Introduction

Simple sequence repeats (SSRs; see glossary, Table 1) can conveniently be used as genetic markers owing to their polymorphism and relative ease of interpretation [1]. SSRs are widely used for, e.g., reconstructing phylogenetic relationships [2], for analysing spatial genetic structure among and within populations [1], and for detecting and explaining patterns linked to habitat fragmentation and gene flow [3]. In many studies, SSR markers demonstrated recent expansions or bottlenecks in various plant [4,5], animal [6–8] and human populations [9,10].

SSRs are often expected to mutate following the stepwise mutation model (SMM) [11], whereby mutations alter the length of the repeat either by adding or by deleting a single repeat unit at a fixed rate [12]. More elaborate mutational processes, such as the two-phase model (TPM; [13]) and the generalized stepwise model (GSM; [14,15]), allow for multi-step mutations. Departures from these models can however occur [16–20]. Moreover, SSR alleles are generally scored as the length (in base pairs) of PCR amplicons (see glossary, Table 1) containing the SSR, with differences in amplicon size taken to represent differences in repeat number in the SSR. Amplicon size includes the length of the flanking region (FR; see glossary, Table 1). These chunks of sequence can be and actually often are polymorphic and may contain both single nucleotide polymorphisms (SNPs) and insertions/deletions (indels; see glossary, Table 1) [18,21–24]. Indels which have been demonstrated to occur in the FR sequences of some SSR alleles clearly contribute to total amplicon size, but they are interpreted under the SMM as (false) variation in SSR length. SNPs contribute to the total sequence variance of SSR loci (see glossary, Table 1) that goes largely unnoticed when only amplicon size is recorded (which is typical of most SSR studies). As a consequence, SSR amplicon size data are prone to a particular form of size homoplasy [15], i.e. equally sized alleles may have different sequences and therefore be evolutionarily different. This casts...
doubt on the evolutionary and population-genetic inferences that can be drawn from estimates of divergence obtained from SSR amplicon size data. It is therefore necessary to examine the kinds of evolutionary information conveyed in SSR alleles, to assess their consequences on fragment size-based evolutionary inferences, and to find ways to take advantage of this so far unexploited source of variability.

Figure 1 succinctly shows how information in FR sequence variation may change one’s view of SSR marker evolutionary information content. Figure 1A shows a “plain” SSR locus, where a mutation (bar on the right branch) changes the number of repeats from n to n+1. The four observed alleles are then linked by a simple genealogy where alleles with the same number of repeats are evolutionarily closer. This is the classical view of SSR allele evolution under repeat-number mutational models. Figure 1B shows four alleles with the same numbers of repeats as in Figure 1A, but in this case information from a SNP in the FR (indicated by the “. . .A/...C” symbols) is added. Here, alleles with different numbers of repeats share the same SNP. Given that SNP mutation rates are generally lower than those of SSRs, the most-likely genealogy is one in which divergence occurs first at the SNP (mutation in the upper right branch) and then in SSRs within flanking-region haplotypes (mutations in the lower branches). In this case, three mutations have occurred, the genealogy is expected to be older than in 1A, and the alleles are grouped based on their FR sequences and not based on the number of repeats. A simulation study [25] revealed that SSR and SNP sources of variation provide independent evolutionary information, so it is reasonable to study the role of SNP variation in SSR flanking regions. We therefore asked the following questions: How much does amplicon size variation reflect variation in SSR repeat number? How important is the contribution of the various sources of sequence variation in total SSR diversity? Do these extra-SSR polymorphisms carry a different population and phylogenetic differentiation signal than the SSRs themselves? If so, how do the different signals combine? How does one best interpret SSR data, once these departures from the SMM have been acknowledged?

Amplicon

<table>
<thead>
<tr>
<th>Glossary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table 1.</strong> Glossary.</td>
</tr>
<tr>
<td><strong>Amplicon</strong></td>
</tr>
<tr>
<td><strong>SSR</strong></td>
</tr>
<tr>
<td><strong>FR</strong></td>
</tr>
<tr>
<td><strong>Indel</strong></td>
</tr>
<tr>
<td><strong>SNP</strong></td>
</tr>
<tr>
<td><strong>Indel</strong></td>
</tr>
<tr>
<td><strong>SSR locus</strong></td>
</tr>
</tbody>
</table>

Materials and Methods

Choice of Data Sets

Three data sets were utilized in this study. The structures of the three data sets analyzed here varied in the following ways: the *Citrus* data represent sequences of individual trees each belonging to a different species, variety or cultivar; the *Jacaranda* data were collected in four different sampling sites considered as four distinct populations; and the oak data came from a single mixed stand of sessile/pedunculate oaks (*Quercus robur, Q. petraea*). The *Citrus*, *Jacaranda*, and *Quercus* data sets were submitted to partially different types of analyses according to the structure of each data set.

Given the taxonomy and prevailing theory on many *Citrus* species being derived by natural hybridization from ancestral species, alleles were chosen to minimize known hybrids from the data set of Barkley et al. [26,27], who examined the genetic diversity of 370 *Citrus* accessions. The individual identity of each accession (as provided by the University of California, Riverside citrus database, http://www.citrusvariety.ucr.edu/) is given in the Supplementary Table 1 of Barkley et al. [26]. The *Jacaranda* data were obtained from three sites in French Guiana (Gounami, Paracou and Saint-Laurent) and one in Brazil (Tapajos) (Table S1). The *Quercus* alleles were taken from genetic analyses of *Q. petraea* and *Q. robur* in a mixed oak stand in Switzerland [28,29].
The SSR markers used for this study were: cAGG9, CCT01 and GT03 for Citrus [26,30]; Jc3H10, Jc3F4 and Jc3H10 for Jacaranda; QrZAG30 for Quercus [31]. Allele sequencing for the Citrus and Quercus data sets was as reported in Barkley et al. [26] and Gugerli et al. [29].

Molecular Methods

Jacaranda markers were obtained following the method developed for SSR markers in tropical crops by Billotte et al. [32]. Ten nanograms of DNA was digested with RsaI, followed by an enrichment step in (GT)\textsubscript{n} and (GA)\textsubscript{n} repeats by hybridization of cleaved fragments with bioin-labelled (CA)\textsubscript{n} and (CT)\textsubscript{n} oligoprobes and capture of the selected sequences with streptavidin-coated magnetic beads. Those fragments were ligated into a pGEM-TEasy vector (Promega, Madison, WI), and plasmids were then used to transform competent XL-1 Blue strain Escherichia coli (Stratagene, La Jolla, CA). Positive colonies were tested using colony polymerase chain reaction (PCR) to check the presence of inserts. Amplified fragments were then transferred from an agarose gel onto a Hybond membrane for SSR presence screening by hybridization with labeled (GT)\textsubscript{15} and (GA)\textsubscript{15} oligoprobes. One hundred and thirteen bacterial clones containing plasmids with inserts that gave strong hybridization signals, with sizes ranging from 500 to 1000bp, were selected and cultured. Plasmids were sequenced with the ABI BigDye\textsuperscript{TM} Terminator V3.1 kit (Applied Biosystems, Foster City, CA) following the manufacturer’s protocol. Cloning products were diluted 1:40, purified with ExoSap-IT (USB Corporation, Cleveland, OH), and separated on an ABI 3130XL capillary sequencer (Applied Biosystems, Foster City, CA). All sequences were edited and aligned using CodonCode Aligner V1.6.3 (CodonCode Corporation, Dedham, MA).

Data Analyses

For each data set (Citrus, Jacaranda, and Quercus), each allele was characterised by the following information: (a) amplicon size variation, (b) SSR variation, (c) FR sequence variation and (d) amplicon sequence variation. Each of these portions of information conveyed by SSR marker data was analysed as a separate source of variation. Results were compared across data sets to assess their evolutionary information content. Indels were coded as SNP mutations to represent them as single mutational events.

Levels of polymorphism (number of alleles \(A\) or haplotypes \(H\), Nei’s genetic diversity \(H_e\), number of SNPs and indels) were recorded for each source of variation. Linkage disequilibrium (LD) was computed for FR sequence variation.

Data were analysed in two ways: (a) for all data sets, at the level of individual alleles; (b) for Jacaranda only, at the population level.

(a) At the individual-allele level, matrices of pairwise genetic distances were computed between individuals and between alleles based on each source of variation. For amplicon length, distance is represented by difference in length; for SSR length, distance is the difference in number of repeat units; for whole amplicon sequence, distance is the total number of differences, including SNPs, indels and number of repeat units; for flanking region sequences, distance is the total number of SNPs and indels. The correlation between matrices was tested by a Mantel test. LD between all pairs of polymorphic sites in FR sequences was tested with a Markov chain Monte Carlo (MCMC) procedure of 10,000 steps and a burn-in phase of 1,000 steps (default values). Pairwise genetic distances for amplicon and FR sequence variation were computed based on the number of mutational steps between sequences. Pairwise genetic distances for SSR

| Table 2. Characteristics of simple sequence repeat (SSR) data sets of the three taxa, Citrus (C), Jacaranda (J) and Quercus (Q). |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Data set       | Type of data    | SSR locus       | Size (bp) \*    | \(N\) \*         | Reference       |
| Set C          | Collection of provenances | cAGG9 | 123 | 34 | [30] |
|                |                 | CCT01 | 167 | 34 | [27] |
|                |                 | GT03  | 177 | 36 | [27] |
| Set J          | Four populations | Jc3A10 | 180 | 100 | This study |
|                |                 | Jc3F4 | 144 | 101 | This study |
|                |                 | Jc3H10 | 215 | 96 | This study |
| Set Q          | One mixed population | QrZAG30 | 255 | 47 | [31] |

\*Total length of the amplicon consensus sequence in base pairs.

For each data set, the number of analysed alleles was as reported in Barkley et al. [26] and Gugerli et al. [29].

Phylogenetic Information in SSR Alleles

The SSR markers used for this study were: cAGG9, CCT01 and GT03 for Citrus [26,30]; Jc3H10, Jc3F4 and Jc3H10 for Jacaranda; QrZAG30 for Quercus [31]. Allele sequences for the Citrus and Quercus data sets was as reported in Barkley et al. [26] and Gugerli et al. [29].

Molecular Methods

Jacaranda markers were obtained following the method developed for SSR markers in tropical crops by Billotte et al. [32]. Ten nanograms of DNA was digested with RsaI, followed by an enrichment step in (GT)\textsubscript{n} and (GA)\textsubscript{n} repeats by hybridization of cleaved fragments with bioin-labelled (CA)\textsubscript{n} and (CT)\textsubscript{n} oligoprobes and capture of the selected sequences with streptavidin-coated magnetic beads. Those fragments were ligated into a pGEM-TEasy vector (Promega, Madison, WI), and plasmids were then used to transform competent XL-1 Blue strain Escherichia coli (Stratagene, La Jolla, CA). Positive colonies were tested using colony polymerase chain reaction (PCR) to check the presence of inserts. Amplified fragments were then transferred from an agarose gel onto a Hybond membrane for SSR presence screening by hybridization with labeled \(\gamma^{32}P\) (GT)\textsubscript{15} and (GA)\textsubscript{15} oligoprobes. One hundred and thirteen bacterial clones containing plasmids with inserts that gave strong hybridization signals, with sizes ranging from 500 to 1000bp, were selected and cultured. Plasmids were sequenced with the ABI BigDye\textsuperscript{TM} Terminator V3.1 kit (Applied Biosystems, Foster City, CA) following the manufacturer’s protocol. Cloning products were diluted 1:40, purified with ExoSap-IT (USB Corporation, Cleveland, OH), and separated on an ABI 3130XL capillary sequencer (Applied Biosystems, Foster City, CA). All sequences were edited and aligned using CodonCode Aligner V1.6.3 (CodonCode Corporation, Dedham, MA).

Data Analyses

For each data set (Citrus, Jacaranda, and Quercus), each allele was characterised by the following information: (a) amplicon size variation, (b) SSR variation, (c) FR sequence variation and (d) amplicon sequence variation. Each of these portions of information conveyed by SSR marker data was analysed as a separate source of variation. Results were compared across data sets to assess their evolutionary information content. Indels were coded as SNP mutations to represent them as single mutational events.

Levels of polymorphism (number of alleles \(A\) or haplotypes \(H\), Nei’s genetic diversity \(H_e\), number of SNPs and indels) were recorded for each source of variation. Linkage disequilibrium (LD) was computed for FR sequence variation.

Data were analysed in two ways: (a) for all data sets, at the level of individual alleles; (b) for Jacaranda only, at the population level.

(a) At the individual-allele level, matrices of pairwise genetic distances were computed between individuals and between alleles based on each source of variation. For amplicon length, distance is represented by difference in length; for SSR length, distance is the difference in number of repeat units; for whole amplicon sequence, distance is the total number of differences, including SNPs, indels and number of repeat units; for flanking region sequences, distance is the total number of SNPs and indels. The correlation between matrices was tested by a Mantel test. LD between all pairs of polymorphic sites in FR sequences was tested with a Markov chain Monte Carlo (MCMC) procedure of 10,000 steps and a burn-in phase of 1,000 steps (default values). Pairwise genetic distances for amplicon and FR sequence variation were computed based on the number of mutational steps between sequences. Pairwise genetic distances for SSR
variation and amplicon size variation were computed as Euclidian distances (genetic distances based on the SSR variation were not computed for *Citrus* loci CCT01 and GT03, which displayed extremely complex repeat sequences; see Figure S1). Most of the above calculations were performed with ARLEQUIN versions 3.11 and 3.5.1.2 [34]. Genetic distance matrices for SSR variation and amplicon size variation data were computed with an ad-hoc routine in R [35], and correlations between genetic distance matrices were computed with the mantel.test function of the R package “ape”.

(b) At the population level, for *Jacaranda*, global and pairwise FST values [36] were computed for all sources of variation; RST values [37] were computed on sources of variability comprising amplicon size variation and SSR variation; NST values [38] were computed on amplicon and FR sequence variations. In NST calculations on amplicon sequence variation, mutations in repeat number were weighted less than substitutions (weight = 0.05) to account for their faster mutation rate. Weight choice was based on the observation that SSRs have, in our *Jacaranda* sample, approximately ten times more alleles than each SNP (see Results section). Considering that rare alleles might be missing from our sample, we estimated a ratio of 1:20 in allele richness between SNPs and SSRs and used this ratio to establish the relative weight. A hierarchical analysis of molecular variance (AMOVA) was performed to partition genetic variance into within- and among-population components. The significance of variance components was tested by 1,000 permutations [39]. These analyses were performed with the R package “ape”.

### Results

#### Distribution of Polymorphism

In order to evaluate homoplasy and the variation in phylogenetic signal from different regions of SSR alleles (repeat region and FR), sequence data produced from three divergent tree genera were evaluated. Forty-six, 101, and 47 samples constituted the *Citrus*, *Jacaranda*, and *Quercus* data sets, respectively. Three, three and one SSRs were genotyped and sequenced for the *Citrus*, *Jacaranda* and *Quercus* data sets, respectively. The lengths of the sequenced fragments varied between 123 and 255 bp (Table 2).

Whenever possible, multiple alleles of the same fragment size were genotyped and sequenced to estimate the degree of sequence divergence among equally sized alleles (“size homoplasy”). This includes both variation in SSR repeat number (that is taken into account by the SMM) and variation in FR sequences, which is the focus of our study.

#### Levels of Polymorphism

- **Citrus**:
  - SSR1: He = 0.772, A = 4, h = 11 (GAG) < sub>g</sub>.
  - SSR2: He = 0.758, A = 4, h = 20 (TCC) < sub>n</sub>.
  - SSR3: He = 0.783, A = 5, h = 12 (GT) < sub>n</sub>.

- **Jacaranda**:
  - Jc3A10: He = 0.865, A = 13, h = 43 (CT) < sub>n</sub>.
  - Jc3F4: He = 0.912, A = 13, h = 32 (GA) < sub>n</sub>.
  - Jc3H10: He = 0.856, A = 13, h = 40 (CT) < sub>n</sub>.

- **Quercus**:
  - QrZAG30: He = 0.966, A = 23, h = 31 (GAC) < sub>n</sub>.

#### Table 3. Polymorphism at each simple sequence repeat (SSR) locus differentiated by segments of the DNA amplicon. SSR1, SSR2 and SSR3: respectively first, second and third SSR occurring in each amplicon (see figure S1 for details on each amplicon’s sequence); FR: flanking regions; He: Nei’s genetic diversity; A: number of alleles; h: number of haplotypes. Data sets: C, *Citrus*; J, *Jacaranda*; Q, *Quercus*.

<table>
<thead>
<tr>
<th>SSR locus (data set)</th>
<th>Amplicon size variation</th>
<th>Amplicon sequence variation</th>
<th>SSR1 variation</th>
<th>SSR2 variation</th>
<th>SSR3 variation</th>
<th>FR sequence variation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>He</td>
<td>A</td>
<td>He</td>
<td>h</td>
<td>Repeat unit</td>
<td>He</td>
</tr>
<tr>
<td>cAGG9 (C)</td>
<td>0.772</td>
<td>4</td>
<td>0.888</td>
<td>11</td>
<td>(GAG) &lt; sub&gt;n&lt;/sub&gt;</td>
<td>0.772</td>
</tr>
<tr>
<td>CCT01 (C)</td>
<td>0.758</td>
<td>4</td>
<td>0.966</td>
<td>20</td>
<td>(TCC) &lt; sub&gt;n&lt;/sub&gt;</td>
<td>–</td>
</tr>
<tr>
<td>GT03 (C)</td>
<td>0.783</td>
<td>5</td>
<td>0.900</td>
<td>12</td>
<td>(GT) &lt; sub&gt;n&lt;/sub&gt;</td>
<td>–</td>
</tr>
<tr>
<td>Jc3A10 (J)</td>
<td>0.856</td>
<td>13</td>
<td>0.937</td>
<td>43</td>
<td>(CT) &lt; sub&gt;n&lt;/sub&gt;</td>
<td>0.769</td>
</tr>
<tr>
<td>Jc3F4 (J)</td>
<td>0.912</td>
<td>13</td>
<td>0.920</td>
<td>32</td>
<td>(GA) &lt; sub&gt;n&lt;/sub&gt;</td>
<td>0.853</td>
</tr>
<tr>
<td>Jc3H10 (J)</td>
<td>0.856</td>
<td>13</td>
<td>0.922</td>
<td>40</td>
<td>(CT) &lt; sub&gt;n&lt;/sub&gt;</td>
<td>0.748</td>
</tr>
<tr>
<td>QrZAG30 (Q)</td>
<td>0.966</td>
<td>23</td>
<td>0.983</td>
<td>31</td>
<td>(GA) &lt; sub&gt;n&lt;/sub&gt;</td>
<td>0.915</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0040699.t003
The highest number of alleles and Nei’s genetic diversity were detected for amplicon sequence variation, then for amplicon size variation and lowest was found for SSR variation (Table 3). FR sequence variation showed the lowest diversity. Indels were observed in six loci out of seven, while SNP density varied between 0.90 and 12.14 SNPs per 100 bp.

An example of the complexity of the polymorphisms observed is shown for markers Jc3H10 and QrZAG30 (Jacaranda and Quercus data set, Figure 2). Three indels of several nucleotides (between 12 and 15 bases) were detected in the FR sequences of the compound marker Jc3H10. One was upstream of the first SSR, the second was between the repeat motifs and the last was detected after the SSR repeats. In the oak marker QrZAG30, in addition to SNPs, a large 66-nucleotide indel was observed, with SNPs occurring among alleles carrying the DNA fragment involved in the indel. Several cases of size homoplasy were detected, such as: (i) indels in the FRs compensating differences in number of repeats, (ii) compound SSRs with the same number of repeats, but composed of different numbers of repeats in different motifs, (iii) SSRs with the same amplicon size and number of repeats but with SNPs in the FR sequences or interrupting the repeat (Figure S1). For the three data sets, 50 alleles with the same amplicon sizes out of 68 (74%) showed size homoplasy (Table S3).

**Linkage Disequilibrium**

LD was tested for all pairs of SNP and indel sites within the FR sequence variation of each marker (Figure 3). For the Citrus data set, only 19 pairs out of 217 (8.8%) showed significant LD. For the three SSR loci, LD was very irregular and there was no relationship between LD and sequence distance (Mantel tests, \( P > 0.05 \) for all loci). For the Jacaranda data set, ten pairs out of 27 (37%) showed significant LD. For marker Jc3F4, the SNP at position 81 was in disequilibrium with two indels and one SNP, all located upstream of the SSR. LD was strongest for Jc3H10, where all three pairs were significant. The relationship between LD and sequence distance was significant (Mantel test, \( P = 0.023 \)). Tests of correlation between LD and distance in bp could not be performed for Jc3A10 and Jc3H10 because only three nucleotide sites were polymorphic (three LD values). For the Quercus data set (marker QrZAG30), 146 pairs of SNPs and indels out of 325 (44.9%) showed significant LD. In particular, 13 or 14 contiguous SNPs and indels located downstream of the SSR at positions 84–107 formed a disequilibrium block, itself in weaker disequilibrium with upstream polymorphic sites at positions 23, 45 and 46 and another SNP downstream the SSR (position 60).

**Correlation between Genetic Distance Matrices**

We tested whether different sources of variation conveyed the same information on relatedness among alleles. To do so, genetic distances were computed for each pair of individuals with each source of variation (SSR variation, amplicon size, FR sequence, and amplicon sequence variation) within each marker, and the correlation between genetic distance matrices was tested by a Mantel test. Among the 36 pairs of matrices, 27 (75%) showed a significant correlation (Figure S2). Genetic distance matrices based on SSR variation and on FR sequence variation were the least correlated (one significant pair out of five). Amplicon sequence variation and amplicon size variation, as well as SSR variation and amplicon size variation, were correlated for all markers. The matrices computed on amplicon sequence variation were correlated to the remaining matrices in 95% of the cases (18 out of 19).
Conversely, FR sequence variation alone predicted other matrices in only 53% of the cases.

**Population Differentiation**

For the *Jacaranda* data set, we tested population genetic differentiation for each source of variation by computing $F_{ST}$, $R_{ST}$ and $N_{ST}$ as appropriate (see Methods; Figure 4). All $F_{ST}$ values were significant. FR sequence variation showed the largest value (0.42) and SSR variation the smallest (0.05). $R_{ST}$ values were very small (0.02) and non-significant when computed on both amplicon size variation and SSR variation. On the contrary, $N_{ST}$ values were large and significant, with values of 0.50 for FR sequence variation and 0.36 for amplicon sequence variation. In the phylogenetic trees obtained from each combination of source of variation/genetic distance measures, $F_{ST}$ and $N_{ST}$ provided the same topology for all cases they were applied to, indicating a more or less close link between pairs of populations Tapajos/Saint-Laurent and Coumani/Paracou (Figure 4). $R_{ST}$ grouped Paracou with Tapajos and Coumani with Saint-Laurent (for amplicon size variation) or separated the three Guiana shield populations from Tapajos (for SSR variation; Figure 4).

**Discussion**

**Distribution of Polymorphism**

SSR polymorphisms are widely used to infer population history and biogeographic patterns. These inferences rely on assumptions about SSR mutational models, and departures from such models are likely to bias divergence and diversity estimates. SSR alleles have been demonstrated to contain other sources of polymorphism in addition to variation in repeat number. In this study we have characterised the contribution of molecular variation occurring outside SSR repeats to SSR marker variability. Whole allele sequencing from SSR markers was analyzed from three tree genera (*Citrus*, *Jacaranda* and *Quercus*). A total of seven (dinucleotide and trinucleotide) SSR markers were targeted, including compound and imperfect markers. More genetic diversity was found for amplicon sequence variation than for amplicon size variation, suggesting that much of data polymorphism is neglected when SSRs are described through amplicon size alone. On average, 7.7 SNPs and 3.4 indels were detected in the flanking sequences of each marker. Genetic diversity ($H_{e}$) ranged between 0.306 and 0.922 for the FRs. LD among SNPs and indels in the FR was generally significant. These results demonstrate that SSR alleles are very often riddled with abundant non-SSR indels, plus vast...
amounts of SNPs both within and near the repeat itself. Similar genetic variation was previously observed at intra- or inter-specific levels [18,42–44] and does not seem to be exceptional in plants. SSR allele mutations are clearly not restricted to the hypervariable tandem repeat region, and polymorphisms outside SSR motifs are also present in animals [22,45,46] and humans [21]. Hidden indels can directly confound the estimation of repeat number from amplicon size ("allele size homoplasy"). While undetected SNPs do not directly influence repeat number estimates, they do cause haplotypes with different evolutionary histories to be treated as if they were identical. Both forms of homoplasy can have profound consequences on the way data should be interpreted.

Correlation of Polymorphism at the Individual-allele Level

Our results show that pieces of genetic information carried by the different portions of an amplicon are only partially correlated. In particular, blocks of linkage disequilibrium never cover the whole amplicon, and matrices of allele-level pairwise genetic distances built on each source of variation are not always well correlated. In other terms, the different sources of variation are not all equally capable to represent the "true" allele genealogy. Not surprisingly, whole-amplicon sequences were the most parsimonious predictors of relatedness described by the different parts of the amplicon sequence ("allele size homoplasy"). While undetected SNPs do not directly influence repeat number estimates, they do cause haplotypes with different evolutionary histories to be treated as if they were identical. Both forms of homoplasy can have profound consequences on the way data should be interpreted.

Figure 4. Phylogenetic trees of Jacaranda populations based on different components of simple sequence repeat (SSR) data. SSR variation (A), amplicon size variation (B), flanking region (FR) sequence variation (C) and amplicon sequence variation (D). Each type of data was analysed according to a pair of suitable genetic distance estimators: $F_{ST}$, suitable for loci following the infinite allele model (IAM); $R_{ST}$, for loci following the stepwise mutation model (SMM); and $N_{ST}$, for loci following the infinite site model (ISM) model. There are four geographic populations: Counami (C), Paracou (P) and Saint-Laurent (S) in French Guiana; Tapajos (T) in Brazil. Note that scales are not the same among trees.

doi:10.1371/journal.pone.0040699.g004

Correlation of Polymorphism at the Population Level

The consequences – and usefulness – of hidden sequence variation are shown when we compare how sources of variation perform in detecting population divergence. This analysis was performed in the Jacaranda data set comprising four populations. In this example, two alternative patterns were expected for population relatedness: (a) a strictly geographical clustering (the three Guiana shield populations cluster and are separated from the Amazonian population), and (b) a pattern derived from independent results on chloroplast DNA divergence (Counami and Paracou populations form a group, Tapajos and Saint-Laurent form another group (Caroline Scotti-Saintagne, INRA, UMR « Ecologie des forêts de Guyane », article under second round of revision for Journal of Biogeography). Our results support the latter hypothesis, as proven by the convergent topology of trees obtained with $F_{ST}$ and $N_{ST}$ independently of the source of variation. The purely geographical hypothesis, on the other hand, is only supported by SSR variation in combination with $R_{ST}$ (although with a non-significant global $R_{ST}$ value). $R_{ST}$ estimates have larger variance than $F_{ST}$ and $N_{ST}$, particularly when based on small numbers of loci, and $R_{ST}$ values obtained here may be unreliable. Nevertheless, the pattern of $R_{ST}$ differentiation closely follows geographical distance. This may indicate that repeat number variation has arisen locally from small numbers of founding alleles and reflects recent population divergence, while variation in flanking regions may follow larger phylogeographic patterns and reflect deeper divergence. If $R_{ST}$ estimated can be trusted, all sources of variation converge with independent, chloroplast-based
results, except SSR variation itself. Therefore, information derived from the “true” nature of SSR data (SSR variation) is actually not representative of the information conveyed by alleles as a whole. This finding strongly hints that other sources of variability, more correctly described by mutation models that apply to DNA sequences, actually provide the bulk of information carried by SSRs. In other words, phylogeographic information carried by amplicon length is not provided by the SSR (which amplicon length is intended to represent) but by variation in flanking regions. Note that this contrasts with analyses carried out at the individual level (see above), where information on genetic distance match when measured as amplicon size and as SSR size. This issue raises the question of the application of the SMM, or other step-based mutational models, to SSRs, as already discussed by Colson and Goldstein [47] and Ellegren [12]. At the same time, these results substantiate the hypothesis of Cornuet et al. [40], who postulated that not all SSRs would evolve under the same mutational model; additional mutational models are required to appropriately handle variation in SSR data produced from different genera. Moreover, variation in the FRs contributed significantly to the phylogenetic signal and sometimes represented the main source of differentiation among individuals and populations, as shown by Jacaranda SSR markers and by the Citrus SSR marker GT03 [27].

Conclusion

Given the complexity of amplicon sequences described here, one may wonder how to properly exploit SSR data. The weight of polymorphisms other than variation in repeat number cannot and should not be overlooked: identity in allele size does not necessarily indicate identity in sequence content or the number of repeats within same sized alleles. When considering individual alleles, amplicon size correlates well with SSR repeat number. Nevertheless, even if repeat number could be obtained directly, it may provide only weak and inconsistent signals of population genetic differentiation, as shown by the Jacaranda data set (but larger numbers of markers may reduce estimation variance and mitigate this problem). Taking sequence variation into account actually adds a significant piece of information to phylogeographic or phylogeographic reconstruction. Population structuring emerges more clearly on the basis of sequence data, and the combination of sequences and SSR variation provides higher resolution. Consequently, genetic distance measures based on the SMM or related models should be restricted to perfect SSRs with invariant FRs. On the other hand, measures such as FST for amplicon size data and NST (for amplicon sequences) seem to be the tool of choice for the analysis of the average, not-so-ideal SSR markers that a population geneticist meets in his or her everyday work.

Supporting Information

Figure S1 Complete alignments of DNA fragments for all SSR markers studied. Marker names are shown above each alignment. (PDF)

Figure S2 Z-statistics and summary of Mantel tests for the correlations of pairs of genetic distance matrices within species. Colouring (shading) indicates the level of significance of the test: green (light grey*), P<0.05; orange (grey*), 0.05<P<0.001; red (dark grey*), P<0.001. Identity of DNA fragment sections is reported: simple sequence repeat (SSR) variation; amplicon size variation, flanking region (FR) sequence variation, and amplicon sequence variation. Marker names are displayed above each panel and data sets refer to Citrus (C), Jacaranda (J) and Quercus (Q). Sequences with single nucleotide polymorphisms (SNPs) within the repeat were excluded from subsequent analyses because of the complexity to describe the mutation model of SSR repeats bracketing SNPs. (PDF)

Table S1 List and details of the sampled sites for the Jacaranda (J) data set. (DOC)

Table S2 Description, amplification conditions and polymorphism of the nuclear simple sequence repeats (SSRs) analysed in Jacaranda copaia. (DOC)

Table S3 List and details of alleles for the three data sets. Amplicon sizes, frequencies and number of associated haplotypes are indicated for each allele. (DOC)

Acknowledgments

We wish to thank Maristerra Lemes for providing a part of the samples, Raphael Leblois and Caroline Scotti-Saintagne for fruitful discussions and for help with preliminary data analyses, and Valérie Troispoux and Saint-Omer Cazal for technical support.

Author Contributions

Conceived and designed the experiments: IS. Performed the experiments: SB FG NB. Analyzed the data: SB IS. Contributed reagents/materials/analysis tools: LM CC. Wrote the paper: SB FG NB IS.

References

15. Estoup A, Jarne P, Cornuet JM (2002) Homoplasy and mutation model at subsequent analyses because of the complexity to describe the mutation model of SSR repeats bracketing SNPs. (PDF)