Title

High-throughput linkage mapping of Australian white cypress pine (*Callitris glaucophylla*) and map transferability to related species

Author name

Shota Sakaguchi[1, 2], Takeshi Sugino[3], Yoshihiko Tsumura[4], Motomi Ito[1], Michael D. Crisp[5], David M. J. S. Bowman[6], Lynda D. Prior[6], Atsushi J. Nagano[7, 8], Mie Honjo[7], Masaki Yasugi[7], Hiroshi Kudo[7], Yu Matsuki[9], Yoshihisa Suyama[9] and Yuji Isagi[3]

Author affiliation

[1] Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, 153-8902 Japan; [2] Research Fellow of the Japan Society for the Promotion of Science; [3] Division of Forest and Biomaterials Science, Graduate School of Agriculture, Kyoto University, Kyoto 6068502, Japan; [4] Faculty of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Ibaraki 3058577, Japan; [5] Research School of Biology, The Australian National University, Canberra, ACT 2601, Australia; [6] School of Biological Sciences, University of Tasmania, Hobart, TAS 7001, Australia; [7] Center for Ecological Research, Kyoto University, Shiga, 5202113 Japan; [8] Independent Researcher in Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency, 9 Division of Biological Resource Sciences, Graduate School of Agriculture, Tohoku University, Osaki, Miyagi, 9896711 Japan

Corresponding author

Shota Sakaguchi

e-mail: sakaguci54@gmail.com

Key words

Ascertainment bias; *Callitris*; megagametophyte; RAD-sequencing; single tree linkage map

Abstract

White cypress pine (*Callitris glaucophylla*) and related species are drought-tolerant evergreen conifers that occur in a wide range of bioclimatic regions in Australia.. To broaden our understanding of its speciation process we applied ecological genomics to identify markers associated with environmental adaptation. . We adopted a single tree linkage mapping approach combined with high-throughput RAD (restriction site associated DNA) sequencing and EST-SSR genotyping to set up a baseline genetic map for *C. glaucophylla*. The generated linkage map was consisted of 4,284 markers positioned on 11 linkage groups, corresponding to the haploid chromosome number of *Callitris* (2n = 22). Map length inflation due to missing observations and errors were controlled by imputing and correcting genotypes, resulting in map length reduction by 76%. Spatial distribution of markers was uneven compared to random expectation with significant clustering in central positions of some linkage groups, which may be associated with recombination cold spots of pericentromere regions. Allelic segregation was shown to be distorted in particular regions of four linkage groups, where selection may have operated on viability genes, leaving allelic distortion in surrounding linked markers. We then tested RAD-SNP marker recovery in and transferability of the linkage map to population genomic data collected for related *Callitris* species. Of the linkage map markers, 1,257 markers (ca. 30%) were recovered in independent RAD-sequencing of *C. glaucophylla* population samples. Genetic diversity and differentiation evaluated using mapped markers reflected ascertainment bias slightly; a decrease in Hs (absolute difference of -0.018) for a related species (*C. gracilis*) and an increase in F_{ST} between *C. glaucophylla* and *C. gracilis* (+0.018) were detected. Although care should be taken given such biases in cross-species transfer, this study demonstrated that the RAD-SNP based linkage map is essentially useful when combined with population genomic analysis of this conifer lineage.

Introduction

Conifers are of immense ecological importance in terrestrial ecosystems (Debreczy and Racz 2006; Gernandt et al. 2011). They are well represented in plant communities that occur in extreme environments including species that form vast coniferous forests in seasonally cold temperate/boreal regions of Northern Hemisphere . The Australian genus *Callitris* (Cupressaceae) (2n=22) is the most speciose and ecologically important conifer group on this predominately arid island continent (Bowman and Harris 1995). While most *Callitris* species are regional endemics, the *C. columellaris* species complex is unusual in terms of its continental-wide distribution extending from humid coastal to arid interior and seasonally dry monsoon environments (Hill and Brodribb 1999). The complex is comprised of five closely related morphospecies (*C. columellaris*, *C. intratropica*, *C. gracilis*, *C. glaucophylla*, and *C. verrucosa*) and shows further genetic differentiation into twelve regional lineages (Sakaguchi et al. 2013). Some of these species have extreme drought-tolerance ( ) and regeneration is via continuous recruitment in wetter regions or in pulses following successive wet years in arid environments (Prior et al. 2011). The species has moderate tolerance to surface fires, but can be killed by high intensity grass fires ( ). Considering its broad distribution and ecological diversification, genetic adaptation along climatic gradients may have played a significant role in the speciation of this *C. columellaris* species complex. Thus genomic analysis of the species complex to identify genomic regions associated with environmental adaptation is expected to broaden our understanding of the diversification of this group, and more generally provide insights into conifer evolution.

One important aspect of ecological genomics is to locate genes or genomic regions underlying adaptive traits in genetic/physical maps. Unlike model species and economically important species, genome sequencing is still not realistic for conifers due to their enormous genome (e.g., genome size of *Callitris* species is estimated to be 8.3-11.2 pg/C in Ohri and Khoshoo 1986, which is 38-51 times larger than *Arabidopsis thaliana*) (Neale et al. 2014; Nystedt et al. 2013). Thus linkage mapping has been the preferred approach to mapping conifer species genomes. Early mapping studies (1990-2005) utilised markers of isozyme, RAPD, ISSR, RFLP (Ritland et al. 2011), and more recently AFLP, SSR, SNP and their combinations have become more popular (Chancerel et al. 2013; Kang et al. 2011; Martínez-García et al. 2013; Moriguchi et al. 2012; Neves et al. 2013; Pavy et al. 2012). Among these, SNPs (single nucleotide polymorphisms) can be used as the most abundant genetic marker for high-resolution mapping because one SNP occurs every 91 base positions on average in conifer gene sequences (González-Martínez et al. 2011). A recently developed RAD-sequencing (restriction site associated DNA sequencing) is an efficient technique to obtain SNP genotype data of population samples even for species with no prior genomic knowledge (Baird et al. 2008; Peterson et al. 2012). This technique has been introduced to linkage mapping studies and shown to be successful in generating high-density linkage maps for non-model organisms (Guo et al. 2015; Kakioka et al. 2013; Talukder et al. 2014; Wu et al. 2014).

This study reports a high-density linkage map of white cypress pine, *Callitris glaucophylla*, which was rapidly constructed by combining RAD-sequencing with single tree mapping (Tulsieram et al. 1992). The single tree mapping analyses haploid DNA samples, using alleles that are randomly segregated from a single diploid individual. Conifers are particularly suitable for this analysis, since we can use abundant open-pollinated seeds to extract haploid megagametophytes, enabling rapid linkage mapping without a need for controlled crosses or mapping progenies. Our analysis focuses on three specific topics. The first is on correction and imputation of SNP genotype data generated by low cost high-throughput sequencing. Although high-throughput sequencing is a powerful way to obtain huge amount of sequence reads, the resultant genotype data often includes many missing sequences and errors, which can hinder precise estimation of map length and marker ordering (Buetow 1991; Hackett and Broadfoot 2003). To compensate for such difficulties, we refined noisy RAD-SNPs data using recently developed imputation and error correction methods (Ward et al. 2013; Wu et al. 2008). Secondly, we will characterise linkage groups by relating our observation of heterogeneous marker density and distorted segregation along linkage groups to biological processes, including variable recombination frequency and natural selection operated on viability genes. Lastly, we will test marker recovery in population samples and transferability to related species. A genetic map is used to illustrate genetic variation (e.g. *H*s and F_{ST}) as a function of marker position along linkage groups, in which information about marker position is sometimes transferred from a different species/population. It matters, in these cases, how many markers are shared between species and whether ascertainment bias has significant impact on population genetic statistics in a species to which map information is transferred (Clark et al. 2005; Luca et al. 2011). To evaluate such uncertainties, we analysed population samples of two closely related *Callitris* species using the same RAD-sequencing protocol but in an independent sequencing run. Using this population data set, we will estimate marker recovery rate and degree of ascertainment bias to consider the utility of our linkage map for population genomic analyses of the *C. columellari*s species complex.

Materials and Methods

Plant materials and DNA extraction

In September 2013, seed cones were collected from a single tree of *Callitris glaucophylla*, corresponding to a widespread lineage (GD in Sakaguchi et al (2013)), on the Balonne Highway, Queensland, Australia (27°59'38"S, 148°21'32"E; supplementary figure 1a). The cones were kept dry in the laboratory under room temperature until the seeds were discharged. After seed collection, they were soaked in water overnight, and megagametophyte tissue was isolated by carefully removing the seed coat and embryo under a stereo microscope. Total DNA was extracted from 88 megagametophytes and a mother tree foliage sample using a hexadecyltrimethylammonium bromide (CTAB) method (Murray and Thompson 1980), after removing polysaccharides with isolation buffer containing 10% polyethylene glycol. DNA concentration was measured using a Qubit® dsDNA BR Assay Kit (Invitrogen, Massachusetts, USA), and adjusted to 10 ng/uL in all samples before genetic experiments.

RAD-sequencing and EST-SSR genotyping

In this study, a double-digest RAD library (Peterson et al. 2012) was prepared for linkage mapping. Briefly, 10 ng of genomic DNA was digested with EcoRI and BglII (New England Biolabs, Ipswich, Massachusetts, USA) and adapters were ligated at 37 ºC overnight in 10 µL volume, which contained 1 µL of 10x NEB buffer 2, 0.1µL of 100x BSA (New England Biolabs), 0.4 µL of 5 µM EcoRI adapter 1 (CTCGTAGACTGCGTACC) and BglII adapter 2 (GATCGACAGTGTACTCTAGTC), 0.1 µL of 100 mM ATP, and 0.5 µL of T4 DNA Ligase (Enzymatics, Beverly, Massachusetts, USA). The reaction solution was then purified with AMPure®XP (Beckman Coulter, California, USA). Next, 3 µL of purified DNA was used in PCR amplification in 10 µL volume, containing 1µL of each 10 µM index and PCR primer 1.0 (5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA-3'), 0.3 µL of KOD-Plus-Neo enzyme and 1 µL of 10x PCR buffer (TOYOBO, Osaka, Japan), 0.6 µL of 25mM $MgSO_4$, 1 µL of 10 mM dNTP. Thermal cycling was initiated with 94 ºC step for 2 min, followed by 20 cycles of 98 ºC for 10 sec, 65 ºC for 30 sec, 68 ºC for 30 sec. The PCR products were pooled and purified again with AMPure®XP. The purified DNA was then loaded to a 2.0 % agarose gel and fragments around 320 bp was retrieved using E-Gel® SizeSelect™ (Life Technologies, Carlsbad, California, USA). After quality measurement using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, California, USA), the library was sequenced with 51 bp single-end reads in one lane of an Illumina HiSeq2000 (Illumina, San Diego, USA) by Macrogen (Seoul, South Korea).

Five EST-SSRs (Ccol_rep_c1953, Ccol_rep_c10619, Ccol_rep_c10836, Ccol_rep_c12796, Ccol_rep_c35787) characterized in Sakaguchi et al (2011) were used for anchoring the RAD-SNPs based linkage map. The PCR reaction was carried out in a final volume of 10 µL, which contained approximately 5 ng of DNA, 5 µL of 2× Multiplex PCR Master Mix (Qiagen, Hilden, Germany), and 0.2 µM of each primer. The PCR thermal profile involved denaturation at 95 ºC for 3 min, followed by 35 cycles of 95 ºC for 30 sec, 53 ºC for 3 min, 68 ºC for 1 min and a final 7 min extension step at 68 ºC. PCR products were loaded onto an auto sequencer (3100 Genetic Analyser; Applied Biosystems, Carlsbad, California, USA), to assess fragment lengths using GeneMapper® software (Applied Biosystems).

Short read processing

RAD-sequencing reads were filtered by Trimmomatic ver 0.32 (Bolger et al. 2014) to remove the adapter and other Illumina-specific sequences and to cut low quality regions based on a quality score threshold of 20, which corresponds to base call accuracy of 99% (parameter used: AVGQUAL:20, LEADING:19, TRAILING:19, SLIDINGWINDOW:30:20). The cleaned reads were mapped to a RAD reference for the *C. columellaris* species complex using Bowtie 2 (Langmead and Salzberg 2012) with a default parameter setting in LOCAL mode. The reference, consisting of 392,320 contigs with N50 length of 163 bp, was constructed by assembling RAD-sequencing reads of 10 individual trees sequenced with a genome sequencer MiSeq (Illumina), using the CLC Genomics Workbench 7.5.1 (CLC bio, Aarhus, Denmark) (parameter used: mismatch cost 3, insertion and deletion cost 2, length fraction 0.5, similarity fraction 0.9) (see more details of the reference assembly in supplementary material 1). Mapping short reads to this reference assembly was intended to allow (i) alignment of gapped-reads using Bowtie 2, (ii) similarity search of longer contigs with a higher probability of blast hit, and (iii) to facilitate use of linkage map information across different RAD-sequencing experiments by tracking contig IDs in the reference. A similarity search of the reference contigs against an EST-library of the *C. columellaris* species complex (Sakaguchi et al. 2011) was performed by a local BLAST algorithm (Altschul et al. 1990), to find the contigs associated with EST sequences. The SAM files produced by Bowtie 2 were loaded to the 'ref_map.pl' implemented in Stacks 1.08 (Catchen et al. 2011) to generate genotype data for megagametophyte samples, with cross type specified as 'DH' (doubled haploid) with the other parameters set as default.

Linkage map construction and genotype imputation

Segregation distortion of each SNP and EST-SSR marker was examined by a Chi-squared test using AntMap (Iwata and Ninomiya 2006). Genome-wide distribution of allelic segregation was analysed with a generalised additive model as a function of missing rate in genotype data and genomic position. A Loess smoothing function was applied to the predictor of marker position on linkage groups. The GAM analyses were performed using 'gam' library in R ver. 3.1.0 (R Development Core Team 2014).

Genotyping by sequencing is generally characterised by high rate of missing data and erroneous SNP calling (Beissinger et al. 2013). Such data can produce apparent double or more recombination events in a single sample, leading to overestimation of map length and ordering errors. In this study, the MSTmap program (Wu et al. 2008) was used for marker grouping and ordering, as the algorithm is shown to outperform other frequently used algorithms, particularly when the input data are noisy or incomplete (Wu et al. 2008). Furthermore, Maskov ver 1.01 (Ward et al. 2013) was used to impute missing data and correct erroneous genotypes, based on the marker orders estimated from initial MSTmap analysis. Samples with more than 70% missing data were removed from the imputation procedure and from further mapping analysis. Subsequently, the imputed genotype data were used to make the final linkage map with MSTmap, with LOD criterion of 8.0 for grouping markers and Kosambi function to convert recombination value to map distance (Kosambi 1943). Marker position was plotted on the linkage map using MapChart ver 2.0 (Voorrips 2002).

Map coverage and marker distribution analysis

Map coverage was estimated as a ratio of observed/expected genome lengths (G_o/G_e) based on the method 4 described in Bishop *et al*. (1983). Spatial distribution of genetic markers on the linkage map was investigated by dividing the map into 1, 5, and 10 cM intervals, and the number of markers within the bins was counted, respectively. To test whether the markers were randomly distributed, expected distributions of marker count under Poisson and negative binomial distributions were compared with the observed data using a Chi-squared test. The Poisson distribution was generated by specifying the observed mean marker density, while the observed mean and variance were used to determine the dispersion parameter to calculate the expected negative binomial distribution, using ‘stats’ library in R ver. 3.1.0.

Testing transferability of a single-tree linkage map to population genomic studies

To investigate transferability of the linkage map derived from the *C. glaucophylla* to population genomics of *C. columellaris* species complex, we additionally sequenced another RAD library, which included population samples of *C. glaucophylla* and *C. gracilis*. The two *Callitris* species are closely related and can hybridize in the areas of overlapping distributions (Sakaguchi et al. 2013) (supplementary figure 1a). A RAD library was prepared using the same method as described above, which included 31 samples of *C. glaucophylla* and *C. gracilis* and 65 *Callitris* samples from a different research project, and these were sequenced with 51 bp reads in a lane of Illumina HiSeq2000 (Illumina). After quality-based trimming using Trimmomatic ver 0.32 (Bolger et al. 2014), the short reads were mapped to the *Callitris* RAD reference assembly using Bowtie 2 (Langmead and Salzberg 2012) with the same parameter setting as used in the linkage map analysis. The ‘ref_map.pl’ pipeline in Stacks 1.08 was used to build RAD locus, and SNP genotype for each individual was exported with a minimum read depth of 8, using the ‘populations’ program (Catchen et al. 2011). The exported genotype data was then processed with PLINK ver 1.07 (Purcell et al. 2007), filtering out markers with low allele frequency (< 0.03), missing individual rate > 0.7, and significant deviation from Hardy-Weinberg equilibrium (P < 0.01).

The R package ‘hierfstat’ (Gouded 2014) was used to calculate summary statistics of expected heterozygosity (*H*s) and genetic differentiation index (F_{ST}) per each marker for *C. glaucophylla* and *C. gracilis*, respectively. To test whether the markers mapped to *C. glaucophylla*’s linkage map show biased genetic variation compared to those in the overall SNP data, 1,000 subsets were randomly sampled to generate distributions of mean values for each summary statistic, using the R function of ‘sample’ in the ‘base’ package. The distributions of mean value calculated from random subsets were compared to observed means to determine statistical significance. A split tree network (Bryant and Moulton 2004) using a Euclidean distance matrix was constructed using SplitsTree4 ver. 4.10 (Huson 1998) to estimate population structure within the samples. In addition, STRUCTURE analysis (Pritchard et al. 2000) was performed under an admixture and allele frequency correlated model (Falush et al. 2003). Using STRUCTURE ver 2.3, twenty independent simulations were run for K = 2 (i.e., assuming two genetic clusters because we analysed genetic structure in two species), with 100,000 burn-in steps followed by 20,000 Markov chain Monte Carlo (MCMC) steps.

Results

SNP discovery by RAD-sequencing

A total of 179.7 million raw single-end reads with 51 bp were obtained, yielding more than 9.1 gigabases. Thirteen samples with less than 0.5 million reads and one sample with an exceptionally high level of heterozygosity were excluded from further analyses to reduce missing and wrongly called SNPs. After quality-based filtering, the average read number for 74 included samples was 1.3 million (max. 2.5 million, min. 0.3 million), with 97.2 % bases having a quality score higher than 30. The filtered reads were then mapped to 124,879 contigs in the RAD reference assembly, and genotypes at 7,560 markers were determined at more than 55 samples. The genotype missing rate was 9.9 % on average, ranging among samples from 0.7 % to 42.6 %. Graphical plotting of *P* values in Chi-squared tests showed apparent spatial trends with genomic position (figure 2). This association was statistically significant at 9 linkage groups (except for linkage groups 3 and 8), even after effects of missing rate were partialled out in GAM modelling (supplementary table 1, supplementary figure 2). The markers that significantly deviated from the expected segregation ratio of 1:1 ($P < 0.05$; black markers in figure 2) were excluded from the linkage map analysis.

A single tree linkage map for Callitris glaucophylla

On the resultant linkage map, 4,284 genetic markers including 4,279 RAD-SNPs and 5 EST-SSRs were located on 11 linkage groups (figure 3, table 1, supplementary material 2), which corresponds to the haploid chromosome number of *Callitris glaucophylla* (Ohri and Khoshoo 1986). A small portion of the SNP markers (=67/4,279, i.e. 1.6%) showed significant hits against EST contigs (table 1). After performing genotype imputation and error correction using Maskov ver 1.01, the observed map length was reduced by 76.1% (from 4,324.9 cM to 1,033.5 cM; table 1), without disturbing marker orders (supplementary figure 3). The imputation procedure also greatly decreased the number of unique positions (from 1,325 to 585) and mean marker interval (from 1.01 cM to 0.24 cM) (table 1). When taking the observed map length of 1,033.5 cM, the linkage map covered more than 99.9% of the estimated genome length of *C. glaucophylla* (1034.0 cM).

These genetic markers showed a non-random distribution along the linkage groups. Distributions of the observed marker counts were more dispersed at every bin size examined (1, 5, 10 cM), compared with the expectation under a Poisson distribution (supplementary figure 4). Chi-squared tests detected significant deviations at every bin size ($P < 0.01$) from both Poisson and negative binomial distributions. The spatial distribution of genetic markers was heterogeneous within the linkage groups. The bins with highest marker density were detected in the middle regions (LG4, 6, 8, 10 in particular) of most linkage groups,, surrounded by regions with sparser markers (figure 4).

Transferability of the linkage map to population genomics of the C. columellaris *species complex*

Population RAD-sequencing of *C. glaucophylla* and *C. gracilis* sampled resulted in 7,472 SNP markers, which met our filtering criteria. Based on these markers, split-network and STRUCTURE analysis showed that the two species are clustered separately (supplementary figure 1), except for one sample collected in Palinyewah, New South Wales, which was genetically intermediate (indicated by a green triangle in supplementary figure 1). Therefore, in the following calculation, the potential hybrid individual was excluded. When population samples were analysed for each species, the number of markers in common with the *C. glaucophylla* linkage map was 1,257 (out of 6,472) for *C. glaucophylla* and 734 (of 6,476) for *C. gracilis*, respectively. Of the 7,472 SNPs detected in the population analysis of the two *Callitris* species, 873 markers were shared in the *C. glaucophylla* linkage map. The number of SNPs mapped to each linkage group was 64 on average, ranging from 45 in LG6 to 107 in LG2. H_S calculated with the mapped markers (0.153) for *C. glaucophylla* was significantly larger than the values based on randomly sampled SNPs sets (0.141), while the observed H_S was significantly smaller in the mapped markers for *C. gracilis* (absolute difference of -0.018) (table 2). A significant difference was also detected between the genetic differentiation estimates; F_{ST} between two species based on mapped markers elevated by 0.018 (table 2).

Discussion

In this study, a nearly saturated linkage map was constructed for *Callitris glaucophylla*, serving as a first genetic map for Callitroideae, which is a Southern Hemisphere cypress lineage with great ecological diversity (Enright and Hill 1995) and long evolutionary history (ca. 150 Mya; Mao et al. 2012. Notably, our map building took only 3.5 months including laboratory work, sequencing and data analysis. The resultant map consists of 4,284 markers over 1,033.5 cM and is one of the most comprehensive maps made for any conifer. .For example, some of the more extensive maps (Ritland et al. 2011) include *Pinus*: 2,841 markers (1,651 cM) and 2,466 markers (1,476 cM) in *P. taeda* L. (Martínez-García et al. 2013; Neves et al. 2013), *Picea*: 1,216 markers (1,865 cM) in *P. mariana* (Mill.) x *P. rubens* Sarg. complex (Kang et al. 2011), and *Cryptomeria*: 1,262 markers (1,405 cM) in *C. japonica* (L.f.) D.Don (Moriguchi et al. 2012)]. Recently, genomes of two economically important Pinaceae conifers (*Picea abies* and *Pinus taeda*) have been sequenced (Neale et al. 2014; Nystedt et al. 2013), thereby opening a new avenue for investigating genomic evolution of these conifers. However, for most other conifer families or genera, full genome sequencing is still not affordable because of their enormous genome sizes and complexity. Alternatively, as demonstrated in this study, single tree mapping combined with high-throughput sequencing can be a time- and cost- effective approach to build dense conifer maps for purposes of evolutionary biology, molecular ecology and tree breeding.

Influence of missing and errors in RAD-derived genotype data

Despite these advantages, linkage mapping based on low-cost, high-throughput sequencing can generate a substantial amount of missing data and errors, reflecting a non-uniform distribution of reads over sequenced regions (Beissinger et al. 2013). In high-resolution mapping, even a low frequency of genotyping error (3% and less) appears as double or multiple recombinants, and can reduce the power to order markers and inflate map length (Buetow 1991; Hackett and Broadfoot 2003). In this study, initial marker ordering was performed by MSTmap program, which can attain high accuracy of ordering (Kendall's $\tau > 0.989$) in simulated genotype data when mapping populations whose size ($n = 100$), missing rate ($\gamma = 0.10$) and error rate ($\eta = 0.10$) (Wu et al. 2008) are comparable to our raw genotype data ($n = 74$, $\gamma = 0.10$). Nevertheless, the estimated map length of *C. glaucophylla* was still too large (4,324.9 cM), compared to those generally reported from other conifer studies (Ritland et al. 2011). To deal with this, we subsequently performed error correction and imputation based on ordered markers using Maskov ver 1.01, which decreased map length greatly (by 76.1 %). The map length became closer to the value (1,405 cM) in the other Cupressaceae conifer, *Cryptomeria japonica* (Moriguchi et al. 2012), which was estimated from genotype data with low missing rate ($\gamma = 0.02$; Y. Moriguchi, personal communication), indicating that our map length was inflated likely due to noise in genotype data. Although it was noticed that maps produced through imputation tend to have fewer unique positions and thus lower resolution (Ward et al. 2013), imputation and error correction is thus an indispensable step to obtain a reliable linkage map using RAD-derived genotype data.

As well as factors such as variation in DNA quality, library preparation, sequencing and assembly errors (Pool et al. 2010), choice of restriction enzyme for efficient RAD-sequencing can have great impacts on the level of genotype missing rate. Prior to this mapping study, we tested four rare-cutter enzymes (EcoRI, MseI, NdeI and PstI) in pairs with BglII to screen for the most efficient enzyme species, with which sufficient read depth per population sample can be obtained. After *de novo* assembling reads from 24 samples representing all the regional lineages of the species complex, it was shown that the number of contigs containing SNPs with a missing rate less than 0.1 varied over two orders of magnitude from 494 (per 147,411contigs for MseI library) to 11,617 (per 200,024 contigs for EcoRI library), which validated our use of the EcoRI-BglII pair for the species. Other considerations on enzyme species would involve use of hypomethylation-sensitive enzymes. It is documented that conifer genomes contain high-copy repeat elements including retrotransposons, which represent 70% and 62% of genomes of *P. taeda* and *P. abies*, respectively (Neale et al. 2014; Nystedt et al. 2013). Since those retrotransposon-rich regions of plant genome are generally heavily methylated (Rabinowicz et al. 2005), hypomethylation-sensitive enzymes can be used to establish reduced representation libraries in order to avoid highly repetitive elements (Larsson et al. 2013; Pegadaraju et al. 2013). Recently, Karam et al. (2014) demonstrated a utility of RAD-sequencing with a hypomethylation-sensitive enzyme to enrich gene-rich regions by sequencing *Cedrus atlantica* Manetti, in which 17% of the contigs coding for proteins were included. This has an important implication for RAD-sequencing of large-genome conifers, in which most SNPs are derived from non-coding regions (98.4 % in this study). While extended linkage disequilibrium in non-coding regions of the conifer genome may make it possible to perform association mapping by anonymous RAD-SNPs (Moritsuka et al. 2012), there is no doubt that EST-SNPs concentrated by RAD-sequencing with hypomethylation-sensitive enzymes are more useful for linking genetic polymorphisms to phenotypic/environmental variation.

Spatial heterogeneity of marker density and segregation pattern over linkage groups

High-density linkage mapping studies sometimes detect strikingly marker-rich regions in linkage groups (Chancerel et al. 2013; Studer et al. 2012; Talukder et al. 2014). We also found a similar pattern of non-random marker distribution in *Callitris* linkage groups. Such spatially heterogeneous marker distribution can be explained with respect to variable recombination rates among genomic regions (Petes 2001). Genome-wide surveys of recombination pattern in model plants have shown that regions of high (hot spots) and low (cold spots) recombination rates are distributed along chromosomes (Gaut et al. 2007), and one obvious cold spot is the heterochromatic pericentromere region where recombination is suppressed (Choi et al. 2013; Wu et al. 2003). In such cold spots, recombination rarely takes place, which leads to distorted genetic distances and marker clustering on the genetic map. Spatial association of a recombination cold spot with a pericentromere may be the case for some linkage groups (e.g. LG4 and 6), where distinct peaks of marker counts were detected at centres of linkage groups, although this cannot be confirmed presently for *C. glaucophylla* without a physical map.

Another spatial trend found for the allelic segregation pattern was that genetic markers showing significant distortions were clustered in particular regions of LG6, 7, 8, 10. Considering that the trend is observed at many linked markers and consistent effects of marker position were detected even when the missing rate controlled in GAM modelling, non-biological factors such as a limited number of sampled loci or missing genotypes is unlikely to account for this observation. Instead, gametic or zygotic selection seems to have operated on viability genes (Gillet and Gregorius 1992), leaving significant allelic distortion in surrounding linked markers. Gametic selection can occur at stages from meiosis to fertilization, in which the process of meiosis itself or differential survival ability among gametophytes influences marker segregation, whereas random segregation is disturbed as a consequence of gametic combination relationships [?] in fertilization. For allogamous forest trees, there has been evidence for substantial inbreeding depression in the early stages of life cycles (Isagi et al. 2007; Naito et al. 2005), and zygotic selection has been suggested in the studies that investigated allelic segregation in conifer megagametophytes (Kuang et al. 1999; Siregar and Yunanto 2008). However, as the megagametophyte samples used in this study are derived from open-pollinated mature seeds, it is currently difficult to exclude occurrence of gametic selection or tease apart a possible interplay of gametic and zygotic selection. Future studies using megagametophyte samples from different developmental stages and samples obtained from selfed and outcrossed seeds would be meaningful for testing these hypotheses.

SNP recovery in independent population sequencings and map transferability

Since genetic diversity and differentiation levels greatly vary among genomic regions in conifers (Eckert et al. 2010; Li et al. 2010; Tsumura et al. 2007; Tsumura et al. 2014), any markers linked to genetic maps are important for obtaining a more detailed picture of genomic evolution. In this study, ca. 30% of SNPs (1,257 markers) in the *C. glaucophylla* linkage map were recovered in independent population RAD-sequencing of the same species. This finding indicates that these markers with known map positions could improve our understanding of the evolution of the *C. columellaris* species complex, by extending a population genetic analysis that used only 30 EST-SSR markers (Sakaguchi et al. 2013).

When applying map information to population analyses, however, we should be aware that the mapped markers may suffer from ascertainment bias. Ascertainment bias arises because usually only a small number of samples are used to identify markers, with which genetic polymorphism is maximized. It follows that the selected markers do not represent the polymorphism preserved in whole populations, and estimated the allele frequency spectrum becomes skewed compared to one obtained from genome sequencing (Albrechtsen et al. 2010). In this study, we found that genetic diversity of *C. gracilis* and differentiation between two *Callitris* species was slightly, but significant statistically, , under- or over-estimated at the mapped SNPs in comparison to the randomly sampled. The detected effects on summary statistics were within expectation as the markers were screened from only a single tree, and showed a pattern consistent with that reported in human studies (Clark et al. 2005; Luca et al. 2011). Another source of bias in summary statistics may have been introduced from the RAD-sequencing technique itself. Because RAD-sequencing uses restriction enzymes for preparing a reduced representation library, a polymorphism in a restriction site can result in allele drop-out where a heterozygous sample appears as a homozygote due to a null allele (Arnold et al. 2013; Gautier et al. 2013). Hence, it can also lead to biased estimates of population summary statistics if allelic drop-out tends to occur at higher probability in populations genetically diverged from the population in which the linkage map was constructed.

Increasing the ascertainment sample size can reduce ascertainment bias (Albrechtsen et al. 2010). In the case of linkage mapping, it would be effective to construct linkage maps from multiple populations/species in *C. columellaris* species complex and join them to make a consensus map with common markers segregating in two or more mapping populations. Such a map would include SNP markers that can capture genetic polymorphism with less ascertainment bias when applied to whole population analysis. It is also desirable to include species in linkage mapping that show the highest levels of genetic divergence, as the number of SNPs shared between congenic species are shown to be inversely related to phylogenetic distances (Pavy et al. 2013). Although we need to deal carefully with ascertainment bias, the RAD-SNPs based linkage map is essentially useful in combined with population genomic analyses of *C. columellaris* species complex. Promising applications of the map information will include detection of linkage disequilibrium arising from genetic admixture (Falush et al. 2003) and identification of genomic regions that are associated with particular adaptive traits and show significant divergence due to natural selection (Andrew and Rieseberg 2013; Chutimanitsakun et al. 2011; Slavov et al. 2014). These factors should have played significant roles in speciation and environmental adaptation of the conifer lineage.

Acknowledgements

We are grateful to NSW State Forests, NSW Department of Environment and Climate Change, Queensland Department of Environment and Resource Management, SA Department of Environment and Heritage, Victorian Department of Sustainability and Environment, the Australian Wildlife Conservancy and many private landholders for help with site selection and permission to sample on their land. We thank N. Nakahama and Y. Unno for their assists in genotyping EST-SSRs and a preliminary analysis, M. Yamasaki for his insightful discussions on statistical analysis, and Y. Moriguchi for kindly providing unpublished data. Funding was provided by Japan Society for the Promotion of Science Grant-in-Aid for JSPS Fellows (13J06059), Grant-in-Aid for Scientific Research (JSPS KAKENHI 24248028 and 26850098), and the Environment Research and Technology Development Fund of the Ministry of the Environment (4-1403).

Table 1

Statistics for of the linage map of *Callitris glaucophylla*.

Linkage group	No. of Markers			Length (cM)		No. of unique position		Mean marker interval (cM)	Mean interval of unique position (cM)
	total	SNPs (EST SNPs)	EST-SSR	without imputation	with imputation	without imputation	with imputation		
LG1	486	486 (4)	0	400.6	109.2	127	59	0.22	1.85
LG2	456	456 (10)	0	425.2	104.3	132	63	0.23	1.66
LG3	468	467 (6)	1	501.4	98.6	160	59	0.21	1.67
LG4	359	358 (1)	1	483.1	95.7	128	46	0.27	2.08
LG5	358	358 (6)	0	424.2	95.6	125	56	0.27	1.71
LG6	351	351 (7)	0	395.5	94.3	117	53	0.27	1.78
LG7	358	358 (8)	0	376.3	94.2	113	49	0.26	1.92
LG8	352	352 (8)	0	309.6	91.5	101	50	0.26	1.83
LG9	344	343 (8)	1	273.0	84.6	86	47	0.25	1.80
LG10	361	361 (3)	0	345.8	84.3	113	53	0.23	1.59
LG11	391	389 (6)	2	390.3	81.2	123	50	0.21	1.62
	4,284	4,279 (67)	5	4,324.9	1,033.5	1,325	585	0.24	1.77

Table 2

Comparison of summary statistics of genetic diversity between the mapped SNP markers to *Callitris glaucophylla* linkage map and the randomly sampled SNP markers.

Summary statistics	Obs. mean of the mapped markers	Mean (99% points) of the sampled markers	Absolute difference in mean values	*P* value
*H*s (*C. glaucophylla*)	0.153	0.128 (0.117, 0.139)	+0.025	< 0.01
*H*s (*C. gracilis*)	0.098	0.116 (0.105, 0.127)	-0.018	< 0.01
F_{ST}	0.064	0.046 (0.037, 0.055)	+0.018	< 0.01

Figure 1

Genome-wide distribution of marker segregation pattern, represented by *P* value in Chi-squared tests against marker position on the eleven linkage groups. Smoothing curves using the loess method with 95% CI are also shown. The dropped markers showing significant deviation from the expected 1:1 segregation ratio (*P* < 0.05) are indicated by black points.

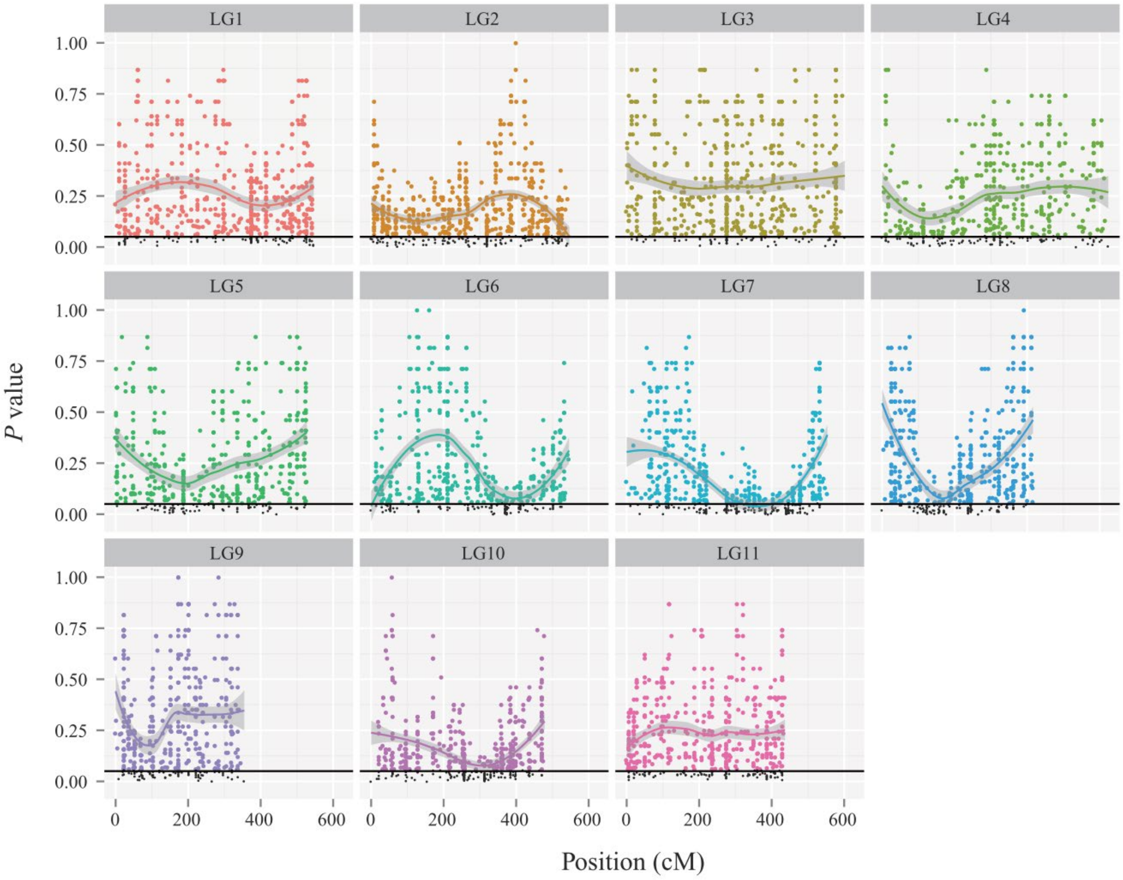


Figure 2

Linkage map for *Callitris glaucophylla* composed of 4,284 genetic markers of 4,279 SNPs (including 67 EST-SNPs) and 5 EST-SSRs. Unique positions estimated after data imputation procedure are shown on the eleven linkage groups, which correspond to the haploid set of 11 chromosomes.

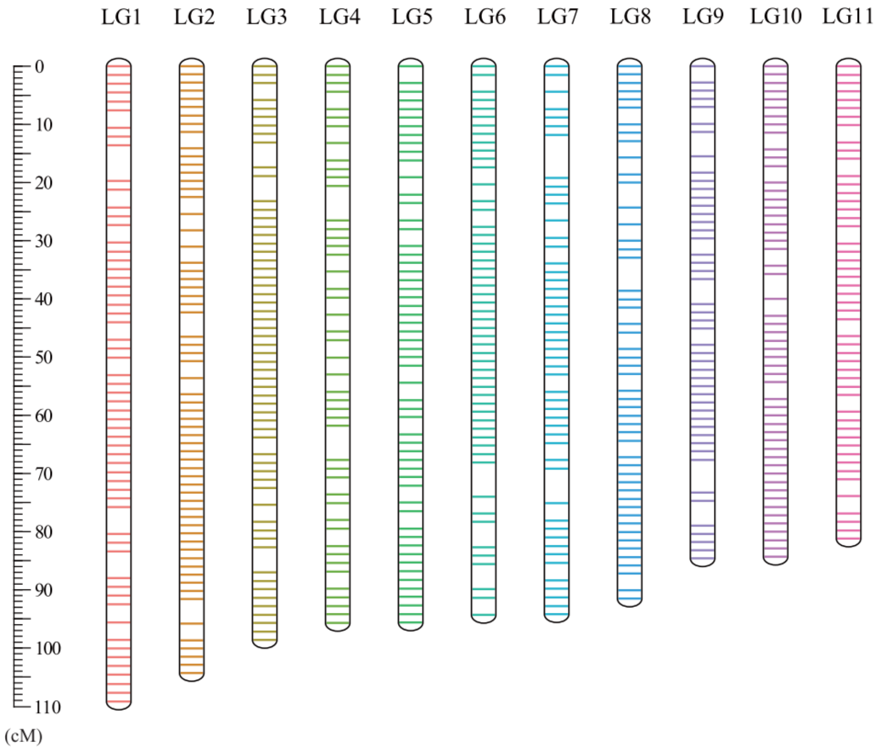
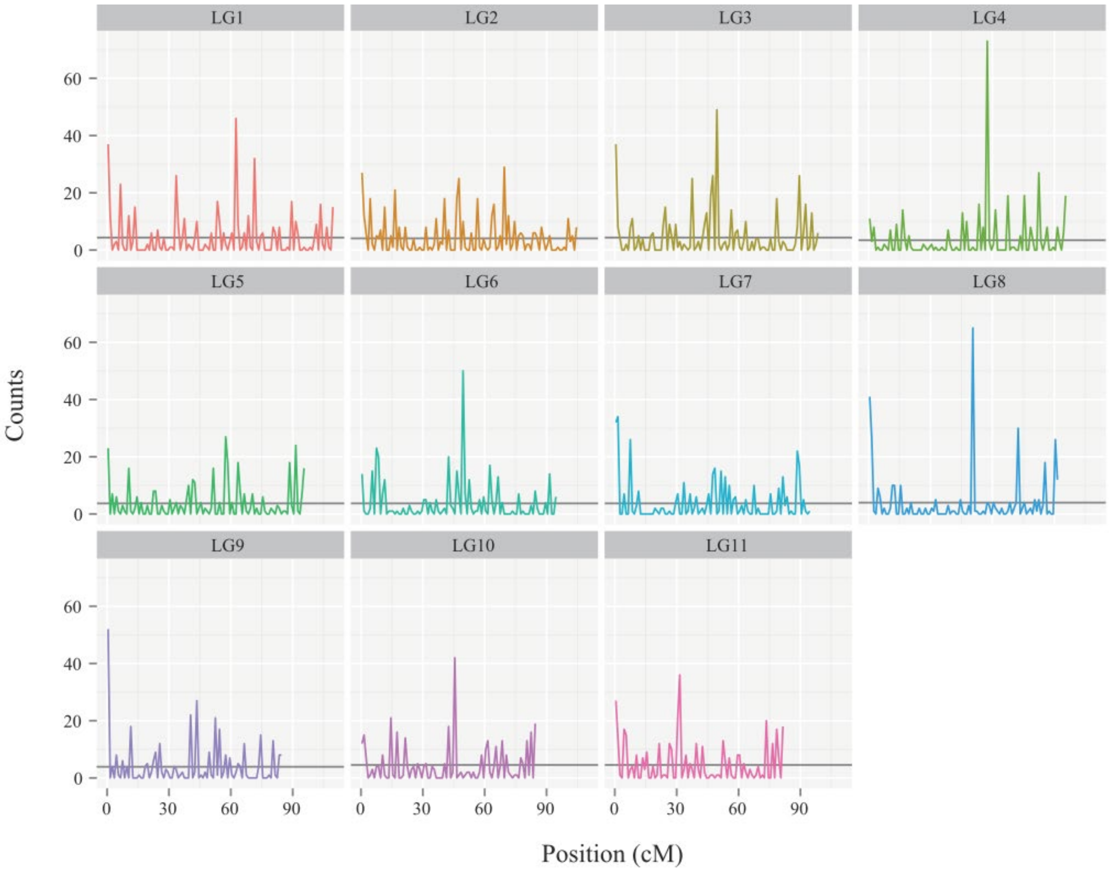


Figure 3

Spatial distribution of marker density as evaluated with a bin width of 1 cM. Marker counts are plotted against marker positions (mid position of each bin). Horizontal lines shows mean marker counts across linkage groups.



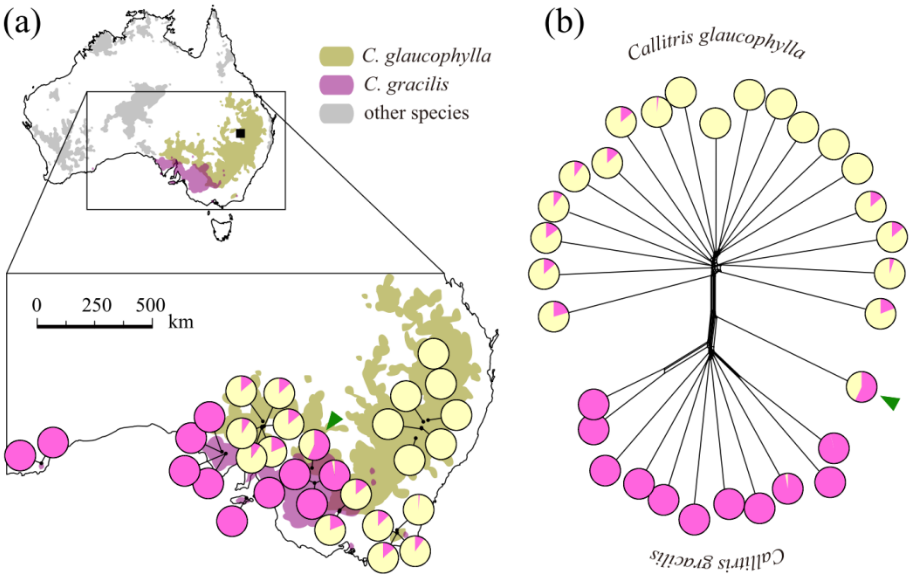
Supplementary table 1

Results of GAM analysis of genetic marker segregation for each linkage group, as a function of missing rate in genotype data and map position.

Linkage group	AIC	Missing rate		Map position	
		F value	*P* value	*F* value	*P* value
LG1	-925.3	1,347.6	< 0.01	19.0	< 0.01
LG2	-929.4	475.6	< 0.01	15.1	< 0.01
LG3	-1,018.3	2,592.1	< 0.01	0.2	0.61
LG4	-717.4	944.4	< 0.01	78.8	< 0.01
LG5	-785.1	1,446.8	< 0.01	30.5	< 0.01
LG6	-617.1	763.2	< 0.01	42.7	< 0.01
LG7	-756.4	524.8	< 0.01	79.4	< 0.01
LG8	-652.6	841.1	< 0.01	0.29	0.58
LG9	-542	1,037.6	< 0.01	26.9	< 0.01
LG10	-833.5	354.6	< 0.01	4.2	0.04
LG11	-860.8	1,115.5	< 0.01	5.6	0.02

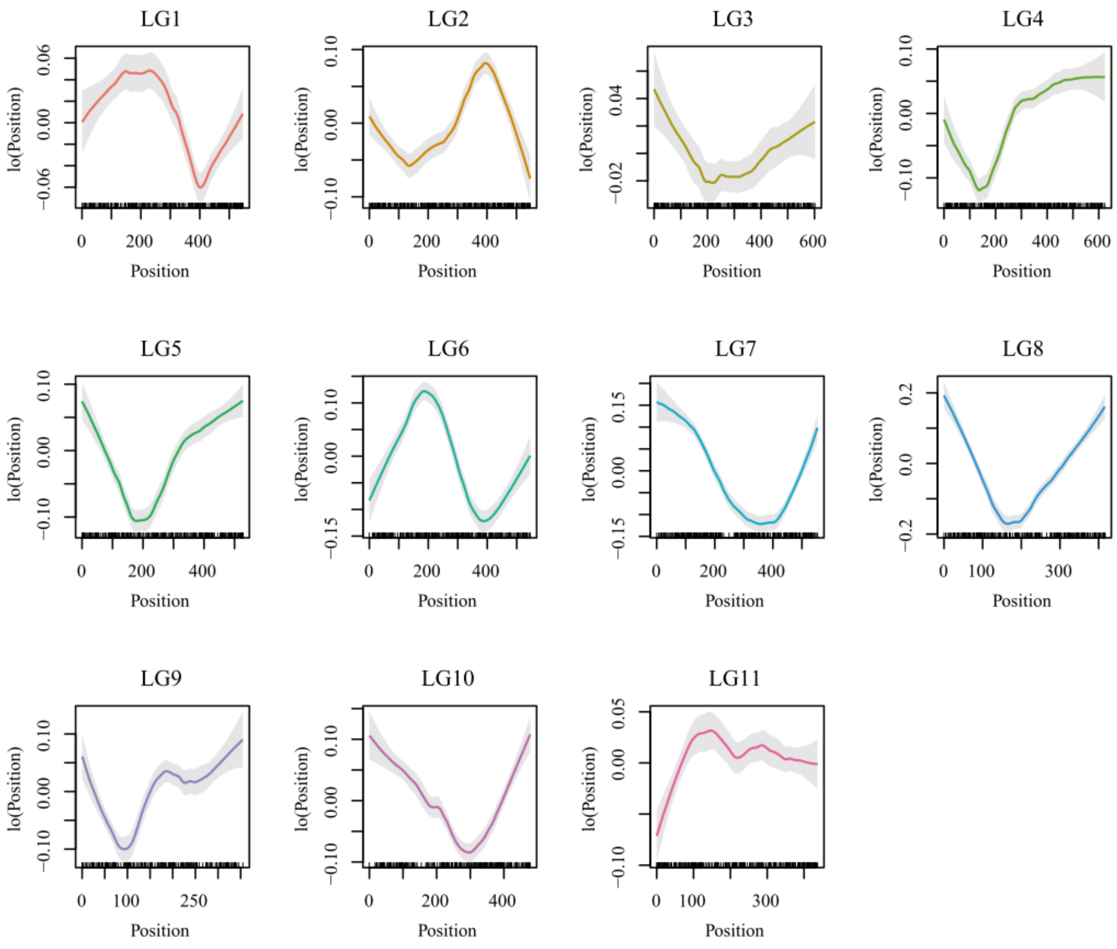
Supplementary figure 1

(a) Distribution of *Callitris columellaris* species complex. Ranges for *C. glaucophylla* (GD lineage) and *C. gracilis* are colored by ochre and pink, respectively. The locality where the seed samples for linkage map construction were collected is indicated by a black square on the smaller map. Superimposed are pie charts illustrating the two genetic clusters, corresponding to the two species, which were detected by STRUCTURE analysis. (b) A split network for 31 individuals of *C. glaucophylla* and *C. gracilis* analysed in this study. Genetic membership estimated from STRUCTURE analysis is placed on the tips. A genetically intermediate individual is indicated by a green triangle.



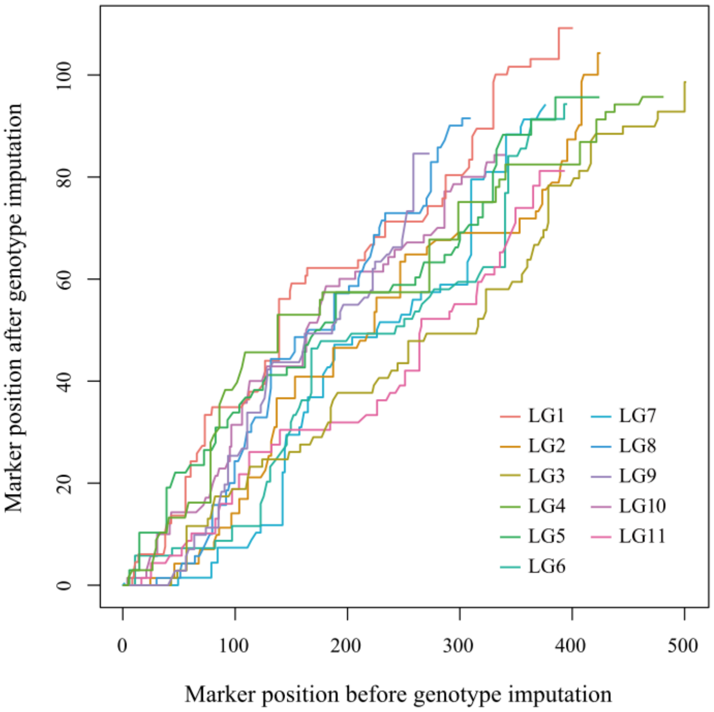
Supplementary figure 2

Graphical results of GAM analysis of genetic marker segregation. Partial effects of genomic position are shown for each linkage group, expressed as fitted loess functions with 95% boot-strapped confidence intervals (gray in color). Ticks in the x-axis represent the location of observations along the predictor.



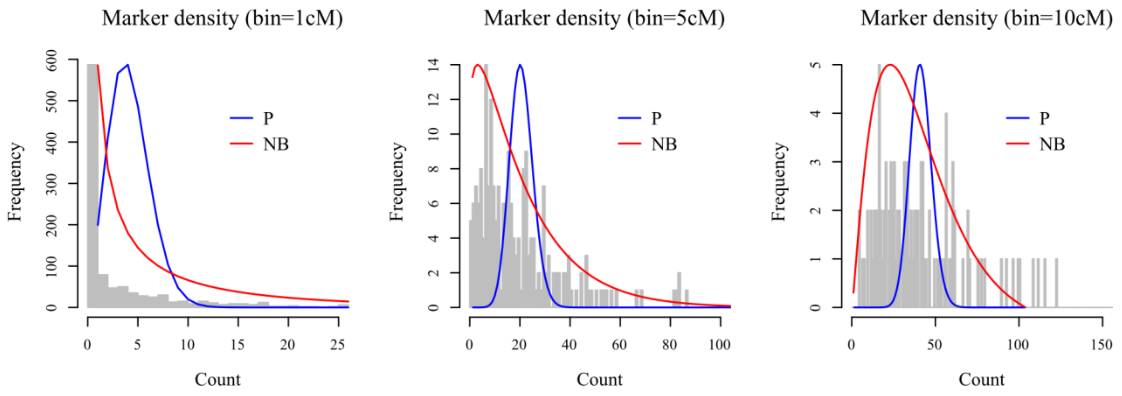
Supplementary figure 3

Relationship between marker positions estimated from genotype data sets with and without imputation and error correction procedures.



Supplementary figure 4

Distribution of genetic marker density calculated based on different bin widths (1, 5, 10 cM). Expected probability curves are estimated using a Poisson distribution (blue) and a negative binomial distribution (red).



References

Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in SNP chips affect measures of population divergence Mol Biol Evol:msq148

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool J Mol Biol 215:403-410

Andrew RL, Rieseberg LH (2013) Divergence is focused on few genomic regions early in speciation: incipient speciation of sunflower ecotypes Evolution 67:2468-2482

Arnold B, Corbett‐Detig R, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling Mol Ecol 22:3179-3190

Baird NA et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers PLoS One 3:e3376

Beissinger TM et al. (2013) Marker density and read depth for genotyping populations using genotyping-by-sequencing Genetics 193:1073-1081

Bishop D, Cannings C, Skolnick M, Williamson J, Weir B (1983) The number of polymorphic DNA clones required to map the human genome. In: Statical analysis of DNA sequence data. Marcel-Dekker, New York, pp 181-200

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data Bioinformatics 30:2114-2120 doi:10.1093/bioinformatics/btu170

Bowman D, Harris S (1995) Conifers of Australia's dry forests and open woodlands. In: Enright N, Hill R (eds) Ecology of Southern Conifers. University of Melbourne, Melbourne, pp 252-270

Bryant D, Moulton V (2004) Neighbor-Net: An agglomerative method for the construction of phylogenetic networks Mol Biol Evol 21:255-265

Buetow KH (1991) Influence of aberrant observations on high-resolution linkage analysis outcomes Am J Hum Genet 49:985

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences G3-Genes Genomes Genetics 1:171-182 doi:10.1534/g3.111.000240

Chancerel E et al. (2013) High-density linkage mapping in a pine tree reveals a genomic region associated with inbreeding depression and provides clues to the extent and distribution of meiotic recombination BMC Biology 11:50

Choi K et al. (2013) *Arabidopsis* meiotic crossover hot spots overlap with H2A. Z nucleosomes at gene promoters Nature genetics 45:1327-1336

Chutimanitsakun Y et al. (2011) Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley Bmc Genomics 12:4

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism Genome Res 15:1496-1502

Debreczy Z, Racz I (2006) Conifers around the world vol II. DendroPress Limited, Budapest

Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, Gonzalez-Martinez SC, Neale DB (2010) Patterns of Population Structure and Environmental Associations to Aridity Across the Range of Loblolly Pine (*Pinus taeda* L., Pinaceae) Genetics 185:969-982 doi:10.1534/genetics.110.115543

Enright N, Hill R (1995) Ecology of Southern Conifers. University of Melbourne, Melbourne

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

Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK (2007) Recombination: an underappreciated factor in the evolution of plant genomes Nat Rev Genet 8:77-84

Gautier M et al. (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations Mol Ecol 22:3165-3178

Gernandt D, Willyard A, Syring JV, Liston A (2011) The Conifers (Pinophyta). In: Plomion C, Bousquet J, Kole C (eds) Genetics, Genomics and Breeding of Conifers. CRC Press, pp 1-39

Gillet E, Gregorius HR (1992) What can be inferred from open-pollination progenies about the source of observed segregation distortion? - a case study in Castanea sativa Mill. Silvae Genetica 41:82-87

González-Martínez S et al. (2011) Patterns of nucleotide diversity and association mapping. In: Genetics, genomics and breeding of conifers. pp 239-275

Gouded J (2014) hierfstat: Estimation and tests of hierarchical F-statistics. R-package.

Guo F et al. (2015) Construction of a SNP-based high-density genetic map for pummelo using RAD sequencing Tree Genetics & Genomes 11:1-11 doi:10.1007/s11295-014-0831-0

Hackett C, Broadfoot L (2003) Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps Heredity 90:33-38

Hill RS, Brodribb TJ (1999) Turner Review No. 2 - Southern conifers in time and space Aust J Bot 47:639-696

Huson DH (1998) SplitsTree: analyzing and visualizing evolutionary data Bioinformatics 14:68-73

Isagi Y, Saito D, Kawaguchi H, Tateno R, Watanabe S (2007) Effective pollen dispersal is enhanced by the genetic structure of an *Aesculus turbinata* population J Ecol 95:983-990 doi:10.1111/j.1365-2745.2007.01272.x

Iwata H, Ninomiya S (2006) AntMap: Constructing genetic linkage maps using an ant colony optimization algorithm Breed Sci 56:371-377 doi:10.1270/jsbbs.56.371

Kakioka R, Kokita T, Kumada H, Watanabe K, Okuda N (2013) A RAD-based linkage map and comparative genomics in the gudgeons (genus *Gnathopogon*, Cyprinidae) Bmc Genomics 14:32

Kang B-Y, Major JE, Rajora OP (2011) A high-density genetic linkage map of a black spruce (*Picea* *mariana*)× red spruce (*Picea rubens*) interspecific hybrid Genome 54:128-143

Karam MJ, Lefèvre F, Dagher‐Kharrat MB, Pinosio S, Vendramin G (2014) Genomic exploration and molecular marker development in a large and complex conifer genome using RADseq and mRNAseq Mol Ecol Resour

Kosambi DD (1943) The estimation of map distances from recombination values AnnEug 12:172-175

Kuang H, Richardson T, Carson S, Wilcox P, Bongarten B (1999) Genetic analysis of inbreeding depression in plus tree 850.55 of Pinus radiata D. Don. I. Genetic map with distorted markers Theor Appl Genet 98:697-703

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2 Nature Methods 9:357-U354 doi:10.1038/nmeth.1923

Larsson H, De Paoli E, Morgante M, Lascoux M, Gyllenstrand N (2013) The Hypomethylated Partial Restriction (HMPR) method reduces the repetitive content of genomic libraries in Norway spruce (*Picea abies*) Tree Genetics & Genomes 9:601-612

Li S, Chen Y, Gao H, Yin T (2010) Potential chromosomal introgression barriers revealed by linkage analysis in a hybrid of *Pinus massoniana* and *P. hwangshanensis* Bmc Plant Biology 10 doi:37
10.1186/1471-2229-10-37

Luca F, Hudson RR, Witonsky DB, Di Rienzo A (2011) A reduced representation approach to population genetic analyses and applications to human evolution Genome Res 21:1087-1098

Mao KS et al. (2012) Distribution of living Cupressaceae reflects the breakup of Pangea Proc Natl Acad Sci U S A 109:7793-7798 doi:10.1073/pnas.1114319109

Martínez-García PJ, Stevens KA, Wegrzyn JL, Liechty J, Crepeau M, Langley CH, Neale DB (2013) Combination of multipoint maximum likelihood (MML) and regression mapping algorithms to construct a high-density genetic linkage map for loblolly pine (*Pinus taeda* L.) Tree Genetics & Genomes 9:1529-1535

Moriguchi Y et al. (2012) The construction of a high-density linkage map for identifying SNP markers that are tightly linked to a nuclear-recessive major gene for male sterility in *Cryptomeria japonica* D. Don Bmc Genomics 13:95

Moritsuka E, Hisataka Y, Tamura M, Uchiyama K, Watanabe A, Tsumura Y, Tachida H (2012) Extended linkage disequilibrium in noncoding regions in a conifer, *Cryptomeria japonica* Genetics 190:1145-1148

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA Nucleic Acids Res 8:4321-4325

Naito Y et al. (2005) Selfing and inbreeding depression in seeds and seedlings of *Neobalanocarpus heimii* (Dipterocarpaceae) J Plant Res 118:423-430

Neale DB et al. (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies Genome biology 15:R59

Neves LG, Davis JM, Barbazuk WB, Kirst M (2013) A high-density gene map of loblolly pine (*Pinus taeda* L.) based on exome sequence capture genotyping G3: Genes| Genomes| Genetics:g3. 113.008714

Nystedt B et al. (2013) The Norway spruce genome sequence and conifer genome evolution Nature 497:579-584 doi:10.1038/nature12211

Ohri D, Khoshoo T (1986) Genome size in gymnosperms Plant Syst Evol 153:119-132

Pavy N et al. (2013) Development of high ‐ density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners Mol Ecol Resour 13:324-336

Pavy N, Pelgas B, Laroche J, Rigault P, Isabel N, Bousquet J (2012) A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers BMC Biol 10:84

Pegadaraju V, Nipper R, Hulke B, Qi L, Schultz Q (2013) De novo sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach Bmc Genomics 14:556

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest RADseq: An Inexpensive Method for *De Novo* SNP Discovery and Genotyping in Model and Non-Model Species PLoS One 7:e37135 doi:10.1371/journal.pone.0037135

Petes TD (2001) Meiotic recombination hot spots and cold spots Nat Rev Genet 2:360-369

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation Genome Res 20:291-300

Prior LD, McCaw WL, Grierson PF, Murphy BP, Bowman DMJS (2011) Population structures of the widespread Australian conifer *Callitris columellaris* are a bio-indicator of continental environmental change For Ecol Manage 262:252-262 doi:10.1016/j.foreco.2011.03.030

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

Purcell S et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses Am J Hum Genet 81:559-575 doi:10.1086/519795

R Development Core Team (2014) R version 3.1.0: A language and environment for statistical computing

Rabinowicz PD et al. (2005) Differential methylation of genes and repeats in land plants Genome Res 15:1431-1440

Ritland K, Krutovsky KV, Tsumura Y, Pelgas B, Isabel N, Bousquet J (2011) Genetic mapping in conifers. In: Genetics, genomics and breeding of conifers. CRC Press, pp 196-238

Sakaguchi S, Bowman DMJS, Prior LD, Crisp MD, Linde CC, Tsumura Y, Isagi Y (2013) Climate, not Aboriginal landscape burning, controlled the historical demography and distribution of fire-sensitive conifer populations across Australia Proceedings of the Royal Society B: Biological Sciences 280 doi:10.1098/rspb.2013.2182

Sakaguchi S et al. (2011) Isolation and characterization of 52 polymorphic EST-SSR markers for *Callitris* *columellaris* (Cupressaceae) Am J Bot 98:E363-E368 doi:10.3732/ajb.1100276

Siregar IZ, Yunanto T (2008) Inference on the Possible Causes of Segregation Distortion from Open Pollination Progenies of Merkus Pine (*Pinus merkusii*) HAYATI Journal of Biosciences 15:173

Slavov GT et al. (2014) Genome-wide association studies and prediction of 17 traits related to phenology, biomass and cell wall composition in the energy grass *Miscanthus sinensis* New Phytol 201:1227-1239 doi:10.1111/nph.12621

Studer B et al. (2012) A transcriptome map of perennial ryegrass (*Lolium perenne* L.) Bmc Genomics 13:140

Talukder ZI, Gong L, Hulke BS, Pegadaraju V, Song Q, Schultz Q, Qi L (2014) A high-density SNP map of sunflower derived from RAD-sequencing facilitating fine-mapping of the rust resistance gene R12 PLoS One 9:e98628

Tsumura Y, Kado T, Takahashi T, Tani N, Ujino-Ihara T, Iwata H (2007) Genome scan to detect genetic structure and adaptive genes of natural populations of *Cryptomeria japonica* Genetics 176:2393-2403

Tsumura Y, Uchiyama K, Moriguchi Y, Kimura MK, Ueno S, Ujino-Ihara T (2014) Genetic Differentiation and Evolutionary Adaptation in *Cryptomeria japonica* G3: Genes| Genomes| Genetics 4:2389-2402

Tulsieram LK, Glaubitz JC, Kiss G, Carlson JE (1992) Single tree genetic linkage mapping in conifers using haploid DNA from megagametophytes Nature Biotechnology 10:686-690

Voorrips RE (2002) MapChart: Software for the graphical presentation of linkage maps and QTLs J Hered 93:77-78 doi:10.1093/jhered/93.1.77

Ward JA et al. (2013) Saturated linkage map construction in *Rubus idaeus* using genotyping by sequencing and genome-independent imputation BMC Genomics 14 doi:2

10.1186/1471-2164-14-2

Wu J et al. (2014) High-density genetic linkage map construction and identification of fruit-related QTLs in pear using SNP and SSR markers Journal of Experimental Botany doi:10.1093/jxb/eru311

Wu J et al. (2003) Physical maps and recombination frequency of six rice chromosomes The Plant Journal 36:720-730

Wu Y, Bhat PR, Close TJ, Lonardi S (2008) Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph PLoS genetics 4:e1000212

Construction of RAD reference assembly of *Callitris columellaris* species complex

DNA samples of 10 individuals (representing 9 regions in Australia and 5 morphospecies of the species complex) were digested by with EcoRI and BglII, and prepared for RAD-sequencing using the same protocol as taken in the linkage mapping analysis. The sample information of samples are summarised as follows.

Sample No.	Species	Region	Population	Latitude	Longitude	Restriction enzymes	No. reads used for *de novo* assembly
1	*C. intratropica*	Kimberley	CintK6	–16° 56'	126° 13'	EcoRI – BglII	993,398
2	*C. intratropica*	The Top End	CintTE5	–13° 13'	132° 39'	EcoRI – BglII	614,007
4	*C. intratropica*	Cape York Peninsula	CintNQ1	–17° 08'	145° 38'	EcoRI – BglII	824,954
5	*C. columellaris*	Central Eastern Coast	CcolQC	–28° 48'	153° 22'	EcoRI – BglII	1,679,270
7	*C. gracilis*	Murray Basin	CgrHKA	–35° 24'	142° 23'	EcoRI – BglII	1,340,960
10	*C. glaucophylla*	Great Dividing Range	CglSQ3	–27° 22'	149° 27'	EcoRI – BglII	1,530,791
16	*C. glaucophylla*	Pilbara	CglHR1	–24° 57'	118° 51'	EcoRI – BglII	2,336,438
14	*C. glaucophylla*	Central Australia	CglCA1	–23° 03'	132° 39'	EcoRI – BglII	1,255,210
19	*C. glaucophylla*	Southwest	CglWA1	–31° 20'	121° 19'	EcoRI – BglII	1,340,860
23	*C. verrucosa*	Southwest	CverLK	–34° 54'	119° 07'	EcoRI – BglII	809,162

The total number of cleaned reads generated by a Miseq sequencer (Illumina, San Diego, USA) was 12,725,050 (ranging from min. 614,007 to max. 2,336,438 per sample). A RAD reference for *Callitris columellaris* species complex was constructed by de novo assembling the reads using CLC Genomics Workbench 7.5.1 (CLC bio, Aarhus, Denmark) (parameter used: mismatch cost 3, insertion and deletion cost 2, length fraction 0.5, similarity fraction 0.9), which resulted in 392,320 contigs with N50 length of 163 bp.