Comparative in silico analysis of EST-SSRs in angiosperm and gymnosperm tree genera

Sonali Sachin Ranade¹, Yao-Cheng Lin², Andrea Zuccolo³, Andrea Zuccolo³, Yves Van de Pee²,⁵ and María del Rosario García-Gil†

Abstract

Background: Simple Sequence Repeats (SSRs) derived from Expressed Sequence Tags (ESTs) belong to the expressed fraction of the genome and are important for gene regulation, recombination, DNA replication, cell cycle and mismatch repair. Here, we present a comparative analysis of the SSR motif distribution in the 5′UTR, ORF and 3′UTR fractions of ESTs across selected genera of woody trees representing gymnosperms (17 species from seven genera) and angiosperms (40 species from eight genera).

Results: Our analysis supports a modest contribution of EST-SSR length to genome size in gymnosperms, while EST-SSR density was not associated with genome size in neither angiosperms nor gymnosperms. Multiple factors seem to have contributed to the lower abundance of EST-SSRs in gymnosperms that has resulted in a non-linear relationship with genome size diversity. The AG/CT motif was found to be the most abundant in SSRs of both angiosperms and gymnosperms, with a relative increase in AT/AT in the latter. Our data also reveals a higher abundance of hexamers across the gymnosperm genera.

Conclusions: Our analysis provides the foundation for future comparative studies at the species level to unravel the evolutionary processes that control the SSR genesis and divergence between angiosperm and gymnosperm tree species.

Keywords: Angiosperms, Gymnosperms, Expressed sequence tags, Simple sequence repeats (SSR), Microsatellites

Background

Microsatellites, also called SSRs (simple sequence repeats) or STRs (short tandem repeats), are 1-6 bp tandem repeat motifs present in both the coding and non-coding fractions of eukaryotic and prokaryotic genomes [1-3]. SSRs are especially abundant in transcribed regions of the genome making them a valuable molecular marker for genetic studies in plants [4]. SSRs result from mutations due to DNA-polymerase slippage during replication and unequal recombination [5]. SSRs are widely used in plant genetic research because of their co-dominant inheritance, relative abundance, multi-allelic nature, high reproducibility and ease of detection [6].

Expressed sequence tags (ESTs) are segments of expressed genes generated by single-pass sequencing of cDNA libraries [7]. In contrast to the genomic SSRs, EST-SSRs represent functional markers located in the coding fractions of the genome and changes in EST-SSRs length can cause a phenotypic effect, irrespective of the mutation site, whether it occurs in 5′- or 3′- UnTranslated Regions (UTRs) or in the Open Reading Frames (ORFs) [8]. The significance of EST-SSRs as a molecular tool in population genetic studies has been known for long [9]. In woody trees, EST-SSRs have been applied in population studies and analysis of genetic diversity in Cycas [10], Picea [11,12], Prunus [13,14], Eucalyptus [15,16] and Populus [17]; in hybrid selection in e.g., Citrus [18]; and also in genetic mapping in Citrus [19], Quercus [20,21] and Pinus [22]. Furthermore, unlike the genomic SSRs, EST-SSRs are easily transferable across species [23], therefore allowing studying polymorphism and genetic diversity in related species [9]. However, EST-SSRs have some disadvantages over genomic SSRs as EST-SSRs are known to be less variable than the genomic SSRs [24] and the amplicon size can also differ from the predicted size due to the effect of presence of introns in the flanking fractions [25].
With the advent of genomics, the availability of ESTs in the public databases, such as NCBI’s dbEST, has increased exponentially allowing for the identification of large numbers of EST-SSRs. For example, characterisation and comparative analysis of EST microsatellites in woody trees have been carried out in Citrus [26-28], Betula [29], Fagus [30], Prunus [31], Quercus [20], Populus [17,32], Eucalyptus [33-35], Cryptomeria [36,37], Taxus [38-40], Ginkgo [41], Picea [5,12] and Pinus [5,42]. However, analysis of SSRs for each individual EST genomic fraction (i.e., 5′- and 3′-UTR, and ORF) has only been carried out in Quercus [20], Cryptomeria [37] and Pinus [43]. Unfortunately, most of the results in those three studies are presented for the entire EST, which can lead to inaccurate results. For example, in Cryptomeria dimers are the most common motif in the 3′-UTR fraction; moreover, when all three EST fractions are considered together, trimers are concluded to be the most frequent motif across the entire EST [37]. Furthermore, AT was shown to be the most frequent dimer motif as an overall result, whereas analysis of each EST fraction separately revealed AG as the most frequent dimer in the ORF fraction [37]. These results demonstrate that SSR characterization on the whole EST sequence as a unit will provide only partial information, which may be misleading and result in discrepancies across studies.

Other discrepancies in EST-SSRs motif abundance and distribution across different plant studies can be attributed to the parameter setup [25], annotation deficiency [44], and the selected EST-SSR analysis algorithm [20]. For example, higher abundance of EST-SSR dimers was reported in Pinus [45,46], whereas Yan et al. [47] reported trimers as the most abundant in the same genus. Thus, comparative EST-SSRs studies will be more reliable when the EST data sets are analysed by applying the same bioinformatics procedure. In this study, we performed a comparative analysis of SSRs in each genomic fraction of EST separately (5′ UTR, ORF and 3′ UTR), across selected angiosperm and gymnosperm genera with a focus on woody trees. The aim was to present highly comparable data on SSR-EST abundance, composition and distribution; for genomes that diverged ~350 Myr [48].

Results
Table 1 shows values for EST-SSRs length and EST-SSR counts per genus across the 5′ UTR, ORF and 3′ UTR fractions (see also Additional file 1: Table S1).

EST-SSR length and complexity
There were no significant differences observed regarding EST-SSRs length between the three genomic fractions within and between taxa. In angiosperms, there was no significant association between genome size and EST-SSRs length for any of the EST fractions. In gymnosperms, however, there was a positive and significant association ($r = 0.6; P$-value < 0.03) between genome size and EST-SSRs motif length for all three EST fractions.

Perfect EST-SSRs were more frequent than compound ones in both taxa and in all three genomic fractions (Additional file 1: Table S2). In angiosperms, Eucalyptus (ORF) had the highest percentage of compound EST-SSR motifs (7.4%), while Cycas (3′ UTR) had the highest percentage of compound SSR motifs (6.8%) in gymnosperms. None of the statistical tests made to compare proportions of complex EST-SSRs within and between taxa were significant. Furthermore, complexity was not significantly associated to genome size.

EST-SSR abundance (motif counts per Mbp)
(i) Overall
In angiosperms, SSR counts showed a wide range across genera, with Prunus having an exceptional high abundance. EST-SSR counts were significantly higher in the 5′ UTR fraction and lower in the ORFs. In gymnosperms, the SSR counts range was narrower than in angiosperms with Zamia and Gnetum having the highest values. EST-SSRs were significantly more abundant in the 3′ UTR fraction, while there was a non-significant difference in abundance between the 5′ UTR and ORF fractions. EST-SSRs were significantly more abundant in angiosperms than in gymnosperms. No association was found between density and genome size in any of the two taxa.

(ii) By motif size
The distribution of counts per Mbp for each of the EST-SSRs, according to motif size, is shown in Table 2. In angiosperms and gymnosperms, dimer motifs showed significantly higher number of counts in all three genomic fractions, followed by trimers, with the exception of Citrus (ORF, trimers > dimers), Cryptomeria (ORF, trimers > dimers) and Gnetum (5′ UTR and ORF, trimers > dimers and trimers > hexamers, respectively). Non-significant differences between dimers and trimers were found in Cryptomeria (5′ UTR) and Gnetum (3′ UTR). In both taxa, the most frequent motif ranking in the ORF was dimer > trimer > hexamer. The same motif ranking was often observed in the UTRs in gymnosperms. Moreover, in angiosperms, hexamers are less often ranked in the third position in the UTRs, supporting a lower representation of hexamers in UTRs in angiosperms. Despite dimers being the motifs with higher number of counts in most of the genera across all three genomic fractions, the proportion of dimers to trimers was clearly lower in the ORF, indicating an enrichment of trimers in the ORF fraction in both taxa. Interestingly, Gnetum was the only genus where dimers rank third when it comes to abundance (ORF, trimers > hexamers > dimers); trimers and
hexamers being relatively abundant across all three fractions. In *Fraxinus* and *Fagus*, trimers and hexamers were also rather abundant.

(iii) By dimer and trimer nucleotide composition

The counts for dimer and trimer nucleotide composition across genomic fractions and genera are shown in Table 3. In angiosperms, the AG/CT dimer motif showed the highest number of counts per Mbp in all genomic fractions and genera, followed by the AT/AT motif, with exception of *Betula* (AT/AT and AG/CT were present in similar numbers), *Citrus* (3’UTR; AT/AT) and *Populus* (3’UTR; AT/AT). In gymnosperms, AT/AT was the most abundant dimer motif in the 3’UTR fraction, with the exception of *Cryptomeria, Cycas* and *Gnetum* where AT/AT and AG/CT were present in similar numbers. In the 5’UTR and ORF fractions in gymnosperms, AG/CT was the most abundant motif in most of the genera, with the exception of *Cycas* (5’UTR), *Ginkgo* (ORF) and *Zamia* (ORF), where AT/AT and AG/CT were present in similar numbers; and *Ginkgo* (5’UTR), *Zamia* (5’UTR) and *Cycas* (ORF), where AT/AT was the most abundant. Overall, AT/AT was often the most abundant dimer in gymnosperms. The dimer motif CG/CG was absent in most of the genera and only present at low density in the ORF of *Populus* and *Quercus*.

In the 3’UTR fraction in angiosperms and gymnosperms AAT/ATT was the most abundant trimer motif in all the genera with the exception of *Eucalyptus* (AAG/CTT, AGG/CTT and CCG/CCG were present in similar numbers), *Fraxinus* (AAT/AAT and ACT/AGT were present in similar numbers), *Prunus* (ACT/AGT most abundant) and *Gnetum* (AAG/CTT most abundant). In the 5’UTR and ORF fractions in angiosperms, AAG/CTT was the most abundant in all genera except in *Betula* (5’UTR; AAC/GTT and ACT/AGT were present in similar numbers), *Betula* (ORF; AAG/CTT, AAC/GTT and ACC/GGT were present in similar numbers) and *Prunus* (ORF; ACT/AGT most abundant). Moreover, in the 5’UTR and ORF in gymnosperms, there was not a single trimer motif that ranked first, instead it varied across genera.

### Discussion

In this study we have investigated the occurrence of EST-SSRs in three EST genomic fractions (5’UTR, ORF and 3’UTR), in a genus-wise analysis in woody trees of two taxa, angiosperms and gymnosperms. Genus-wise EST-SSRs analysis for EST genomic fractions separately supports the unequal distribution of EST-SSR motifs across the EST sequences. EST-SSR length is positively associated with genome size in gymnosperms (i.e. larger genomes have longer EST-SSRs). However, EST-SSR density is not proportional to genome size; instead other factors seem to have contributed to the EST-SSR density in gymnosperms. We observed two main differences...
Table 2 Counts per Mbp of different SSR motifs in each genomic fraction in: (a) Angiosperms and (b) Gymnosperms

| (a) | Populus | Eucalyptus | Betula | Fagus | Quercus | Citrus | Prunus | Fraxinus |
|-----|---------|------------|--------|-------|---------|--------|--------|---------|
|      | 5'UTR  | ORF  | 3'UTR | 5'UTR | ORF  | 3'UTR | 5'UTR | ORF  | 3'UTR | 5'UTR | ORF  | 3'UTR | 5'UTR | ORF  | 3'UTR | 5'UTR | ORF  | 3'UTR |
| Dimer | 948    | 272  | 379    | 1821   | 699  | 459    | 1131   | 649  | 880    | 1304   | 230  | 397    | 2193   | 530  | 832    | 318    | 96   | 122    |
| Trimer | 250    | 209  | 146    | 232    | 412  | 91     | 151    | 181  | 0      | 172    | 161  | 77     | 232    | 286  | 126    | 190    | 104  | 43     |
| Tetramer | 85     | 16   | 42     | 77     | 27   | 27     | 47     | 10   | 41     | 65     | 7    | 21     | 97     | 15   | 49     | 32     | 7    | 15     |
| Pentamer | 97     | 16   | 35     | 49     | 15   | 23     | 0      | 34   | 0      | 39     | 3    | 35     | 88     | 15   | 45     | 24     | 4    | 9      |
| Hexamer | 68     | 54   | 28     | 43     | 67   | 17     | 0      | 8    | 0      | 60     | 56   | 49     | 70     | 91   | 26     | 17     | 27   | 7      |
| Heptamer | 27     | 7    | 18     | 29     | 14   | 14     | 57     | 6    | 24     | 52     | 2    | 41     | 50     | 7    | 25     | 16     | 5    | 9      |
| Octamer | 6      | 2    | 2      | 2      | 4    | 5      | 0      | 0    | 0      | 3      | 1    | 0      | 3      | 1    | 3      | 4      | 1    | 2      |
| Novamer | 1      | 3    | 1      | 7      | 5    | 1      | 0      | 6    | 0      | 0      | 3    | 0      | 1      | 3    | 1      | 1      | 2    | 1      |
| Decamer | 3      | 2    | 2      | 7      | 6    | 2      | 19     | 0    | 0      | 4      | 2    | 1      | 5      | 1    | 2      | 2      | 2    | 2      |

| (b) | Picea | Pinus | Cryptomeria | Gnetum | Cycas | Zamia | Ginkgo |
|-----|-------|-------|-------------|--------|-------|-------|-------|
|      | 5'UTR | ORF  | 3'UTR | 5'UTR | ORF  | 3'UTR | 5'UTR | ORF  | 3'UTR | 5'UTR | ORF  | 3'UTR | 5'UTR | ORF  | 3'UTR | 5'UTR | ORF  | 3'UTR | 5'UTR | ORF  | 3'UTR | 5'UTR | ORF  | 3'UTR | 5'UTR | ORF  | 3'UTR |
| Dimer | 183   | 128  | 199    | 169    | 121  | 140    | 46     | 58   | 116    | 133    | 104  | 182    | 118    | 84   | 354    | 503    | 504  | 578    | 319    | 164  | 483    |
| Trimer | 14    | 41   | 12     | 5      | 30   | 8      | 43     | 85   | 47     | 260    | 355  | 169    | 10     | 13   | 12     | 35     | 143  | 60     | 52     | 24   | 12     |
| Tetramer | 6     | 1    | 10     | 6      | 2    | 11     | 17     | 1    | 9      | 78     | 15   | 69     | 17     | 1    | 12     | 36     | 18   | 52     | 8      | 4    | 13     |
| Pentamer | 24    | 5    | 12     | 12     | 4    | 8      | 41     | 8    | 20     | 45     | 31   | 41     | 9      | 1    | 6      | 11     | 9    | 21     | 7      | 1    | 17     |
| Hexamer | 9     | 26   | 6      | 14     | 23   | 8      | 27     | 54   | 22     | 111    | 154  | 74     | 10     | 7    | 13     | 10     | 17   | 3      | 0      | 17   | 14     |
| Heptamer | 8     | 1    | 8      | 7      | 2    | 6      | 37     | 8    | 22     | 0      | 2    | 5      | 7      | 2    | 2      | 15     | 8    | 19     | 0      | 0    | 0      |
| Octamer | 2     | 1    | 1      | 1      | 1    | 1      | 8      | 1    | 1      | 0      | 0    | 0      | 0      | 0    | 0      | 0      | 0    | 0      |
| Novamer | 1     | 3    | 1      | 1      | 2    | 2      | 0      | 3    | 0      | 2      | 0    | 0      | 0      | 0    | 0      | 1      | 0    | 0      |
| Decamer | 1     | 1    | 1      | 1      | 1    | 2      | 4      | 1    | 2      | 5      | 0    | 8      | 0      | 0    | 0      | 2      | 0    | 0      |

Ranade et al. BMC Plant Biology 2014, 14:220 Page 4 of 10
http://www.biomedcentral.com/1471-2229/14/220
Table 3 Counts per Mbp of dimer and trimer motifs in all three genomic fractions in: (a) Angiosperms and (b) Gymnosperms

(a)  
| Motif  | Populus | Eucalyptus | Betula | Fagus | Quercus | Citrus | Prunus | Fraxinus |
|--------|----------|------------|--------|-------|---------|--------|--------|---------|
|        | 5'UTR | ORF | 3'UTR | 5'UTR | ORF | 3'UTR | 5'UTR | ORF | 3'UTR | 5'UTR | ORF | 3'UTR | 5'UTR | ORF | 3'UTR | 5'UTR | ORF | 3'UTR | 5'UTR | ORF | 3'UTR | 5'UTR | ORF | 3'UTR |
| AC/GT  | 53    | 22  | 53    | 27    | 7   | 12    | -     | -    | 98   | 61    | 11   | 7     | 103   | 27   | 43    | 27    | 8   | 18    | 165   | 48   | 101   | 91    | 6    | 34    |
| AG/CT  | 822   | 185 | 148   | 1788  | 684 | 431   | 1131  | 649  | 350  | 1173  | 181  | 262   | 1885  | 439  | 471   | 230   | 73   | 47    | 5992  | 2226 | 2655  | 431   | 113  | 109   |
| AT/AT  | 73    | 57  | 178   | 7     | 8   | 15    | -     | -    | -    | 69    | 38   | 128   | 205   | 63   | 317   | 60    | 16   | 57    | 697   | 129  | 811   | -     | 6    | -     |
| CG/GG  | -     | 8   | -     | -     | -   | -     | -     | -    | -    | -     | 1    | -     | -     | -     | -     | -     | -    | -     | -     | -    | -     | -     | -    | -     |
| ACG/CG  | 27   | 42  | 14    | 29    | 63  | 11    | -     | -    | -    | 6     | 10   | -     | 5     | 18   | 1     | 6     | 17   | 2     | 15    | 63   | 9     | -     | -    | -     |
| ACT/AG  | 23   | 23  | 14    | 17    | 4   | 9     | 66    | 34   | -    | 13    | 24   | 5     | 27    | 38   | 29    | 6     | 9    | 4     | 29    | 114  | 136   | -     | 4    | 21    |
| AAC/GTT | 10   | 12  | 9     | 5     | 1   | -     | 85    | 57   | -    | 15    | 30   | 10    | 39    | 55   | 17    | 3     | 10   | 1     | 40    | 43   | 22    | -     | -    | -     |
| AAG/CTT | 93   | 46  | 43    | 98    | 63  | 28    | -     | 77   | -    | 111   | 52   | 22    | 125   | 91   | 31    | 38    | 28   | 12    | 168   | 90   | 86    | -     | 4    | 11    |
| AAT/ATT | 30   | 11  | 51    | -     | 5   | 3     | -     | 14   | -    | 13    | 11   | 36    | 24    | 14   | 42    | 29    | 16   | 21    | 34    | 26   | 41    | -     | 4    | 25    |
| ACC/GGT | 26   | 37  | 1    | 52    | 183 | 12    | -     | -    | -    | 13    | 22   | -     | 5     | 20   | 3     | 2     | 7    | 1     | 36    | 33   | 26    | -     | -    | -     |
| AGG/CCT | 33   | 32  | 7     | 26    | 63  | 12    | -     | -    | -    | 13    | 22   | -     | 5     | 20   | 3     | 2     | 7    | 1     | 36    | 33   | 26    | -     | -    | -     |
| CCG/CCG | 4    | 7   | 1     | 52    | 183 | 12    | -     | -    | -    | 1     | 1    | 6     | -     | 2    | 6    | 1     | -    | 5     | -     | -    | -     | -     | -    | -     |

(b)  
| Motif  | Picea | Pinus | Cryptomeria | Gnetum | Cycas | Zamia | Ginkgo |
|--------|-------|-------|-------------|--------|-------|-------|--------|
|        | 5'UTR | ORF | 3'UTR | 5'UTR | ORF | 3'UTR | 5'UTR | ORF | 3'UTR | 5'UTR | ORF | 3'UTR | 5'UTR | ORF | 3'UTR | 5'UTR | ORF | 3'UTR | 5'UTR | ORF | 3'UTR | 5'UTR | ORF | 3'UTR |
| AC/GT  | 3     | 3    | 4     | 1      | 2    | 1     | 19    | -    | 9     | 36    | -    | -     | 11    | 51   | 79    | 116  | 88    | 40    | 44   | 17    |
| AG/CT  | 95    | 93   | 37    | 101    | 80   | 40    | 19    | 46   | 55    | 76    | 60   | 91    | 54    | 33   | 143   | 170   | 194  | 182   | 120   | 60   | 120   |
| AT/AT  | 85    | 31   | 157   | 67     | 39   | 100   | 8     | 12   | 53    | 20    | 45   | 91    | 64    | 40   | 160   | 254   | 194  | 308   | 346   | 60   | 346   |
| CG/GG  | -     | -    | -     | -      | -    | -     | -     | -    | -     | -     | -    | -     | -     | -    | -     | -     | -    | -     | -     | -    | -     | -     | -    | -     |
| ACG/CG  | 1     | 10   | 1     | -      | 8    | -     | -     | 14   | 7     | 76    | 172  | 48    | -     | 2    | -     | 38    | -    | 10    | -     | -    | -     | -     | -    | -     |
| ACT/AG  | 1     | 1    | 2     | -      | 3    | -     | 9     | 8    | 7     | 38    | 31   | -     | -     | -    | -     | -     | 14   | 11   | -     | -     | -    | -     | -     | -    | -     |
| AAC/GTT | 1     | 3    | 1     | 2      | 1    | 5     | 5     | 13   | 13    | 12    | -    | -     | -     | -    | -     | -     | -    | -     | -     | -    | -     | -     | -    | -     |
| AAG/CTT | 1     | 6    | 2     | 1      | 7    | -     | 6     | 26   | 10    | 76    | 57   | 86    | 10    | 9    | 8     | 32    | 16   | -     | 5     | -    | -     | -     | -    | -     |
| AAT/ATT | 4     | 3    | 5     | 2      | 2    | 5     | 5     | 3    | 20    | -     | 4    | 12    | -     | 2    | 5     | 18    | 31   | 33    | 40    | 8    | 12    |
| ACC/GGT | 1     | 2    | -     | 1      | 2    | -     | 11    | -    | -     | 27    | -    | -     | -     | -    | -     | -     | 9    | -    | -     | -     | -    | -     | -     | -    | -     |
| AGG/CCT | 5     | 13   | 1     | -      | 4    | 1     | 8     | 19   | -     | 29    | 37   | 12    | -     | 6    | 8     | 14    | -    | 12    | -     | -    | -     | -     | -    | -     |
| CCG/CCG | 1     | 4    | -     | 2      | -    | -     | 10    | 4    | 14    | -     | -    | -     | -     | -    | -     | -     | -    | -     | -     | -    | -     | -     | -    | -     |
between angiosperm and gymnosperm genera, which may reflect evolutionary differences following their divergence 350 Myr [48], such as the increased presence of hexamers and AT-rich motifs in the gymnosperm genera.

**Low contribution of EST-SSRs to genome size diversity**

Our EST-SSRs length values are in accordance with those previously reported in the literature [5,27,45]. In gymnosperms, we observe a positive and significant association between the EST-SSRs length and genome size. Thus, the largest genomes (Pinus and Picea) also have, on average, the longest EST-SSRs. Although this suggests a higher relaxation towards genome enlargement in those two genera, the yet small differences in length between the studied gymnosperm genera suggests that EST-SSRs length contribution to Pinus and Picea genome obesity may be only modest. Instead, EST-SSRs length has been suggested to be mainly the result of a balance between slippage events and point mutation [8], which have resulted in a rather homogeneous EST-SSRs length, as suggested before [45]. Unlike in gymnosperms, our analysis does not support an association between the EST-SSRs length and genome size in angiosperms. A potential association however could be masked by the multiple polyploidization events and their role in genome size diversification in angiosperms [49]. Although other factors may have played a role in genome size diversity in angiosperms; transposable element (TE) expansion seems to be the most determinant factor [50]. Conifer genome expansion can also be attributed to a large extent to TE expansion [51,52], although its role in genome size diversification is yet to be proven within the gymnosperm taxon.

Our values for percentage of perfect and compound EST-SSRs in Gnetum and Pinus agree with those reported by Victoria et al. [46] and are not correlated with genome size in any of the taxa. Our data also does not support the contribution of overall EST-SSRs abundance to genome size expansion. Instead, angiosperm genera with smaller genomes compared to those in gymnosperms show, on average a significantly higher abundance (four order of magnitude higher) of EST-SSRs. The lower density of EST-SSRs in gymnosperm compared to angiosperm species is in agreement with previous reports [5,45,47] and does not support a possible constant abundance of SSRs in the transcribed portions of the genome across species as suggested by Morgante et al. [4]. Several studies have concluded that EST-SSRs abundance is inversely related to the genome size [5,37], while others attribute EST-SSRs abundance partly to the action of selection and the effectiveness of mechanisms for regulating slippage errors [44,53]. Our more extensive investigation however does not support a simple linear relationship between EST-SSR abundance and genome size. For example, two gymnosperm genera such as Gnetum and Zamia have similar or even higher frequencies of SSRs than angiosperm genera such as Citrus, which has a smaller genome size. This suggests that other factors affecting genome evolution in both taxa need to be considered to explain EST-SSR abundance diversity in the plant kingdom.

EST-SSR abundance across EST fractions also differs between gymnosperm and angiosperms. In angiosperms, EST-SSRs are significantly more abundant in the 5′ UTR fraction, while in gymnosperms there is on an average a higher abundance of EST-SSRs in the 3′ UTR fraction. In angiosperms, a higher density of EST-SSRs in the UTR fractions has been reported previously [4,20,54,55]; while other studies support a higher abundance in the ORF fraction [44]. A higher EST-SSR abundance in the 5′ UTR could be attributed to a regulatory role [56,57]. In Cryptomeria, a higher density of EST-SSRs in the ORF fraction has also been shown [37]. However, due to the limited number of studies performed on each EST fraction separately, a generalization on the relative abundance of SSRs across those fractions warrants further investigation.

**Motif size: while dimers dominate, hexamers are more common in the gymnosperm EST sequences**

Our study reveals an overall higher abundance of dimers across all three genomic fractions (with six exceptions). In an EST-SSRs analysis that included lower and upper plant species, Victoria et al. [46] reported that trimers are more frequent in the majority of groups of higher plants; while individual studies in angiosperm trees have shown dimers as the most abundant motif in genera such as Populus [17,45] and Eucalyptus [16,34]. In Quercus, trimers were reported as the most abundant motif in the ORF fraction, while dimers were more frequent in the UTR fractions [20]. Trimers were the most common motif in Citrus according to some studies [19,27] whereas Palmieri et al. [28] described dimers as the most abundant motifs in the same genus. In gymnosperms, a higher abundance of EST-SSR dimers has previously reported in Pinus, Picea, and Ginkgo [5,24,45,46]; while Yan et al. [47] reported trimers as the most abundant in Pinus. Similarly, trimers were the most frequent in the ORF in Pinus, while dimers were the most common in the 3′ UTR fraction [43]. In agreement with our study, increased representation of trimers in the ORF was shown before in Cryptomeria [37]. Trimers and hexamers were reported to be more common in the ORF compared to the UTRs in Quercus [20] and Cryptomeria [37]. Similarly, we also observe trimers and hexamers as common in both taxa with reference to ORF.

Our data shows that despite the fact that dimers are the most frequent repeats in majority of the genera in all
the three genomic fractions, the proportion of dimers to trimers (dimers/trimers) decreases significantly in the ORF fraction. Predominance of trimers in the coding regions was reported previously in animals and plants [58]. ORF enrichment in trimers is expected considering that dimers alter the frameshift (i.e., nucleotide triplet or codon is the unit for translation), which should be avoided if the correct translation of the ORF into a protein should be maintained. Presence of SSR dimers in the ORF fraction can potentially affect gene amino acid sequences consequently altering their function due to frameshift mutations, while SSRs in the UTR fractions will affect transcription, translation or splicing of gene products [8]. Moreover, if the number of dimer repeats is divisible by three, it will result in the alternation of two amino acids (e.g., (AT)₆: ATA-TAT-ATA-TAT: Ile-Tyr-Ile-Tyr), thus potentially leaving the reading frame un-altered, as previously suggested by Kantety et al. in cereal species [59].

Dimer/Trimer nucleotide composition: AT-rich motifs are common in gymnosperms

Our study reveals a low abundance of AC/GT motif in all studied genera. Unlike as in mammals, the AC/GT motif is known to occur at low frequency in plants [4,60]. The difference between plants and mammals has been attributed to differences in methylation patterns. AC/GT abundance in animals was suggested as the result of transition of methylated C residue to T (CG → AC/GT), while the absence of a C-hotspot in plants could have prevented the predominance of AC/GT repeats [4,60]. In agreement with previous works, the CG/CG motif (which creates CpG islands acting as regulatory elements through methylation) is almost absent in all our studied genera across all three genomic fractions. There is however an overall predominance of AG/CT (all three genomic fractions) and AAG/CTT (5′UTR and ORF) motifs in angiosperms, which are also target for methylation in plants [61]. In gymnosperms, AG/CT is also the most abundant motif in the 5′UTR and ORF fractions (with few genera where AT/AT is more abundant). In the 3′UTR regions, there is predominance of AT/AT (gymnosperms) and AAT/ATT (both taxa), which are not the target for methylation [62]. An increased content in A + T nucleotides in the 3′UTR fraction has been reported before in vertebrates [63], mammals [64], yeast [65] and Arabidopsis [4], which seems to be related to the UTR processing signal composition.

An overall predominance of AG/CT and AT/AT dimer motifs in EST sequences was supported by previous studies in angiosperms [20,34,47] and gymnosperms [5,46,47]. In angiosperms, AG/CT was reported as the most abundant in Eucalyptus [16,34,47], Citrus [26-28] and Populus [45,47,66]. In Quercus, AC/GT was shown as the most abundant dimer [20]. In agreement with an overall enrichment in AT/AT motif gymnosperms (specially in the 3′ UTR fraction), other studies have also reported AT/AT as the most frequent dimer in Pinus [5,43,45-47], Picea [5,24,45] and Ginkgo [45]. Berube et al. [5] also demonstrate a similar finding with a higher abundance of AT/AT dimers in the 3′ sequenced ESTs in Pinus and Picea. The motif AG/CT was shown to be the most abundant in Cycas [45] and Gnetum [46]; the latter being also supported by our data. In Cryptomeria, AT/AT was shown to be the most abundant in the UTR fractions, while AG/CT was the most abundant in the ORF [37].

In agreement with our results, previous studies also support a higher abundance of the AAG/CTT motif in angiosperms. In gymnosperms, our study reveals predominance of the AAT/ATT motif in the 3′ UTR fraction; moreover, trimer predominance in the other two fractions seems genus dependent. In angiosperms, AAG/CTT was ranked first in frequency in Eucalyptus [16,47], Citrus [26-28] and Poplar [45,47,66]. In Eucalyptus, other studies reported AGG/CCT [34] as the most abundant trimer motifs. In Quercus, AAT/ATT was shown to be the most common trimer motif [20]. In gymnosperms, AAT/ATT was shown to be the most abundant trimer in Pinus [45]. Other studies report AAG/CTT as the most common trimer in Pinus [43,47], Picea [24] and Cycas [45]. Also ACG/CGT was presented as the most abundant trimer in Pinus and Picea [5]. In Cryptomeria, our trimer motif dominance across the EST fractions corresponds with that reported by [37] (i.e., AGG, 5′UTR; AAG, ORF; AAT, 3′UTR).

Conclusions

Our EST-SSR comparative analysis in eight angiosperm genera and seven gymnosperm genera has revealed interesting differential features among both taxa. While dimers dominate, hexamers are more common in the gymnosperm EST sequences than the angiosperms, and AT-rich motifs among the dimers are the most abundant in gymnosperms. These results provide the foundation for future comparative studies at the species level to unravel the evolutionary processes that control the SSR genesis and divergence between angiosperm and gymnosperm tree species.

Methods

Genomic resources and bioinformatics

Description of the EST resources analysed in this study is represented in Additional file 1: Table S1. ESTs from 40 species from eight genera in angiosperms and 17 species from seven genera in gymnosperms were considered for the EST-SSR analysis in this study. EST sequences of the selected species were retrieved from the dbEST database of the NCBI. The criterion for species selection,
analysis and the results presented in this work was based on the availability of the sequence data in the EST data-
bank. To remove redundancy, EST sequences were as-
sembled into contigs and singlets, species-wise, using the sequence assembly program CAP3 with its default setting [67]. For each genus, the species-wise assembled contigs and singlets were pooled together and the se-
quence redundancy at genus level was removed using
CD-HIT [68] with a cut off value of 90% (ensuring 90% sequence identity). The ORF detection is based on the
same principle as the generic eukaryotic gene prediction
program used for searching the coding regions from a
given nucleotide sequence. Based on the coding poten-
tial profiles trained from Angiosperms (Arabidopsis) and
Gymnosperms (Norway spruce) protein coding genes, we used AUGUSTUS [69] to distinguish the coding and
the UTR regions, and the coding direction of a given transcript sequence. The main feature in detecting ORF
on transcript sequence is that the ORF is located in an intron-less, single exon coding region. However, due to
the unexpected higher coding potential in the UTR re-

gion, one transcript might contain more than one ORF.
In such cases, we have selected the longest ORF as the
the true coding region and the adjacent nucleotide sequence
as the UTR region. Thus the longest ORF was selected from each of the EST sequence from the genus-wise
collection of sequences and the 5′UTR and 3′UTR fractions of the sequence were assigned based on the
coordinate direction of the ORF. Three groups of se-
quences were thus created with reference to each genus,
namely 5′UTR, ORF and 3′UTR. SSRLocator v.1 [70] was used to retrieve the SSR information at the genus
level from each of the three groups derived. SSRLocator was used with the following settings, SSR repeat motifs
and number of repeats shown respectively, dimer-10,
trimer-7, tetramer-5, pentamer-4, hexamer-4, heptamer-3,
octamer-3, nonamer-3, decamer-2. The space between
compound SSRs was set to 100 bp. Thus repetitions that
occurred in the adjacent regions lower than 100 bp, were
considered as compound SSRs. These settings are in com-
pliance with the search parameters for repetitive elements
in class I (20 bp) described as more efficient molecular
markers followed by Temnykh et al. [71]. Mononucleotide
repeats can be difficult to accurately assay and are gen-
erally eliminated from the SSR analysis [45,72–74] and
consequently these repeats were excluded from this
study. Therefore, in this article we discuss the occur-
rence of microsatellites specific to 5′UTR, ORF or 3′
UTR fractions of the ESTs. While recording the count of a
particular repeat motif, circular permutations and/or re-
verse complements of each other were clustered together
(e.g. AC = GT = CA = TG, ACG = CGA = GCA = TGC =
GCT = CGT = AGC = TCG = CAG = GTC = TGC = GAC
and AAC = ACA = CAA = TTG = TGT = GTT) [5]. We
also screened for perfect and compound SSRs. Perfect
SSRs are the repeat motifs that are simple tandem se-
quence, without any interruptions within the repeat
(e.g. TATATATATATATA or [TA]n); while a compound
SSR consists of the sequence containing two adjacent dist-
tinct SSRs separated by none to any number of base pairs
(e.g. TATATATAGTGTGTGTGT or [TA]n-[GT]n).

Statistical analysis
A non-parametric Tukey HSD test was carried to com-
pare the means of EST-SSRs length between all categories.
We carried out a 2 × 3 contingency χ2 test for hetero-
genicity of microsatellite counts (motif counts/total EST-
fraction in Mbp) among the three EST genomic regions.
Statistical analyses were all carried out using the R software
package [75].

Additional file

Additional file 1: Table S1. EST database size, number of nucleotides
used for SSR analysis and counts of repeat motifs per Mbp in each
fraction: (a) Angiosperms and (b) Gymnosperms. Table S2 SSR motif
complexity in: (a) Angiosperms and (b) Gymnosperms.

Abbreviations
SSR: Simple sequence repeats; EST: Expressed sequence tags; UTR: Untranslated
region; ORF: Open reading frame; Myr: Million years; TE: Transposable element.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
SSR was involved in the design of the study and manuscript writing. SSR
performed the bioinformatics analysis. MRGG was involved in the design
of the study and manuscript writing. MRGG was responsible of the statistical
analyses. YCL, AZ and YvDP contributed to the bioinformatics work. All authors
read and approved the final manuscript.

Acknowledgements
SSR salary was supported by the Faculty of Forest Science, SLU, Umeå, Sweden.
Travel cost for SSR was covered by the travel grant from FORMAS. YCL was
supported by the Wallenbergs Stiftelse, Norway spruce genome project. YCL and
YvDP were supported by Ghent University Multidisciplinary Research Partnerships
“Bioinformatics: from nucleotides to networks”. Authors acknowledge the support
of computational resources from Norway spruce genome consortium.

Author details
1Umeå Plant Science Centre (UPSC), Department of Forest Genetics and
Plant Physiology, Swedish University of Agricultural Sciences, SE-901-83
Umeå, Sweden. 2Department of Plant Systems Biology (VIB) and Department
of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark
927, 9052 Ghent, Belgium. 3Istituto di Genomica Applicata, Via J. Linussio 51,
33100 Udine, Italy. 4Institute of Life Sciences, Scuola Superiore Sant’Anna,
56127 Pisa, Italy. 5Genomics Research Institute, University of Pretoria, Hatfield
Campus, Pretoria 0028, South Africa.

Received: 7 April 2014 Accepted: 5 August 2014
Published: 21 August 2014

References
1. Tautz D, Renz M: Simple sequences are ubiquitous repetitive components
of eukaryotic genomes. Nucleic Acids Res 1984, 12(10):4127–4138.
2. Zane L, Bargelloni L, Pattarino T: Strategies for microsatellite isolation: a
review. Mol Ecol 2002, 11(1):1–16.
3. Gupta M, Chyi YS, Romerosewnętron J, Owen JL: Amplification of DNA markers from evolutionarily diverse genomes using single primers of simple-sequence repeats. Theor Appl Genet 1994, 89(7–8):998–1006.

4. Morgante M, Hanafey M, Powell W: Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 2002, 30(2):194–200.

5. Berube Y, Zhuang J, Rungis D, Ralph S, Bohlmann J, Ritland K: Characterization of EST SSRs in loblolly pine and spruce. Tree Genet Genomes 2007, 3(3):251–259.

6. Powell W, Mackay GC, Provan J: Polymorphism revealed by simple sequence repeat patterns. Trends Plant Sci 1996, 1(5):215–222.

7. Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC: Genomic sequence: initial sequencing and analysis of Pseudomonas aeruginosa. Nat Genet 1995, 9(4):373–383.

8. Li YC, Koral AB, Fahima T, Nevo E: Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 2004, 21(9):1997–2001.

9. Chen CX, Bowman KD, Choi YA, Dang PM, Rao MN, Huang S, Soneji JR: Identification of microsatellites in the coding regions of the Populus Genome. Mol Biol Evol 2007, 24(3):618–627.

10. Cibrian-Jaramillo A, Marler TE, DeSalle R, Brenner ED: Identification and characterization of EST-derived microsatellites in the inner bark cDNA library of Fagus crenata (Fagaceae). Conserv Genet Resour 2009, 3(1):137–148.

11. Aleksić J, Schwertler S, Mengi G, Geburek T: EST-SSRS developed for other Picea species amplify in Picea abies and reveal high genetic variation in two natural populations. Bel J Bot 2009, 142(1):89–95.

12. Fluch S, Kojanchev D, Homann GA: Characterization of variable EST SSR markers for Norway spruce (Picea abies). BMC Res Notes 2011, 4:4001.

13. Xie H, Sui Y, Zhang FG, Xu Y, Ma RC: SSR allelic variation in almond (Prunus dulcis Mill.). Theor Appl Genet 2006, 112(2):366–372.

14. Rehner A, Fatadi R, Elaab A, Taghavi H, Hassan D, Gratzel T, Folta K, Chaparro J: Genetic diversity of some wild almonds and related Prunus species revealed by SSR and EST-SSR molecular markers. Plant Syst Evol 2012, 298(1):173–192.

15. Cupteno FB, Leal JB, Correa RA, GaitOil FA: Genetic diversity of Eucalyptus hybrids estimated by genomic and EST microsatellite markers. Biol Plantarum 2011, 55(2):379–382.

16. Yasodha R, Sumathri R, Chezhian P, Kathwa S, Ghosh M: Eucalyptus microsatellites mined in silico: survey and evaluation. J Genet 2008, 87(1):21–25.

17. Xinye Z, Congwen S, Yadong Z, Yanling Y, Minren H: Development of Amplification of DNA microsatellites from ESTs of Citrus. Genet Mol Biol 2007, 30(3):1109–1118.

18. Lu Y, Li H, Jia Q, Huang H, Tong Z: Identification of SSR loci in Betula luminifera using birch EST data. J For Res 2011, 22(2):201–204.

19. Ueno S, Taguchi Y, Tomaru N, Tsumura Y: Development of EST-SSR markers from an inner bark cDNA library of Fagus crenata (Fagaceae). Conserv Genet 2009, 10(5):1477–1485.

20. Vendramin E, Dettori MT, Giovannini J, Miceli S, Quarta R, Verde I: A set of EST-SSRs isolated from peach fruit transcriptome and their transportability across Prunus species. Mol Ecol Notes 2007, 7(3):307–310.

21. Xiong X, Yin TM, Wang MX, Tuskan GA: Characterization of microsatellites in the coding regions of the Populus Genome. Mol Biol Evol 2011, 27(1):59–66.

22. Elbelo E, de Souza AN, Sarto D, Tass SM: In silico characterization of microsatellites in Eucalyptus spp.: abundance, length variation and transposition associations. Genet Mol Biol 2005, 28(3):582–588.

23. Ceresini PC, Silva CLSP, Misio FF, Souza EC, Fischer CN, Guillerme R, Gregorio I, da Silva EHT, Cricerli RMB, da Silva MTA, Garcia JF, Avelar GA, Neto LR, Marcon AR, Bacci M, Marini DC: Satellite analysis and database of microsatellites from ESTs of eucalyptus. Genet Mol Biol 2005, 28(3):589–600.

24. Faria DA, Marmo BM, Pappar MRR, Pappas GJ, Grattapaglia D: A selected set of EST-derived microsatellites, polymorphic and transferable across 6 species of Eucalyptus. J Hered 2010, 101(4):512–520.

25. Moriguchi Y, Ueno S, Ujino-ihara T, Futamuna N, Matsumoto A, Shinohara K, Tsumura Y: Characterization of EST-SSRs from Cryptomeria japonica. Conserv Genet Resour 2009, 3(1):373–376.

26. Ueno S, Moriguchi Y, Uchijama Y, Ujino-ihara T, Futamuna N, Sakurai T, Shinohara K, Tsumura Y: A second generation framework for the analysis of microsatellites in expressed sequence tags and the development of EST-SSR markers for a conifer, Cryptomeria japonica. BMC Genomics 2012, 13:36.

27. Zhang FM, Su T, Yang Y, Zhai YH, Ji YH, Chen ST: Development of seven novel Est-SSR markers from Cryptus parvishahuensis (cycadaceae). Ann Bot 2009, 103(2):E159–E161.

28. Yang Y, Li Y, Li LF, Ge XG, Gong X: Isolation and characterization of microsatellite markers for Cycas deboensis Y. C. Zhong et C. J. Chen (Cycadaceae). Mol Ecol Res 2008, 9(4):913–915.

29. Wang ZF, Ye WH, Cao HL, Li ZC, Peng SL: Identification and characterization of EST-SSRs and cpSSRs in endangered Cycas hainanensis. Conserv Genet 2008, 9(6):1079–1081.

30. HongHong F, JingChun L, ZhengPeng Y, Li YL, YongPing C: Characteristics of EST-SSR distribution in Ginkgo ESTs. Genom Appl Biol 2009, 28(5):869–873.

31. Liewalaksanevanwong C, Ritland CE, El-Kassaby Y, Ritland K: Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. Theor Appl Genet 2004, 109(2):361–369.

32. Chaunde PR, Universal P, Aamiri A, Collada C, Guvarna A, Cervantes MT, Vendramin GG, Garcia V, Frietigo JMM, Echt C, Richardson T, Plomion C: Cross-species transferability and mapping of genomic and cDNA SSRs in pines. Theor Appl Genet 2004, 109(6):1204–1214.

33. da Maia LC, de Souza VQ, Kopp MM, de Carvalho FF, de Oliveira AC: Tandem repeat distribution of gene transcripts in three plant families. Genet Mol Biol 2009, 32(4):822–833.

34. von Stackelberg M, Reinsing SA, Reski R: Identification of genic moss SSR markers and a comparative analysis of twenty-four algal and plant gene indices reveal species-specific rather than group-specific characteristics of microsatellites. BMC Plant Biol 2006, 6:9.

35. Victoria FC, da Maia LC, de Oliveira AC: In silico comparative analysis of SSR markers in plants. BMC Plant Biol 2011, 11:5.
50. Tenaillon MI, Hollister JD, Gaut BS: A triptych of the evolution of plant transposable elements. Trends Plant Sci 2010, 15(8):471–478.

51. Morse AH, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, Garcia SA, Kubisiak TJ, Amerson HV, Carlsson JE, Nelson CD, Davis JW: Evolution of genome size and complexity in pinus. Plos One 2009, 4(2):e4332.

52. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, da Maia LC, Palmieri DA, de Souza VQ, Kopp MM, de Carvalho FI, Costa de Oliveira A: SSR Locator: tool for simple sequence repeat discovery integrated with primer design and PCR simulation. Int J Plant Genomics 2008, 2008:41296.

53. Hancock JM: Genomic size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. Genetica 2002, 115(1):93–103.

54. Luro FL, Costantino G, Terol J, Argout X, Allario T, Wincker P, Talon M, Metzgar D, Bytof J, Wills C: Selection against frameshift mutations limits microsatellite expansion in coding DNA. Mol Genet Genomics 2008, 277(7):469–480.

55. Kantety RV, La Rota M, Matthews DE, Sorrells ME: The abundance of various sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 2002, 48(5–6):501–510.

56. Shabalina SA, Ogurtsov AY, Lipman DJ, Kondrashov AS: Comparative analysis of microsatellites in Arabidopsis lyrata and A. thaliana. Nucleic Acids Res 2004, 32(12):3819–3826.

57. Grover A, Ashihvarya V, Sharma PC: Biased distribution of microsatellite motifs in the rice genome. Mol Genet Genomics 2007, 277(5):469–480.

58. Law JA, Jacobsen SE: Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 2010, 11(2):104–120.

59. He XJ, Chen T, Zhu JK: Regulation and function of DNA methylation in plants and animals. Cell Res 2011, 21(3):442–465.

60. Pesole G, Bernardi G, Saccone C: Isochore specificity of AUG initiator context of human genes. FEBS Lett 1999, 464(1–2):60–62.

61. Shabalina SA, Ogurtsov AY, Lipman DJ, Kondrashov AS: Patterns in interspecies similarity correlate with nucleotide composition in mammalian 3′UTRs. Nucleic Acids Res 2003, 31(18):5433–5439.

62. Sonah H, Deshmukh RK, Sharma A, Singh VP, Gupta DK, Gacche RN, Kaur JS, Khattra J: Cluster analysis of microsatellite genotyping data for gene finding in eukaryotes. Nucleic Acids Res 2004, 32(Web Server issue): W309–W312.

63. Pesole G, Bernardi G, Saccone C: Isochore specificity of AUG initiator context of human genes. FEBS Lett 1999, 464(1–2):60–62.

64. He XJ, Chen T, Zhu JK: Regulation and function of DNA methylation in plants and animals. Cell Res 2011, 21(3):442–465.

65. Pesole G, Bernardi G, Saccone C: Isochore specificity of AUG initiator context of human genes. FEBS Lett 1999, 464(1–2):60–62.

66. Shabalina SA, Ogurtsov AY, Lipman DJ, Kondrashov AS: Patterns in interspecies similarity correlate with nucleotide composition in mammalian 3′UTRs. Nucleic Acids Res 2003, 31(18):5433–5439.

67. Tenenkvkh S, Seiner K, Lukashova A, Lipovich L, Cartinhour S, McCouch S: Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposition associations, and genetic marker potential. Genome Res 2001, 11(8):1441–1452.

68. Guichoux E, Lagache L, Wagner S, Chaumeil P, Leger P, Legpais O, Lepoitevin C, Malauta T, Revardel E, Salin F, Petit RJ: Current trends in microsatellite genotyping. Mol Ecol Resour 2011, 11(6):591–611.

69. Kim TS, Booth JG, Gauch HG, Sun Q, Park J, Lee YH, Lee K: Simple sequence repeats in Neurospora crassa: distribution, polymorphism and evolutionary inference. BMC Genomics 2008, 9:31.

70. Sun JX, Mullikin JC, Patterson N, Reich DE: Microsatellites are molecular clocks that support accurate inferences about history. Mol Biol Evol 2009, 26(10):1017–1027.