Article (refereed)

Wachowiak, W.; **Salmela, M.J.**; Ennos, R.A.; Iason, G.; **Cavers, S.**. 2011 High genetic diversity at the extreme range edge: nucleotide variation at nuclear loci in Scots pine (Pinus sylvestris L.) in Scotland. *Heredity*, 106. 775-787. 10.1038/hdy.2010.118

Title: High genetic diversity at the extreme range edge: nucleotide variation at nuclear loci in Scots pine (*Pinus sylvestris* L.) in Scotland

Authors: Witold Wachowiak[1,2], Matti J. Salmela[1,3], Richard A. Ennos[3], Glenn Iason[4], Stephen Cavers[1]

[1] Centre for Ecology and Hydrology Edinburgh, Bush Estate, Penicuik, Midlothian EH26 0QB, UK

[2] Institute of Dendrology, Polish Academy of Sciences, Parkowa 5, 62-035 Kórnik, Poland

[3] Institute of Evolutionary Biology, School of Biological Sciences, Ashworth Laboratories, University of Edinburgh, Edinburgh EH9 3JT, UK

[4] Macaulay Land Use Research Institute, Craigiebuckler, Aberdeen AB15 8QH, UK

Corresponding author: Stephen Cavers (scav@ceh.ac.uk) Centre for Ecology and Hydrology Edinburgh, Bush Estate, Penicuik, Midlothian EH26 0QB, UK. Phone: +44 (0) 131 4458552, Fax: +44 (0) 131 4453943

Keywords: adaptation, bottleneck, nucleotide diversity, population differentiation, linkage disequilibrium, recolonisation

Running title: Nucleotide diversity in Scots pine

Abstract:

Nucleotide polymorphism at twelve nuclear loci was studied in Scots pine populations across an environmental gradient in Scotland, to evaluate the impacts of demographic history and selection on genetic diversity. At eight loci, diversity patterns were compared between Scottish and continental European populations. At these loci, a similar level of diversity (θ_{sil}=~0.01) was found in Scottish vs. mainland European populations contrary to expectations for recent colonisation, however less rapid decay of linkage disequilibrium was observed in the former (ρ=0.0086±0.0009, ρ=0.0245±0.0022 respectively). Scottish populations also showed a deficit of rare nucleotide variants (multilocus Tajima's *D*=0.316 vs. *D*=-0.379) and differed significantly from mainland populations in allelic frequency and/or haplotype structure at several loci. Within Scotland, western populations showed slightly reduced nucleotide diversity (π_{tot}=0.0068) compared to those from the south and east (0.0079 and 0.0083, respectively) and about three times higher recombination to diversity ratio (ρ / θ=0.71 versus 0.15 and 0.18, respectively). By comparison with results from coalescent simulations, the observed allelic frequency spectrum in the western populations was compatible with a relatively recent bottleneck (0.00175 x $4N_e$ generations) that reduced the population to about 2% of the present size. However heterogeneity in the allelic frequency distribution among geographical regions in Scotland suggests that subsequent admixture of populations with different demographic histories may also have played a role.

Introduction

Nucleotide polymorphism is influenced by several factors including mutation, migration, selection and random genetic drift. In tree species, the current increase in sequence data gathered from nuclear gene loci has been driven mostly by the search for the molecular signature of natural selection (Achaz, 2009; Neale and Ingvarsson, 2008; Savolainen and Pyhäjärvi, 2007). Selection can leave its traces as deviations from neutrality in the level of nucleotide diversity, allele frequency distribution or correlation between polymorphic sites (linkage disequilibrium) (Achaz, 2009). However, the capability to detect selection at individual loci is heavily dependent on the assumptions of the neutral model (e.g. constant long term population size, random mating), the strength of, and time since, selection and the number of loci involved (and their relative effect) in selectively-influenced traits (Wright and Gaut, 2005). Therefore, prior to testing for selection, datasets must be evaluated for violations of neutral model assumptions. Such processes, e.g. historical changes in population size and distribution, may drive deviations from neutrality that mimic the effect of selection. However, these effects are expected to be genome-wide and so can be distinguished from selective influences by simultaneous assessment of data from multiple loci. Although the patterns of variation in the majority of nuclear loci studied to date obey neutral expectations and the signature of selection has been elusive (Savolainen and Pyhäjärvi, 2007) polymorphisms at nuclear loci provide highly valuable insights into evolutionary history (Heuertz *et al*, 2006; Pyhäjärvi *et al*, 2007).

All northern European tree populations have experienced substantial historical changes in distribution. For example, palynological and phylogeographic data indicate that during the last glacial maximum (25-18 000 years ago (ya)), most species were confined to the southern peninsulas (Iberia, Italy and the Balkans) and some parts of eastern and central Europe (Cheddadi *et al*, 2006; Pyhäjärvi *et al*, 2008; Willis and van Andel, 2004) and only reached their most northerly limits around 9000 ya. The recolonization history of forest trees, accompanied by adaptation to local environments, has potentially influenced the pattern of nucleotide diversity both among locally adapted populations and between range edge populations and putative refugial populations. In theory, population bottlenecks reduce nucleotide diversity in range-edge populations relative to that in source populations, although this is dependent on the timing and severity of the bottleneck. In contrast, admixture of populations due, for example, to recolonization from different refugia, may increase diversity (Petit~~i~~ et al. 2003). However, recent studies in continental European populations of Scots pine (Pyhäjärvi *et al*, 2007) and Norway spruce (Heuertz *et al*, 2006) found little evidence at the nucleotide level for the effects of recent (post-glacial) population size changes during migration and suggested bottlenecks in the mid-to-late Pleistocene. In addition, similar to other predominantly outcrossing tree species with highly efficient long distance gene flow via pollen (Hamrick *et al*, 1992), neutral genetic differentiation between Scots pine populations is low. For instance, marginal population differentiation was reported for neutral markers between Finnish populations (Karhu *et al*, 1996), between Scandinavian and eastern parts of the range (Wang *et al*, 1991) and, at several candidate gene loci for growth phenology and cold tolerance, among populations along a latitudinal cline in continental Europe (Dvornyk *et*

al, 2002; García-Gil *et al*, 2003). The large population sizes of forest trees and capability for maintenance of high levels of genetic variation within populations seems to further buffer against rapid changes in genetic diversity, but causes difficulties in detection of recent demographic processes. If the migrations following the most recent glaciations are to have left any signature at all in contemporary populations of forest trees, it seems likely to be detectable only where populations have experienced severe bottlenecks or became rapidly isolated.

In Scotland, Scots pine (*Pinus sylvestris* L.*)* is at the extreme north-western edge of its vast distribution, which reaches across Europe and Asia and is the largest of any pine species (Critchfield and Little, 1965). Pines first colonized the land that became the British Isles about 10 000 ya, at around the time that Ireland became isolated, and reached northern Scotland by about 9000 years ago (Huntley and Birks, 1983; Svendsen *et al*, 1999). According to fossil data in Scotland, pine first appeared in the Wester Ross region in the northwest, and then shortly afterwards in the Cairngorms in the east (Birks, 1989). The subsequent formation of the English Channel (c.6000 ya) and competition from broadleaved species in southern Britain left Scottish pinewoods physically separated by at least 500 km from mainland populations in continental Europe. Nowadays, native pinewoods in Scotland cover about 18 000 hectares, in 84 differently-sized fragments patchily distributed within a ~ 200 x 200 km area across significant environmental gradients in altitude, soil type, growing season length and annual rainfall mainly in the east-west direction (e.g. annual rainfall varies from 700 to 3000 mm across 160 km) (Mason *et al*, 2004). Small-scale provenance experiments have shown genetic variation

between Scottish populations from different locations, e.g. in root frost hardiness and growth in seedlings (Perks and McKay, 1997) and differentiation among populations at several quantitative traits (Perks and Ennos, 1999). There is reasonable evidence from pollen (Birks, 1989), allozymes, monoterpenes and *mt*DNA (Kinloch *et al*, 1986; Sinclair *et al*, 1998) suggesting a west/east population subdivision within Scotland and that populations from these regions may have different origins (Ballantyne and Harris, 1994; Bennett, 1995). Given the iconic status of Scots pine in Scotland and the severe fragmentation of the population, there is considerable interest in evaluating its population history.

In this study, we focus on the Scottish Scots pine population as a unique and isolated oceanic fragment at the northwest extreme of the distribution to assess whether recent demographic processes have influenced patterns of nucleotide variation. We analysed patterns of nucleotide diversity, allele frequency and linkage disequilibrium in a multilocus nuclear gene dataset in samples gathered from multiple locations within putatively divergent regions within Scotland and compared our data to those from samples from northern and central Europe, Turkey and Spain. Using this data and coalescent simulation analysis, we aimed to assess whether Scottish populations show the molecular signature of demographic history and the extent to which they are differentiated from those in continental Europe.

Materials and Methods

Sampling and DNA extraction

Seed samples from 21 locations in Scotland were included in the study (Figure 1.). The trees were sampled across an environmental gradient related to differences in altitude, length of growing season, annual rainfall and average mean temperature in winter (Supplementary Table S1). Cones were collected from mature trees in recognised old-growth Scots pine forest; at these sites trees are typically over 150 years old and often much older (Steven and Carlisle, 1959). Trees were separated by at least 50 m to minimise sampling of closely related individuals. Sampling included the seven currently adopted seed zones of the species in Scotland, from each of which 3 locations were sampled, 2 individuals per location.

For most of the between-population analyses the samples were grouped according to climatic characteristics into three geographical locations – western, southern and eastern, represented by eighteen, twelve and twelve individuals, respectively (Figure 1, Supplementary Table S1). The western group has the lowest mean altitude (~142m), the longest growing season (~ 240 days), highest mean temperature in winter (~ 2^0C) and high annual rainfall (~ 2000mm). The eastern group has the highest mean altitude (~372m), the shortest growing season (~175 days), and is the coldest (-0.1^0C) and driest (~1050mm) part of the distribution, whilst southern group was intermediate between these extremes except for annual rainfall (~2130mm). Field trials have demonstrated genetic differences in phenology and growth rate among provenances originating within these groups (Perks and Ennos, 1999).

Genomic DNA was extracted from haploid megagametophyte, maternal tissue which surrounds the embryo in the seed. As DNA samples were haploid, the haplotypes

could be determined by direct sequencing. In total, 42 DNA extracts were prepared, representing two different trees from each location. Seeds were germinated for a few days in moisturized petri dishes and then extracted following a standard CTAB (cetyltrimethylammonium bromide) protocol with addition of PVP to 1% concentration in the lysis buffer.

Loci studied

In total, sixteen nuclear loci were analysed. This included several dehydrin genes that were identified in expression studies in Scots pine (Joosen *et al*, 2006). Based on the number and position of the conserved segments (Close 1997), we analysed the class SK4 of dehydrins (*dhn1*), SK2 (*dhn2*) and a group of K2 genes (*dhn3* and *dhn7*). We analysed also SK type of dehydrin upregulated by water stress in *Pinus taeda* roots (Eveno *et al*, 2007) and a putative dehydrin (*dhy-like*) described for Scots pine (Pyhäjärvi *et al*, 2007). Other loci described in more detail in original papers include abscissic acid responsive protein (*abaR)* (Wachowiak *et al*, 2009); early response to dehydration 3 protein (*erd3*), abscissic acid, water dehydrative stress and ripening induced gene family members 1 and 3 (*lp3-1, lp3-3*), Caffeoyl CoA *O*-methyltransferase (*ccoaomt*), putative arabinogalactan/ proline-rich protein (*PR-AGP4-1*) and putative arabinogalactan/ glycin-rich protein (*grp3*) (Eveno *et al*, 2007); ABI3-interacting protein 2 (*a3ip2*), alcohol dehydrogenase C (*adhC*) and chalcone synthase (*chcs*) (Pyhäjärvi *et al*, 2007).

In previous work, ten loci (*dhn1,2,3,7*, *dhy-like*, *dhy2PP*, *abaR*, *a3ip2, adhC, chcs*) were analysed in Scots pine from the continental European range including fifteen samples from Northern Europe (populations from Northern and Southern Finland and

Sweden), fifteen from Central Europe (Poland, Austria and France), and five from each of Turkey and Spain (Pyhäjärvi *et al*, 2007; Wachowiak *et al*, 2009). The reference sequences of eight loci in total (excluding *dhy-like* and *adhC,* see below) were compared with those from Scottish populations. The samples from the Iberian Peninsula and Turkey were treated separately in between-region comparisons as they display specific mitochondrial types not observed in mainland European distribution of the species which suggests different histories and no contribution to recolonization after last glaciation (Pyhäjärvi *et al*, 2008; Soranzo *et al*, 2000).

PCR amplification and sequencing

PCR-amplification was performed with PTC-200 (MJ Research) and carried out in a total volume of 25µl containing about 10ng of haploid template DNA, 50µM of each of dNTP, 0.2µM of each primer and 0.25U *Taq* DNA polymerase with the respective 1x PCR buffer (NovaZyme, Poland). PCR followed standard amplification procedures with $MgCl_2$ concentration optimised for each primer pair as described in Supplementary Table S2. PCR fragments were purified using QIAquick^{TM}PCR Purification Kit (Qiagen). About 20 ng of PCR product was used as a template in 10 µl sequencing reactions with the Big Dye Terminator DNA Sequencing Kit (Applied Biosystems) performed by the GenePool sequencing service, University of Edinburgh. All samples were sequenced in both directions. CodonCode Aligner software was used for editing and assembling of the sequence chromatograms to produce alignments based on nucleotide sequence from both DNA strands. Haplotype sequences of each locus reported in this paper are deposited in the EMBL sequence database under accession numbers GQ262040 – GQ262490.

204

205 ***Sequence analysis***

206 High quality sequences were obtained for most of the samples at twelve loci (Table 1).
207 PCR amplification or sequencing failed in most of the samples at *dhy-like*, *adhC* and *grp3*
208 and these loci, together with *PR-AGP4-1* which was monomorphic across all 42 samples,
209 were excluded from further analysis. Nucleotide sequence alignments were constructed in
210 ClustalX and were further manually adjusted using GenDoc. All sequence
211 polymorphisms were visually rechecked from chromatograms edited with BioEdit.
212 Coding and noncoding regions (introns, UTRs) were annotated based on the NCBI
213 (http://www.ncbi.nlm.nih.gov/) sequence information at each locus and web-based gene
214 identification tool at PlantGDB (http://www.plantgdb.org/cgibin/PlantGDB/GeneSeqer/
215 PlantGDBgs.cgi). The influence of demography on the multilocus pattern of variation and
216 locus specific effects were assessed by looking at the amount of nucleotide diversity,
217 correlation between polymorphic sites and allelic frequency distribution between
218 different geographical locations in Scotland and in comparison to mainland populations
219 of the species and by comparing observed statistics with simulated values under a range
220 of demographic scenarios. Neutrality tests at intraspecific level were applied to search for
221 departures from a neutral model of evolution. Sequences from *Pinus pinaster* were used
222 as an outgroup for intraspecific comparisons to test for a signal of longer term selection.

223

224 ***Nucleotide diversity***

225 Two measures of nucleotide diversity were applied: 1) an average number of nucleotide
226 differences per nucleotide site between two sequences π, (Nei, 1987) calculated with

227 DNAsp 4.0) and 2) Watterson's (1975) estimate of the population mutation parameter,
228 theta (θ_W, equal to $4N_e\mu$, where N_e is the effective population size and μ is the mutation
229 rate per nucleotide site per generation), computed based on the number of segregating
230 sites and the length of each locus using MCMC simulation under a Bayesian model as
231 previously described (Pyhäjärvi *et al*, 2007). The estimates of nucleotide diversity were
232 conducted for all samples combined and separately for south, east and west regional
233 groups of Scottish populations. Scottish and continental European populations were
234 compared at eight loci for which informative data was available (Pyhäjärvi *et al*, 2007;
235 Wachowiak *et al*, 2009). Exceptionally high nucleotide diversity was found at *lp3-3* locus
236 compared to other loci in our dataset. Due to the size of the conifer genome and the
237 occurrence of multigene families (Ahuja and Neale, 2005), erroneous co-amplification of
238 different loci from the same family is possible and may account for unusual diversity
239 estimates at specific loci. Therefore, locus *lp3-3* was excluded from multilocus or average
240 estimates reported in the study to avoid bias and ensure that estimates were conservative;
241 the locus was included in population structure analysis and coalescent simulations.
242
243 ***Linkage disequilibrium and haplotype diversity***
244 The level of linkage disequilibrium was measured as the correlation coefficient r^2
245 (Hill and Robertson, 1968) using informative sites. Indels and sites with three nucleotide
246 variants identified in *dhn1* (3), *dhy2PP* (1) were excluded from the analysis. Under
247 mutation-drift-equilibrium model, the decay of linkage disequilibrium with physical
248 distance was estimated using non-linear regression of r^2 between polymorphic sites and
249 the distance (in base pairs) between sites as detailed in (Wachowiak *et al*, 2009). The

non-linear least-squares estimate of ρ ($\rho = 4N_e c$, where N_e is effective population size, c is the recombination rate) between adjacent sites was fitted by the nls-function implemented in the R statistical package (http://www.r-project.org). The overall and group specific least-squares estimates of ρ were computed and compared to other estimates in Scots pine (Pyhäjärvi *et al*, 2007; Wachowiak *et al*, 2009).

The number of haplotypes and haplotype diversity (H_d) were estimated for each gene using DNAsp. Insertions and deletions were included in all estimates. Coalescence simulations with locus specific or average ρ for six loci and without recombination were used to assess whether there are more or fewer haplotypes than expected and whether haplotype diversity is higher or lower than expected given the number of segregating sites. The number of haplotypes and haplotype diversity were calculated for all samples combined and separately for the three regional groups of Scots pine in Scotland.

Neutrality tests

Deviations of particular genes from the frequency distribution spectrum under the standard neutral model of evolution were assessed with Tajima's *D* test (Tajima 1989) and Fay and Wu's *H* (Fay and Wu 2000). Negative values of Tajima's *D* indicate an excess of low frequency polymorphisms consistent with positive directional selection or recent population expansion, whereas positive values indicate an excess of intermediate frequency polymorphism potentially due to balancing selection or population contraction. Fay and Wu's *H* test measures departures from neutrality based on high-frequency derived alleles. An excess of high frequency derived alleles compared to neutral expectations may result from recent positive selection or strong population structure with

uneven sampling from populations. The distribution of test statistics was investigated for each locus for all populations combined and separately for the three regional groups. Multilocus estimates of Tajima's *D* were assessed with HKA software (http://lifesci.rutgers.edu/~heylab). The estimates were also calculated along the sequence of each locus by a sliding window of 100 sites with successive displacement of 25 sites. As lack of recombination makes the *D* test overly conservative (Thornton 2005), the significance of locus specific and multilocus Tajima's *D* was also evaluated by coalescent simulations dependent on population mutation and recombination rate (MANVa software www.ub.edu/softevol/manva, based on coalescent program *ms*, Hudson, 2002). Different estimates of ρ including locus specific estimates, lowest and highest value across loci and average value for six loci were used in coalescent simulations. As similar probability values for multilocus *D* statistics were observed in simulations with different recombination rate estimates, the results based on the average values of ρ at the analysed loci are reported unless otherwise stated.

For tests based on nucleotide variation between species we used reference sequence data from *P. pinaster* for outgroup comparison. To assess the correlation between the level of nucleotide polymorphism and divergence at each locus we applied 1) the McDonald and Kreitman, (1991) test, based on comparison of the pattern of within-species polymorphism and between-species divergence at synonymous and nonsynonymous sites in a gene, and 2) HKA test (Hudson *et al*, 1987) which allows the detection of loci that demonstrate unusual patterns of polymorphism compared to divergence across genes. Comparison of multilocus polymorphism and divergence at all sites was assessed using HKA software (http://lifesci.rutgers.edu/ ~heylab). The ratio of

296 nonsynonymous (*K*a) and synonymous site (*K*s) nucleotide divergence from the outgroup
297 species (Hughes and Nei, 1988) was calculated using DnaSP.
298
299 ***Population structure***
300 To check if there was a geographical difference in allelic frequency spectra, regional
301 groups of Scottish populations were compared to each other and to previously analyzed
302 continental European populations from northern and central Europe, Spain and Turkey
303 (Pyhäjärvi *et al*, 2007; Wachowiak *et al*, 2009). Genetic differentiation between the
304 regions was studied locus by locus at both haplotype and SNP/Indel level and also by
305 averaging pairwise F_{ST} over all polymorphic sites across loci. The significance of genetic
306 differentiation was evaluated by 1000 permutations of the samples between groups using
307 Arlequin ver. 3.0 (Excoffier *et al*, 2005). Population structure from the haplotypic data
308 was tested by S_{nn} and K_{ST}^{*} statistics (Hudson *et al*, 1992; Lynch and Crease, 1990), which
309 are more appropriate for sequence-based haplotype data where diversity may be high and
310 sample size low, rendering frequency-based approaches problematic. Their significance
311 was evaluated using 1000 permutations, where samples were randomly assigned into
312 different groups (Hudson, 2000). Genetic clustering of the individuals based on both full
313 sequence data and all segregating sites and indels at 12 loci (for Scottish populations) and
314 at 8 loci (for Scottish and mainland European populations) was conducted using BAPS
315 5.2 (Corander and Tang, 2007). Polymorphic sites from each locus were treated as linked
316 molecular data to account for dependence between segregating sites in the gene.
317 Completely linked sites (r^2=1) were excluded from the analysis.
318
319 **Coalescent simulations**

To further infer the demographic history of Scottish Scots pine populations we compared the observed distribution of average Tajima's *D* and Fay and Wu's *H* at the candidate loci separately in the western, southern and eastern group and in all geographic regions combined to the simulated values under several demographic scenarios including the standard neutral model (constant population size), growth model and bottleneck model followed by exponential growth (Supplementary Figure S4). Regional groups of populations were analysed separately as detailed aspects of the frequency spectrum may differ between groups that are not differentiated based on genetic clustering methods (Pyhäjärvi et al 2007). Coalescent simulations were run independently for each locus and various demographic scenarios using the program *ms* (Hudson 2002) and the approach described by Haddrill *et al* (2005). In each case, 5000 replicates were simulated for each locus. The analyses were performed with recombination using the locus specific (when available) or average value of ρ per site for the analysed loci in each geographic group (Table 2, Supplementary Table S3). We tested various bottleneck scenarios of different age and severity. The time from the end of the bottleneck (measured in units of $4N_0$ generations from the present) ranged from 0.0002 to 0.05 and bottleneck severities (measured in units of the current population size) from 0.001 to 0.5. Assuming for instance, N_e of 200000 and generation time of 25 years, the time range corresponds to between 4 000 and 1 million years and severity from 0.1 to 50% of the current population size. In most bottleneck models tested, the ancestral and current effective population sizes were assumed to be equal, bottleneck duration (f) was fixed to f=0.0015 (units of $4N_0$ generations from the present) and the growth rate of 10 was constant across simulations as in previous studies (Heuertz et al. 2006). A subset of simulations were run also with

343 *f*=0.006 and corresponding equal or doubled ancestral population size as compared to the
344 current one, and also separately for a set of 11 and 9 loci (excluding *lp3-3* and *dhn1* and
345 *abaR*, respectively as the later showed some evidence of selection). A schematic
346 representation of the simulated bottleneck model is shown in Supplementary Figure S4.
347 The simulation results for each demographic scenario were summarized using the
348 program analyser HKA. The perl script multitest_pop1.pl was used to perform multilocus
349 tests of *ms*-generated genealogies (including *P*-values of the observed mean values of
350 Tajima's *D* and Fay and Wu's *H* statistics) summarized using analyser HKA. The
351 programs are available from http://genomics.princeton.edu/
352 AndolfattoLab/Andolfatto_Lab.html.
353
354 **Results**
355 **Nucleotide polymorphism and divergence**
356 The average total nucleotide diversity (π) in Scottish populations at eleven loci was π_{tot} =
357 0.0078 and at nonsynonymous sites was π_{ns} = 0.0031 (Table 2). Slightly lower average
358 nucleotide diversity was found in the west (π_{tot} = 0.0068) as compared to southern and
359 eastern regional groups (π_{tot} = 0.0079 and 0.0083, respectively) and similar values were
360 found at nonsynonymous sites (π_{ns}= ~ 0.003) (Supplementary Table S3). Multilocus
361 estimates of silent Watterson theta was θ_{sil}=0.0095 (with 95% credibility intervals of
362 0.0074-0.0122) for all Scottish populations combined, θ_{sil}=0.0086 (0.0063-0.0117) in the
363 west, θ_{sil}=0.0111 (0.0080-0.0152) in the south and θ_{sil}=0.0103 (0.0074-0.0143) in the
364 east. In comparisons between Scottish vs. mainland European populations at eight loci,
365 similar but slightly higher average values of total nucleotide diversity (π_{tot} = 0.0070 vs

366 0.0062) and silent multilocous theta (θ_{sil}=0.0108 vs. 0.0093) were found in Scottish
367 populations (Table 3).
368
369 **Linkage disequilibrium and haplotype polymorphisms**
370 Rapid decay of linkage disequilibrium between pairs of parsimony informative sites at
371 eleven loci was found in Scottish populations, with $\rho = 0.0085 \pm 0.0009$ (Table 2) and
372 expected r^2 values of 0.2 at a distance of about 400 bp. The decay of linkage
373 disequilibrium in the western group ($\rho = 0.0074 \pm 0.0008$) was more rapid as compared to
374 the south (0.0025 ± 0.0004) and east (0.0024 ± 0.0006) (Figure 2) and the pattern was
375 constant at most loci (Supplementary Table S3). Overall, Scottish populations had about
376 three times slower decay of linkage disequilibrium as compared to mainland populations
377 at the same set of eight loci of similar sample size ($\rho = 0.0086 \pm 0.0009$ vs
378 0.0245 ± 0.0022, respectively) (Supplementary Figure S1). However, the rate of decay of
379 LD and the relative level of recombination to diversity (ρ / θ ratio) were similar between
380 western Scottish and north and central European regions (Table 3) but these parameters
381 were over three times smaller in southern and eastern groups of Scotland.
382 The average number of haplotypes per gene was 12 and haplotype diversity was very
383 high (H_d=0.789±0.042). Similar haplotype diversity was found in western (H_d
384 =0.754±0.077), southern (H_d=0.819±0.088) and eastern (H_d=0.800±0.090) groups
385 (Supplementary Table S3). Haplotype diversity was slightly higher than mainland
386 European populations at the same set of eight loci (H_d=0.831±0.038 vs H_d=0.795±0.051)
387 and also compared to previous estimates for Scots pine (H_d=0.683±0.059, Wachowiak *et*
388 *al.* 2009). Locus *Lp3-3* contained two sets of haplotypes (each of 18 samples equally

389 distributed across three geographical groups) with highly reduced levels of nucleotide
390 polymorphism (π_{tot} = 0.0090 and 0.0074, respectively) as compared to the whole gene
391 estimate (π_{tot} = 0.0370) and a ten-fold difference in the level of divergence (K_{sil1}=0.013 vs
392 K_{sil2}=0.116) (Supplementary Table S4 and Supplementary Figure S2). A neutral
393 coalescence process, compatible with a constant-size neutral model without
394 recombination or erroneous coamplifications of different gene family members could
395 potentially generate such a pattern. However, no reading-frame shifts or premature stop
396 codons, which would suggest the presence of non-functional alleles, were found at the
397 locus.
398
399 **Neutrality tests**
400 Tendency towards an excess of old over recent mutations across genes was detected by
401 multilocus Tajima's *D* at eleven loci in the total data set (*D*=0.118) (Table 2), in the
402 western (*D*=0.364), southern (*D*=0.103) and eastern (*D*=0.260) groups (Supplementary
403 Table S3). Significant excess of intermediate frequency mutations was found at *dhn2*
404 (*D*=1.968, P<0.05) and *lp3-3* (*D*=2.846, P<0.01). Statistically significant positive values
405 of Tajima's *D* were identified in sliding window analyses in a few regions within *dhn2* (*D*
406 = 2.36-2.48 at 307–449 bp), *a3ip2* (*D* = 2.22 at 401-501 bp) and *lp3-3* (*D* = 2.13-3.18 at
407 51–454 bp) loci. Overall, an excess of high-frequency derived variants indicated by
408 negative mean values of Fay and Wu's *H* statistics was found in all Scottish populations
409 (*H*= -0.494) (Table 2), in the west (*H*= -0.447) and east (*H*= -0.145) groups, but slightly
410 positive values were found in the south (*H*= 0.144) (Supplementary Table S3). The
411 aggregated Scottish populations show a deficit of rare variants (multilocus Tajima's

412 *D*=0.316) as compared to mainland European populations (*D*=-0.379). Both geographical
413 regions show negative mean value of Fay and Wu's *H* statistics (*H*=-0.564 and -1.240,
414 respectively) indicating an excess of high-frequency derived SNPs (Table 3).

415 An excess of fixed nonsynonymous over fixed synonymous substitutions and
416 polymorphic sites was found at *dhn1* locus in McDonald-Kreitman test (Fisher's exact
417 test, $P = 0.05$), as previously found in European mainland populations (Wachowiak *et al*,
418 2009). An excess of nonsynonymous sites as compared to synonymous sites was found at
419 *abaR* (Supplementary Table S5). The level of divergence was similar across all sites and
420 at silent sites only (~4%), and was slightly lower than previous estimates for Scots pine
421 (K=~0.05, Wachowiak *et al.* 2009). Overall, positive correlation between polymorphism
422 and divergence (HKA test) was found at eleven loci combined.

423

424 **Population differentiation**

425 *Differentiation between Scottish populations*

426 Significant differentiation measured as an average over all polymorphic sites was found
427 between southern and eastern groups at *dhn1* (F_{ST}=0.034, P<0.05) and between southern
428 and eastern as compared to the western group at *ccoaomt* (F_{ST}=0.149, P<0.05 and F_{ST}
429 =0.102, P<0.01, respectively) and *lp3-1* (F_{ST} =0.100, P<0.05 and F_{ST} =0.197, P<0.001,
430 respectively) (Supplementary Table S6). A difference in frequency of indel
431 polymorphisms at *dhn1*, four silent substitutions and indel polymorphisms at *lp3-1* and
432 absence of four silent polymorphisms in the western group as compared to the others at
433 *ccoaomt* locus contributed the most to the differentiation between groups. Based on
434 haplotype differentiation, the western group differed from the southern group at *a3ip*

(S_{nn}=0.629, P<0.05), *lp3-1* (S_{nn}=0.758, P<0.01) and at *ccoaomt* and *lp3-1* based on K_{ST} statistics (K_{ST} = 0.066 and 0.051, P<0.05, respectively). They also differ from the east group at *lp3-1* locus (K_{ST}=0.075, P<0.05). Significant F_{ST} statistics based on haplotype frequency were found for *lp3-1* in the south and east as compared to western group (P<0.05) and nearly significant values for *ccoaomt* between south and west groups (P=0.06) (Supplementary Table S6). No difference between west-south, west-east and south-east groups were found based on average F_{ST} over all polymorphic sites and indels combined across the loci (F_{ST} =-0.013, -0.013, and 0.01, respectively).

Differentiation between Scottish vs European continental populations

Based on allele frequency and/or haplotype diversity statistics Scottish populations were differentiated from continental European populations at six out of eight loci analysed (Supplementary Table S7). Significant population differentiation (F_{ST}), measured both as an average over polymorphic sites and at the haplotype level, was found at *dhn2*, *dhn7*, *abaR* and *chcs*. Based on the average proportion of nearest-neighbor haplotypes that are present in the same locality (S_{nn}) both groups were differentiated at *dhn2*, *dhn7*, *dhy2PP* and *a3iP* (P<0.001-0.05). Two loci, *dhn2* and *dhn7*, also showed high similarity between pairs of sequences derived from each region (K_{ST} = 0.098 and 0.067, respectively, P<0.01).

Significant differentiation was found between Scottish populations versus continental European populations measured as an average of F_{ST} values over all polymorphic sites detected (Table 4). The only exceptions include southern Scottish populations as

457 compared to northern and central Europe, eastern Scottish compared to northern
458 European and western Scottish compared to Spanish populations.
459 Analysis of genetic clustering with full sequence data gave the best support for all
460 individuals from European mainland and Scottish populations at eight loci and for
461 individuals from Scottish populations at 12 loci belonging to one genetic cluster. At all
462 polymorphic sites and indels at both eight and twelve loci, the best support was obtained
463 for four clusters, but without clear pattern of geographical distribution (Supplementary
464 Figure S3).
465

466 **Coalescent simulations**

467 For each geographic group of Scottish Scots pine populations the observed pattern of the
468 frequency distribution spectrum was not compatible with either the standard neutral or
469 growth models. In simulations under the SNM and growth model the mean Tajima's *D*
470 was significantly lower and Fay and Wu's *H* significantly higher than the observed values
471 except for the southern group, the only one with positive mean *H* values (Table 5,
472 Supplementary Table S8). Among the 20 different bottleneck models tested the most
473 compatible for the western group was a relatively recent bottleneck (t=0.00125) that
474 reduced the population to 2% of the present size followed by moderate population growth
475 (Table 5, Supplementary Table 9). This model also held for the eastern group but was
476 always rejected for the southern group, where different bottleneck scenarios never lead to
477 positive values for both Tajima's *D* and Fay and Wu's *H* statistics (Supplementary Table
478 9). In general, the simulations indicate heterogeneity in the allelic frequency distribution
479 among geographic regions in Scotland.

Discussion

Multilocus signatures of population history

The Scottish populations showed clear molecular signatures of different demographic histories. Across all regions, the allele frequency distribution was skewed towards intermediate frequency polymorphisms, and the rate of decline of linkage disequilibrium was reduced and nucleotide diversity levels were equivalent to or higher than continental European populations of the species. The skew of allelic frequency distribution, apparent as positive values of Tajima's *D*, was in clear contrast to previous reports for this species in continental Europe (Palmé *et al*, 2008; Wachowiak *et al*, 2009) and for published studies of other species (North American Douglas fir, Eckert *et al*, 2009; *P. taeda* González-Martínez *et al*, 2006a; other conifer species Savolainen and Pyhäjärvi, 2007; European *Quercus petraea* Derory *et al*, 2009; *Populus tremula* Ingvarsson, 2005) where negative values of Tajima's *D* have been found. In these species, the excess of low frequency derived mutations has been ascribed to the influence of postglacial range expansion (Brown *et al*, 2004; Pyhäjärvi *et al*, 2007) or potentially the influence of recurrent selective sweeps (e.g. Eckert *et al*, 2009). In contrast, rather than range expansion, the bias towards intermediate-frequency polymorphisms in Scottish populations suggests the influence of a bottleneck although, as shown in recent simulation studies, a skew of allelic frequency variants may also result from pooling local samples with different demographic histories (Städler *et al*, 2009). However, the bottleneck hypothesis was also supported by the overall pattern of linkage disequilibrium (LD), which showed a reduced rate of decline relative to continental European populations of the species. In coalescent simulations, the bottleneck scenario fits best for

503 western populations and the data were compatible with a relatively recent, severe
504 bottleneck. Depending on the effective population size and generation time assumed, this
505 bottleneck ended a maximum of a few tens of thousands of years ago (e.g about 25 000
506 ya assuming N_e=200 000). Bottlenecking is expected to increase association (correlation
507 among sites with distance) of alleles and polymorphic sites across loci. In Scottish
508 populations, the decay of LD was almost three times slower than that in mainland
509 populations. Reduced decay of LD has also been observed in populations of American *P.*
510 *taeda* that had probably experienced bottlenecks (Brown *et al*, 2004; González-Martínez
511 *et al*, 2006a) and contrasting allele frequency distributions were observed between
512 northern populations and recently bottlenecked southern populations of *Quercus crispula*
513 in Japan (where the latter showed positive Tajima's *D,* Quang *et al*, 2008).
514
515 Although there are exceptions (Grivet *et al*, 2009), it is expected that bottlenecks should
516 have a stronger impact on the allele frequency distribution spectrum and LD than on the
517 overall level of diversity (Wright *et al*, 2005). Long-lived, wind-pollinated tree species
518 should be capable of maintaining genetic diversity even during range shifts; i.e. they are
519 buffered against rapid changes in genetic variation due to fluctuations in population size
520 (Austerlitz *et al*, 2000). Indeed, relative to mainland European populations of Scots pine,
521 Scottish populations did not show a decline in nucleotide diversity, as is expected where
522 colonisation has been relatively recent (Nei *et al*, 1975; Pannell and Dorken, 2006). In
523 fact, genetic variation in Scottish populations seems to be slightly higher than in
524 mainland populations (θ_{sil}=0.011 vs 0.009, respectively) and relative to previous
525 estimates for the species (θ_{sil}= 0.005 at 16 loci with some related to timing of bud set

526 (Pyhäjärvi *et al*, 2007) and θ_{sil}=0.0089 at 14 cold tolerance candidate loci (Wachowiak *et*
527 *al*, 2009)). Compared to estimates in other forest tree species, overall diversity in Scottish
528 populations (π_{tot}=0.0078) is only lower than that in broadleaved *Populus tremula* (0.0111,
529 Ingvarsson, 2005) and is higher than that in *Q. crispula*, (0.0069, Quang and Harada
530 2008), *Q. petraea* (0.0062, Derory *et al.* 2009), *P. pinaster* (0.0055, Eveno et al. 2008),
531 *P. taeda* (0.0040, Brown *et al,* 2004), *Picea abies* (0.0039, Heuertz *et al*, 2007) and other
532 conifers (Savolainen and Pyhäjärvi, 2007). The diversity estimate for Scottish
533 populations is compatible with the patterns of genetic variation observed in previous
534 studies (monoterpenes Forrest, 1980; Forrest, 1982), allozymes Kinloch *et al*, 1986),
535 chloroplast DNA microsatellite markers Provan *et al*, 1998).
536
537 Although it seems clear that bottlenecking has been an influence on Scottish populations,
538 estimation of the timing of the event is heavily dependent on various assumptions
539 including the effective population size and generation time estimates. For instance, in
540 continental populations of Norway spruce and Scots pine, simulation studies suggested a
541 rather ancient bottleneck that ended several hundred thousand to more than one million
542 years ago, respectively (Lascoux *et al*, 2008). In our data, coalescent simulation of
543 various demographic scenarios supported the conclusion that bottlenecking had occurred,
544 but suggested more recent timing. A similar signal, suggesting bottlenecking on a
545 timescale related to the most recent glaciation, was detected in Italian populations of
546 Aleppo pine (Grivet *et al*, 2009). Furthermore, the severity of the bottleneck experienced
547 by Scottish populations appears to have been strong enough to account for the observed

548 discrepancy in allelic frequency distributions and decay of LD, in contrast to continental
549 European tree populations (Lascoux *et al*, 2008).
550
551 However, as we observed heterogeneity in the pattern of nucleotide diversity among
552 regions within Scotland, it seems likely that different parts of the population have
553 experienced different demographic histories. The ratio of recombination to diversity and
554 the level of linkage disequilibrium in western Scottish populations were similar to those
555 in mainland European populations but about three times higher than those in southern and
556 eastern Scottish groups. Various bottleneck scenarios could be clearly rejected for the
557 southern group in our coalescent simulation analysis. The homogenizing effects of gene
558 flow on genetic diversity are well known for highly outcrossing wind pollinated species,
559 and there is evidence for historically high gene flow among Scottish populations from
560 work using chloroplast markers (Provan *et al*, 1998). In addition, molecular and isozyme
561 studies provide no suggestion of a difference in outcrossing rates between regions that
562 could account for a difference in spatial distribution of polymorphism (Kinloch *et al*,
563 1986). As, until recently, Scots pine covered large parts of Scotland, differentiation
564 between regional groups due to genetic drift also seems unlikely, as this should be most
565 significant for small populations (Pannell and Dorken, 2006). Inter-regional differences
566 also seem unlikely to be the result of selection. If this was the case, we would expect
567 differences in the frequency distribution spectrum between groups or at least reduced
568 diversity levels at selected loci. However, the observed dominance of intermediate
569 frequency variants in all groups together with very rapid decay of linkage disequilibrium
570 (within a few hundred base pairs) excludes a selective sweep as an explanation.

571 Furthermore, nucleotide and haplotype diversity is at least as high in southern and eastern
572 groups as in the western group, whereas directional selection should reduce diversity.
573 Therefore, overall, historical changes in population size and distribution seem a more
574 plausible explanation for the pattern of nucleotide variation in Scottish populations and,
575 as a single migration and bottleneck event cannot account for the observed pattern of
576 diversity, it seems that heterogeneity within the Scottish population is most likely to be
577 the result of admixture of populations from different origins.
578
579 Compared to continental Europe, southern and eastern groups of Scottish pines showed
580 no overall difference in allele frequency distribution at polymorphic sites from north or
581 central European populations, but differentiation from Spanish and Turkish populations.
582 On the other hand, the western group was significantly differentiated from all mainland
583 populations except those from Spain. In previous studies, populations from the west of
584 Scotland were more closely related to southern European populations in monoterpene
585 composition and isozyme frequency (Forrest, 1982) or geographically structured *mt*DNA
586 variation (Sinclair *et al*, 1998) than to populations from north-central Europe, which were
587 more similar to the southern and eastern Scottish pinewoods. Similarities between
588 western Scotland and south European Scots pine could simply be stochastic, due to
589 homogenising selection for similar environments or, alternatively, could reflect common
590 ancestry of the populations. Genetic similarity at *mt*DNA markers (maternally transmitted
591 in pines) suggests the latter. However, as Iberian populations did not contribute to the
592 most recent recolonization of central and northern Europe (Prus-Glowacki and Stephan,
593 1994; Pyhäjärvi *et al*, 2008; Soranzo *et al*, 2000; Tobolski and Hanover, 1971), this

genetic similarity would reflect a common origin predating the last glacial period. Therefore, contemporary Scottish populations may originate from western populations that survived the last glaciation in southwestern parts of the British Isles, western continental Europe (Ballantyne and Harris, 1994; Bennett, 1995) or now-submerged parts of the continental shelf. Future genetic studies at more loci (including new *mt*DNA markers) and in more populations would allow more precise assessment of the spatial distribution of haplotypes in Scottish and mainland populations and testing of colonisation hypotheses. This should soon be feasible as new genomic resources for pine, including multiple nuclear and *mt*DNA loci, are currently being developed (e.g. through the EVOLTREE Network of Excellence).

Effects of selection at individual loci

At mutation-drift equilibrium, genetic drift and gene flow influence the level of differentiation between populations for selectively neutral markers (Kawecki and Ebert, 2004; Savolainen *et al*, 2007). Little differentiation between Scottish and mainland European populations of Scots pine at neutral markers (Kinloch *et al*, 1986; Provan *et al*, 1998; Prus-Glowacki and Stephan, 1994) but divergence at quantitative traits for characters of adaptive importance (e.g. phenology, growth and survival rates, Ennos *et al*, 1998; Worrell, 1992, Hurme et al. 1997) suggests that selection is driving adaptive differentiation in both geographical regions. As they differ significantly in climatic, edaphic and biotic conditions, it is possible that observed nucleotide and/or haplotype differentiation at *dhn2* and *dhn7* and some differences in the allele-frequency spectrum at *dhy2PP*, *abaR*, *a3iP2* and *chcs* may be due to selection. Similarly, reduced nucleotide

617 and haplotype diversity and a difference in the frequency and distribution of
618 polymorphism found at *lp3-1* and *ccoaomt* in the western as compared to the southern
619 and eastern groups of Scottish populations could have been affected by diversifying
620 selection at the range edge where populations are under direct oceanic influence. In
621 contrast, the haplotype dimorphism at *lp3-3* could potentially result from the long-term
622 action of balancing selection, maintaining variation across geographical regions.
623 However, as admixture at *lp3-3* cannot be ruled out, a study of nucleotide polymorphism
624 in mainland European populations would be necessary to verify whether or not balancing
625 selection has been an influence at this locus.
626
627 Some of the loci analysed showed distinct nucleotide diversity patterns relative to genetic
628 background in other species (e.g. *lp3-1* and *ccoaomt* in *P. pinaster* Eveno *et al*, 2008,
629 *ccoaomt* in *P. taeda,* González-Martínez *et al*, 2006a). Although there is accumulating
630 evidence on the polygenic character of adaptive traits from QTL studies (Buckler *et al*,
631 2009; Howe *et al*, 2003), it remains unclear whether or not there are genes of major effect
632 that contribute to adaptive variation in conifers. In the case of Scottish pinewoods,
633 adaptation was probably driven by postglacial migration from a predominantly
634 continental to an oceanic environment over the past ~7000 yrs. For long-lived conifers,
635 adaptive differentiation would be expected to occur over several dozens of generations
636 after vicariance. However, even though selection can be very effective in species with
637 large population sizes, the time since the last glaciation seems too short for pine species
638 to have accumulated new mutations that could be rapidly fixed by selection. Adaptive
639 divergence is therefore more likely to result from selection acting on standing variation,

640 which may have arisen in endemic populations that survived last glaciations in Western
641 Europe or the British Isles. Moreover, as differentiation at the trait level in forest trees is
642 likely to result from allelic associations among large numbers of loci, rather than changes
643 in allelic frequencies at individual loci, the signature of selection may be more readily
644 detectable as covariance of allele frequencies at multiple loci (Derory *et al*, 2009; Latta,
645 2004; Le Corre and Kremer, 2003). Therefore many more loci, including regulatory
646 regions (to date, generally omitted from analyses of nucleotide variation in conifers),
647 would need to be studied in parallel before the influence of selection could be verified.
648 Scottish populations, which show considerable ecological, phenotypic and genetic
649 diversity over short geographic distances, represent an excellent study system for
650 multilocus analysis of complex trait variation (González-Martínez *et al*, 2006b; Neale and
651 Savolainen, 2004). Such studies will, however, have to take into account the potential
652 role of recent population history in shaping patterns of nucleotide diversity, and therefore
653 ensure that sampling is conducted at sufficient density to control for historical influences.
654 Association studies of allelic variants and adaptive variation at quantitative traits between
655 individuals from different, locally-adapted populations could also better validate the
656 signatures of selection and the functional role of the nuclear genes studied.
657

Acknowledgments

WW acknowledges financial support from the Polish Ministry of Science (grant nr 3653/B/P01/2008/35), NERC and EU Network of Excellence EVOLTREE (mobility grant). MJS is a Ph.D. student supported by the Scottish Forestry Trust. We thank Dave Sim, Joan Beaton and Ben Moore (Macaulay Institute) for making the seed collections, the owners of the woodlands for their cooperation and Joan Cottrell (Forest Research) and anonymous reviewers for constructive comments on the manuscript.

Conflict of interest statement

The authors declare that there are no conflicts of interest.

References

Achaz G (2009). Frequency Spectrum Neutrality Tests: One for All and All for One. *Genetics* **183**(1)**:** 249-258.

Ahuja MR, Neale DB (2005). Evolution of genome size in conifers. *Silvae Genet* **54**(3)**:** 126-137.

Austerlitz F, Mariette S, Machon N, Gouyon PH, Godelle B (2000). Effects of colonization processes on genetic diversity: Differences between annual plants and tree species. *Genetics* **154**(3)**:** 1309-1321.

Ballantyne CK, Harris C (1994). *The Periglaciation of Great Britain*. Cambridge University Press: Cambridge, 330pp.

Bennett KD (1995). Post-glacial dynamics of pine (*Pinus sylvestris*) and pinewoods in Scotland. In: Aldhous JR (ed) *Scottish Natural Heritage*. Forestry Commission, The Royal Society for the Protection of Birds: Edinburgh, pp 23-39.

Birks HJB (1989). Holocene isochrone maps and patterns of tree-spreading in the British Isles. *Journal of Biogeography* **16**(6)**:** 503-540.

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004). Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proc Natl Acad Sci U S A* **101**(42)**:** 15255-15260.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C *et al* (2009). The Genetic Architecture of Maize Flowering Time. *Science* **325**(5941)**:** 714-718.

Corander J, Tang J (2007). Bayesian analysis of population structure based on linked molecular information. *Mathematical Biosciences* **205**(1)**:** 19-31.

Critchfield WB, Little E. (1965). U.S. Department of Agriculture, pp 97.

Derory J, Scotti-Saintagne C, Bertocchi E, Le Dantec L, Graignic N, Jauffres A *et al* (2009). Contrasting relationships between the diversity of candidate genes and variation of bud burst in natural and segregating populations of European oaks. *Heredity*, **104:** 438-448

Dvornyk V, Sirvio A, Mikkonen M, Savolainen O (2002). Low nucleotide diversity at the pal1 locus in the widely distributed Pinus sylvestris. *Mol Biol Evol* **19**(2)**:** 179-188.

Eckert AJ, Wegrzyn JL, Pande B, Jermstad KD, Lee JM, Liechty JD *et al* (2009). 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**(1)**:** 289-298.

Ennos RA, Worrell R, Malcolm DC (1998). The genetic management of native species in Scotland. *Forestry* **71**(1)**:** 1-23.

Eveno E, Collada C, Guevara MA, Leger V, Soto A, Diaz L *et al* (2008). Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. *Mol Biol Evol* **25**(2)**:** 417-437.

Excoffier L, Laval G, Schneider S (2005). Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1:** 47-50.

Fay JC, Wu C-I (2000). Hitchhiking Under Positive Darwinian Selection. *Genetics* **155**(3)**:** 1405-1413.

Forrest GI (1980). Genotypic variation among native Scots pine populations in Scotland based on monoterpene analysis. *Forestry* **53**(2)**:** 101-128.

Forrest GI (1982). Relationship of some European Scots pine populations to native Scottish woodlands based on monoterpene analyses. *Forestry* **55**(1)**:** 19-37.

García-Gil MR, Mikkonen M, Savolainen O (2003). Nucleotide diversity at two phytochrome loci along a latitudinal cline in Pinus sylvestris. *Molecular Ecology* **12**(5)**:** 1195-1206.

González-Martínez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB (2006a). DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. *Genetics* **172**(3)**:** 1915-1926.

González-Martínez SC, Krutovsky KV, Neale DB (2006b). Forest-tree population genomics and adaptive evolution. *New Phytologist* **170**(2)**:** 227-238.

Grivet D, Sebastiani F, González-Martínez SC, Vendramin GG (2009). Patterns of polymorphism resulting from long-range colonization in the Mediterranean conifer Aleppo pine. *New Phytologist* **184:** 1016-1028.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P (2005). Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations. *Genome Res* **15**(6)**:** 790-799.

Hamrick JL, Godt MJW, Sherman-Broyles SL (1992). Factors influencing levels of genetic diversity in woody plants species. *New For* **6:** 95-124.

Heuertz M, De Paoli E, Kallman T, Larsson H, Jurman I, Morgante M *et al* (2006). Multilocus Patterns of Nucleotide Diversity, Linkage Disequilibrium and Demographic History of Norway Spruce [*Picea abies* (L.) Karst]. *Genetics* **174**(4)**:** 2095-2105.

772 Hill WG, Robertson A (1968). Linkage disequilibrium in finite populations. *Theoretical*
773 *and Applied Genetics* **38:** 226-231.
774
775 Howe GT, Aitken SN, Neale DB, Jermstad KD, Wheeler NC, Chen THH (2003). From
776 genotype to phenotype: unraveling the complexities of cold adaptation in forest trees.
777 *Canadian Journal of Botany-Revue Canadienne De Botanique* **81**(12)**:** 1247-1266.
778
779 Hudson RR (2000). A New Statistic for Detecting Genetic Differentiation. *Genetics*
780 **155**(4)**:** 2011-2014.
781
782 Hudson RR (2002). Generating samples under a Wright-Fisher neutral model of genetic
783 variation. *Bioinformatics* **18**(2)**:** 337-338.
784
785 Hudson RR, Boos DD, Kaplan NL (1992). A statistical test for detecting geographic
786 subdivision. *Mol Biol Evol* **9**(1)**:** 138-151.
787
788 Hudson RR, Kreitman M, Aguade M (1987). A Test of Neutral Molecular Evolution
789 Based on Nucleotide Data. *Genetics* **116**(1)**:** 153-159.
790
791 Hughes AL, Nei M (1988). Pattern of nucleotide substitution at major histocompatibility
792 complex class I loci reveals overdominant selection. *Nature* **335**(6186)**:** 167-170.
793
794 Huntley B, Birks HJB (1983). *An Atlas of Past and Present Pollen Maps for Europe: 0-*
795 *13000 Years Ago*. Cambridge University Press: Cambridge, 667pp.
796
797 Ingvarsson PK (2005). Nucleotide polymorphism and linkage disequilbrium within and
798 among natural populations of European Aspen (Populus tremula L., Salicaceae). *Genetics*
799 **169**(2)**:** 945 - 953.
800
801 Joosen RVL, Lammers M, Balk PA, Bronnum P, Konings MCJM, Perks M *et al* (2006).
802 Correlating gene expression to physiological parameters and environmental conditions
803 during cold acclimation of Pinus sylvestris, identification of molecular markers using
804 cDNA microarrays. *Tree Physiol* **26**(10)**:** 1297-1313.
805
806 Karhu A, Hurme P, Karjalainen M, Karvonen P, Kärkkäinen K, Neale D *et al* (1996). Do
807 molecular markers reflect patterns of differentiation in adaptive traits of conifers? *Theor*
808 *Appl Genet* **93**(1-2)**:** 215-221.
809
810 Kawecki TJ, Ebert D (2004). Conceptual issues in local adaptation. *Ecology Letters*
811 **7**(12)**:** 1225-1241.
812
813 Kinloch BB, Westfall RD, Forrest GI (1986). Caledonian Scots pine - origins and genetic
814 structure. *New Phytologist* **104**(4)**:** 703-729.
815

Lascoux M, Pyhäjärvi T, Källman T, Savolainen O (2008). Past demography in forest trees: what can we learn from nuclear DNA sequences that we do not already know? *Plant Ecology & Diversity* **1**(2)**:** 209 - 215.

Latta RG (2004). Relating processes to patterns of genetic variation across landscapes. *Forest Ecology and Management* **197**(1-3)**:** 91-102.

Le Corre V, Kremer A (2003). Genetic variability at neutral markers, quantitative trait loci and trait in a subdivided population under selection. *Genetics* **164**(3)**:** 1205-1219.

Lynch M, Crease TJ (1990). The analysis of population survey data on DNA sequence variation. *Mol Biol Evol* **7**(4)**:** 377-394.

Ma XF, Szmidt AE, Wang XR (2006). Genetic structure and evolutionary history of a diploid hybrid pine Pinus densata inferred from the nucleotide variation at seven gene loci. *Mol Biol Evol* **23**(4)**:** 807 - 816.

Mason WL, Hampson A, Edwards C (2004). *Managing the Pinewoods of Scotland*. Forestry Commission: Edinburgh.

McDonald JH, Kreitman M (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351:** 652 - 654

Neale DB, Ingvarsson PK (2008). Population, quantitative and comparative genomics of adaptation in forest trees. *Curr Opin Plant Biol* **11**(2)**:** 149-155.

Neale DB, Savolainen O (2004). Association genetics of complex traits in conifers. *Trends in Plant Science* **9**(7)**:** 325-330.

Nei M (1987). Molecular Evolutionary Genetics. *New York, Columbia University Press*.

Nei M, Maruyama T, Chakraborty R (1975). The Bottleneck Effect and Genetic Variability in Populations. *Evolution* **29**(1)**:** 1-10.

Padmanabhan V, Dias DMAL, Newton RJ (1997). Expression analysis of a gene family in loblolly pine (Pinus taeda L.) induced by water deficit stress. *Plant Molecular Biology* **35**(6)**:** 801-807.

Palmé AE, Wright M, Savolainen O (2008). Patterns of divergence among conifer ESTs and polymorphism in *Pinus sylvestris* identify putative selective sweeps. *Mol Biol Evol***, 25**: 2567 - 2577.

Pannell J, Dorken M (2006). Colonisation as a common denominator in plant metapopulations and range expansions: effects on genetic diversity and sexual systems. *Landscape Ecology* **21**(6)**:** 837-848.

862 Perks MP, Ennos RA (1999). Analysis of genetic variation for quantitative characters
863 between and within four native populations of Scots pine (*Pinus sylvestris*). *Botanical*
864 *Journal of Scotland* **51:** 103-110.
865
866 Perks MP, McKay HM (1997). Morphological and physiological differences in Scots
867 pine seedlings of six seed origins. *Forestry* **70:** 223-232.
868
869 Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, Brewer S, Cheddadi R *et al* (2003).
870 Glacial refugia: Hotspots but not melting pots of genetic diversity. *Science* **300**(5625)**:**
871 1563-1565.
872
873 Provan J, Soranzo N, Wilson NJ, McNicol JW, Forrest GI, Cottrell J *et al* (1998). Gene-
874 pool variation in Caledonian and European Scots pine (*Pinus sylvestris* L.) revealed by
875 chloroplast simple-sequence repeats. *Proceedings of the Royal Society of London Series*
876 *B-Biological Sciences* **265**(1407)**:** 1697-1705.
877
878 Prus-Glowacki W, Stephan BR (1994). Genetic variation of Pinus sylvestris from Spain
879 in relation to other European populations. *Silvae Genet* **43**(1)**:** 7-14.
880
881 Pyhäjärvi T, Garcia-Gil MR, Knürr T, Mikkonen M, Wachowiak W, Savolainen O
882 (2007). Demographic History Has Influenced Nucleotide Diversity in European *Pinus*
883 *sylvestris* Populations. *Genetics* **177**(3)**:** 1713-1724.
884
885 Pyhäjärvi T, Salmela MJ, Savolainen O (2008). Colonization routes of *Pinus sylvestris*
886 inferred from distribution of mitochondrial DNA variation. *Tree Genetics & Genomes*
887 **4**(2)**:** 247-254.
888
889 Quang ND, Ikeda S, Harada K (2008). Nucleotide variation in Quercus crispula Blume.
890 *Heredity* **101**(2)**:** 166-174.
891
892 Savolainen O, Pyhäjärvi T (2007). Genomic diversity in forest trees. *Curr Opin Plant*
893 *Biol* **10**(2)**:** 162-167.
894
895 Savolainen O, Pyhäjärvi T, Knürr T (2007). Gene Flow and Local Adaptation in Trees.
896 *Annual Review of Ecology, Evolution, and Systematics* **38**(1)**:** 595-619.
897
898 Sinclair WT, Morman JD, Ennos RA (1998). Multiple origins for Scots pine (*Pinus*
899 *sylvestris* L.) in Scotland: evidence from mitochondrial DNA variation. *Heredity* **80:** 233-
900 240.
901
902 Soranzo N, Alia R, Provan J, Powell W (2000). Patterns of variation at a mitochondrial
903 sequence-tagged-site locus provides new insights into the postglacial history of European
904 *Pinus sylvestris* populations. *Molecular Ecology* **9**(9)**:** 1205-1211.
905
906

907 Stadler T, Haubold B, Merino C, et al. (2009) The impact of sampling schemes on the
908 site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**(1): 205-
909 216.
910
911 Steven HM, Carlisle A. *The Native Pinewoods of Scotland*. (1959) Edinburgh: Oliver
912 and Boyd.
913
914 Svendsen JI, Astakhov VI, Bolshiyanov DY, Demidov I, Dowdeswell JA, Gataullin V *et*
915 *al* (1999). Maximum extent of the Eurasian ice sheets in the Barents and Kara Sea region
916 during the Weichselian. *Boreas* **28**(1)**:** 234 - 242.
917
918 Tobolski JJ, Hanover JW (1971). Genetic Variation in the Monoterpenes of Scotch pine.
919 *Forest Science* **17**(3)**:** 293-299.
920
921 Wachowiak W, Balk P, Savolainen O (2009). Search for nucleotide diversity patterns of
922 local adaptation in dehydrins and other cold-related candidate genes in Scots pine (Pinus
923 sylvestris L.). *Tree Genetics & Genomes* **5**(1)**:** 117-132.
924
925 Wang J-T, Gould J, Padmanabhan V, Newton R (2003). Analysis And Localization of the
926 Water-Deficit Stress-Induced Gene (lp3). *Journal of Plant Growth Regulation* **21**(4)**:**
927 469-478.
928
929 Wang XR, Szmidt AE, Lindgren D (1991). Allozyme differentiation among populations
930 of *Pinus sylvestris* L. from Sweden and China. *Hereditas* **114**(3)**:** 219-226.
931
932 Watterson GA (1975). On the number of segregating sites in genetical models without
933 recombination. *Theor Popul Biol* **7**(2)**:** 256 - 276.
934
935 Worrell R (1992). A Comparison Between European Continental and British Provenances
936 of Some British Native Trees: Growth, Survival and Stem Form. *Forestry* **65**(3)**:** 253-
937 280.
938
939 Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD *et al* (2005).
940 The Effects of Artificial Selection on the Maize Genome. *Science* **308**(5726)**:** 1310-1314.
941
942 Wright SI, Gaut BS (2005). Molecular population genetics and the search for adaptive
943 evolution in plants. *Mol Biol Evol* **22**(3)**:** 506-519.
944
945
946
947

948

949

950

951

952

953

Titles and legends to figures

Figure 1. Main map: location of 21 Scots pine populations from Scotland (divided for most between-population analyses into groups: ● West, ▲ South and + East). Inset shows locations of the 8 mainland European populations with which comparisons were made and location of main map (highlighted). See Material and Methods for details.

Figure 2. Scatter plot of the squared correlation coefficient of allele frequencies (r^2) as a function of distance in base pairs between pairs of polymorphic sites in western (A), southern (B) and eastern (C) groups at all loci combined. Decline in linkage disequilibrium is shown by nonlinear fitting curve of the mutation-recombination-drift model (see material and methods section for details). Recombination rate parameter ρ (standard error in parenthesis) for western group is ρ= 0.0074 (0.0008), for southern group is ρ =0.0025 (0.0004) and ρ =0.0024 (0.0006) for the east.

A. WEST **B. SOUTH** **C. EAST**

Tables and Figures

Table1. Loci included in nucleotide diversity analyses

			Base pairs screened				
Gene	**Protein / Function**	**n**	**Total**	**Coding[a]**	**Intron**	**UTRs[b]**	**Indels[c]**
dhn1	dehydrin 1 – dehydrative stress response	40	1265	489	423	353	10 (194)
dhn2	dehydrin 2 – dehydrative stress response	33	449	235 (2)	119	95	2 (6)
dhn3	dehydrin 3 – dehydrative stress response	34	428	330 (1)	-	98	2 (69)
dhn7	dehydrin 7 – dehydrative stress response	38	364	264 (1)	-	100	1 (2)
dhy2PP	dehydrin – dehydrative stress response	42	485	381 (2)	95	9	1 (8)
abaR	abscisic acid responsive protein	40	419	334 (1)	85	-	4 (23)
a3ip2	ABI3-interacting protein 2	39	882	169 (2)	120	593	1 (21)
ccoaomt	caffeoyl CoA *O*-methyltransferase	41	563	316 (3)	247	-	0
chcs	chalcone synthase	35	331	85 (1)	-	246	1 (1)
erd3	early responsive to dehydration 3	38	583	379 (3)	204	-	0
lp3-1	ABA and WDS induced gene-1	35	438	168 (1)	-	270	1 (8)
lp3-3	ABA and WDS induced gene-3	36	463	232 (2)	231	-	6 (154)
Total		**451**	**6670**	**3382 (19)**	**1524**	**1764**	**29 (486)**

n – haploid sample size, [a] number of exons in parenthesis; [b] untranslated region (5'UTR); [c] number of indels and length range in parenthesis;

Table 2. Summary statistics of nucleotide and haplotype variation and frequency distribution spectrum of polymorphism at analysed genes in Scottish populations of Scots pine. Silent sites variation reported separately for west, south and east geographical groups, otherwise average values for all samples combined.

			Nucleotide diversity											Haplotype diversity
			Nonsynonym.		Silent[a]									
Locus	L	π	SNPs	π	SNPs	π	π_{West}	π_{South}	π_{East}	ρ[b] (SE)	*D*[c]	*H*[d]	N	H_d (SD)
dhn1	1071	0.0144	8 (4)	0.0044	54 (12)	0.0203	0.0215	0.0225	0.0158	0.0113 (0.0012)	0.039	-0.285	24	0.964 (0.014)
dhn2	442	0.0074	1	0.0024	7 (0)	0.0112	0.0107	0.0110	0.0113	0.0779 (0.0379)	1.968[*1,2]	0.458	11	0.888 (0.028)
dhn3	359	0.0198	10 (2)	0.0137	13 (2)	0.0291	0.0325	0.0226	0.0366	-	0.911	1.733	8	0.829 (0.03)
dhn7	362	0.0042	3 (1)	0.0042	4 (3)	0.0043	0.0033	0.0036	0.0061	-	-0.223	-2.339	6	0.711 (0.039)
dhy2PP	476	0.0101	5 (3)	0.0011	12 (2)	0.0244	0.0240	0.0223	0.0257	0.0680 (0.0228)	0.103	1.738	21	0.954 (0.015)
abaR	396	0.0048	5 (0)	0.0052	3 (2)	0.0043	0.0035	0.0048	0.0046	-	0.052	-2.323	10	0.755 (0.057)
a3iP2	861	0.0043	2 (2)	0.0008	12 (5)	0.0049	0.0051	0.0045	0.0054	0.0022 (0.0016)	0.360	-2.231	11	0.779 (0.052)
ccoaomt	563	0.0018	1 (1)	0.0000	5 (0)	0.0032	0.0011	0.0053	0.0038	-	-0.711	-0.746	5	0.348 (0.092)
chcs	330	0.0075	1 (1)	0.0008	12 (6)	0.0094	0.0066	0.0137	0.0088	-	-0.682	-1.267	9	0.766 (0.065)
erd3	583	0.0021	3 (3)	0.0006	7 (4)	0.0037	0.0036	0.0029	0.0046	-	-1.464[*2]	-1.351	9	0.73 (0.052)
lp3-1	430	0.0095	1 (1)	0.0004	12 (1)	0.0137	0.0107	0.0144	0.0135	0.0195 (0.0111)	0.944	1.175	20	0.955 (0.017)
lp3-3	309	0.0370	8 (0)	0.0177	18 (1)	0.0660	0.0692	0.0683	0.0676	0.0018 (0.0011)	2.846[***1,2]	1.600	19	0.949 (0.019)
Total	**6182**		**48(18)**		**159 (38)**									
Mean[e]	**515**	**0.0078**		**0.0031**		**0.0117**	**0.0111**	**0.0116**	**0.0124**	**0.0085 (0.0009)**[f]	**0.118**	**-0.494**	**12.18**	**0.789 (0.042)**

L – length of sequence in base pairs excluding indels; π – nucleotide diversity (Nei 1987); N – number of haplotypes; H_d – haplotype diversity (standard deviation); [a] synonymous and noncoding positions; [b] least-squares estimate of recombination parameter; [c] *D* test (Tajima 1989); [d] *H* test (Fay and Wu 2000); [e] average values at 11 loci excluding *lp3-3*; [f] stimates based on informative sites at all loci excluding *lp3*-3; "-" not estimated due to low number of informative sites; [*] statistically significant values based on coalescence simulations (1) without recombination and (2) with average recombination rate at six loci [*]P<0.05; [***] P<0.01

Table 3. Descriptive statistics for nucleotide variation at eight loci in Scottish and continental European populations of Scots pine. Description of regional groups in Scotland as in Figure 1 and Supplementary Table S1.

Groups		θ^a	C.I. (95%)[b]	ρ (SE) [c]	ρ / θ	*D*[d]	*H*[e]
Scottish	**West**	0.0103	0.0072 – 0.0147	0.0073 (0.0008)	0.71	0.580*	-0.400
	South	0.0130	0.0089 – 0.0188	0.0020 (0.0004)	0.15	0.107	0.066
	East	0.0117	0.0080 – 0.0170	0.0021 (0.0006)	0.18	0.499*	-0.128
	All	0.0108	0.0081 – 0.0145	0.0085 (0.0009)	0.79	0.316	-0.564
Continental European	**North**[f]	0.0095	0.0065 – 0.0137	0.0062 (0.0010)	0.65	-0.143	-0.750
	Central[g]	0.0103	0.0072 – 0.0147	0.0090 (0.0009)	0.87	-0.359	-1.077
	North+Central	0.0096	0.0070 – 0.0131	0.0214 (0.0019)	0.45	-0.316	-1.116
	Spain	0.0098	0.0058 – 0.0167	-	-	-0.539	-0.371
	Turkey	0.0055	0.0030 – 0.0099	-	-	-0.279	-0.792
	All	0.0093	0.0068 – 0.0125	0.0245 (0.0002)	2.69	-0.379	-1.240

[a] median for silent sites;
[b] 95% credibility intervals for θ;
[c] least-squares estimate of ρ;
[d] Tajima's *D* test based on all sites; *P<0.05, statistical significance determined by coalescent simulations with and without recombination (see material and methods);
[e] Fay and Wu *H* test;
[f] North: Finland North, Finland South, Sweden;
[g] Central: Poland, France, Austria; " -" not estimated due to low sample size (~5 for each locus) and low number of informative sites from each population.

Table 4. Differentiation between Scottish and continental European populations of Scots pine measured as average F_{ST} over all polymorphic sites and indels at 8 loci combined.

	North	Central	Spain	Turkey	North+Central	All[a]
West	0.032***	0.026**	0.02	0.091***	0.029*	0.022*
South	0.009	0.011	0.053**	0.112***	0.010	0.011
East	0.019	0.040***	0.072**	0.145***	0.037*	0.039*
All Scottish	0.023**	0.035**	0.035*	0.095***	0.028*	0.025*

[a] all continental European populations combined; *P<0.05, **P<0.01, ***P<0.001;

Table 5. Alternative demographic models tested against total and regional groups of populations in Scotland

Group	Observed[a]		SN[b]		Growth[c]		Bottleneck[d]	
	Mean *D*	Mean *H*	Mean *D*	Mean *H*	Mean *D*	Mean *H*	Mean *D*	Mean *H*
West	0.364	-0.447	-0.057 (0.578)	0.001 (0.406)	-0.059 (0.585)	-0.018 (0.411)	0.371 (0.116)	-0.504 (0.630)
South	0.103	0.144	-0.066 (0.588)	-0.009 (0.419)	-0.056 (0.575)	0.023 (0.403)	0.310 (0.161)	-0.494 (0.636)
East	0.260	-0.145	-0.056 (0.589)	0.003 (0.419)	-0.065 (0.596)	-0.013 (0.434)	0.311 (0.165)	-0.487 (0.645)
All	-0.015	-0.494	-0.072 (0.602)	0.028 (0.407)	-0.072 (0.605)	0.026 (0.406)	0.661 (0.020)	-0.495 (0.613)

[a] observed mean values of Tajima's *D* and Fay and Wu's *H* statistics at 11 loci

[b] standard neutral model

[c] results for exponential growth of rate 10 starting 0.00125 x 4*Ne* before present

[d] results shown are for bottleneck of severity s=0.02 that started 0.00175 x $4N_e$ generations before present. Duration of bottleneck was set up to 0.0015 and population growth rate to 10. Assuming e.g. N_e =200000 and generation time of 25 years, the bottleneck ended about 25 thousand years ago. Current and the ancestral population size were assumed to be qual. In parenthesis are the *P*-values for the observed means of each parameter.

Title: High genetic diversity at the extreme range edge: nucleotide variation at nuclear loci in Scots pine (*Pinus sylvestris* L.) in Scotland

Authors: Witold Wachowiak[1,2], Matti J. Salmela[1,3], Richard A. Ennos[3], Glenn Iason[4], Stephen Cavers[1]

[1] Centre for Ecology and Hydrology Edinburgh, Bush Estate, Penicuik, Midlothian EH26 0QB, UK

[2] Institute of Dendrology, Polish Academy of Sciences, Parkowa 5, 62-035 Kórnik, Poland

[3] Institute of Evolutionary Biology, School of Biological Sciences, Ashworth Laboratories, University of Edinburgh, Edinburgh EH9 3JT, UK

[4] Macaulay Land Use Research Institute, Craigiebuckler, Aberdeen AB15 8QH, UK

Corresponding author: Stephen Cavers (scav@ceh.ac.uk) Centre for Ecology and Hydrology Edinburgh, Bush Estate, Penicuik, Midlothian EH26 0QB, UK. Phone: +44 (0) 131 4458552, Fax: +44 (0) 131 4453943

Keywords: adaptation, bottleneck, nucleotide diversity, population differentiation, linkage disequilibrium, recolonisation

Running title: Nucleotide diversity in Scots pine

Abstract:

Nucleotide polymorphism at twelve nuclear loci was studied in Scots pine populations across an environmental gradient in Scotland, to evaluate the impacts of demographic history and selection on genetic diversity. At eight loci, diversity patterns were compared between Scottish and continental European populations. At these loci, a similar level of diversity (θ_{sil}=~0.01) was found in Scottish vs. mainland European populations contrary to expectations for recent colonisation, however less rapid decay of linkage disequilibrium was observed in the former (ρ=0.0086±0.0009, ρ=0.0245±0.0022 respectively). Scottish populations also showed a deficit of rare nucleotide variants (multilocus Tajima's *D*=0.316 vs. *D*=-0.379) and differed significantly from mainland populations in allelic frequency and/or haplotype structure at several loci. Within Scotland, western populations showed slightly reduced nucleotide diversity (π_{tot}=0.0068) compared to those from the south and east (0.0079 and 0.0083, respectively) and about three times higher recombination to diversity ratio (ρ / θ=0.71 versus 0.15 and 0.18, respectively). By comparison with results from coalescent simulations, the observed allelic frequency spectrum in the western populations was compatible with a relatively recent bottleneck (0.00175 x $4N_e$ generations) that reduced the population to about 2% of the present size. However heterogeneity in the allelic frequency distribution among geographical regions in Scotland suggests that subsequent admixture of populations with different demographic histories may also have played a role.

Introduction

Nucleotide polymorphism is influenced by several factors including mutation, migration, selection and random genetic drift. In tree species, the current increase in sequence data gathered from nuclear gene loci has been driven mostly by the search for the molecular signature of natural selection (Achaz, 2009; Neale and Ingvarsson, 2008; Savolainen and Pyhäjärvi, 2007). Selection can leave its traces as deviations from neutrality in the level of nucleotide diversity, allele frequency distribution or correlation between polymorphic sites (linkage disequilibrium) (Achaz, 2009). However, the capability to detect selection at individual loci is heavily dependent on the assumptions of the neutral model (e.g. constant long term population size, random mating), the strength of, and time since, selection and the number of loci involved (and their relative effect) in selectively-influenced traits (Wright and Gaut, 2005). Therefore, prior to testing for selection, datasets must be evaluated for violations of neutral model assumptions. Such processes, e.g. historical changes in population size and distribution, may drive deviations from neutrality that mimic the effect of selection. However, these effects are expected to be genome-wide and so can be distinguished from selective influences by simultaneous assessment of data from multiple loci. Although the patterns of variation in the majority of nuclear loci studied to date obey neutral expectations and the signature of selection has been elusive (Savolainen and Pyhäjärvi, 2007) polymorphisms at nuclear loci provide highly valuable insights into evolutionary history (Heuertz *et al*, 2006; Pyhäjärvi *et al*, 2007).

All northern European tree populations have experienced substantial historical changes in distribution. For example, palynological and phylogeographic data indicate that during the last glacial maximum (25-18 000 years ago (ya)), most species were confined to the southern peninsulas (Iberia, Italy and the Balkans) and some parts of eastern and central Europe (Cheddadi *et al*, 2006; Pyhäjärvi *et al*, 2008; Willis and van Andel, 2004) and only reached their most northerly limits around 9000 ya. The recolonization history of forest trees, accompanied by adaptation to local environments, has potentially influenced the pattern of nucleotide diversity both among locally adapted populations and between range edge populations and putative refugial populations. In theory, population bottlenecks reduce nucleotide diversity in range-edge populations relative to that in source populations, although this is dependent on the timing and severity of the bottleneck. In contrast, admixture of populations due, for example, to recolonization from different refugia, may increase diversity (Petit~~i~~ et al. 2003). However, recent studies in continental European populations of Scots pine (Pyhäjärvi *et al*, 2007) and Norway spruce (Heuertz *et al*, 2006) found little evidence at the nucleotide level for the effects of recent (post-glacial) population size changes during migration and suggested bottlenecks in the mid-to-late Pleistocene. In addition, similar to other predominantly outcrossing tree species with highly efficient long distance gene flow via pollen (Hamrick *et al*, 1992), neutral genetic differentiation between Scots pine populations is low. For instance, marginal population differentiation was reported for neutral markers between Finnish populations (Karhu *et al*, 1996), between Scandinavian and eastern parts of the range (Wang *et al*, 1991) and, at several candidate gene loci for growth phenology and cold tolerance, among populations along a latitudinal cline in continental Europe (Dvornyk *et*

al, 2002; García-Gil *et al*, 2003). The large population sizes of forest trees and capability for maintenance of high levels of genetic variation within populations seems to further buffer against rapid changes in genetic diversity, but causes difficulties in detection of recent demographic processes. If the migrations following the most recent glaciations are to have left any signature at all in contemporary populations of forest trees, it seems likely to be detectable only where populations have experienced severe bottlenecks or became rapidly isolated.

In Scotland, Scots pine (*Pinus sylvestris* L.*)* is at the extreme north-western edge of its vast distribution, which reaches across Europe and Asia and is the largest of any pine species (Critchfield and Little, 1965). Pines first colonized the land that became the British Isles about 10 000 ya, at around the time that Ireland became isolated, and reached northern Scotland by about 9000 years ago (Huntley and Birks, 1983; Svendsen *et al*, 1999). According to fossil data in Scotland, pine first appeared in the Wester Ross region in the northwest, and then shortly afterwards in the Cairngorms in the east (Birks, 1989). The subsequent formation of the English Channel (c.6000 ya) and competition from broadleaved species in southern Britain left Scottish pinewoods physically separated by at least 500 km from mainland populations in continental Europe. Nowadays, native pinewoods in Scotland cover about 18 000 hectares, in 84 differently-sized fragments patchily distributed within a ~ 200 x 200 km area across significant environmental gradients in altitude, soil type, growing season length and annual rainfall mainly in the east-west direction (e.g. annual rainfall varies from 700 to 3000 mm across 160 km) (Mason *et al*, 2004). Small-scale provenance experiments have shown genetic variation

between Scottish populations from different locations, e.g. in root frost hardiness and growth in seedlings (Perks and McKay, 1997) and differentiation among populations at several quantitative traits (Perks and Ennos, 1999). There is reasonable evidence from pollen (Birks, 1989), allozymes, monoterpenes and *mt*DNA (Kinloch *et al*, 1986; Sinclair *et al*, 1998) suggesting a west/east population subdivision within Scotland and that populations from these regions may have different origins (Ballantyne and Harris, 1994; Bennett, 1995). Given the iconic status of Scots pine in Scotland and the severe fragmentation of the population, there is considerable interest in evaluating its population history.

In this study, we focus on the Scottish Scots pine population as a unique and isolated oceanic fragment at the northwest extreme of the distribution to assess whether recent demographic processes have influenced patterns of nucleotide variation. We analysed patterns of nucleotide diversity, allele frequency and linkage disequilibrium in a multilocus nuclear gene dataset in samples gathered from multiple locations within putatively divergent regions within Scotland and compared our data to those from samples from northern and central Europe, Turkey and Spain. Using this data and coalescent simulation analysis, we aimed to assess whether Scottish populations show the molecular signature of demographic history and the extent to which they are differentiated from those in continental Europe.

Materials and Methods

Sampling and DNA extraction

Seed samples from 21 locations in Scotland were included in the study (Figure 1.). The trees were sampled across an environmental gradient related to differences in altitude, length of growing season, annual rainfall and average mean temperature in winter (Supplementary Table S1). Cones were collected from mature trees in recognised old-growth Scots pine forest; at these sites trees are typically over 150 years old and often much older (Steven and Carlisle, 1959). Trees were separated by at least 50 m to minimise sampling of closely related individuals. Sampling included the seven currently adopted seed zones of the species in Scotland, from each of which 3 locations were sampled, 2 individuals per location.

For most of the between-population analyses the samples were grouped according to climatic characteristics into three geographical locations – western, southern and eastern, represented by eighteen, twelve and twelve individuals, respectively (Figure 1, Supplementary Table S1). The western group has the lowest mean altitude (~142m), the longest growing season (~ 240 days), highest mean temperature in winter (~ 2^0C) and high annual rainfall (~ 2000mm). The eastern group has the highest mean altitude (~372m), the shortest growing season (~175 days), and is the coldest (-0.1^0C) and driest (~1050mm) part of the distribution, whilst southern group was intermediate between these extremes except for annual rainfall (~2130mm). Field trials have demonstrated genetic differences in phenology and growth rate among provenances originating within these groups (Perks and Ennos, 1999).

Genomic DNA was extracted from haploid megagametophyte, maternal tissue which surrounds the embryo in the seed. As DNA samples were haploid, the haplotypes

could be determined by direct sequencing. In total, 42 DNA extracts were prepared, representing two different trees from each location. Seeds were germinated for a few days in moisturized petri dishes and then extracted following a standard CTAB (cetyltrimethylammonium bromide) protocol with addition of PVP to 1% concentration in the lysis buffer.

Loci studied

In total, sixteen nuclear loci were analysed. This included several dehydrin genes that were identified in expression studies in Scots pine (Joosen *et al*, 2006). Based on the number and position of the conserved segments (Close 1997), we analysed the class SK4 of dehydrins (*dhn1*), SK2 (*dhn2*) and a group of K2 genes (*dhn3* and *dhn7*). We analysed also SK type of dehydrin upregulated by water stress in *Pinus taeda* roots (Eveno *et al*, 2007) and a putative dehydrin (*dhy-like*) described for Scots pine (Pyhäjärvi *et al*, 2007). Other loci described in more detail in original papers include abscissic acid responsive protein (*abaR)* (Wachowiak *et al*, 2009); early response to dehydration 3 protein (*erd3*), abscissic acid, water dehydrative stress and ripening induced gene family members 1 and 3 (*lp3-1, lp3-3*), Caffeoyl CoA *O*-methyltransferase (*ccoaomt*), putative arabinogalactan/ proline-rich protein (*PR-AGP4-1*) and putative arabinogalactan/ glycin-rich protein (*grp3*) (Eveno *et al*, 2007); ABI3-interacting protein 2 (*a3ip2*), alcohol dehydrogenase C (*adhC*) and chalcone synthase (*chcs*) (Pyhäjärvi *et al*, 2007).

In previous work, ten loci (*dhn1,2,3,7*, *dhy-like*, *dhy2PP*, *abaR*, *a3ip2, adhC, chcs*) were analysed in Scots pine from the continental European range including fifteen samples from Northern Europe (populations from Northern and Southern Finland and

Sweden), fifteen from Central Europe (Poland, Austria and France), and five from each of Turkey and Spain (Pyhäjärvi *et al*, 2007; Wachowiak *et al*, 2009). The reference sequences of eight loci in total (excluding *dhy-like* and *adhC,* see below) were compared with those from Scottish populations. The samples from the Iberian Peninsula and Turkey were treated separately in between-region comparisons as they display specific mitochondrial types not observed in mainland European distribution of the species which suggests different histories and no contribution to recolonization after last glaciation (Pyhäjärvi *et al*, 2008; Soranzo *et al*, 2000).

PCR amplification and sequencing

PCR-amplification was performed with PTC-200 (MJ Research) and carried out in a total volume of 25µl containing about 10ng of haploid template DNA, 50µM of each of dNTP, 0.2µM of each primer and 0.25U *Taq* DNA polymerase with the respective 1x PCR buffer (NovaZyme, Poland). PCR followed standard amplification procedures with $MgCl_2$ concentration optimised for each primer pair as described in Supplementary Table S2. PCR fragments were purified using QIAquick^{TM}PCR Purification Kit (Qiagen). About 20 ng of PCR product was used as a template in 10 µl sequencing reactions with the Big Dye Terminator DNA Sequencing Kit (Applied Biosystems) performed by the GenePool sequencing service, University of Edinburgh. All samples were sequenced in both directions. CodonCode Aligner software was used for editing and assembling of the sequence chromatograms to produce alignments based on nucleotide sequence from both DNA strands. Haplotype sequences of each locus reported in this paper are deposited in the EMBL sequence database under accession numbers GQ262040 – GQ262490.

204

205 ***Sequence analysis***

206 High quality sequences were obtained for most of the samples at twelve loci (Table 1).
207 PCR amplification or sequencing failed in most of the samples at *dhy-like*, *adhC* and *grp3*
208 and these loci, together with *PR-AGP4-1* which was monomorphic across all 42 samples,
209 were excluded from further analysis. Nucleotide sequence alignments were constructed in
210 ClustalX and were further manually adjusted using GenDoc. All sequence
211 polymorphisms were visually rechecked from chromatograms edited with BioEdit.
212 Coding and noncoding regions (introns, UTRs) were annotated based on the NCBI
213 (http://www.ncbi.nlm.nih.gov/) sequence information at each locus and web-based gene
214 identification tool at PlantGDB (http://www.plantgdb.org/cgibin/PlantGDB/GeneSeqer/
215 PlantGDBgs.cgi). The influence of demography on the multilocus pattern of variation and
216 locus specific effects were assessed by looking at the amount of nucleotide diversity,
217 correlation between polymorphic sites and allelic frequency distribution between
218 different geographical locations in Scotland and in comparison to mainland populations
219 of the species and by comparing observed statistics with simulated values under a range
220 of demographic scenarios. Neutrality tests at intraspecific level were applied to search for
221 departures from a neutral model of evolution. Sequences from *Pinus pinaster* were used
222 as an outgroup for intraspecific comparisons to test for a signal of longer term selection.

223

224 ***Nucleotide diversity***

225 Two measures of nucleotide diversity were applied: 1) an average number of nucleotide
226 differences per nucleotide site between two sequences π, (Nei, 1987) calculated with

227 DNAsp 4.0) and 2) Watterson's (1975) estimate of the population mutation parameter,
228 theta (θ_W, equal to $4N_e\mu$, where N_e is the effective population size and μ is the mutation
229 rate per nucleotide site per generation), computed based on the number of segregating
230 sites and the length of each locus using MCMC simulation under a Bayesian model as
231 previously described (Pyhäjärvi *et al*, 2007). The estimates of nucleotide diversity were
232 conducted for all samples combined and separately for south, east and west regional
233 groups of Scottish populations. Scottish and continental European populations were
234 compared at eight loci for which informative data was available (Pyhäjärvi *et al*, 2007;
235 Wachowiak *et al*, 2009). Exceptionally high nucleotide diversity was found at *lp3-3* locus
236 compared to other loci in our dataset. Due to the size of the conifer genome and the
237 occurrence of multigene families (Ahuja and Neale, 2005), erroneous co-amplification of
238 different loci from the same family is possible and may account for unusual diversity
239 estimates at specific loci. Therefore, locus *lp3-3* was excluded from multilocus or average
240 estimates reported in the study to avoid bias and ensure that estimates were conservative;
241 the locus was included in population structure analysis and coalescent simulations.
242
243 ***Linkage disequilibrium and haplotype diversity***
244 The level of linkage disequilibrium was measured as the correlation coefficient r^2
245 (Hill and Robertson, 1968) using informative sites. Indels and sites with three nucleotide
246 variants identified in *dhn1* (3), *dhy2PP* (1) were excluded from the analysis. Under
247 mutation-drift-equilibrium model, the decay of linkage disequilibrium with physical
248 distance was estimated using non-linear regression of r^2 between polymorphic sites and
249 the distance (in base pairs) between sites as detailed in (Wachowiak *et al*, 2009). The

non-linear least-squares estimate of ρ ($\rho = 4N_e c$, where N_e is effective population size, c is the recombination rate) between adjacent sites was fitted by the nls-function implemented in the R statistical package (http://www.r-project.org). The overall and group specific least-squares estimates of ρ were computed and compared to other estimates in Scots pine (Pyhäjärvi *et al*, 2007; Wachowiak *et al*, 2009).

The number of haplotypes and haplotype diversity (H_d) were estimated for each gene using DNAsp. Insertions and deletions were included in all estimates. Coalescence simulations with locus specific or average ρ for six loci and without recombination were used to assess whether there are more or fewer haplotypes than expected and whether haplotype diversity is higher or lower than expected given the number of segregating sites. The number of haplotypes and haplotype diversity were calculated for all samples combined and separately for the three regional groups of Scots pine in Scotland.

Neutrality tests

Deviations of particular genes from the frequency distribution spectrum under the standard neutral model of evolution were assessed with Tajima's *D* test (Tajima 1989) and Fay and Wu's *H* (Fay and Wu 2000). Negative values of Tajima's *D* indicate an excess of low frequency polymorphisms consistent with positive directional selection or recent population expansion, whereas positive values indicate an excess of intermediate frequency polymorphism potentially due to balancing selection or population contraction. Fay and Wu's *H* test measures departures from neutrality based on high-frequency derived alleles. An excess of high frequency derived alleles compared to neutral expectations may result from recent positive selection or strong population structure with

uneven sampling from populations. The distribution of test statistics was investigated for each locus for all populations combined and separately for the three regional groups. Multilocus estimates of Tajima's *D* were assessed with HKA software (http://lifesci.rutgers.edu/~heylab). The estimates were also calculated along the sequence of each locus by a sliding window of 100 sites with successive displacement of 25 sites. As lack of recombination makes the *D* test overly conservative (Thornton 2005), the significance of locus specific and multilocus Tajima's *D* was also evaluated by coalescent simulations dependent on population mutation and recombination rate (MANVa software www.ub.edu/softevol/manva, based on coalescent program *ms*, Hudson, 2002). Different estimates of ρ including locus specific estimates, lowest and highest value across loci and average value for six loci were used in coalescent simulations. As similar probability values for multilocus *D* statistics were observed in simulations with different recombination rate estimates, the results based on the average values of ρ at the analysed loci are reported unless otherwise stated.

For tests based on nucleotide variation between species we used reference sequence data from *P. pinaster* for outgroup comparison. To assess the correlation between the level of nucleotide polymorphism and divergence at each locus we applied 1) the McDonald and Kreitman, (1991) test, based on comparison of the pattern of within-species polymorphism and between-species divergence at synonymous and nonsynonymous sites in a gene, and 2) HKA test (Hudson *et al*, 1987) which allows the detection of loci that demonstrate unusual patterns of polymorphism compared to divergence across genes. Comparison of multilocus polymorphism and divergence at all sites was assessed using HKA software (http://lifesci.rutgers.edu/ ~heylab). The ratio of

296 nonsynonymous (*K*a) and synonymous site (*K*s) nucleotide divergence from the outgroup
297 species (Hughes and Nei, 1988) was calculated using DnaSP.
298
299 ***Population structure***
300 To check if there was a geographical difference in allelic frequency spectra, regional
301 groups of Scottish populations were compared to each other and to previously analyzed
302 continental European populations from northern and central Europe, Spain and Turkey
303 (Pyhäjärvi *et al*, 2007; Wachowiak *et al*, 2009). Genetic differentiation between the
304 regions was studied locus by locus at both haplotype and SNP/Indel level and also by
305 averaging pairwise F_{ST} over all polymorphic sites across loci. The significance of genetic
306 differentiation was evaluated by 1000 permutations of the samples between groups using
307 Arlequin ver. 3.0 (Excoffier *et al*, 2005). Population structure from the haplotypic data
308 was tested by S_{nn} and K_{ST}^{*} statistics (Hudson *et al*, 1992; Lynch and Crease, 1990), which
309 are more appropriate for sequence-based haplotype data where diversity may be high and
310 sample size low, rendering frequency-based approaches problematic. Their significance
311 was evaluated using 1000 permutations, where samples were randomly assigned into
312 different groups (Hudson, 2000). Genetic clustering of the individuals based on both full
313 sequence data and all segregating sites and indels at 12 loci (for Scottish populations) and
314 at 8 loci (for Scottish and mainland European populations) was conducted using BAPS
315 5.2 (Corander and Tang, 2007). Polymorphic sites from each locus were treated as linked
316 molecular data to account for dependence between segregating sites in the gene.
317 Completely linked sites (r^2=1) were excluded from the analysis.
318
319 **Coalescent simulations**

To further infer the demographic history of Scottish Scots pine populations we compared the observed distribution of average Tajima's *D* and Fay and Wu's *H* at the candidate loci separately in the western, southern and eastern group and in all geographic regions combined to the simulated values under several demographic scenarios including the standard neutral model (constant population size), growth model and bottleneck model followed by exponential growth (Supplementary Figure S4). Regional groups of populations were analysed separately as detailed aspects of the frequency spectrum may differ between groups that are not differentiated based on genetic clustering methods (Pyhäjärvi et al 2007). Coalescent simulations were run independently for each locus and various demographic scenarios using the program *ms* (Hudson 2002) and the approach described by Haddrill *et al* (2005). In each case, 5000 replicates were simulated for each locus. The analyses were performed with recombination using the locus specific (when available) or average value of ρ per site for the analysed loci in each geographic group (Table 2, Supplementary Table S3). We tested various bottleneck scenarios of different age and severity. The time from the end of the bottleneck (measured in units of $4N_0$ generations from the present) ranged from 0.0002 to 0.05 and bottleneck severities (measured in units of the current population size) from 0.001 to 0.5. Assuming for instance, N_e of 200000 and generation time of 25 years, the time range corresponds to between 4 000 and 1 million years and severity from 0.1 to 50% of the current population size. In most bottleneck models tested, the ancestral and current effective population sizes were assumed to be equal, bottleneck duration (f) was fixed to f=0.0015 (units of $4N_0$ generations from the present) and the growth rate of 10 was constant across simulations as in previous studies (Heuertz et al. 2006). A subset of simulations were run also with

343 *f*=0.006 and corresponding equal or doubled ancestral population size as compared to the
344 current one, and also separately for a set of 11 and 9 loci (excluding *lp3-3* and *dhn1* and
345 *abaR*, respectively as the later showed some evidence of selection). A schematic
346 representation of the simulated bottleneck model is shown in Supplementary Figure S4.
347 The simulation results for each demographic scenario were summarized using the
348 program analyser HKA. The perl script multitest_pop1.pl was used to perform multilocus
349 tests of *ms*-generated genealogies (including *P*-values of the observed mean values of
350 Tajima's *D* and Fay and Wu's *H* statistics) summarized using analyser HKA. The
351 programs are available from http://genomics.princeton.edu/
352 AndolfattoLab/Andolfatto_Lab.html.
353
354 **Results**
355 **Nucleotide polymorphism and divergence**
356 The average total nucleotide diversity (π) in Scottish populations at eleven loci was π_{tot} =
357 0.0078 and at nonsynonymous sites was π_{ns} = 0.0031 (Table 2). Slightly lower average
358 nucleotide diversity was found in the west (π_{tot} = 0.0068) as compared to southern and
359 eastern regional groups (π_{tot} = 0.0079 and 0.0083, respectively) and similar values were
360 found at nonsynonymous sites (π_{ns}= ~ 0.003) (Supplementary Table S3). Multilocus
361 estimates of silent Watterson theta was θ_{sil}=0.0095 (with 95% credibility intervals of
362 0.0074-0.0122) for all Scottish populations combined, θ_{sil}=0.0086 (0.0063-0.0117) in the
363 west, θ_{sil}=0.0111 (0.0080-0.0152) in the south and θ_{sil}=0.0103 (0.0074-0.0143) in the
364 east. In comparisons between Scottish vs. mainland European populations at eight loci,
365 similar but slightly higher average values of total nucleotide diversity (π_{tot} = 0.0070 vs

366 0.0062) and silent multilocous theta (θ_{sil}=0.0108 vs. 0.0093) were found in Scottish
367 populations (Table 3).
368
369 **Linkage disequilibrium and haplotype polymorphisms**
370 Rapid decay of linkage disequilibrium between pairs of parsimony informative sites at
371 eleven loci was found in Scottish populations, with $\rho = 0.0085 \pm 0.0009$ (Table 2) and
372 expected r^2 values of 0.2 at a distance of about 400 bp. The decay of linkage
373 disequilibrium in the western group ($\rho = 0.0074 \pm 0.0008$) was more rapid as compared to
374 the south (0.0025 ± 0.0004) and east (0.0024 ± 0.0006) (Figure 2) and the pattern was
375 constant at most loci (Supplementary Table S3). Overall, Scottish populations had about
376 three times slower decay of linkage disequilibrium as compared to mainland populations
377 at the same set of eight loci of similar sample size ($\rho = 0.0086 \pm 0.0009$ vs
378 0.0245 ± 0.0022, respectively) (Supplementary Figure S1). However, the rate of decay of
379 LD and the relative level of recombination to diversity (ρ / θ ratio) were similar between
380 western Scottish and north and central European regions (Table 3) but these parameters
381 were over three times smaller in southern and eastern groups of Scotland.
382 The average number of haplotypes per gene was 12 and haplotype diversity was very
383 high (H_d=0.789±0.042). Similar haplotype diversity was found in western (H_d
384 =0.754±0.077), southern (H_d=0.819±0.088) and eastern (H_d=0.800±0.090) groups
385 (Supplementary Table S3). Haplotype diversity was slightly higher than mainland
386 European populations at the same set of eight loci (H_d=0.831±0.038 vs H_d=0.795±0.051)
387 and also compared to previous estimates for Scots pine (H_d=0.683±0.059, Wachowiak *et*
388 *al.* 2009). Locus *Lp3-3* contained two sets of haplotypes (each of 18 samples equally

389 distributed across three geographical groups) with highly reduced levels of nucleotide
390 polymorphism (π_{tot} = 0.0090 and 0.0074, respectively) as compared to the whole gene
391 estimate (π_{tot} = 0.0370) and a ten-fold difference in the level of divergence (K_{sil1}=0.013 vs
392 K_{sil2}=0.116) (Supplementary Table S4 and Supplementary Figure S2). A neutral
393 coalescence process, compatible with a constant-size neutral model without
394 recombination or erroneous coamplifications of different gene family members could
395 potentially generate such a pattern. However, no reading-frame shifts or premature stop
396 codons, which would suggest the presence of non-functional alleles, were found at the
397 locus.
398
399 **Neutrality tests**
400 Tendency towards an excess of old over recent mutations across genes was detected by
401 multilocus Tajima's *D* at eleven loci in the total data set (*D*=0.118) (Table 2), in the
402 western (*D*=0.364), southern (*D*=0.103) and eastern (*D*=0.260) groups (Supplementary
403 Table S3). Significant excess of intermediate frequency mutations was found at *dhn2*
404 (*D*=1.968, P<0.05) and *lp3-3* (*D*=2.846, P<0.01). Statistically significant positive values
405 of Tajima's *D* were identified in sliding window analyses in a few regions within *dhn2* (*D*
406 = 2.36-2.48 at 307–449 bp), *a3ip2* (*D* = 2.22 at 401-501 bp) and *lp3-3* (*D* = 2.13-3.18 at
407 51–454 bp) loci. Overall, an excess of high-frequency derived variants indicated by
408 negative mean values of Fay and Wu's *H* statistics was found in all Scottish populations
409 (*H*= -0.494) (Table 2), in the west (*H*= -0.447) and east (*H*= -0.145) groups, but slightly
410 positive values were found in the south (*H*= 0.144) (Supplementary Table S3). The
411 aggregated Scottish populations show a deficit of rare variants (multilocus Tajima's

412 *D*=0.316) as compared to mainland European populations (*D*=-0.379). Both geographical
413 regions show negative mean value of Fay and Wu's *H* statistics (*H*=-0.564 and -1.240,
414 respectively) indicating an excess of high-frequency derived SNPs (Table 3).

415 An excess of fixed nonsynonymous over fixed synonymous substitutions and
416 polymorphic sites was found at *dhn1* locus in McDonald-Kreitman test (Fisher's exact
417 test, $P = 0.05$), as previously found in European mainland populations (Wachowiak *et al*,
418 2009). An excess of nonsynonymous sites as compared to synonymous sites was found at
419 *abaR* (Supplementary Table S5). The level of divergence was similar across all sites and
420 at silent sites only (~4%), and was slightly lower than previous estimates for Scots pine
421 (K=~0.05, Wachowiak *et al.* 2009). Overall, positive correlation between polymorphism
422 and divergence (HKA test) was found at eleven loci combined.

423

424 **Population differentiation**

425 *Differentiation between Scottish populations*

426 Significant differentiation measured as an average over all polymorphic sites was found
427 between southern and eastern groups at *dhn1* (F_{ST}=0.034, P<0.05) and between southern
428 and eastern as compared to the western group at *ccoaomt* (F_{ST}=0.149, P<0.05 and F_{ST}
429 =0.102, P<0.01, respectively) and *lp3-1* (F_{ST} =0.100, P<0.05 and F_{ST} =0.197, P<0.001,
430 respectively) (Supplementary Table S6). A difference in frequency of indel
431 polymorphisms at *dhn1*, four silent substitutions and indel polymorphisms at *lp3-1* and
432 absence of four silent polymorphisms in the western group as compared to the others at
433 *ccoaomt* locus contributed the most to the differentiation between groups. Based on
434 haplotype differentiation, the western group differed from the southern group at *a3ip*

(S_{nn}=0.629, P<0.05), *lp3-1* (S_{nn}=0.758, P<0.01) and at *ccoaomt* and *lp3-1* based on K_{ST} statistics (K_{ST} = 0.066 and 0.051, P<0.05, respectively). They also differ from the east group at *lp3-1* locus (K_{ST}=0.075, P<0.05). Significant F_{ST} statistics based on haplotype frequency were found for *lp3-1* in the south and east as compared to western group (P<0.05) and nearly significant values for *ccoaomt* between south and west groups (P=0.06) (Supplementary Table S6). No difference between west-south, west-east and south-east groups were found based on average F_{ST} over all polymorphic sites and indels combined across the loci (F_{ST} =-0.013, -0.013, and 0.01, respectively).

Differentiation between Scottish vs European continental populations

Based on allele frequency and/or haplotype diversity statistics Scottish populations were differentiated from continental European populations at six out of eight loci analysed (Supplementary Table S7). Significant population differentiation (F_{ST}), measured both as an average over polymorphic sites and at the haplotype level, was found at *dhn2*, *dhn7*, *abaR* and *chcs*. Based on the average proportion of nearest-neighbor haplotypes that are present in the same locality (S_{nn}) both groups were differentiated at *dhn2*, *dhn7*, *dhy2PP* and *a3iP* (P<0.001-0.05). Two loci, *dhn2* and *dhn7*, also showed high similarity between pairs of sequences derived from each region (K_{ST} = 0.098 and 0.067, respectively, P<0.01).

Significant differentiation was found between Scottish populations versus continental European populations measured as an average of F_{ST} values over all polymorphic sites detected (Table 4). The only exceptions include southern Scottish populations as

457 compared to northern and central Europe, eastern Scottish compared to northern
458 European and western Scottish compared to Spanish populations.
459 Analysis of genetic clustering with full sequence data gave the best support for all
460 individuals from European mainland and Scottish populations at eight loci and for
461 individuals from Scottish populations at 12 loci belonging to one genetic cluster. At all
462 polymorphic sites and indels at both eight and twelve loci, the best support was obtained
463 for four clusters, but without clear pattern of geographical distribution (Supplementary
464 Figure S3).
465

466 **Coalescent simulations**

467 For each geographic group of Scottish Scots pine populations the observed pattern of the
468 frequency distribution spectrum was not compatible with either the standard neutral or
469 growth models. In simulations under the SNM and growth model the mean Tajima's *D*
470 was significantly lower and Fay and Wu's *H* significantly higher than the observed values
471 except for the southern group, the only one with positive mean *H* values (Table 5,
472 Supplementary Table S8). Among the 20 different bottleneck models tested the most
473 compatible for the western group was a relatively recent bottleneck (t=0.00125) that
474 reduced the population to 2% of the present size followed by moderate population growth
475 (Table 5, Supplementary Table 9). This model also held for the eastern group but was
476 always rejected for the southern group, where different bottleneck scenarios never lead to
477 positive values for both Tajima's *D* and Fay and Wu's *H* statistics (Supplementary Table
478 9). In general, the simulations indicate heterogeneity in the allelic frequency distribution
479 among geographic regions in Scotland.

Discussion

Multilocus signatures of population history

The Scottish populations showed clear molecular signatures of different demographic histories. Across all regions, the allele frequency distribution was skewed towards intermediate frequency polymorphisms, and the rate of decline of linkage disequilibrium was reduced and nucleotide diversity levels were equivalent to or higher than continental European populations of the species. The skew of allelic frequency distribution, apparent as positive values of Tajima's *D*, was in clear contrast to previous reports for this species in continental Europe (Palmé *et al*, 2008; Wachowiak *et al*, 2009) and for published studies of other species (North American Douglas fir, Eckert *et al*, 2009; *P. taeda* González-Martínez *et al*, 2006a; other conifer species Savolainen and Pyhäjärvi, 2007; European *Quercus petraea* Derory *et al*, 2009; *Populus tremula* Ingvarsson, 2005) where negative values of Tajima's *D* have been found. In these species, the excess of low frequency derived mutations has been ascribed to the influence of postglacial range expansion (Brown *et al*, 2004; Pyhäjärvi *et al*, 2007) or potentially the influence of recurrent selective sweeps (e.g. Eckert *et al*, 2009). In contrast, rather than range expansion, the bias towards intermediate-frequency polymorphisms in Scottish populations suggests the influence of a bottleneck although, as shown in recent simulation studies, a skew of allelic frequency variants may also result from pooling local samples with different demographic histories (Städler *et al*, 2009). However, the bottleneck hypothesis was also supported by the overall pattern of linkage disequilibrium (LD), which showed a reduced rate of decline relative to continental European populations of the species. In coalescent simulations, the bottleneck scenario fits best for

503 western populations and the data were compatible with a relatively recent, severe
504 bottleneck. Depending on the effective population size and generation time assumed, this
505 bottleneck ended a maximum of a few tens of thousands of years ago (e.g about 25 000
506 ya assuming N_e=200 000). Bottlenecking is expected to increase association (correlation
507 among sites with distance) of alleles and polymorphic sites across loci. In Scottish
508 populations, the decay of LD was almost three times slower than that in mainland
509 populations. Reduced decay of LD has also been observed in populations of American *P.*
510 *taeda* that had probably experienced bottlenecks (Brown *et al*, 2004; González-Martínez
511 *et al*, 2006a) and contrasting allele frequency distributions were observed between
512 northern populations and recently bottlenecked southern populations of *Quercus crispula*
513 in Japan (where the latter showed positive Tajima's *D,* Quang *et al*, 2008).
514
515 Although there are exceptions (Grivet *et al*, 2009), it is expected that bottlenecks should
516 have a stronger impact on the allele frequency distribution spectrum and LD than on the
517 overall level of diversity (Wright *et al*, 2005). Long-lived, wind-pollinated tree species
518 should be capable of maintaining genetic diversity even during range shifts; i.e. they are
519 buffered against rapid changes in genetic variation due to fluctuations in population size
520 (Austerlitz *et al*, 2000). Indeed, relative to mainland European populations of Scots pine,
521 Scottish populations did not show a decline in nucleotide diversity, as is expected where
522 colonisation has been relatively recent (Nei *et al*, 1975; Pannell and Dorken, 2006). In
523 fact, genetic variation in Scottish populations seems to be slightly higher than in
524 mainland populations (θ_{sil}=0.011 vs 0.009, respectively) and relative to previous
525 estimates for the species (θ_{sil}= 0.005 at 16 loci with some related to timing of bud set

526 (Pyhäjärvi *et al*, 2007) and θ_{sil}=0.0089 at 14 cold tolerance candidate loci (Wachowiak *et*
527 *al*, 2009)). Compared to estimates in other forest tree species, overall diversity in Scottish
528 populations (π_{tot}=0.0078) is only lower than that in broadleaved *Populus tremula* (0.0111,
529 Ingvarsson, 2005) and is higher than that in *Q. crispula*, (0.0069, Quang and Harada
530 2008), *Q. petraea* (0.0062, Derory *et al.* 2009), *P. pinaster* (0.0055, Eveno et al. 2008),
531 *P. taeda* (0.0040, Brown *et al,* 2004), *Picea abies* (0.0039, Heuertz *et al*, 2007) and other
532 conifers (Savolainen and Pyhäjärvi, 2007). The diversity estimate for Scottish
533 populations is compatible with the patterns of genetic variation observed in previous
534 studies (monoterpenes Forrest, 1980; Forrest, 1982), allozymes Kinloch *et al*, 1986),
535 chloroplast DNA microsatellite markers Provan *et al*, 1998).
536
537 Although it seems clear that bottlenecking has been an influence on Scottish populations,
538 estimation of the timing of the event is heavily dependent on various assumptions
539 including the effective population size and generation time estimates. For instance, in
540 continental populations of Norway spruce and Scots pine, simulation studies suggested a
541 rather ancient bottleneck that ended several hundred thousand to more than one million
542 years ago, respectively (Lascoux *et al*, 2008). In our data, coalescent simulation of
543 various demographic scenarios supported the conclusion that bottlenecking had occurred,
544 but suggested more recent timing. A similar signal, suggesting bottlenecking on a
545 timescale related to the most recent glaciation, was detected in Italian populations of
546 Aleppo pine (Grivet *et al*, 2009). Furthermore, the severity of the bottleneck experienced
547 by Scottish populations appears to have been strong enough to account for the observed

548 discrepancy in allelic frequency distributions and decay of LD, in contrast to continental
549 European tree populations (Lascoux *et al*, 2008).
550
551 However, as we observed heterogeneity in the pattern of nucleotide diversity among
552 regions within Scotland, it seems likely that different parts of the population have
553 experienced different demographic histories. The ratio of recombination to diversity and
554 the level of linkage disequilibrium in western Scottish populations were similar to those
555 in mainland European populations but about three times higher than those in southern and
556 eastern Scottish groups. Various bottleneck scenarios could be clearly rejected for the
557 southern group in our coalescent simulation analysis. The homogenizing effects of gene
558 flow on genetic diversity are well known for highly outcrossing wind pollinated species,
559 and there is evidence for historically high gene flow among Scottish populations from
560 work using chloroplast markers (Provan *et al*, 1998). In addition, molecular and isozyme
561 studies provide no suggestion of a difference in outcrossing rates between regions that
562 could account for a difference in spatial distribution of polymorphism (Kinloch *et al*,
563 1986). As, until recently, Scots pine covered large parts of Scotland, differentiation
564 between regional groups due to genetic drift also seems unlikely, as this should be most
565 significant for small populations (Pannell and Dorken, 2006). Inter-regional differences
566 also seem unlikely to be the result of selection. If this was the case, we would expect
567 differences in the frequency distribution spectrum between groups or at least reduced
568 diversity levels at selected loci. However, the observed dominance of intermediate
569 frequency variants in all groups together with very rapid decay of linkage disequilibrium
570 (within a few hundred base pairs) excludes a selective sweep as an explanation.

571 Furthermore, nucleotide and haplotype diversity is at least as high in southern and eastern
572 groups as in the western group, whereas directional selection should reduce diversity.
573 Therefore, overall, historical changes in population size and distribution seem a more
574 plausible explanation for the pattern of nucleotide variation in Scottish populations and,
575 as a single migration and bottleneck event cannot account for the observed pattern of
576 diversity, it seems that heterogeneity within the Scottish population is most likely to be
577 the result of admixture of populations from different origins.
578
579 Compared to continental Europe, southern and eastern groups of Scottish pines showed
580 no overall difference in allele frequency distribution at polymorphic sites from north or
581 central European populations, but differentiation from Spanish and Turkish populations.
582 On the other hand, the western group was significantly differentiated from all mainland
583 populations except those from Spain. In previous studies, populations from the west of
584 Scotland were more closely related to southern European populations in monoterpene
585 composition and isozyme frequency (Forrest, 1982) or geographically structured *mt*DNA
586 variation (Sinclair *et al*, 1998) than to populations from north-central Europe, which were
587 more similar to the southern and eastern Scottish pinewoods. Similarities between
588 western Scotland and south European Scots pine could simply be stochastic, due to
589 homogenising selection for similar environments or, alternatively, could reflect common
590 ancestry of the populations. Genetic similarity at *mt*DNA markers (maternally transmitted
591 in pines) suggests the latter. However, as Iberian populations did not contribute to the
592 most recent recolonization of central and northern Europe (Prus-Glowacki and Stephan,
593 1994; Pyhäjärvi *et al*, 2008; Soranzo *et al*, 2000; Tobolski and Hanover, 1971), this

genetic similarity would reflect a common origin predating the last glacial period. Therefore, contemporary Scottish populations may originate from western populations that survived the last glaciation in southwestern parts of the British Isles, western continental Europe (Ballantyne and Harris, 1994; Bennett, 1995) or now-submerged parts of the continental shelf. Future genetic studies at more loci (including new *mt*DNA markers) and in more populations would allow more precise assessment of the spatial distribution of haplotypes in Scottish and mainland populations and testing of colonisation hypotheses. This should soon be feasible as new genomic resources for pine, including multiple nuclear and *mt*DNA loci, are currently being developed (e.g. through the EVOLTREE Network of Excellence).

Effects of selection at individual loci

At mutation-drift equilibrium, genetic drift and gene flow influence the level of differentiation between populations for selectively neutral markers (Kawecki and Ebert, 2004; Savolainen *et al*, 2007). Little differentiation between Scottish and mainland European populations of Scots pine at neutral markers (Kinloch *et al*, 1986; Provan *et al*, 1998; Prus-Glowacki and Stephan, 1994) but divergence at quantitative traits for characters of adaptive importance (e.g. phenology, growth and survival rates, Ennos *et al*, 1998; Worrell, 1992, Hurme et al. 1997) suggests that selection is driving adaptive differentiation in both geographical regions. As they differ significantly in climatic, edaphic and biotic conditions, it is possible that observed nucleotide and/or haplotype differentiation at *dhn2* and *dhn7* and some differences in the allele-frequency spectrum at *dhy2PP*, *abaR*, *a3iP2* and *chcs* may be due to selection. Similarly, reduced nucleotide

617 and haplotype diversity and a difference in the frequency and distribution of
618 polymorphism found at *lp3-1* and *ccoaomt* in the western as compared to the southern
619 and eastern groups of Scottish populations could have been affected by diversifying
620 selection at the range edge where populations are under direct oceanic influence. In
621 contrast, the haplotype dimorphism at *lp3-3* could potentially result from the long-term
622 action of balancing selection, maintaining variation across geographical regions.
623 However, as admixture at *lp3-3* cannot be ruled out, a study of nucleotide polymorphism
624 in mainland European populations would be necessary to verify whether or not balancing
625 selection has been an influence at this locus.
626
627 Some of the loci analysed showed distinct nucleotide diversity patterns relative to genetic
628 background in other species (e.g. *lp3-1* and *ccoaomt* in *P. pinaster* Eveno *et al*, 2008,
629 *ccoaomt* in *P. taeda,* González-Martínez *et al*, 2006a). Although there is accumulating
630 evidence on the polygenic character of adaptive traits from QTL studies (Buckler *et al*,
631 2009; Howe *et al*, 2003), it remains unclear whether or not there are genes of major effect
632 that contribute to adaptive variation in conifers. In the case of Scottish pinewoods,
633 adaptation was probably driven by postglacial migration from a predominantly
634 continental to an oceanic environment over the past ~7000 yrs. For long-lived conifers,
635 adaptive differentiation would be expected to occur over several dozens of generations
636 after vicariance. However, even though selection can be very effective in species with
637 large population sizes, the time since the last glaciation seems too short for pine species
638 to have accumulated new mutations that could be rapidly fixed by selection. Adaptive
639 divergence is therefore more likely to result from selection acting on standing variation,

640 which may have arisen in endemic populations that survived last glaciations in Western
641 Europe or the British Isles. Moreover, as differentiation at the trait level in forest trees is
642 likely to result from allelic associations among large numbers of loci, rather than changes
643 in allelic frequencies at individual loci, the signature of selection may be more readily
644 detectable as covariance of allele frequencies at multiple loci (Derory *et al*, 2009; Latta,
645 2004; Le Corre and Kremer, 2003). Therefore many more loci, including regulatory
646 regions (to date, generally omitted from analyses of nucleotide variation in conifers),
647 would need to be studied in parallel before the influence of selection could be verified.
648 Scottish populations, which show considerable ecological, phenotypic and genetic
649 diversity over short geographic distances, represent an excellent study system for
650 multilocus analysis of complex trait variation (González-Martínez *et al*, 2006b; Neale and
651 Savolainen, 2004). Such studies will, however, have to take into account the potential
652 role of recent population history in shaping patterns of nucleotide diversity, and therefore
653 ensure that sampling is conducted at sufficient density to control for historical influences.
654 Association studies of allelic variants and adaptive variation at quantitative traits between
655 individuals from different, locally-adapted populations could also better validate the
656 signatures of selection and the functional role of the nuclear genes studied.
657

Acknowledgments

WW acknowledges financial support from the Polish Ministry of Science (grant nr 3653/B/P01/2008/35), NERC and EU Network of Excellence EVOLTREE (mobility grant). MJS is a Ph.D. student supported by the Scottish Forestry Trust. We thank Dave Sim, Joan Beaton and Ben Moore (Macaulay Institute) for making the seed collections, the owners of the woodlands for their cooperation and Joan Cottrell (Forest Research) and anonymous reviewers for constructive comments on the manuscript.

Conflict of interest statement

The authors declare that there are no conflicts of interest.

References

Achaz G (2009). Frequency Spectrum Neutrality Tests: One for All and All for One. *Genetics* **183**(1)**:** 249-258.

Ahuja MR, Neale DB (2005). Evolution of genome size in conifers. *Silvae Genet* **54**(3)**:** 126-137.

Austerlitz F, Mariette S, Machon N, Gouyon PH, Godelle B (2000). Effects of colonization processes on genetic diversity: Differences between annual plants and tree species. *Genetics* **154**(3)**:** 1309-1321.

Ballantyne CK, Harris C (1994). *The Periglaciation of Great Britain*. Cambridge University Press: Cambridge, 330pp.

Bennett KD (1995). Post-glacial dynamics of pine (*Pinus sylvestris*) and pinewoods in Scotland. In: Aldhous JR (ed) *Scottish Natural Heritage*. Forestry Commission, The Royal Society for the Protection of Birds: Edinburgh, pp 23-39.

Birks HJB (1989). Holocene isochrone maps and patterns of tree-spreading in the British Isles. *Journal of Biogeography* **16**(6)**:** 503-540.

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004). Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proc Natl Acad Sci U S A* **101**(42)**:** 15255-15260.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C *et al* (2009). The Genetic Architecture of Maize Flowering Time. *Science* **325**(5941)**:** 714-718.

Corander J, Tang J (2007). Bayesian analysis of population structure based on linked molecular information. *Mathematical Biosciences* **205**(1)**:** 19-31.

Critchfield WB, Little E. (1965). U.S. Department of Agriculture, pp 97.

Derory J, Scotti-Saintagne C, Bertocchi E, Le Dantec L, Graignic N, Jauffres A *et al* (2009). Contrasting relationships between the diversity of candidate genes and variation of bud burst in natural and segregating populations of European oaks. *Heredity*, **104:** 438-448

Dvornyk V, Sirvio A, Mikkonen M, Savolainen O (2002). Low nucleotide diversity at the pal1 locus in the widely distributed Pinus sylvestris. *Mol Biol Evol* **19**(2)**:** 179-188.

Eckert AJ, Wegrzyn JL, Pande B, Jermstad KD, Lee JM, Liechty JD *et al* (2009). 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**(1)**:** 289-298.

Ennos RA, Worrell R, Malcolm DC (1998). The genetic management of native species in Scotland. *Forestry* **71**(1)**:** 1-23.

Eveno E, Collada C, Guevara MA, Leger V, Soto A, Diaz L *et al* (2008). Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. *Mol Biol Evol* **25**(2)**:** 417-437.

Excoffier L, Laval G, Schneider S (2005). Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1:** 47-50.

Fay JC, Wu C-I (2000). Hitchhiking Under Positive Darwinian Selection. *Genetics* **155**(3)**:** 1405-1413.

Forrest GI (1980). Genotypic variation among native Scots pine populations in Scotland based on monoterpene analysis. *Forestry* **53**(2)**:** 101-128.

Forrest GI (1982). Relationship of some European Scots pine populations to native Scottish woodlands based on monoterpene analyses. *Forestry* **55**(1)**:** 19-37.

García-Gil MR, Mikkonen M, Savolainen O (2003). Nucleotide diversity at two phytochrome loci along a latitudinal cline in Pinus sylvestris. *Molecular Ecology* **12**(5)**:** 1195-1206.

González-Martínez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB (2006a). DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. *Genetics* **172**(3)**:** 1915-1926.

González-Martínez SC, Krutovsky KV, Neale DB (2006b). Forest-tree population genomics and adaptive evolution. *New Phytologist* **170**(2)**:** 227-238.

Grivet D, Sebastiani F, González-Martínez SC, Vendramin GG (2009). Patterns of polymorphism resulting from long-range colonization in the Mediterranean conifer Aleppo pine. *New Phytologist* **184:** 1016-1028.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P (2005). Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations. *Genome Res* **15**(6)**:** 790-799.

Hamrick JL, Godt MJW, Sherman-Broyles SL (1992). Factors influencing levels of genetic diversity in woody plants species. *New For* **6:** 95-124.

Heuertz M, De Paoli E, Kallman T, Larsson H, Jurman I, Morgante M *et al* (2006). Multilocus Patterns of Nucleotide Diversity, Linkage Disequilibrium and Demographic History of Norway Spruce [*Picea abies* (L.) Karst]. *Genetics* **174**(4)**:** 2095-2105.

772 Hill WG, Robertson A (1968). Linkage disequilibrium in finite populations. *Theoretical*
773 *and Applied Genetics* **38:** 226-231.
774
775 Howe GT, Aitken SN, Neale DB, Jermstad KD, Wheeler NC, Chen THH (2003). From
776 genotype to phenotype: unraveling the complexities of cold adaptation in forest trees.
777 *Canadian Journal of Botany-Revue Canadienne De Botanique* **81**(12)**:** 1247-1266.
778
779 Hudson RR (2000). A New Statistic for Detecting Genetic Differentiation. *Genetics*
780 **155**(4)**:** 2011-2014.
781
782 Hudson RR (2002). Generating samples under a Wright-Fisher neutral model of genetic
783 variation. *Bioinformatics* **18**(2)**:** 337-338.
784
785 Hudson RR, Boos DD, Kaplan NL (1992). A statistical test for detecting geographic
786 subdivision. *Mol Biol Evol* **9**(1)**:** 138-151.
787
788 Hudson RR, Kreitman M, Aguade M (1987). A Test of Neutral Molecular Evolution
789 Based on Nucleotide Data. *Genetics* **116**(1)**:** 153-159.
790
791 Hughes AL, Nei M (1988). Pattern of nucleotide substitution at major histocompatibility
792 complex class I loci reveals overdominant selection. *Nature* **335**(6186)**:** 167-170.
793
794 Huntley B, Birks HJB (1983). *An Atlas of Past and Present Pollen Maps for Europe: 0-*
795 *13000 Years Ago*. Cambridge University Press: Cambridge, 667pp.
796
797 Ingvarsson PK (2005). Nucleotide polymorphism and linkage disequilbrium within and
798 among natural populations of European Aspen (Populus tremula L., Salicaceae). *Genetics*
799 **169**(2)**:** 945 - 953.
800
801 Joosen RVL, Lammers M, Balk PA, Bronnum P, Konings MCJM, Perks M *et al* (2006).
802 Correlating gene expression to physiological parameters and environmental conditions
803 during cold acclimation of Pinus sylvestris, identification of molecular markers using
804 cDNA microarrays. *Tree Physiol* **26**(10)**:** 1297-1313.
805
806 Karhu A, Hurme P, Karjalainen M, Karvonen P, Kärkkäinen K, Neale D *et al* (1996). Do
807 molecular markers reflect patterns of differentiation in adaptive traits of conifers? *Theor*
808 *Appl Genet* **93**(1-2)**:** 215-221.
809
810 Kawecki TJ, Ebert D (2004). Conceptual issues in local adaptation. *Ecology Letters*
811 **7**(12)**:** 1225-1241.
812
813 Kinloch BB, Westfall RD, Forrest GI (1986). Caledonian Scots pine - origins and genetic
814 structure. *New Phytologist* **104**(4)**:** 703-729.
815

Lascoux M, Pyhäjärvi T, Källman T, Savolainen O (2008). Past demography in forest trees: what can we learn from nuclear DNA sequences that we do not already know? *Plant Ecology & Diversity* **1**(2)**:** 209 - 215.

Latta RG (2004). Relating processes to patterns of genetic variation across landscapes. *Forest Ecology and Management* **197**(1-3)**:** 91-102.

Le Corre V, Kremer A (2003). Genetic variability at neutral markers, quantitative trait loci and trait in a subdivided population under selection. *Genetics* **164**(3)**:** 1205-1219.

Lynch M, Crease TJ (1990). The analysis of population survey data on DNA sequence variation. *Mol Biol Evol* **7**(4)**:** 377-394.

Ma XF, Szmidt AE, Wang XR (2006). Genetic structure and evolutionary history of a diploid hybrid pine Pinus densata inferred from the nucleotide variation at seven gene loci. *Mol Biol Evol* **23**(4)**:** 807 - 816.

Mason WL, Hampson A, Edwards C (2004). *Managing the Pinewoods of Scotland*. Forestry Commission: Edinburgh.

McDonald JH, Kreitman M (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351:** 652 - 654

Neale DB, Ingvarsson PK (2008). Population, quantitative and comparative genomics of adaptation in forest trees. *Curr Opin Plant Biol* **11**(2)**:** 149-155.

Neale DB, Savolainen O (2004). Association genetics of complex traits in conifers. *Trends in Plant Science* **9**(7)**:** 325-330.

Nei M (1987). Molecular Evolutionary Genetics. *New York, Columbia University Press*.

Nei M, Maruyama T, Chakraborty R (1975). The Bottleneck Effect and Genetic Variability in Populations. *Evolution* **29**(1)**:** 1-10.

Padmanabhan V, Dias DMAL, Newton RJ (1997). Expression analysis of a gene family in loblolly pine (Pinus taeda L.) induced by water deficit stress. *Plant Molecular Biology* **35**(6)**:** 801-807.

Palmé AE, Wright M, Savolainen O (2008). Patterns of divergence among conifer ESTs and polymorphism in *Pinus sylvestris* identify putative selective sweeps. *Mol Biol Evol***, 25**: 2567 - 2577.

Pannell J, Dorken M (2006). Colonisation as a common denominator in plant metapopulations and range expansions: effects on genetic diversity and sexual systems. *Landscape Ecology* **21**(6)**:** 837-848.

862 Perks MP, Ennos RA (1999). Analysis of genetic variation for quantitative characters
863 between and within four native populations of Scots pine (*Pinus sylvestris*). *Botanical*
864 *Journal of Scotland* **51:** 103-110.
865
866 Perks MP, McKay HM (1997). Morphological and physiological differences in Scots
867 pine seedlings of six seed origins. *Forestry* **70:** 223-232.
868
869 Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, Brewer S, Cheddadi R *et al* (2003).
870 Glacial refugia: Hotspots but not melting pots of genetic diversity. *Science* **300**(5625)**:**
871 1563-1565.
872
873 Provan J, Soranzo N, Wilson NJ, McNicol JW, Forrest GI, Cottrell J *et al* (1998). Gene-
874 pool variation in Caledonian and European Scots pine (*Pinus sylvestris* L.) revealed by
875 chloroplast simple-sequence repeats. *Proceedings of the Royal Society of London Series*
876 *B-Biological Sciences* **265**(1407)**:** 1697-1705.
877
878 Prus-Glowacki W, Stephan BR (1994). Genetic variation of Pinus sylvestris from Spain
879 in relation to other European populations. *Silvae Genet* **43**(1)**:** 7-14.
880
881 Pyhäjärvi T, Garcia-Gil MR, Knürr T, Mikkonen M, Wachowiak W, Savolainen O
882 (2007). Demographic History Has Influenced Nucleotide Diversity in European *Pinus*
883 *sylvestris* Populations. *Genetics* **177**(3)**:** 1713-1724.
884
885 Pyhäjärvi T, Salmela MJ, Savolainen O (2008). Colonization routes of *Pinus sylvestris*
886 inferred from distribution of mitochondrial DNA variation. *Tree Genetics & Genomes*
887 **4**(2)**:** 247-254.
888
889 Quang ND, Ikeda S, Harada K (2008). Nucleotide variation in Quercus crispula Blume.
890 *Heredity* **101**(2)**:** 166-174.
891
892 Savolainen O, Pyhäjärvi T (2007). Genomic diversity in forest trees. *Curr Opin Plant*
893 *Biol* **10**(2)**:** 162-167.
894
895 Savolainen O, Pyhäjärvi T, Knürr T (2007). Gene Flow and Local Adaptation in Trees.
896 *Annual Review of Ecology, Evolution, and Systematics* **38**(1)**:** 595-619.
897
898 Sinclair WT, Morman JD, Ennos RA (1998). Multiple origins for Scots pine (*Pinus*
899 *sylvestris* L.) in Scotland: evidence from mitochondrial DNA variation. *Heredity* **80:** 233-
900 240.
901
902 Soranzo N, Alia R, Provan J, Powell W (2000). Patterns of variation at a mitochondrial
903 sequence-tagged-site locus provides new insights into the postglacial history of European
904 *Pinus sylvestris* populations. *Molecular Ecology* **9**(9)**:** 1205-1211.
905
906

907 Stadler T, Haubold B, Merino C, et al. (2009) The impact of sampling schemes on the
908 site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**(1): 205-
909 216.
910
911 Steven HM, Carlisle A. *The Native Pinewoods of Scotland*. (1959) Edinburgh: Oliver
912 and Boyd.
913
914 Svendsen JI, Astakhov VI, Bolshiyanov DY, Demidov I, Dowdeswell JA, Gataullin V *et*
915 *al* (1999). Maximum extent of the Eurasian ice sheets in the Barents and Kara Sea region
916 during the Weichselian. *Boreas* **28**(1)**:** 234 - 242.
917
918 Tobolski JJ, Hanover JW (1971). Genetic Variation in the Monoterpenes of Scotch pine.
919 *Forest Science* **17**(3)**:** 293-299.
920
921 Wachowiak W, Balk P, Savolainen O (2009). Search for nucleotide diversity patterns of
922 local adaptation in dehydrins and other cold-related candidate genes in Scots pine (Pinus
923 sylvestris L.). *Tree Genetics & Genomes* **5**(1)**:** 117-132.
924
925 Wang J-T, Gould J, Padmanabhan V, Newton R (2003). Analysis And Localization of the
926 Water-Deficit Stress-Induced Gene (lp3). *Journal of Plant Growth Regulation* **21**(4)**:**
927 469-478.
928
929 Wang XR, Szmidt AE, Lindgren D (1991). Allozyme differentiation among populations
930 of *Pinus sylvestris* L. from Sweden and China. *Hereditas* **114**(3)**:** 219-226.
931
932 Watterson GA (1975). On the number of segregating sites in genetical models without
933 recombination. *Theor Popul Biol* **7**(2)**:** 256 - 276.
934
935 Worrell R (1992). A Comparison Between European Continental and British Provenances
936 of Some British Native Trees: Growth, Survival and Stem Form. *Forestry* **65**(3)**:** 253-
937 280.
938
939 Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD *et al* (2005).
940 The Effects of Artificial Selection on the Maize Genome. *Science* **308**(5726)**:** 1310-1314.
941
942 Wright SI, Gaut BS (2005). Molecular population genetics and the search for adaptive
943 evolution in plants. *Mol Biol Evol* **22**(3)**:** 506-519.
944
945
946
947

948

949

950

951

952

953

Titles and legends to figures

Figure 1. Main map: location of 21 Scots pine populations from Scotland (divided for most between-population analyses into groups: ● West, ▲ South and + East). Inset shows locations of the 8 mainland European populations with which comparisons were made and location of main map (highlighted). See Material and Methods for details.

Figure 2. Scatter plot of the squared correlation coefficient of allele frequencies (r^2) as a function of distance in base pairs between pairs of polymorphic sites in western (A), southern (B) and eastern (C) groups at all loci combined. Decline in linkage disequilibrium is shown by nonlinear fitting curve of the mutation-recombination-drift model (see material and methods section for details). Recombination rate parameter ρ (standard error in parenthesis) for western group is ρ= 0.0074 (0.0008), for southern group is ρ =0.0025 (0.0004) and ρ =0.0024 (0.0006) for the east.

A. WEST **B. SOUTH** **C. EAST**

Tables and Figures

Table1. Loci included in nucleotide diversity analyses

			Base pairs screened				
Gene	**Protein / Function**	**n**	**Total**	**Coding[a]**	**Intron**	**UTRs[b]**	**Indels[c]**
dhn1	dehydrin 1 – dehydrative stress response	40	1265	489	423	353	10 (194)
dhn2	dehydrin 2 – dehydrative stress response	33	449	235 (2)	119	95	2 (6)
dhn3	dehydrin 3 – dehydrative stress response	34	428	330 (1)	-	98	2 (69)
dhn7	dehydrin 7 – dehydrative stress response	38	364	264 (1)	-	100	1 (2)
dhy2PP	dehydrin – dehydrative stress response	42	485	381 (2)	95	9	1 (8)
abaR	abscisic acid responsive protein	40	419	334 (1)	85	-	4 (23)
a3ip2	ABI3-interacting protein 2	39	882	169 (2)	120	593	1 (21)
ccoaomt	caffeoyl CoA *O*-methyltransferase	41	563	316 (3)	247	-	0
chcs	chalcone synthase	35	331	85 (1)	-	246	1 (1)
erd3	early responsive to dehydration 3	38	583	379 (3)	204	-	0
lp3-1	ABA and WDS induced gene-1	35	438	168 (1)	-	270	1 (8)
lp3-3	ABA and WDS induced gene-3	36	463	232 (2)	231	-	6 (154)
Total		**451**	**6670**	**3382 (19)**	**1524**	**1764**	**29 (486)**

n – haploid sample size, [a] number of exons in parenthesis; [b] untranslated region (5'UTR); [c] number of indels and length range in parenthesis;

Table 2. Summary statistics of nucleotide and haplotype variation and frequency distribution spectrum of polymorphism at analysed genes in Scottish populations of Scots pine. Silent sites variation reported separately for west, south and east geographical groups, otherwise average values for all samples combined.

			Nucleotide diversity											Haplotype diversity
			Nonsynonym.		Silent[a]									
Locus	L	π	SNPs	π	SNPs	π	π_{West}	π_{South}	π_{East}	ρ[b] (SE)	*D*[c]	*H*[d]	N	H_d (SD)
dhn1	1071	0.0144	8 (4)	0.0044	54 (12)	0.0203	0.0215	0.0225	0.0158	0.0113 (0.0012)	0.039	-0.285	24	0.964 (0.014)
dhn2	442	0.0074	1	0.0024	7 (0)	0.0112	0.0107	0.0110	0.0113	0.0779 (0.0379)	1.968[*1,2]	0.458	11	0.888 (0.028)
dhn3	359	0.0198	10 (2)	0.0137	13 (2)	0.0291	0.0325	0.0226	0.0366	-	0.911	1.733	8	0.829 (0.03)
dhn7	362	0.0042	3 (1)	0.0042	4 (3)	0.0043	0.0033	0.0036	0.0061	-	-0.223	-2.339	6	0.711 (0.039)
dhy2PP	476	0.0101	5 (3)	0.0011	12 (2)	0.0244	0.0240	0.0223	0.0257	0.0680 (0.0228)	0.103	1.738	21	0.954 (0.015)
abaR	396	0.0048	5 (0)	0.0052	3 (2)	0.0043	0.0035	0.0048	0.0046	-	0.052	-2.323	10	0.755 (0.057)
a3iP2	861	0.0043	2 (2)	0.0008	12 (5)	0.0049	0.0051	0.0045	0.0054	0.0022 (0.0016)	0.360	-2.231	11	0.779 (0.052)
ccoaomt	563	0.0018	1 (1)	0.0000	5 (0)	0.0032	0.0011	0.0053	0.0038	-	-0.711	-0.746	5	0.348 (0.092)
chcs	330	0.0075	1 (1)	0.0008	12 (6)	0.0094	0.0066	0.0137	0.0088	-	-0.682	-1.267	9	0.766 (0.065)
erd3	583	0.0021	3 (3)	0.0006	7 (4)	0.0037	0.0036	0.0029	0.0046	-	-1.464[*2]	-1.351	9	0.73 (0.052)
lp3-1	430	0.0095	1 (1)	0.0004	12 (1)	0.0137	0.0107	0.0144	0.0135	0.0195 (0.0111)	0.944	1.175	20	0.955 (0.017)
lp3-3	309	0.0370	8 (0)	0.0177	18 (1)	0.0660	0.0692	0.0683	0.0676	0.0018 (0.0011)	2.846[***1,2]	1.600	19	0.949 (0.019)
Total	**6182**		**48(18)**		**159 (38)**									
Mean[e]	**515**	**0.0078**		**0.0031**		**0.0117**	**0.0111**	**0.0116**	**0.0124**	**0.0085 (0.0009)**[f]	**0.118**	**-0.494**	**12.18**	**0.789 (0.042)**

L – length of sequence in base pairs excluding indels; π – nucleotide diversity (Nei 1987); N – number of haplotypes; H_d – haplotype diversity (standard deviation); [a] synonymous and noncoding positions; [b] least-squares estimate of recombination parameter; [c] *D* test (Tajima 1989); [d] *H* test (Fay and Wu 2000); [e] average values at 11 loci excluding *lp3-3*; [f] stimates based on informative sites at all loci excluding *lp3*-3; "-" not estimated due to low number of informative sites; [*] statistically significant values based on coalescence simulations (1) without recombination and (2) with average recombination rate at six loci [*]P<0.05; [***] P<0.01

Table 3. Descriptive statistics for nucleotide variation at eight loci in Scottish and continental European populations of Scots pine. Description of regional groups in Scotland as in Figure 1 and Supplementary Table S1.

Groups		θ^a	C.I. (95%)[b]	ρ (SE) [c]	ρ / θ	*D*[d]	*H*[e]
Scottish	**West**	0.0103	0.0072 – 0.0147	0.0073 (0.0008)	0.71	0.580*	-0.400
	South	0.0130	0.0089 – 0.0188	0.0020 (0.0004)	0.15	0.107	0.066
	East	0.0117	0.0080 – 0.0170	0.0021 (0.0006)	0.18	0.499*	-0.128
	All	0.0108	0.0081 – 0.0145	0.0085 (0.0009)	0.79	0.316	-0.564
Continental European	**North**[f]	0.0095	0.0065 – 0.0137	0.0062 (0.0010)	0.65	-0.143	-0.750
	Central[g]	0.0103	0.0072 – 0.0147	0.0090 (0.0009)	0.87	-0.359	-1.077
	North+Central	0.0096	0.0070 – 0.0131	0.0214 (0.0019)	0.45	-0.316	-1.116
	Spain	0.0098	0.0058 – 0.0167	-	-	-0.539	-0.371
	Turkey	0.0055	0.0030 – 0.0099	-	-	-0.279	-0.792
	All	0.0093	0.0068 – 0.0125	0.0245 (0.0002)	2.69	-0.379	-1.240

[a] median for silent sites;
[b] 95% credibility intervals for θ;
[c] least-squares estimate of ρ;
[d] Tajima's *D* test based on all sites; *P<0.05, statistical significance determined by coalescent simulations with and without recombination (see material and methods);
[e] Fay and Wu *H* test;
[f] North: Finland North, Finland South, Sweden;
[g] Central: Poland, France, Austria; " -" not estimated due to low sample size (~5 for each locus) and low number of informative sites from each population.

Table 4. Differentiation between Scottish and continental European populations of Scots pine measured as average F_{ST} over all polymorphic sites and indels at 8 loci combined.

	North	Central	Spain	Turkey	North+Central	All[a]
West	0.032***	0.026**	0.02	0.091***	0.029*	0.022*
South	0.009	0.011	0.053**	0.112***	0.010	0.011
East	0.019	0.040***	0.072**	0.145***	0.037*	0.039*
All Scottish	0.023**	0.035**	0.035*	0.095***	0.028*	0.025*

[a] all continental European populations combined; *P<0.05, **P<0.01, ***P<0.001;

Table 5. Alternative demographic models tested against total and regional groups of populations in Scotland

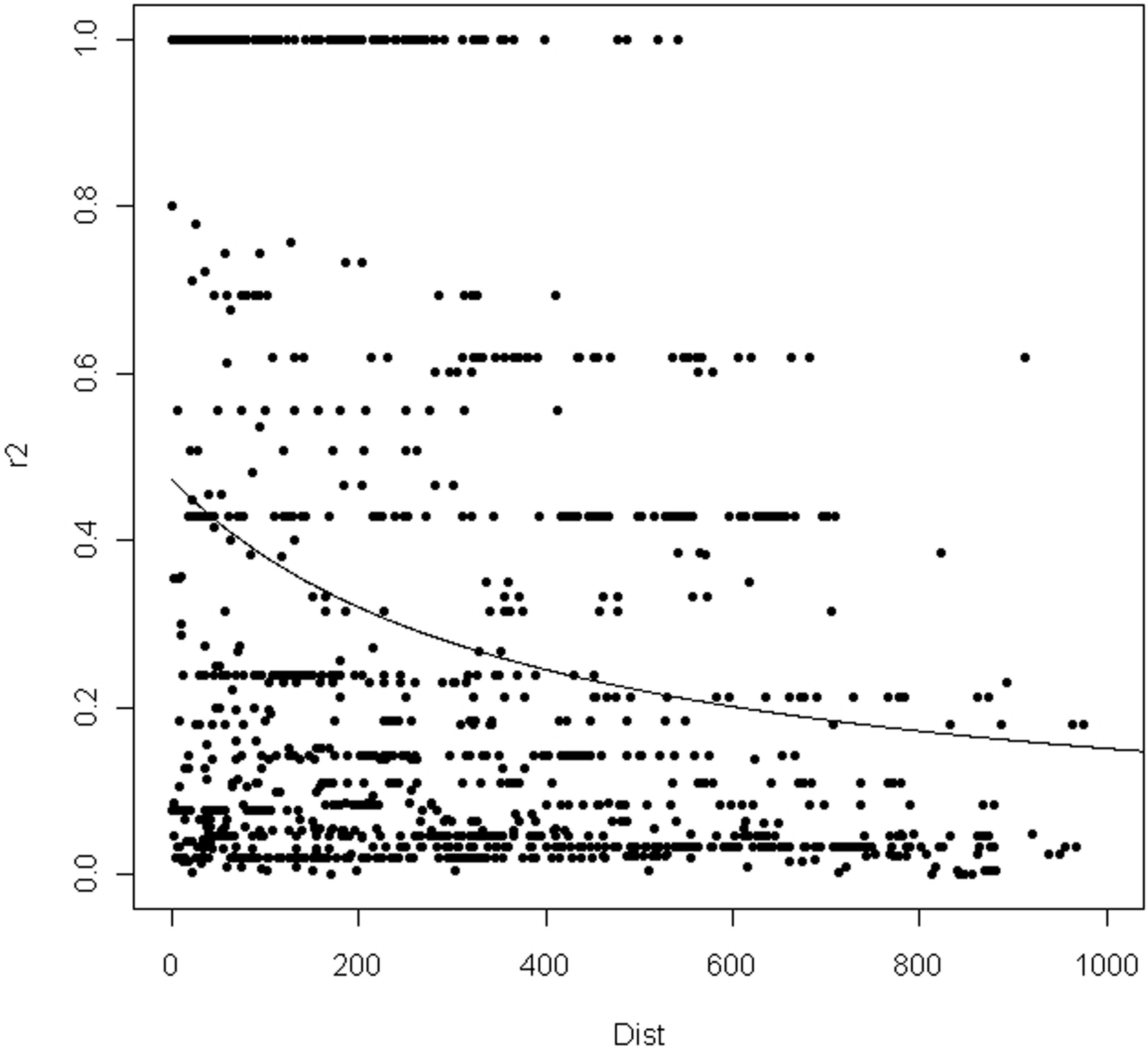
Group	Observed[a]		SN[b]		Growth[c]		Bottleneck[d]	
	Mean *D*	Mean *H*	Mean *D*	Mean *H*	Mean *D*	Mean *H*	Mean *D*	Mean *H*
West	0.364	-0.447	-0.057 (0.578)	0.001 (0.406)	-0.059 (0.585)	-0.018 (0.411)	0.371 (0.116)	-0.504 (0.630)
South	0.103	0.144	-0.066 (0.588)	-0.009 (0.419)	-0.056 (0.575)	0.023 (0.403)	0.310 (0.161)	-0.494 (0.636)
East	0.260	-0.145	-0.056 (0.589)	0.003 (0.419)	-0.065 (0.596)	-0.013 (0.434)	0.311 (0.165)	-0.487 (0.645)
All	-0.015	-0.494	-0.072 (0.602)	0.028 (0.407)	-0.072 (0.605)	0.026 (0.406)	0.661 (0.020)	-0.495 (0.613)

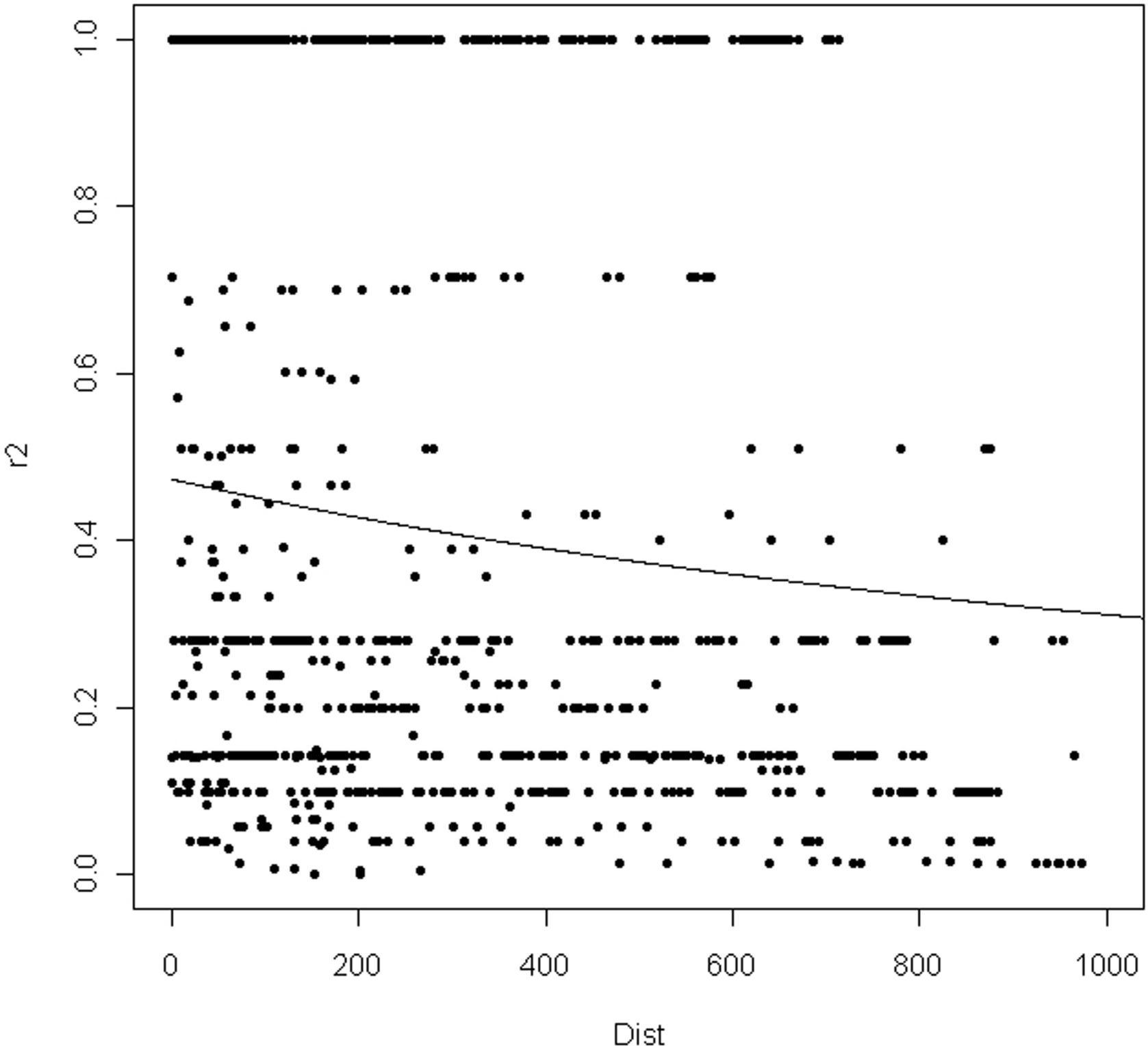
[a] observed mean values of Tajima's *D* and Fay and Wu's *H* statistics at 11 loci

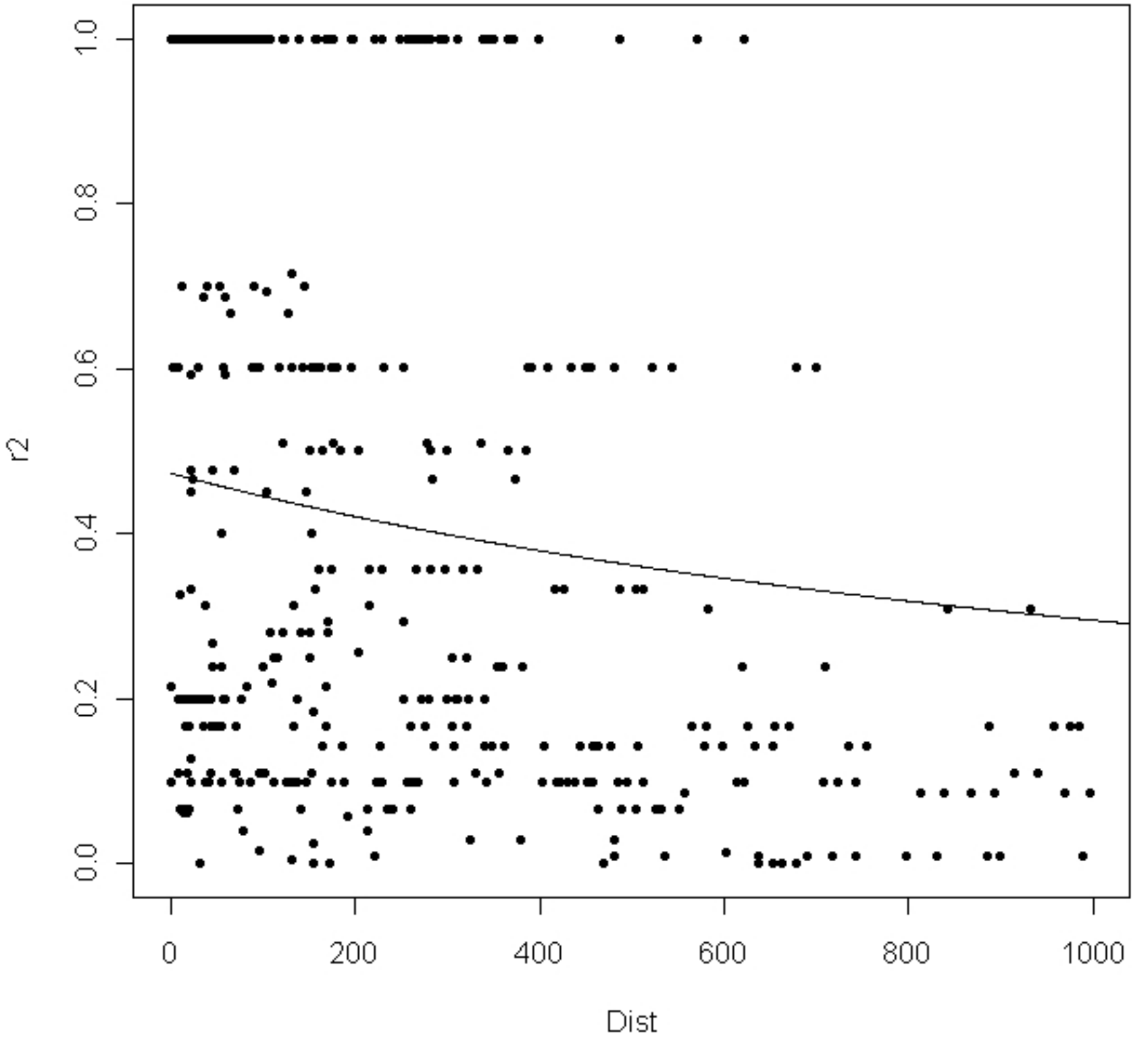
[b] standard neutral model

[c] results for exponential growth of rate 10 starting 0.00125 x 4*Ne* before present

[d] results shown are for bottleneck of severity s=0.02 that started 0.00175 x $4N_e$ generations before present. Duration of bottleneck was set up to 0.0015 and population growth rate to 10. Assuming e.g. N_e =200000 and generation time of 25 years, the bottleneck ended about 25 thousand years ago. Current and the ancestral population size were assumed to be qual. In parenthesis are the *P*-values for the observed means of each parameter.


r2
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
Dist


r2
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
Dist


r2
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
Dist