata* hybrids, known or supposed tangors (hybrids between *C. reticulata* and *C. sinensis*) and tangelos (hybrids between *C. reticulata* and *C. paradisi*). The clementine, a variety selected from a seedling of "Common mandarin" one century ago in Algeria, is the most popular variety of mandarin in the Mediterranean Basin. Most of its genome is inherited from *C. reticulata*, but it seems to have been introgressed in small part from *C. maxima* (6%). The allelic constitution of clementine is in agreement with the hypothesis of a "Common mandarin" × *C. sinensis* hybridisation (Deng et al. 1996; Nicolosi et al. 2000). In addition, the "Temple", "Ellendale", "Murcott" and "King" varieties have been considered as tangor. These varieties showed close to 90% contribution of the *C. reticulata* genome and 10% contribution of the *C. maxima* genome, as expected for hybrids between *C. reticulata* and *C. sinensis*. Moreover, they shared most of their alleles with these two species. Our results confirm the hypothesis of Swingle (1943) Coletta Filho et al. (1998) and Nicolosi et al. (2000) regarding the origin of King. As expected, tangelos had a greater contribution of *C. maxima* than tangors (approximately 20%).

Of the genotypes of uncertain origin, we found that *C. daoxianensis* is mostly of *C. reticulata* origin (94%). This result is in agreement with Li et al. (1992), who considered *C. daoxianensis* to be a wild mandarin. Rhobs el Arsa was considered by Federici et al. (1998) to be a cross between *C. aurantium* and *C. medica*, as are lemons. Our results are in agreement with this hypothesis. The origin of Kadu Mul has not been reported previously. Our results prompt the hypothesis that Kadu Mul arose from a cross *C. medica* × *C. reticulata*, as we found that Kadu Mul exhibits 42.3 and 54.9% contribution from *C. medica* and *C. reticulata*, respectively.

This study showed that the ancestral *C. reticulata* group contributes to a great proportion of the genomes of secondary species and recent hybrids. The facultative apomixis exhibited by all secondary species probably arose from the *C. reticulata* germplasm.

Cultivated citrus: a highly structured gene pool with generalised linkage disequilibrium that is not favourable for global association genetic studies

Previous molecular studies (Herrero et al. 1996; Federici et al. 1998; Nicolosi et al. 2000; Luro et al. 2001; Ollitrault et al. 2003; Barkley et al. 2006; Liang et al. 2007) have provided evidence of a strong diversification between the ancestral taxa of all cultivated forms. Therefore, the analysis of the organisation of cultivated citrus and the study of the LD organisation of the genome were necessary to estimate how association studies should be conducted in *Citrus*.

Our analysis of F_{stat} parameters in the subset of the three basic taxa genotypes (*C. reticulata*, *C. medica* and *C. maxima*) with non-significant F_{is} value but high F_{it} and F_{st} values confirms the important structure of the allelic diversity between these taxa. The interspecific differentiation was particularly high using InDel markers. Eleven of 50 SSR markers and 7 of 12 InDel markers displayed significant deficits of heterozygous genotypes in the whole sample. This indicates a strong population subdivision (Hartl and Clark 1997) and, therefore, a low gene flow between *C. medica*, *C. reticulata* and *C. maxima*. The differentiation between these sexually compatible taxa can be explained by the foundation effect in three geographic zones and by an initial allopatric evolution. *C. maxima* originated in the Malay Archipelago and Indonesia, *C. medica* evolved in Northeastern India and the nearby region of Burma and China and *C. reticulata* diversification

occurred over a region including Vietnam, Southern China and Japan (Webber et al. 1967; Scora 1975). Later on, human activity facilitated migration and hybridization among the differentiated gene pools of the basic taxa. However, the partial apomixis observed in most of the secondary species has strongly limited the interspecific gene flow.

Using 50 mapped SSR markers, we found that the LD decay was very slow as the distance increased in a same linkage group. Moreover, a similar distribution of LD was found when considering LD within or between linkage groups (65.69 and 53.68% of the D' values >0.5, respectively). 99.3% of significant p values (<0.05) were observed both within and between linkage groups. This LD structure confirms that the history of cultivated *Citrus* (initial allopatric differentiation of basic taxa followed by a limited number of interspecific meiosis) is not a favourable situation for association genetic studies. Indeed, significant LD between polymorphisms on different chromosomes may produce associations between a marker and a phenotype, even though the marker is not physically linked to the locus responsible for the phenotypic variation. Similar population structures exist in many crops where the complex breeding history and limited gene flow found in most wild plants have created complex stratification (Flint-Garcia et al. 2003; Abdurakhmonov and Abdukarimov 2008). LD between unlinked loci primarily happens due to the occurrence of distinct allele frequencies with different ancestry in an admixed or structured population when predominant parents exist in germplasm groups. This was the case in our sample representative of the cultivated *Citrus* genus. Statistical methodologies have been developed to properly interpret the results of association tests when using such structured populations (Pritchard et al. 2000; Reich and Goldstein 2001; Price et al. 2006; Yu et al. 2006). However, to be applied properly, these methods require that a significant part of the structured population results from recombination between the ancestral genomes with sufficient meiosis events to reduce the initial extent of LD, whereas the actual cultivated citrus germplasm arises from a limited number of such inter-ancestry meiosis. This result precludes LD-based association study at the genus level without developing additional interspecific hybrids, such as BC1 or F2, between ancestral taxa or hybrids of the secondary species. In addition, the potential use of genetic association studies within basic species should be explored, particularly in *C. reticulata* where useful polymorphisms (resistance to biotic and abiotic constraints and some quality factors) have been identified. Moreover, markers with a higher rate of identity-by-descent, such as InDels or SNPs, should be more useful than SSRs for genetic association studies.

Conclusions

This work achieves for the first time in citrus, the development of InDel markers as an important tool for diversity and phylogenetic studies in citrus. InDel markers appear to be better phylogenetic markers for tracing the contributions of the three ancestral species to the secondary species and modern cultivars, whereas SSR markers are more useful for intraspecific diversity analysis. Most of the genetic organisation of the *Citrus* gene pool is related to the differentiation between *C. reticulata*, *C. maxima* and *C. medica*. High and generalised LD was observed, probably due to the initial differentiation between the basic species and a limited number of interspecific meiosis. This structure precludes association genetic studies at the genus level without developing additional recombinant populations from interspecific hybrids. Association genetic studies should also be affordable at intraspecific level in a less structured pool such as *C. reticulata*.

Acknowledgments This research was jointly financed by the grant AGL2008-00596-Ministry of Science and Innovation of Spain-Fondo Europeo de Desarrollo Regional (FEDER) and the grant Prometeo/2008/121 from the Generalitat Valenciana, Spain.

References

Abdurakhmonov IY, Abdukarimov A (2008) Application of Association Mapping to Understanding the Genetic Diversity of Plant Germplasm Resources. Int J Plant Genomics 2008:574927. doi:10.1155/2008/574927

Aleza P, Froelicher Y, Schwarz S, Agusti M, Hernandez M, Juarez J, Luro F, Morillon R, Navarro L, Ollitrault P (2011) Tetraploidization events by chromosome doubling of nucellar cells are frequent in apomictic citrus and are dependent on genotype and environment. Ann Bot. doi:10.1093/aob/mcr099

Bachtrog D, Agis M, Imhof M, Schlotterer C (2000) Microsatellite variability differs between dinucleotide repeat motifs-evidence from *Drosophila melanogaster*. Mol Biol Evol 17:1277–1285

Barkley NA, Roose ML, Krueger RR, Federici CT (2006) Assessing genetic diversity and population structure in a citrus germplasm collection utilizing simple sequence repeat markers (SSRs). Theor App Genet 112:1519–1531

Barkley NA, Krueger RR, Federici CT, Roose ML (2009) What phylogeny and gene genealogy analyses reveal about homoplasy in citrus microsatellite alleles. Plant Syst Evol 282:71–86

Barrett HC, Rhodes AM (1976) A numerical taxonomic study of affinity relationships in cultivated Citrus and its close relatives. Syst Bot 1:105–136

Bretó MP, Ruiz C, Pina JA, Asins MJ (2001) The diversification of *Citrus clementina* Hort. ex Tan., a vegetatively propagated crop species. Mol Phylogenet Evol 21:285–293

Britten RJ, Rowen L, Williams J, Cameron RA (2003) Majority of divergence between closely related DNA samples is due to indels. PNAS 100:661–4665

Calabrese F (1992) The history of citrus in the Mediterranean countries and Europe. Proc Int Soc Citricult 1:35–38

Casa AM, Pressoir G, Brown PJ, Mitchell SE, Rooney WL, Tuinstra MR, Francks CD, Kresovich S (2008) Community resources and strategies for association mapping in sorghum. Crop Sci 48:30–40

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:19

Coletta Filho HD, Machado MA, Targon MLPN, Moreira MCPQDG, Pompeu J Jr (1998) Analysis of the genetic diversity among mandarins (*Citrus* spp.) using RAPD markers. Euphytica 102: 133–139

Cuenca J, Froelicher Y, Aleza P, Juárez J, Navarro L, Ollitrault P (2011) Multilocus half-tetrad analysis and centromere mapping in *Citrus*: evidence of SDR mechanism for 2n megagametophyte production and partial chromosome interference in mandarin cv. 'Fortune'. Heredity 107:462–470

De Moraes A, dos Santos Soares Filho W, Guerra M (2007) Karyotype diversity and the origin of grapefruit. Chromosome Res 1:115–121

Deng ZN, Gentile A, Nicolosi E, Continella G, Tribulato E (1996) Parentage determination of some *citrus* hybrids by molecular markers. Proc Int Soc Citricul 2:849–854

Deu M, Glaszmann JC (2004) Linkage disequilibrium in sorghum. In: Plant & animal genomes XII conference, 10–14 January, Town & Country Convention Center, San Diego, CA, W10

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Fanciullino AL, Dhuique-Mayer C, Luro F, Casanova J, Morillon R, Ollitrault P (2006) Carotenoid diversity in cultivated *Citrus* is highly influenced by genetic factors. J Agric Food Chem 54:4397–4406

Fanciullino AL, Dhuique-Mayer C, Luro F, Morillon R, Ollitrault P (2007) Carotenoid biosynthetic pathway in the *Citrus* genus: number of copies and phylogenetic diversity of seven genes. J Agric Food Chem 55:7405–7417

Federici CT, Fang DQ, Scora RW, Roose ML (1998) Phylogenetic relationships within the genus *Citrus* (Rutaceae) and related genera as revealed by RFLP and RAPD analysis. Theor Appl Genet 96:812–822

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

Forment J et al (2005) Development of a *citrus* genome-wide EST collection and cDNA microarray as resources for genomic studies. Plant Mol Biol 57:375–391

Froelicher Y, Dambier D, Costantino G, Lotfy S, Didout C, Beaumont V, Brottier P, Risterucci AM, Luro F, Ollitrault P (2008) Characterization of microsatellite markers in *Citrus reticulata* Blanco. Mol Ecol Resour 8:119–122

Froelicher Y, Mouhaya W, Bassene JB, Costantino G, Kamiri M, Luro F, Morillon R, Ollitrault P (2011) New universal mitochondrial PCR markers reveal new information on maternal *citrus* phylogeny. Tree Genet Genomes 7:49–61

Garris AJ, McCouch SR, Kresovich S (2003) Population structure and its effects on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice *Oryza sativa* L. Genetics 165:759–769

Green RM, Vardi A, Galun E (1986) The plastome of *Citrus*. Physical map, variation among *Citrus* cultivars and species and comparison with related genera. Theor Appl Genet 72:170–177

Gulsen O, Roose ML (2001) Chloroplast and nuclear genome analysis of the parental lemons. J Am Soc Hort Sci 126:210–215

Gupta P, Rustgi S, Kulwal P (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol Biol 57:461–485

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95–98

Hartl DL, Clark AG (1997) Principles of population genetics. Sinauer Associates Incorporated, Sunderland

Hayashi K, Yoshida H, Ashikawa I (2006) Development of PCR-based allele-specific and InDel marker sets for nine rice blast resistance genes. Theor Appl Genet 113:251–260

Herrero R, Asins MJ, Carbonell AE, Navarro L (1996) Genetic diversity in the orange subfamily Aurantioideae. I. Intraspecies and intragenus genetic variability. Theor Appl Genet 92:599–609

Ingvarsson P (2005) Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European Aspen (*Populus tremula* L., Salicaceae). Genetics 169:945–953

Jarne P, Lagoda PJL (1996) Microsatellites, from molecules to populations and back. Trends Ecol Evol 11:424–429

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11:94

Kamiri M, Stift M, Srairi I, Costantino G, El Moussadik A, Hmyene A, Bakry F, Ollitrault P, Froelicher Y (2011) Evidence for non-disomic inheritance in a *Citrus* interspecific tetraploid somatic between *C. reticulata* and *C. lemon* hybrid using SSR markers and cytogenetic analysis. Plant Cell Rep 30:1415–1425

Kijas JMH, Thomas MR, Fowler JCS, Roose ML (1997) Integration of trinucleotide microsatellites into a linkage map of *citrus*. Theor Appl Genet 94:701–706

Krueger RR, Navarro L (2007) *Citrus* germplasm resources. In: Khan I (ed) Citrus genetics, breeding and biotechnology. CABI Publishing, CAB International, Wallington, pp 45–140

Li WB, Liu GF, He SW (1992) Leaf isozymes of mandarin. Proc Int Soc Citricult 1:217–220

Liang G, Xiong G, Guo Q, He Q, Li X (2007) AFLP analysis and the taxonomy of *Citrus*. Acta Hortic 760:137–142

Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21: 2128–2129

Luro F, Rist D, Ollitrault P (2001) Evaluation of genetic relationships in *Citrus* genus by means of sequence tagged microsatellites. Acta Hortic 546:237–242

Luro F, Costantino G, Terol J, Argout X, Allario T, Wincker P et al (2008) Transferability of the EST-SSRs developed on Nules clementine (*Citrus clementina* Hort ex Tan) to other *Citrus* species and their effectiveness for genetic mapping. BMC Genomics 9:287

Mabberley DJ (1997) A classification for edible *Citrus* (Rutaceae). Telopea 7:167–172

Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16:1182–1190

Morales M, Roig E, Monforte AJ, Arús P, Garcia-Mas J (2004) Single-nucleotide polymorphisms detected in expressed sequence tags of melon (*Cucumis melo* L.). Genome 47:352–360

Nicolosi E (2007) Origin and taxonomy. In: Khan I (ed) *Citrus* genetics, breeding and biotechnology. CABI Publishing, CAB International, Wallington, pp 19–43

Nicolosi E, Deng ZN, Gentile A, La Malfa S, Continella G, Tribulato E (2000) Citrus phylogeny and genetic origin of important species as investigated by molecular markers. Theor Appl Genet 100:1155–1166

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, Stahl EA, Weigel D (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 30:190–193

Ollitrault P, Jacquemond C, Dubois C, Luro F (2003) Genetic diversity of cultivated tropical plants. In: Hamon P, Seguin M, Perrier X, Glaszmann J-C (eds) Montpellier. CIRAD, pp 193–217

Ollitrault F, Terol J, Pina JA, Navarro L, Talon M, Ollitrault P (2010) Development of SSR markers from *Citrus clementina* (Rutaceae) BAC end sequences and interspecific transferability in Citrus. Am J Bot 97:124–129

Ollitrault P, Terol J, Chen C, Federici CT, Lofty S, Hippolyte I, Ollitrault F, Berard A, Chauveau A, Constantino G, Kacar Y, Mu L, Cuenca J, Garcia-Lor A, Froelicher Y, Aleza P, Boland A, Billot C, Navarro L, Luro F, Roose ML, Gmitter FG, Talon M, Brunel D (2011) A reference linkage map of *C. clementina* based on SNPs, SSRs and Indels. In: Plant & Animal genomes XIX conference, San Diego, CA, USA, p 477

Park S, Yu HJ, Mun JH, Lee SC (2010) Genome-wide discovery of DNA polymorphism in *Brassica rapa*. Mol Genet Genomics 283:135–145

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol Ecol Notes 6:288–295

Perrier X, Jacquemoud-Collet JP (2006) DARwin software. http://darwin.cirad.fr/darwin

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Rafalski A, Morgante M (2004) Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet 20:103–111

Raman H, Raman R, Wood R, Martin P (2006) Repetitive indel markers within the ALMT1 gene conditioning aluminium tolerance in wheat (*Triticum aestivum* L.). Mol Breed 18:171–183

Reich DE, Goldstein DB (2001) Detecting association in a case–control study while correcting for population stratification. Genet Epidemiol 20:4–16

Remington DL, Thornsberry JM, Matsouka Y, Wilson LM, Whitt SR (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci USA 98:11479–11484

Roose ML, Federici CT, Mu L, Kwok K, Vu C (2009) Map-based ancestry of sweet orange and other Citrus variety groups. In: Second international citrus biotechnology symposium, Catania, Italy, p 28

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Scora RW (1975) On the history and origin of Citrus. Bull Torr Bot Club 102:369–375

Scora RW, Kumamoto J, Soost RK, Nauer EM (1982) Contribution to the origin of the grapefruit *Citrus paradisi* (Rutaceae). Syst Bot 7:170–177

Swingle WT (1943) The botany of Citrus and its wild relatives in the orange subfamily. In: Webber HJ, Batchelor DL (eds) The citrus industry, vol 1, pp 128–474

Swingle WT, Reece PC (1967) The botany of Citrus and its wild relatives. In: Reuther W, Webber HJ, Batchelor LD (eds) The citrus industry, 2nd edn. University of California, Berkeley vol 1, pp 190–430

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

Terol J, Conesa A, Colmenero JM, Cercos M, Tadeo FR, Agustí J, Alós E, Andres F, Soler G, Brumos J, Iglesias DJ, Götz S, Legaz F, Argout X, Courtois B, Ollitrault P, Dossat C, Wincker P, Morillon R, Talon M (2007) Analyses of 13000 unique Citrus clusters associated with fruit quality, production and salinity tolerance. BMC Genomics 8:31

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

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet 28:286–289

Torres AM, Soost RK, Diedenhofen U (1978) Leaf isozymes as genetic markers in citrus. Am J Bot 65:869–881

Uzun A, Yesilogu T, Aka-kacar Y, Tuzcu O, Gulsen O (2009) Genetic diversity and relationships within Citrus and related genera based on sequence related amplified polymorphism markers. Sci Hortic 121:306–312

Väli Ü, Brandström M, Johansson M, Ellegren H (2008) Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. BMC Genet 9:8

Vasemägi A, Gross R, Palm D, Paaver T, Primmer CR (2010) Discovery and application of insertion-deletion (INDEL) polymorphisms for QTL mapping of early life-history traits in *Atlantic salmon*. BMC Genomics 11:156

Watkins WS, Ricker CE, Bamshad MJ, Carroll ML, Nguyen SV, Batzer MA, Harpending HC, Rogers AR, Jorde LB (2001) Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms. Am J Hum Genet 68:738–752

Webber HJ, Reuther W, Lawton HW (1967) History and development of the citrus industry. In: Reuther W, Webber HJ, Batchelor LD (eds) The citrus industry, vol 1. University of California Press, Berkeley, pp 1–39

Weber JL, Wrong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the analysis of population structure. Evolution 38:1358–1370

Wright S (1969) Evolution and the genetics of populations. The theory of gene frequencies, vol 2. The University of Chicago Press, Chicago

Wright S (1978) Evolution and the genetics of population, variability within and among natural populations. The University of Chicago Press, Chicago

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. Plant Genome 1:5–20