Extended linkage disequilibrium in non-coding regions in a conifer, *Cryptomeria japonica*

Etsuko Moritsuka *,1, Yosuke Hisataka §,1, Miho Tamura §, Kentaro Uchiyama **, Atsushi Watanabe §§, Yoshihiko Tsumura **, and Hidenori Tachida *

* Department of Biology, Faculty of Sciences, Kyushu University, Fukuoka, Fukuoka 812-8581, Japan, § Graduate School of Systems Life Sciences, Kyushu University, Fukuoka, Fukuoka 812-8581, Japan, **Department of Forest Genetics, Forestry and Forest Products Research Institute, Tsukuba, Ibaraki 305-8687, Japan, §§ Tree Breeding Center, Forestry and Forest Products Research Institute, Hitachi, Ibaraki 319-1301, Japan

1 These authors contributed equally to this study.
Running title:
Linkage disequilibrium in non-coding region

Keywords:
BAC, conifer, DNA polymorphism, linkage disequilibrium, non-coding region

Corresponding author:
Hidenori Tachida
Department of Biology, Faculty of Sciences, Kyushu University,
6-10-1 Hakozaki, Fukuoka, 812-8581, Japan.
Tel.: 81-92-642-3899
Fax: 81-92-642-3899
Email: htachida@kyudai.jp
ABSTRACT

We measured linkage disequilibrium in mostly non-coding regions of *Cryptomeria japonica*, a conifer belonging to Cupressaceae. Linkage disequilibrium was extensive and did not decay even at a distance of 100 kb. The average estimate of the population recombination rate per base pair was $1.55 \times 10^{-5}$, and was less than $1/70$ of that in the coding regions. We discuss the impact of low recombination rates in a large part of the genome on association studies.
Linkage disequilibrium (LD) in the coding genes and their surrounding regions of conifers has been reported to extend to only several hundred to a few thousand base pairs (Brown et al. 2004; Heuertz et al. 2006; Pyhäjärvi et al. 2007; but see Pyhäjärvi et al. 2011 for LD extending to several kb). This observation led to the conclusion that conifers are not suitable for finding associations with traits using a genomic scan (Neale and Savolainen 2004). However, if the genome-wide extent of LD is not the same as that in the regions studied thus far, the strategy may change. Indeed, the extents of LD differ between genomic regions in maize possibly due to differences in the recombination rate (Gore et al. 2009). Furthermore, conifers generally have lower recombination rates than angiosperms (Jaramillo-Correa et al. 2010). Thus, it is important to know the extents of LD in regions other than the coding genes and their surrounding regions. However, the extent of LD in such regions has not been investigated in conifers to our knowledge possibly because of the lack of sequence information in those regions (but see Hamberger et al. 2009; Keeling et al. 2010; Kovach et al. 2010).

We recently constructed a bacterial artificial chromosome (BAC) library for Cryptomeria japonica, a conifer belonging to Cupressaceae sensu lato (Gadek et al. 2000; Kusumi et al. 2000) and determined the sequences of eight random clones from the library. We used these sequences to design primers and resequenced fragments in those regions using the same tree samples used by Fujimoto et al. (2008) for the study of the coding regions. From these data, we investigated the extent of LD at a scale of up to 100 kb in the mostly non-coding regions covered by those BACs (see supporting information, File S1, for more details on materials and methods).

First, we give a summary of the sequence analysis of the BAC clones relevant to this study. Details of the analysis will be published elsewhere (M. Tamura, A. Watanabe, K. Uchiyama, N. Futamura, K. Shinohara, Y. Tsumura, H. Tachida, unpublished results).
Eight random clones (BAC1-BAC8) from the BAC library developed by The Forestry and Forest Products Research Institute in Japan were sequenced. The total length of the sequences was about 800 kb. These sequences were mostly non-coding and at least 51% of them consisted of known (the majority being LTR retrotransposons) or unknown repeats. Only part (four exons) of one putative protein-coding gene, homologous to genes coding for calcium-dependent protein kinases, was found in BAC6. The ratio of nonsynonymous to synonymous divergence between the exons and the corresponding regions in *Taxodium distichum*, a close relative of *C. japonica*, was 0.086 suggesting the gene was active when the two species diverged. The size of the largest intron of this gene was 70 kb (see Figure 1).

In the resequencing experiment, in total, 19 fragments including the coding regions in BAC6 could be sequenced for at least 40 samples, and their total length was 15517 bp (Figure 1). We will call them the BAC regions hereafter. The fragment size ranged from 509 to 2016 bp, and the maximum lengths between pairs of sites in the sequenced regions in BAC3, BAC6, and BAC7 were 65409, 110613 and 103427 bp, respectively.

We aligned those 40 or more sequences along with the corresponding sequence of the BACs. Thus, the total number of samples was at least 41. We computed various statistics of genetic diversity excluding sites involving indel variation (Table S2) using DNAsp 5.0 (Rozas et al. 2003). The average nucleotide diversity across all sites was 0.35%, which was similar to those at the silent sites in the mostly coding regions of the same species studied by Kado et al. (2003) and Fujimoto et al. (2008).

Figure 2 shows the squared allelic correlation coefficient, $r^2$, against the number of bp separating the two sites. It decayed very little in all three BACs. Thus, LD extended to sites separated by up to 110 kb. We also estimated the population recombination rate, $4N_e r$, where $N_e$ and $r$ are the effective size of the population and the recombination rate, respectively, in the BAC regions using the composite likelihood method (Hudson 2001;
McVean et al. 2002) implemented in LDhat 2.1. The results for the BAC regions are shown in Table 1, along with those for the five coding genes previously studied by Fujimoto et al. (2008). The composite likelihood curves for the dataset are shown in Figure S1 (the BACs) and Figure S2 (the coding genes). The average $r/u$ and $4N_e r$ per base were $5.04 \times 10^{-3}$ and $1.55 \times 10^{-5}$, respectively, for the BAC regions and $4.46 \times 10^{-1}$ and $1.18 \times 10^{-3}$, respectively, for the five coding genes. Thus, the population recombination rate seems to be less than $1/70$ in the BAC regions compared to that in the coding genes studied by Fujimoto et al. (2008).

One explanation for the difference of LD between the BACs and the coding regions is difference in the recombination rate. In angiosperms, gene density and the recombination rate are positively correlated, possibly due to low gene density and low recombination rates in heterochromatic regions (Gaut et al. 2007). Therefore, if the genome of *C. japonica* is mostly heterochromatic, we expect to observe low recombination rates and low gene density in random BAC clones. Alternatively, the effective size, $N_e$, in the BAC regions might be smaller than that in the coding genes because of genetic draft (Gillespie 2000), background selection (Charlesworth et al. 1993) and/or the weak selection Hill-Robertson effect (McVean and Charlesworth 2000). However, because the levels of the silent nucleotide diversity in the two regions were comparable, this explanation seemed unlikely.

The slow decay of LD in the BAC regions is quite different from the rapid decay of LD within 500-2000 bp observed in Pinaceae and poplars (Brown et al. 2004; Ingvarsson 2005, 2008; Ingvarsson et al. 2008; Krutovsky and Neale 2005; González-Martínez et al. 2006; Heuertz et al. 2006; Pyhäjärvi et al. 2007; Li et al. 2010) although more extensive LD has been found in a few cases (Pyhäjärvi et al. 2011; Eckert et al. 2010). *Cryptomeria japonica* is distantly related to these species (Chaw et al. 2000), and the difference may be ascribed to differences in the species studied;
however, a more plausible explanation would be the differences in the regions studied. The other studies examined relatively narrow regions that included coding genes (except for the 80-kb gene-rich region surrounding the *phytochrome B2* locus in aspen studied by Ingvarsson *et al.* 2008), while our BAC regions were in mostly non-coding regions. Indeed, the extensive linkage disequilibrium found here might be consistent with the inter-locus linkage disequilibrium found in coastal Douglas fir (Eckert *et al.* 2009) and in *Pinus taeda* (Eckert *et al.* 2010).

Based on the rapid decay of linkage disequilibrium in conifers, Neale and Savolainen (2004) concluded that genome-wide association studies would not be possible in conifers because of the enormous SNP marker density required (see also Neale and Kremer 2011). However, the number of necessary SNP markers may not be as large if most of the genome is segregating as blocks, as found in the BAC regions studied here. Thus, genome-wide association studies may be feasible in conifers whose genome sizes are generally large (Ohri and Khoshoo 1986), although it would be difficult to locate the causal polymorphism if LD is extensive (Atwell *et al.* 2010). Nevertheless, identifying causal polymorphisms is not necessary for genomic selection (Meuwissen *et al.* 2001). Of course, we need to choose the positions of SNP markers carefully so that they are spaced appropriately in terms of recombination rates by identifying recombination-poor and -rich regions (see Flint-Garcia *et al.* 2003). If heterochromatic regions are recombination-poor in conifers as they are in angiosperms (Gaut *et al.* 2007), then identifying heterochromatic regions is important for designing efficient markers.

In the present paper, we showed that linkage disequilibrium in the BAC regions was much more extensive than that in the coding regions. However, our data were restricted to those from only three random BAC clones. We need to test the generality of these features in future studies.
Acknowledgement

We thank two anonymous referees for their helpful comments. This study was partially supported by the Program for promotion of Basic and Applied Researches for Innovations in Bio-oriented Industry (BRAIN) and by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (nos. 22370083).

LITERATURE CITED

Atwell, S., Y. S. Huang, B. J. Vilhjlmsson, G. Willems, M. Horton, et al., 2010


Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold hardiness in coastal Douglas fir (Pseudotsuga menziesii var. menziesii). Genetics 183: 289-298.


Heuertz, M., E. De Paoli, T. Källman, H. Larsson, I. Jurman, et al., 2006 Multilocus
patterns of nucleotide diversity, linkage disequilibrium and demographic


Ingvarsson, P. K., 2005 Nucleotide polymorphism and linkage disequilibrium within
and among natural populations of European Aspen (*Populus tremula*, L.,

Ingvarsson, P. K., 2008 Multilocus patterns of nucleotide polymorphism and the
demographic history of *Populus tremula*. Genetics **180**: 329-340.

Nucleotide polymorphism and phenotypic associations within and around the
Genetics **178**: 2217-2226.

of recombination to heterozygosity differs among plant evolutionary lineages

Kado, T., H. Yoshimaru, Y. Tsumura and H. Tachida, 2003 DNA variation in a conifer,
*Cryptomeria japonica* (Cupressaceae sensu lato). Genetics **164**: 1547-1559.

and functional characterization of monofunctional *ent*-Copalyl Diphosphate
and *ent*-Kaurene Synthases in white spruce reveal different patterns for
diterpene synthase evolution for primary and secondary metabolism in

taeda* genome is characterized by diverse and highly diverged repetitive
sequences. BMC Genomics **11**: 420.


Pyhäjärvi, T., M. R. García-Gil, T. Knürr, M. Mikkonen, W. Wachowiak, et al., 2007 Demographic history has influenced nucleotide diversity in European Pinus
sylvestris populations. Genetics 177: 1713-1724.


<table>
<thead>
<tr>
<th>Region</th>
<th>Length</th>
<th>$R_M$</th>
<th>$4N_r$</th>
<th>$4N_r$/bp</th>
<th>$4Nu$/bp</th>
<th>$r/u$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BACs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BAC3</td>
<td>65411</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00659</td>
<td>0</td>
</tr>
<tr>
<td>BAC6</td>
<td>110278</td>
<td>1</td>
<td>3</td>
<td>2.72 $\times 10^{-5}$</td>
<td>0.00317</td>
<td>0.00858</td>
</tr>
<tr>
<td>BAC7</td>
<td>103427</td>
<td>7</td>
<td>2</td>
<td>1.93 $\times 10^{-5}$</td>
<td>0.00296</td>
<td>0.00653</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td>1.55 $\times 10^{-5}$</td>
<td>0.00346</td>
<td>0.00504</td>
</tr>
<tr>
<td>Coding genes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCED</td>
<td>2125</td>
<td>2</td>
<td>3</td>
<td>0.00141</td>
<td>0.00213</td>
<td>0.66280</td>
</tr>
<tr>
<td>AMT</td>
<td>2682</td>
<td>2</td>
<td>2</td>
<td>0.00075</td>
<td>0.00093</td>
<td>0.80184</td>
</tr>
<tr>
<td>Cal</td>
<td>3541</td>
<td>2</td>
<td>3</td>
<td>0.00085</td>
<td>0.00454</td>
<td>0.18661</td>
</tr>
<tr>
<td>AQU</td>
<td>1787</td>
<td>1</td>
<td>4</td>
<td>0.00223</td>
<td>0.00469</td>
<td>0.47727</td>
</tr>
<tr>
<td>Cryj2</td>
<td>4257</td>
<td>6</td>
<td>3</td>
<td>0.00070</td>
<td>0.00681</td>
<td>0.10348</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td>0.00118</td>
<td>0.00382</td>
<td>0.44640</td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

FIGURE 1. Positions of the regions in BAC3, BC6 and BAC7 used for the study.

The regions sequenced for at least 41 samples are shown by | |, and the positions of the exons of a putative coding gene are indicated by black boxes. We expected to amplify only one copy of each fragment if it existed in a single copy in the genome because we used haploid genomic DNA from megagametophytes. However, because the genome of *C. japonica* contained many copies of very similar fragments, only sequences of several fragments each in BAC3, BAC6, and BAC7 indicated in the figure could be determined for at least 40 samples by direct sequencing. We used long PCR and gel electrophoresis to confirm the distances between the sites in the sequenced fragments. The lengths of the PCR products approximately agreed with those calculated from the assembled BAC sequences (data not shown).

FIGURE 2. The levels of linkage disequilibrium in BAC3, BAC6 and BAC7.

The squared allelic correlation coefficient, $r^2$, for two sites are plotted against the number of bases separating them for BAC3 (A), BAC6 (B) and BAC7 (C).