

1 Natural selection and recombination rate variation shape nucleotide 2 polymorphism across the genomes of three related *Populus* species

3

4 Jing Wang^{*}, Nathaniel R. Street[†], Douglas G. Scofield^{**§}, Pär K. Ingvarsson^{*}

5

6 ^{*} Department of Ecology and Environmental Science, Umeå University, Umeå, SE
7 90187, Sweden

8 [†] Umeå Plant Science Centre, Department of Plant Physiology, Umeå University,
9 Umeå, SE 90187, Sweden

10 [‡] Department of Ecology and Genetics: Evolutionary Biology, Uppsala University,
11 Uppsala, SE 75105, Sweden

12 [§] Uppsala Multidisciplinary Center for Advanced Computational Science, Uppsala
13 University, Uppsala, SE 75105, Sweden

14

15 **Running title:** Population genomics of *Populus*

16

17 **Keywords:** *Populus*, whole genome re-sequencing, nucleotide polymorphism,
18 recombination, natural selection

19

20 **Corresponding author:**

21 Dr Pär K. Ingvarsson, Department of Ecology and Environmental Science, Umeå
22 University, Umeå, SE 90187, Sweden. Phone: +46907867414; Fax: +46-(0)-90-786-
23 6705; E-mail: par.ingvarsson@umu.se

24 **Abstract**

25 A central aim of evolutionary genomics is to identify the relative roles that various
 26 evolutionary forces have played in generating and shaping genetic variation within
 27 and among species. Here we use whole-genome re-sequencing data to characterize
 28 and compare genome-wide patterns of nucleotide polymorphism, site frequency
 29 spectrum and population-scaled recombination rates in three species of *Populus*: *P.*
 30 *tremula*, *P. tremuloides* and *P. trichocarpa*. We find that *P. tremuloides* has the
 31 highest level of genome-wide variation, skewed allele frequencies and population-
 32 scaled recombination rates, whereas *P. trichocarpa* harbors the lowest. Our findings
 33 highlight multiple lines of evidence suggesting that natural selection, both due to
 34 purifying and positive selection, has widely shaped patterns of nucleotide
 35 polymorphism at linked neutral sites in all three species. Differences in effective
 36 population sizes and rates of recombination are largely explaining the disparate
 37 magnitudes and signatures of linked selection we observe among species. The present
 38 work provides the first phylogenetic comparative study at genome-wide scale in forest
 39 trees. This information will also improve our ability to understand how various
 40 evolutionary forces have interacted to influence genome evolution among related
 41 species.

42

43

44

45

46

47

48 **Introduction**

49 A major goal in evolutionary genetics is to understand how genomic variation is
50 established and maintained within and between species (Nordborg *et al.* 2005; Begun
51 *et al.* 2007), and different evolutionary forces have substantial impacts in shaping
52 genetic variation throughout the genome (Hellmann *et al.* 2005). Under the neutral
53 theory, genetic variation is the manifestation of the balance between mutation and
54 genetic drift (Kimura 1983). Demographic fluctuations, such as population expansion
55 and/or bottlenecks, can cause patterns of genome-wide variation deviating from
56 standard neutral model in various ways (Li and Durbin 2011). It is now clear,
57 however, that natural selection, via positive selection favoring beneficial mutations
58 (genetic hitchhiking) and/or purifying selection against deleterious mutations
59 (background selection), plays an important role in moulding the landscape of
60 nucleotide polymorphism in many species (Begun and Aquadro 1992; Begun *et al.*
61 2007; Cutter and Choi 2010; Mackay *et al.* 2012).

62 If natural selection is pervasive across the genome, patterns of genetic
63 variation at linked neutral sites can be influenced by selection in a number of ways.
64 First, positive correlations between levels of neutral polymorphism and recombination
65 rates are expected since linked selection is expected to remove more neutral
66 polymorphism in low-recombination regions compared to high-recombination regions
67 and such a pattern is unlikely to be generated by demographic processes alone (Begun
68 and Aquadro 1992; Kulathinal *et al.* 2008; McGaugh *et al.* 2012; Campos *et al.* 2014;
69 Charlesworth and Campos 2014). Second, besides influencing the level of neutral
70 variability, recombination rate can affect the efficacy of selection through the process
71 known as Hill-Robertson interference (HBI) (Hill and Robertson 1966). If HRI is

operating, genetic linkage effects in regions of low recombination will reduce the local effective population size (N_e), and accordingly reduce the efficacy of selection ($N_e s$), since the effects of selection are determined by the product of N_e and the selection coefficient on a mutation (s) (Kimura 1983). We would therefore expect both a reduced fixation of favorable mutations and an increased frequency of deleterious mutations in these regions (Hill and Robertson 1966; Haddrill *et al.* 2007; Campos *et al.* 2014). Third, signatures and magnitudes of linked selection are sensitive to the density of functional important sites (e.g. gene density) within specific genomic regions (Flowers *et al.* 2012). In accordance with the view that genes represent the most likely targets of natural selection, regions with a high density of genes are expected to have undergone stronger effects of linked selection and exhibit lower levels of neutral polymorphism (Nordborg *et al.* 2005; Flowers *et al.* 2012). Therefore, a positive or negative co-variation of recombination rate and gene density would act to either obscure or strengthen the signatures of linked selection across the genome (Cutter and Payseur 2003; Cutter and Choi 2010; Flowers *et al.* 2012). Lastly, a distinctive signature of recurrent selective sweeps is the local reduction of linked neutral polymorphism in regions experiencing frequent adaptive substitutions (Andolfatto 2007). A substantial number of adaptive substitutions are likely composed of amino acid substitutions and a negative correlation between neutral polymorphism and non-synonymous divergence can thus be particularly informative of the prevalence of selective sweeps (Macpherson *et al.* 2007). With the advance of next-generation sequencing technology, sufficient genome-wide data among multiple related species are becoming available (Luikart *et al.* 2003; Ellegren 2014). Phylogenetic comparative approaches will thus place us in a stronger position to

96 understand how various evolutionary forces have interacted to shape the
97 heterogeneous patterns of nucleotide polymorphism across the genome (Hufford *et al.*
98 2012; Cutter and Payseur 2013; Lawrie and Petrov 2014).

99 Thus far, genome-wide comparative studies have largely dealt with
100 experimental model species, mammals, and cultivated plants of either agricultural or
101 horticultural interest (Locke *et al.* 2011; Hufford *et al.* 2012; Liu *et al.* 2014). Forest
102 trees, as a group, are characterized by extensive geographical distributions and are of
103 high ecological and economic value (Neale and Kremer 2011). Most forest trees have
104 largely persisted in an undomesticated state and, until quite recently, without
105 anthropogenic influence (Neale and Kremer 2011). Accordingly, in contrast to crop
106 and livestock lineages that have been through strong domestication bottlenecks, most
107 extant populations of forest trees harbor a wealth of genetic variation and they are
108 thus excellent model systems for dissecting the dominant evolutionary forces that
109 sculpt patterns of variation throughout the genome (González-Martínez *et al.* 2006;
110 Neale and Kremer 2011). Among forest tree species, the genus *Populus* represents a
111 particularly attractive choice because of its wide geographic distribution, important
112 ecological role in a wide variety of habitats, multiple economic uses in wood and
113 energy products, and relatively small genome size (Eckenwalder 1996; Jansson and
114 Douglas 2007). Here, we studied three *Populus* species which differ in morphology,
115 geographic distribution, population size and phylogenetic relationship (Figure S1)
116 (Jansson *et al.* 2010; Wang *et al.* 2014). *P. tremula* and *P. tremuloides* (collectively
117 ‘aspens’) have wide native ranges across Eurasia and North America respectively, are
118 closely related, and belong to the same section of the genus (section *Populus*)
119 (Jansson *et al.* 2010). In contrast, *P. trichocarpa* belongs to a different section of the

genus (section *Tacamahaca*) that is reproductively isolated from members of the *Populus* section (Jansson *et al.* 2010). The distribution of *P. trichocarpa* is restricted to western regions of North America and its distribution range is considerably smaller than the two aspen species (Dickmann and Kuzovkina 2008). Importantly, *P. trichocarpa* also represents the first tree species to have its genome published (Tuskan *et al.* 2006) and the genome sequence and annotation have undergone continual improvement [<http://phytozome.jgi.doe.gov>]. This enables us to provide important context for our genome comparisons. The phylogenetic relationship of the three species ((*P. tremula*–*P. tremuloides*) *P. trichocarpa*) is well established by both chloroplast and nuclear DNA sequences (Hamzeh and Dayanandan 2004; Wang *et al.* 2014).

In this study, we used datasets generated by Next-Generation Sequencing (NGS) to characterize, compare and contrast genome-wide patterns of nucleotide diversity, site frequency spectrum, recombination rate, and to infer contextual patterns of selection throughout the genomes for all three species.

Materials and Methods

Samples and sequencing

Leaf samples were collected from 24 genotypes of *P. tremula* and 24 genotypes of *P. tremuloides* (Table S1). Genomic DNA was extracted from leaf samples, and paired-end sequencing libraries with insert sizes of 650bp were constructed for all genotypes. Whole-genome sequencing with a minimum expected depth of 20 × was performed on the Illumina HiSeq 2000 platform at the Science for Life Laboratory, Stockholm, Sweden and 2×100-bp paired-end reads were generated for all genotypes. Two

144 samples of *P. tremuloides* failed to yield the expected coverage and were therefore
 145 removed from subsequent analyses. We obtained publicly available short read
 146 Illumina data of 24 *P. trichocarpa* individuals from NCBI SRA (Table S1).
 147 Individuals were selected to have a similar read depth as the samples of the two aspen
 148 species. The accession numbers of *P. trichocarpa* samples can be found in Evans *et al.*
 149 2014. All analyses are thus based on data from 24 *P. tremula*, 22 *P. tremuloides* and
 150 24 *P. trichocarpa* genotypes.

151

152 ***Raw read filtering, read alignment and post-processing alignment***

153 Prior to read alignment, we used Trimmomatic (Lohse *et al.* 2012) to remove adapter
 154 sequences from reads. Since the quality of reads always drops towards the end of
 155 reads, we used Trimmomatic to cut off bases from the start and/or end of reads when
 156 the quality values were smaller than 20. If the length of the processed reads was
 157 reduced to below 36 bases after trimming, reads were completely discarded. FastQC
 158 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) was used to check and
 159 compare the per-base sequence quality between the raw sequence data and the filtered
 160 data. After quality control, all paired-end and orphaned single-end reads from each
 161 sample were mapped to the *P. trichocarpa* version 3 (v3.0) genome (Tuskan *et al.*
 162 2006) using the BWA-MEM algorithm with default parameters in bwa-0.7.10 (Li
 163 2013).

164 Several post-processing steps of alignments were performed to minimize the
 165 number of artifacts in downstream analysis: First, we performed indel realignment
 166 since mismatching bases were usually found in regions with insertions and deletions
 167 (indels) (Wang *et al.* 2015). The RealignerTargetCreator in GATK (The Genome

Analysis Toolkit) (DePristo *et al.* 2011) was first used to find suspicious-looking intervals that were likely in need of realignment. Then, the IndelRealigner was used to run the realigner over those intervals. Second, as reads resulting from PCR duplicates can arise during the sequencing library preparation, we used the MarkDuplicates methods in the Picard package (<http://picard.sourceforge.net>) to remove those reads or read pairs having identical external coordinates and the same insert length. In such cases only the single read with the highest summed base qualities was kept for downstream analysis. Third, in order to exclude genotyping errors caused by paralogous or repetitive DNA sequences where reads were poorly mapped to the reference genome, or by other genome feature differences between *P. trichocarpa* and *P. tremula* or *P. tremuloides*, we removed sites with extremely low and extremely high read depths after investigating the empirical distribution of read coverage. We filtered out sites with a total coverage less than $100 \times$ or greater than $1200 \times$ across all samples per species. When reads were mapped to multiple locations in the genome, they were randomly assigned to one location with a mapping score of zero by BWA-MEM. In order to account for such misalignment effects, we removed those sites if there were more than 20 mapped reads with mapping score equaling to zero across all individuals in each species. Lastly, because the short read alignment is generally unreliable in highly repetitive genomic regions, we filtered out sites that overlapped with known repeat elements as identified by RepeatMasker (Tarailo-Graovac and Chen 2009). In the end, the subset of sites that passed all these filtering criteria in the three *Populus* species were used in downstream analyses.

190

191 ***Single nucleotide polymorphism (SNP) and genotype calling***

192 We implemented two complementary bioinformatics approaches: First, many studies
193 have pointed out the bias inherent in population genetic estimates using genotype
194 calling approach from NGS data (Nielsen *et al.* 2011; Nevado *et al.* 2014). Single- or
195 multiple-sample genotype calling can result in a bias in the estimation of site
196 frequency spectrum (SFS), as the former usually leads to overestimation of rare
197 variants, whereas the latter often leads to the opposite (Nielsen *et al.* 2011).
198 Therefore, in this study we employed a method, implemented in the software package
199 - Analysis of Next-Generation Sequencing Data (ANGSD v0.602) (Korneliussen *et*
200 *al.* 2014), to estimate the SFS and all population genetic statistics derived from the
201 SFS without calling genotypes. Second, for those analyses that require accurate SNP
202 and genotype calls, we performed SNP calling with HaplotypeCaller of the GATK
203 v3.2.2 (DePristo *et al.* 2011), which called SNPs and indels simultaneously via local
204 re-assembly of haplotypes for each individual and created single-sample gVCFs.
205 GenotypeGVCFs in GATK was then used to merge multi-sample records together,
206 correct genotype likelihoods, and re-genotype the newly merged record and perform
207 re-annotation. Several filtering steps were then used to reduce the number of false
208 positive SNPs and retain high-quality SNPs: (1) We removed all SNPs that
209 overlapped with sites excluded by all previous filtering criteria. (2) We only retained
210 bi-allelic SNPs with a distance of more than 5 bp away from any indels. (3) We
211 treated genotypes with quality score (GQ) lower than 10 as missing and then removed
212 those SNPs with genotype missing rate higher than 20%. (4) We removed SNPs that
213 showed significant deviation from Hardy-Weinberg Equilibrium ($P < 0.001$). After all
214 filtering, 8,502,169 SNPs were detected among the three *Populus* species and were
215 used in downstream analyses.

216

217 ***Population structure***

218 We used 4-fold synonymous SNPs with minor allele frequency >0.1 to perform
219 population structure analyses with ADMIXTURE (Alexander *et al.* 2009). We ran
220 ADMIXTURE on all the sampled individuals among species and on the samples
221 within each species separately. The number of genetic clusters (K) was varied from 1
222 to 6. The most likely number of genetic cluster was selected by minimizing the cross-
223 validation error in ADMIXTURE.

224

225 ***Diversity and divergence - related summary statistics***

226 For nucleotide diversity and divergence estimates, only the reads with
227 mapping quality above 30 and the bases with quality score higher than 20 were used
228 in all downstream analyses with ANGSD (Korneliussen *et al.* 2014). First, we used
229 the -doSaf implementation in ANGSD to calculate the site allele frequency likelihood
230 based on the SAMTools genotype likelihood model (Li *et al.* 2009). Then, we used
231 the -realSFS implementation in ANGSD to obtain an optimized folded global SFS
232 using Expectation Maximization (EM) algorithm for each species. Based on the
233 global SFS, we used the -doThetas function in ANGSD to estimate the per-site
234 nucleotide diversity from posterior probability of allele frequency based on a
235 maximum likelihood approach (Kim *et al.* 2011). Two standard estimates of
236 nucleotide diversity, the average pairwise nucleotide diversity (Θ_π) (Tajima 1989) and
237 the proportion of segregating sites (Θ_w) (Watterson 1975), and one neutrality statistic
238 test Tajima's D (Tajima 1989) were summarized along all 19 chromosomes using
239 non-overlapping sliding windows of 100 kilobases (Kbp) and 1 megabases (Mbp).

240 Windows with less than 10% of covered sites left from previous quality filtering steps
241 were excluded. In the end, 3340 100-Kbp and 343 1-Mbp windows, with an average
242 of 50,538 and 455,910 covered bases per window, were respectively included.

243 All these statistics were also calculated for each type of functional element (0-
244 fold non-synonymous, 4-fold synonymous, intron, 3' UTR, 5' UTR, and intergenic
245 sites) over non-overlapping 100-Kbp and 1-Mbp windows in all three *Populus*
246 species. The category of gene models followed the gene annotation of *P. trichocarpa*
247 version 3.0 (Tuskan *et al.* 2006). For protein-coding genes, we only included genes
248 with at least 90% of covered sites left from previous filtering steps to ensure that the
249 three species have the same gene structures. For regions overlapped by different
250 transcripts in each gene, we classified each site according to the following hierarchy
251 (from highest to lowest): Coding regions (CDS), 3'UTR, 5'UTR, Intron. Thus, if a
252 site resides in a 3'UTR in one transcript and CDS for another, the site was classified
253 as CDS. We used the transcript with the highest content of protein-coding sites to
254 categorize synonymous and non-synonymous sites within each gene. A respective of
255 16.52 Mbp, 3.4 Mbp, 7.19 Mbp, 4.02 Mbp, 31.89 Mbp and 73.46 Mbp were
256 partitioned into 0-fold non-synonymous (where all DNA sequence changes lead to
257 protein sequence changes), 4-fold synonymous (where all DNA sequence changes
258 lead to the same protein sequences), 3'UTR, 5'UTR, intron, and intergenic categories.

259

260 ***Linkage disequilibrium (LD) and population-scaled recombination rate (ρ)***

261 A total of 1,409,377 SNPs, 1,263,661 SNPs and 710,332 SNPs with minor allele
262 frequency higher than 10% were used for the analysis of LD and ρ in *P. tremula*, *P.*
263 *tremuloides* and *P. trichocarpa*, respectively. To estimate and compare the rate of LD

264 decay among the three *Populus* species, we firstly used PLINK 1.9 (Purcell *et al.*
265 2007) to randomly thin the number of SNPs to 100,000 in each species. We then
266 calculated the squared correlation coefficients (r^2) between all pairs of SNPs that were
267 within a distance of 50 Kbp using PLINK 1.9. The decay of LD against physical
268 distance was estimated using nonlinear regression of pairwise r^2 vs. the physical
269 distance between sites in base pairs (Remington *et al.* 2001).

270 We estimated the population-scaled recombination rate ρ using the Interval
271 program of LDhat 2.2 (McVean *et al.* 2004) with 1,000,000 MCMC iterations
272 sampling every 2,000 iterations and a block penalty parameter of five. The first
273 100,000 iterations of the MCMC iterations were discarded as a burn-in. We then
274 calculated the scaled value of ρ in each 100-Kbp and 1-Mbp window by averaging
275 over all SNPs in that window. Only windows with more than 10,000 (in 100 Kbp
276 windows) and 100,000 sites (in 1 Mbp windows) and 100 SNPs left from previous
277 filtering steps were used for the estimation of ρ .

278

279 ***Estimating the distribution of fitness effects of new amino acid mutations (DFE)***
280 ***and the proportion of adaptive amino acid substitutions (α)***

281 We generated the folded SFS in each species for a class of selected sites (0-fold non-
282 synonymous sites) and a class of putatively neutral reference sites (4-fold
283 synonymous sites) from SNPs data using a custom Perl script. We employed a
284 maximum likelihood (ML) approach as implemented in the program DFE-alpha
285 (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009) to fit a
286 demographic model with a step of population size change to the neutral SFS. Fitness
287 effects of new deleterious mutations at the selected site class were sampled from a

gamma distribution after incorporating the estimated parameters for the demographic model. This method assumes that fitness effects of new mutations at neutral sites are zero and unconditionally deleterious at selected sites since it assumes that advantageous mutations are too rare to contribute to polymorphism (Keightley and Eyre-Walker 2007). We report the proportion of amino acid mutations falling into different effective strengths of selection ($N_e s$) range: 0-1, 1-10, >10, respectively.

From the estimated DFE, the proportion (α) and the relative rate (ω) of adaptive substitution at 0-fold non-synonymous sites were estimated using the method of Eyre-Walker and Keightley (2009). This method explicitly account for past changes in population size and the presence of slightly deleterious mutations. Among the total of 8,502,169 SNPs detected by GATK, on average less than 1% were shared between either of the two aspen species and *P. trichocarpa* (Figure S2). We therefore used the aspen species and *P. trichocarpa* as each other's outgroup species to calculate between-species nucleotide divergence at 4-fold synonymous and 0-fold non-synonymous sites since it is unlikely to be influenced by shared ancestral polymorphisms. Jukes-Cantor multiple hits correction was applied to the divergence estimates (Jukes and Cantor 1969). For the parameters of $N_e s$, α and ω , we generated 200 bootstrap replicates by resampling randomly across all SNPs in each site class using R (R Development Core Team 2014). We excluded the top and bottom 2.5% of bootstrap replicates and used the remainder to represent the 95% confidence intervals for each parameter.

309

310 ***Genomic correlates of diversity***

311 In order to examine the factors influencing levels of neutral polymorphism in all three
 312 *Populus* species, we firstly assume that the 4-fold synonymous sites in genic regions
 313 were selectively neutral as every possible mutation at 4-fold degenerate sites is
 314 synonymous. In the following we refer to the pairwise nucleotide diversity at 4-fold
 315 synonymous sites ($\theta_{4\text{-fold}}$) as “neutral polymorphism”. As a comparison to genic
 316 region, we also estimated levels of nucleotide diversity at intergenic sites ($\theta_{\text{Intergenic}}$).
 317 Then, we tabulated several other genomic features within each 100 Kbp and 1 Mbp
 318 window that may correlate with patterns of polymorphism. First, we summarized
 319 population-scaled recombination rate (ρ) as described above for each species. Second,
 320 we tabulated GC content as the fraction of bases where the reference sequence (*P.*
 321 *trichocarpa* v3.0) was a G or a C. Third, we measured the gene density as the number
 322 of functional genes within each window according to the gene annotation of *P.*
 323 *trichocarpa* version 3.0. Any portion of a gene that fell within a window was counted
 324 as a full gene. Fourth, we accounted for the variation of mutation rate by calculating
 325 the number of fixed differences per neutral site (either 4-fold synonymous sites or
 326 intergenic sites) between aspen and *P. trichocarpa* within each window, which was
 327 performed in the ngsTools (Fumagalli *et al.* 2014). The reason why we used
 328 divergence between aspen and *P. trichocarpa* to measure mutation rate is because
 329 they are distantly related (Wang *et al.* 2014), and the estimate of divergence are
 330 unlikely to be influenced by shared ancestral polymorphisms between species as
 331 shown above. Fifth, we tabulated the number of covered bases in each window as
 332 those left from all previous filtering criteria.

333 We used Spearman's rank-order correlation test to examine pairwise
334 correlations between the variables of interest. In order to account for the
335 autocorrelation between variables, we further calculated partial correlations between
336 the variables of interest by removing the confounding effects of other variables (Kim
337 and Soojin 2007). All statistical tests were performed using R version 3.2.0 unless
338 stated otherwise.

339

340 **Data Availability**

341 All newly generated Illumina reads of 24 *P. tremula* and 22 *P. tremuloides* from this
342 study have been submitted to the Short Read Archive (SRA) at NCBI. All accession
343 IDs can be found in Table S1.

344

345 **Results**

346 We generated whole-genome sequencing data for 24 *P. tremula* and 22 *P. tremuloides*
347 (Table S1) with all samples sequenced to relatively high depth (24.2×-69.2×; Table
348 S2). We also downloaded whole-genome re-sequencing data for 24 samples of *P.*
349 *trichocarpa* from the NCBI Short Read Archive (Evans et al. 2014). After adapter
350 removal and quality trimming, 949.2 Gb of high quality sequence data remained
351 (Figure S3; Table S2). The mean mapping rate of reads to the *P. trichocarpa*
352 reference genome were 89.8% for *P. tremula*, 91.1% for *P. tremuloides*, and 95.2%
353 for *P. trichocarpa* (Table S2). On average, the genome-wide coverage of uniquely
354 mapped reads was more than 20 × for each species (Table S2). After excluding sites
355 with extreme coverage, low mapping quality, or those overlapping with annotated
356 repetitive elements (see Materials and Methods), 42.8% of collinear genomic

357 sequences remained for downstream analyses. 54.9% of these sites were found within
358 gene boundaries, covering 70.1% of all genic regions predicted from the *P.*
359 *trichocarpa* assembly. The remainder sites (45.1%) were located in intergenic regions.

360

361 ***Genome-wide patterns of polymorphism, site frequency spectrum and***
362 ***recombination among the three Populus species***

363 When analyzing population structure between species, we found the model exhibited
364 the lowest cross-validation error when the number of ancestral populations (K) = 3
365 (Figure S4b), which clearly subdivided the three species into three distinct clusters
366 (Figure S4a). When we analyzed population structure within each species separately,
367 we found that the cross-validation error increased linearly with increasing K , with
368 $K=1$ minimizing the cross-validation error for all three species (Figure S4c-e). Our
369 results are slightly different from several earlier studies that have documented
370 population subdivision in these species (De Carvalho *et al.* 2010; Evans *et al.* 2014)
371 but it is likely due to the small sample sizes used in this study (22-24 individuals). In
372 addition, it could also be caused by the low power of model-based approaches in
373 detecting population structure when it is very weak (Alexander *et al.* 2009). More
374 generally, population structure is expected to be weak in *Populus* given the great
375 dispersal capabilities of both pollen and seeds (Eckenwalder 1996; Jansson and
376 Douglas 2007).

377 The two aspen species harbor substantial levels of nucleotide diversity across
378 the genome ($\Theta_{\pi}=0.0133$ in *P. tremula*; $\Theta_{\pi}=0.0144$ in *P. tremuloides*), approximately
379 two to three-fold higher than diversity in *P. trichocarpa* ($\Theta_{\pi}=0.0059$) (Figure 1; Table
380 S3). Among various genomic contexts, we found the levels of nucleotide diversity

381 were highest at intergenic sites, followed by 4-fold synonymous sites, 3'UTRs,
 382 5'UTRs, introns and were lowest at 0-fold non-synonymous sites (Figure S5; Table
 383 S3). In accordance with the view that the large majority of amino acid mutations are
 384 selected against (Larracuente *et al.* 2008), we found significantly lower Tajima's D at
 385 0-fold non-synonymous sites compared to 4-fold synonymous sites ($P < 0.001$, Mann-
 386 Whitney U test) (Figure S6; Table S3). In addition, we observed significantly positive
 387 correlations of Θ_{Π} between each pair of the three species across the whole genome
 388 (Figure 2a). The overall nucleotide diversity estimated in *P. trichocarpa* was slightly
 389 higher than the value reported in Evans *et al.* 2014 ($\Theta_{\Pi} = 0.0041$), but this likely only
 390 reflects differences between the methods used in the two studies. In this study, we
 391 utilized the full information of the filtered data and estimated the population genetic
 392 statistics directly from genotype likelihoods, which take statistical uncertainty of SNP
 393 and genotype calling into account and should give more accurate estimates (Kim *et al.*
 394 2011; Nielsen *et al.* 2011).

395 Compared to patterns of polymorphism, we observed much weaker
 396 correlations of the site frequency spectrum, summarized using the Tajima's D statistic
 397 (Tajima 1989), between species (Figure 2b). *P. tremuloides* (average Tajima's D =
 398 -1.169) showed substantially greater negative values of Tajima's D along all
 399 chromosomes compared to both *P. trichocarpa* (average Tajima's D = 0.064) and *P.*
 400 *tremula* (average Tajima's D = -0.272) (Figure S7; Table S3), reflecting a large excess
 401 of low-frequency polymorphisms segregating in this species. Furthermore, the three
 402 *Populus* species showed different extents of genome-wide LD decay (Figure S8), with
 403 LD decaying fastest in *P. tremuloides* and slowest in *P. trichocarpa* (Figure S8). This
 404 reflects the rank order of their population-scaled recombination rates ($\rho = 4N_e c$)

(Figure S9), for which the mean ρ over 100 Kbp non-overlapping windows was highest in *P. tremuloides* (8.42 Kbp⁻¹), followed by *P. tremula* (3.23 Kbp⁻¹), and lowest in *P. trichocarpa* (2.19 Kbp⁻¹). Intermediate correlations of recombination rates were observed between species (Figure 2c). In addition, concordant values of Θ_{Π} , Tajima's D and ρ for all three species were also observed in 1 Mbp windows (Figure S10). For populations under drift-mutation-recombination equilibrium, $\rho = 4N_e c$ (where N_e is the effective population size and c is the recombination rate) and $\theta_w = 4N_e \mu$ (where N_e is the effective population size and μ is the mutation rate). In order to compare the relative contribution of recombination (c) and mutation (μ) in shaping genomic variation, we measured the ratio of population recombination rate to the nucleotide diversity (ρ/θ_w) across the genome (Figure S11). The mean c/μ in *P. tremula*, *P. tremuloides* and *P. trichocarpa* was 0.22, 0.39 and 0.38 respectively.

Distribution of fitness effects and proportion of adaptive amino acid substitutions

We quantified the efficacy of both purifying and positive selection using the information of polymorphism and divergence among the three species. The estimated distribution of fitness effects of new 0-fold non-synonymous mutations indicate that the majority of new amino acid mutations were strongly deleterious ($N_e s > 10$) and likely to be under strong levels of purifying selection in all three species (Figure 3; Table S4). There was a greater proportion of amino acid mutations under moderate levels of purifying selection ($1 < N_e s < 10$) in *P. tremuloides* (~31%), compared to *P. tremula* (~16%) and *P. trichocarpa* (~10%). In comparison, we found a higher proportion of weakly deleterious mutations that behave as effectively neutral ($N_e s < 1$)

in *P. trichocarpa* (~31%) relative to *P. tremula* (~23%) and *P. tremuloides* (~16%) (Figure 3; Table S4).

Using 4-fold synonymous sites as a neutral reference, we employed an extension of the McDonald-Kreitman test (Eyre-Walker and Keightley 2009) to estimate the fraction of adaptive amino acid substitutions (α) and the rate of adaptive substitution relative to the rate of neutral substitution (ω) in all three species. Both α and ω were highest in *P. tremuloides* (α : ~65% [95% CI: 63.6%-65.8%]; ω : ~0.24 [95% CI: 0.231-0.242]), intermediate in *P. tremula* (α : ~43% [95% CI: 41.9%-43.5%]; ω : ~0.16 [95% CI: 0.151-0.159]) and lowest in *P. trichocarpa* (α : ~20% [95% CI: 18.8%-31.1%]; ω : ~0.07 [95% CI: 0.068-0.112]) (Figure 3; Table S4).

Neutral polymorphism, but not divergence, is positively correlated with recombination rate

If natural selection (either purifying or positive selection) occurs throughout the genome at similar rates, they should leave a stronger imprint on patterns of neutral polymorphism in regions experiencing low recombination (Begun and Aquadro 1992). In accordance with this expectation, we found significantly positive correlations between levels of neutral polymorphism ($\theta_{4\text{-fold}}$) and population recombination rates in both aspen species (Table 1), with correlations being stronger in *P. tremula* than in *P. tremuloides*. In *P. trichocarpa*, however, we found either no or weak correlation between diversity and recombination rate (Table 1). Compared to 100 Kbp windows, correlations were stronger for 1 Mbp windows in all species, which most likely results from the higher signal-to-noise ratio provided by larger genomic windows (Table 1). In the remainder of this paper we thus focus our analyses primarily on data generated

with a 1 Mbp window size. When performing simple linear regression analysis between diversity and recombination rate over 1 Mbp windows, the recombination rate explained 45.8%, 21.3%, and 3.9% of the amount of neutral genetic variation in *P. tremula*, *P. tremuloides* and *P. trichocarpa*, respectively (Figure 4). If the positive relationship between diversity and recombination rate was merely caused by the mutagenic effect of recombination, similar patterns should also be observed between divergence and recombination rate (Kulathinal *et al.* 2008). However, no such correlations were observed in any of the three species (Table 1; Figure 4). The correlations between neutral diversity and recombination rate were slightly lower, but still significant, after using partial correlations to control for possible confounding factors such as GC content, gene density, divergence at neutral sites, and the number of neutral bases covered by sequencing data (Table 1).

In accordance with the view that genes represent the most likely targets of natural selection (Lohmueller *et al.* 2011), the correlations between intergenic diversity and recombination rate were substantially weaker than those correlations in genic regions (Table 1). Only 7.3% of the variation in intergenic nucleotide diversity in *P. tremula* could be explained by variation in the recombination rate, whereas the impact of recombination rate variation on intergenic diversity in *P. tremuloides* and *P. trichocarpa* was negligible (<1% Figure S12; Table 1). However, after using partial correlation analyses to control for possible confounding factors, the correlations between intergenic diversity and recombination rate became significant in all species. Compared to genic regions these correlations were slightly higher in *P. trichocarpa*, of similar magnitude in *P. tremuloides* and weak in *P. tremula* (Table 1).

476 ***The effect of recombination on the efficacy of natural selection***

477 We characterized the ratio of non-synonymous to synonymous polymorphism ($\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$) and divergence ($d_{0\text{-fold}}/d_{4\text{-fold}}$) to assess whether there was a relationship
478 between the efficacy of natural selection and the rate of recombination (Table 2).
479 Once GC content, gene density and the number of 4-fold synonymous and 0-fold non-
480 synonymous sites were taken into account, we found no correlation between
481 recombination rate and $d_{0\text{-fold}}/d_{4\text{-fold}}$ in any of the three species (Table 2). We also did
482 not observe any significant correlations between recombination rate and $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$
483 over 1 Mbp windows after controlling for confounding factors (Table 2). However,
484 when using 100 Kbp windows, we found significantly negative correlations between
485 recombination rate and $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ in *P. tremula* and *P. tremuloides*, but not in *P.*
486 *trichocarpa* (Table 2).

488 489 ***Inconsistent effect of gene density on patterns of polymorphism in genic vs.*** 490 ***intergenic regions***

491 We measured gene density as the number of protein-coding genes in each 1 Mbp
492 window, which in turn was highly correlated with the proportion of coding bases in
493 each window (Figure S13). For all three species, we found significantly positive
494 correlations between population recombination rate and gene density (Figure 5a;
495 Table 3). However, rather than being linear, the relationships between recombination
496 rate and gene density was curvilinear with a significant positive correlation observed
497 only in regions of low gene density (gene number smaller than ~85 within each 1Mbp
498 window) (Table 3). For regions of high gene density (gene number greater than ~85
499 within each 1Mbp window) we found no correlations between recombination rate and

gene density in both aspen species, and only a weak, positive correlation in *P. trichocarpa* (Figure 5a; Table 3). After controlling for GC content and the number of bases covered by sequencing data, the correlation became significant in regions of high gene density for *P. tremula*, but remained non-significant for *P. tremuloides* (Table 3).

We then examined the relationship between neutral polymorphism and gene density. Compared to the prediction of lower diversity in regions with higher functional density (Payseur and Nachman 2002), we found that the correlation pattern between gene density and levels of neutral polymorphism in genic regions ($\theta_{4\text{-fold}}$) was highly consistent with the pattern found in recombination rate, where significantly positive correlations were found in regions of low gene density and either no or weak negative correlation was found in regions of high gene density (Figure 5b; Table 3). After again controlling for potential confounding variables, the positive correlations remained significant in regions of low gene density among all three species (Table 3), as well as in high gene-density regions in *P. tremuloides* and *P. trichocarpa* (Table 3). We did not find any significant relationships between neutral divergence and gene density in any of the three species (Figure S14).

Compared with genic regions, correlations between intergenic diversity and gene density followed a different pattern in the three species (Figure 5c). In intergenic regions nucleotide diversity and gene density were positively correlated in regions of low gene density but negatively correlated in regions of high gene density (Figure 5c; Table 3). These correlations remained significant even after controlling for possible confounding variables (Table 3). No relationship between intergenic divergence and gene density was found in any species (Figure S14).

524

525 *Negative correlations between synonymous diversity and non-synonymous*
 526 *divergence at small physical scales*

527 A negative relationship between synonymous diversity and non-synonymous
 528 divergence has been suggested to be a strong evidence of the occurrence of recurrent
 529 selective sweeps (Andolfatto 2007), and such a pattern has previously been observed
 530 in *P. tremula* using data from a small number of candidate genes (Ingvarsson 2010).
 531 Here, however, we found either no or very weak negative correlations between neutral
 532 polymorphism ($\theta_{4\text{-fold}}$) and the rate of non-synonymous substitutions ($d_{0\text{-fold}}$) in all
 533 three species for both 100 Kbp and 1 Mbp windows, and these correlations did not
 534 change after controlling for possible confounding factors (Table 4). However, the
 535 effects of recurrent selective sweeps on synonymous nucleotide diversity are thought
 536 to be high localized within genes (Andolfatto 2007), and we therefore examined the
 537 association between $\theta_{4\text{-fold}}$ and $d_{0\text{-fold}}$ at smaller physical scales, using data from
 538 20,759 genes that retained more than 90% of bases after all filtering steps. In contrast
 539 to the lack of correlations observed across larger scales (100 Kbp or 1 Mbp), we
 540 found a significantly negative correlation between $\theta_{4\text{-fold}}$ and $d_{0\text{-fold}}$ in all three species
 541 when assessed within genes (Table 4). After accounting for the possible influence of
 542 mutation rate variation among genes by normalizing $\theta_{4\text{-fold}}$ by neutral divergence rate
 543 ($d_{4\text{-fold}}$), the negative correlations became stronger in all species (Figure S15; Table 4).

544

545 **Discussion**

546

547 ***Genome-wide patterns of nucleotide polymorphism, site frequency spectrum and*** 548 ***recombination***

549 We have characterized and compared genome-wide patterns of nucleotide
550 polymorphism, site frequency spectra and recombination rates in three species of
551 *Populus*: *P. tremula*, *P. tremuloides* and *P. trichocarpa*. Although levels of nucleotide
552 diversity varied greatly throughout the genome in all three species, we find strong
553 genome-wide correlations of nucleotide diversity among species. It likely reflects
554 conserved variation in mutation rates and/or shared selective constraints across the
555 genomes in these closely related species during the time since their last common
556 ancestor (Hudson *et al.* 1987; Charlesworth *et al.* 1993). Levels of nucleotide
557 diversity are slightly higher in *P. tremuloides* than in *P. tremula*, and the two aspen
558 species collectively harbor greater than two-fold levels of genome-wide diversity
559 compared to *P. trichocarpa*. In accordance with the larger current census population
560 size and substantially more extensive geographic range (Eckenwalder 1996), the
561 higher genetic diversity in both aspen species most likely reflects their larger effective
562 population sizes (N_e) compared to *P. trichocarpa*. Nevertheless, interspecific variation
563 in the mutation rate also deserves further study, particularly in light of recent results
564 showing a feed-forward effect of genome-wide levels of heterozygosity and mutation
565 rates (Lynch 2015; Yang *et al.* 2015). Compared to the consistent pattern of
566 nucleotide diversity between species, the weak correlations in the allele frequency
567 spectrum (Tajima's D) likely reflect different demographic histories for the three
568 species during the Quaternary ice ages (Ingvarsson 2008; Callahan *et al.* 2013; Zhou
569 *et al.* 2014). For instance, the genome-wide excess of rare frequency alleles we

observe in *P. tremuloides* is likely explained by a recent, substantial population expansion that was specific to this species.

In contrast to the mutation rate, recombination rates are only partially conserved among the three species (Figure 2c). The genome-wide average of the ratio of recombination to mutation (ρ/θ_w or c/μ) was similar in *P. tremuloides* (0.39) and *P. trichocarpa* (0.38), but substantially smaller in *P. tremula* (0.22). If mutation rates are indeed unchanged between species, as suggested above, the lower estimate of c/μ in *P. tremula* indicates considerably lower recombination rates in *P. tremula* relative to the other two species. These discrepant results obtained from patterns of polymorphism and recombination between *P. tremula* and *P. tremuloides* likely stems from different effects of effective population size on nucleotide diversity and linkage disequilibrium (Tenesa *et al.* 2007). These effects are known to operate over different time-scales and are likely therefore differentially affected by temporal variation in the effective population size (Tenesa *et al.* 2007; Cutter *et al.* 2013). The recent population size expansion that we infer to have taken place in *P. tremuloides* can thus also explain why its recombination rate is seemingly higher than in *P. tremula*, even if they share similar levels of genome-wide polymorphism. Finally, the c/μ estimates we have estimated for *Populus* are in line with recent genome-wide estimates from several other plant species, such as *Medicago truncatula* (0.29) (Branca *et al.* 2011), *Mimulus guttatus* (0.8) (Hellsten *et al.* 2013) and *Eucalyptus grandis* (0.65) (Silva-Junior and Grattapaglia 2015).

Pervasive signatures of purifying and positive selection across the Populus genome

593 In line with results from most other plant species (Gossmann *et al.* 2010), a majority
 594 (>50%-60%) of new amino acid altering mutations are subject to strong purifying
 595 selection (defined as $N_e > 10$) in *Populus*. We find that the efficacy of purifying
 596 selection on weakly deleterious mutations is positively correlated with the inferred N_e ,
 597 with purifying selection acting more efficiently in *P. tremuloides* that has the largest
 598 N_e compared to the other two species. The same pattern is also found for rates of
 599 adaptive evolution, where estimates of the proportion of amino acid substitutions
 600 driven to fixation by positive selection are highest in *P. tremuloides* (65%), lowest in
 601 *P. trichocarpa* (20%) and intermediate in *P. tremula* (43%). The prevalence of
 602 adaptive evolution in *Populus* contrasts markedly with the estimates in most plant
 603 species, where little evidence of widespread adaptive evolution is found (Gossmann *et*
 604 *al.* 2010). However, *Populus* is not unique among plants showing high rates of
 605 adaptive evolution, and similar estimates have recently been reported in both *Capsella*
 606 *grandiflora* (Slotte *et al.* 2010; Williamson *et al.* 2014) and a number of *Helianthus*
 607 species (Strasburg *et al.* 2011). Most earlier studies doing such estimation have been
 608 based on subsets of genes rather than genome-wide data, and more estimates from
 609 other plant species would be valuable to assess whether the high rate of adaptive
 610 evolution we find in *Populus* is widespread or exceptional.

611 Patterns of genomic variation contain abundant information on the relative
 612 importance of natural selection versus neutral processes in the evolutionary process
 613 (Cutter and Payseur 2013). We find that 0-fold non-synonymous sites exhibit
 614 significantly lower levels of polymorphism compared to 4-fold synonymous sites, and
 615 combined with an excess of rare variants found at 0-fold non-synonymous sites, our
 616 results suggest that the vast majority of amino acid mutations in *Populus* are under

617 purifying selection (Larracuente *et al.* 2008). In addition, introns and 5' UTR sites are
 618 also under some degree of selective constraint, although this constraint is much
 619 weaker than what we observe at non-synonymous sites. 3' UTR sites seem to be
 620 either neutral or at least under comparable extents of selective constraint as 4-fold
 621 synonymous sites are (Andolfatto 2005). In contrast to genic categories, we find
 622 substantially higher levels of polymorphism in intergenic regions in all three species.
 623 Although an artifact of mapping errors due to a greater fraction of repetitive
 624 sequences in intergenic regions could not be entirely excluded, the markedly increase
 625 in diversity may also reflect either higher mutation rates or relaxed selective
 626 constraint in these regions. Future investigations are required to assess the relative
 627 contribution of these alternative factors (Kimura 1983; Begun *et al.* 2007).

628 Apart from strong selective constraints on protein-coding genes, multiple lines
 629 of evidence suggest that genome-wide patterns of polymorphism have been shaped by
 630 widespread natural selection in all three *Populus* species. First, we find significantly
 631 positive correlations between neutral polymorphism and population-scaled
 632 recombination rate in both genic and intergenic regions, even after controlling for
 633 confounding variables such as GC content, gene density, mutation rate and the
 634 number of covered sites by the data. While such a pattern is indicative of the action of
 635 natural selection, it could be explained by either background selection or selective
 636 sweeps. Both of these selective forces affect neutral sites through linkage, and the
 637 impact of selection on linked neutral diversity is more drastic and extensive in regions
 638 of low recombination (Begun and Aquadro 1992; McGaugh *et al.* 2012; Slotte 2014).
 639 The differences in the strength of the association between recombination and levels of
 640 neutral polymorphism likely reflect differences in the effective population size

641 between species (Cutter and Payseur 2013; Corbett-Detig *et al.* 2015), as we observe
 642 substantially stronger signatures of linked selection in *P. tremula* and *P. tremuloides*
 643 compared to *P. trichocarpa*, matching the larger N_e inferred for these species.
 644 However, the impact of natural selection at linked sites also depends greatly on the
 645 local environment of recombination (Cutter and Payseur 2013; Slotte 2014), and in
 646 line with this we observe the strongest signatures of linked selection in *P. tremula*
 647 instead of *P. tremuloides*, consistent with the lower levels of genome-wide
 648 recombination rates we find in *P. tremula*. Different magnitudes of linked selection
 649 provide one of the major explanations for the disparate patterns of genomic variation
 650 among even closely related species (Corbett-Detig *et al.* 2015) and we find that this
 651 also holds true for the three species of *Populus* we have investigated.

652 Second, we find slightly negative correlations between recombination rate and
 653 the ratio of non-synonymous- to synonymous- polymorphism, but not divergence, in
 654 *P. tremula* and *P. tremuloides*, a pattern that suggests a reduced efficacy of purifying
 655 selection at eliminating weakly deleterious mutations in low recombination regions
 656 (Hill and Robertson 1966; Cutter and Choi 2010). The reduction of the efficacy of
 657 natural selection in regions of low recombination, known as Hill-Robertson
 658 interference, may help to understand patterns of partially positive correlations
 659 between gene density and recombination rate in these species (Gaut *et al.* 2007).
 660 Given the relaxed efficacy of purifying selection in regions of low recombination
 661 where weakly deleterious mutations are more likely to accumulate at a high rate,
 662 important functional elements are unlikely to cluster in these regions, as has already
 663 been shown in several other plant species (Anderson *et al.* 2006; Branca *et al.* 2011;
 664 Flowers *et al.* 2012) Consistent with this prediction (Haddrill *et al.* 2007), we find

665 positive association between gene density and recombination rate in regions that
666 experience low rates of recombination. In high-recombination regions where selection
667 is more effective at eliminating slightly deleterious mutations, the association
668 becomes much weaker in all three species. However, it remains unclear whether it is
669 the recombination gradients that drive the functional organization of genomes in
670 response to selection, or whether it is the gradients of functional genomic elements
671 that in turn modify the evolution of recombination rates in *Populus*.

672 Third, by examining the relationship of neutral polymorphism, recombination
673 rate and gene density, we find that levels of neutral polymorphism in genic regions
674 are primarily driven by local rates of recombination, regardless of the density of
675 functional genes. In contrast, we observe a more complex pattern in intergenic regions
676 where levels of intergenic polymorphism are mainly driven by recombination rates in
677 regions of low gene density, while in regions of high gene density levels of intergenic
678 diversity are primarily shaped by the density of nearby genes. As we find that gene
679 density and recombination rates co-vary in all three species, the signatures of linked
680 selection associated with gene density could thus become obscured by rates of
681 recombination, especially in regions of low gene density (Flowers *et al.* 2012). As
682 shown in most plants studied so far (Nordborg *et al.* 2005; Slotte 2014), a negative
683 relationship between gene density and levels of neutral polymorphism is more likely
684 attributed to more intense purifying selection against deleterious mutations in regions
685 of greater gene density, and the magnitude of such effects depends on the strength of
686 purifying selection (Sella *et al.* 2009). In accordance with this expectation, most new
687 mutations in genic regions are strongly deleterious and would be eliminated too
688 quickly to remove large amounts of genetic variation at linked neutral loci. Thus even

in regions of high gene density, we do not find negative correlations between gene density and genetic diversity in genic regions. However, background selection due to deleterious mutations of moderate effect in intergenic regions could account for the negative association we observe between levels of intergenic polymorphism and gene density in regions of high gene density. It is apparent that the extent to which natural selection is acting on noncoding regions of the genome in *Populus* (e.g. intergenic regions) will be an interesting avenue for future studies.

Finally, in all three *Populus* species we find significantly negative correlations between levels of synonymous polymorphism and the rate of amino acid substitution at the scale of single genes. This pattern could be driven by either recurrent selective sweeps or background selection (Charlesworth *et al.* 1993; Andolfatto 2007). However, background selection reduce local N_e due to the removal of weakly deleterious mutations, and is therefore expected to result in both reduced levels of nucleotide polymorphism and an increase of the fixation rate of slightly deleterious mutations (Charlesworth *et al.* 1993). Background selection is thus expected to affect the rates of both synonymous and non-synonymous substitutions equally, but when variation in the rates of synonymous substitution is taken into account, we find a substantially stronger (rather than weaker) negative correlation between levels of synonymous polymorphism and the rate of protein evolution. This suggest that the negative relationship we observe between non-synonymous substitution rate and levels of variation at synonymous sites is most likely driven by effects of recurrent selective sweeps in all three species (Andolfatto 2007; Sella *et al.* 2009). Furthermore, the physical scale at which these signatures of natural selection are detected carries valuable information about the strength of positive selection at the genomic level

(Macpherson *et al.* 2007). Since the signatures of recurrent selective sweeps are only detectable on a genic scale, it mostly reflects relatively weak selection on the majority of adaptive amino acid substitutions and may thus explain why we do not observe the effects at either 100-Kbp or 1-Mbp scales (Macpherson *et al.* 2007; Sella *et al.* 2009).

Conclusion and perspectives

In summary, our findings highlight multiple lines of evidence suggesting that natural selection, both due to purifying and positive selection, has shaped patterns of nucleotide polymorphism at linked neutral sites in all three *Populus* species. Compared to the predictions of the Neutral Theory which suggest that adaptations contribute negligibly to divergence between species (Kimura 1983), we find that around 20% - 65% of all amino acid substitutions are driven to fixation by adaptive evolution in *Populus*. These estimates are in accordance with the results from a number of other organisms with large effective population sizes, such as *Drosophila* (Sella *et al.* 2009), mammalian (Halligan *et al.* 2010; Carneiro *et al.* 2012) and a few plant species (Slotte *et al.* 2010; Strasburg *et al.* 2011), but substantially higher than in species with relatively small effective population sizes, such as humans and most other plant species, where little evidence of adaptive evolution has been detected (Eyre-Walker and Keightley 2009; Gossmann *et al.* 2010). Given that all three *Populus* species share similar life-cycle characteristics, such as outcrossing mating system, relatively large N_e and limited population subdivision, future studies from other long-lived forest trees are needed to investigate whether these are characteristics more generally influencing genome-wide patterns of selection in plants (Hough *et al.* 2013). Furthermore, differences in N_e and rates of recombination among the three

737 *Populus* species are largely explaining differences in the magnitude of linked
738 selection we observe between them.

739 Our analyses suggest pervasive adaptive evolution in all three species of
740 *Populus* and although alternative hypotheses such as demographic effects could lead
741 to spurious evidence of natural selection (Fay *et al.* 2001), the presence of linked
742 selection could also bias inferences of demographic history (Slotte 2014). Due to the
743 pervasive effects of linked selection we have documented in these species, our
744 findings suggest that more attention should be paid to the process of choosing neutral
745 sites for demographic inferences. Alternatively, new methods that allow for the joint
746 estimation of demography and selection from genome-wide data are urgently needed.

747

748 **Acknowledgements**

749

750 We are grateful to Rick Lindroth for providing access to the samples of *P.*
751 *tremuloides* used in this study. We thank Carin Olofsson for extracting DNA for all
752 samples used in this study and Robert J. Williamson for sharing a data analysis script
753 with us. We also thank both the editor and two anonymous referees for useful
754 comments on the manuscript. The research has been funded through grants from
755 Vetenskapsrådet and a Young Researcher Award from Umeå University to PKI. JW
756 was supported by a scholarship from the Chinese Scholarship Council.

757

758 **Literature Cited**

759

760 Alexander, D.H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of
761 ancestry in unrelated individuals. *Genome Res.* 19: 1655-1664.

762 Anderson, L.K., A. Lai, S.M. Stack, C. Rizzon, and B.S. Gaut, 2006 Uneven
763 distribution of expressed sequence tag loci on maize pachytene chromosomes.
764 *Genome Res.* 16: 115-122.

765 Andolfatto, P., 2005 Adaptive evolution of non-coding DNA in *Drosophila*. *Nature*
766 437:1149-1152.

767 Andolfatto, P., 2007 Hitchhiking effects of recurrent beneficial amino acid
768 substitutions in the *Drosophila melanogaster* genome. *Genome Res.* 17: 1755-1762.

769 Begun, D.J., and C.F. Aquadro, 1992 Levels of naturally occurring DNA
770 polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356:
771 519 -520.

772 Begun, D.J., A.K. Holloway, K. Stevens, L.W. Hillier, Y.-P. Poh *et al.*, 2007
773 Population genomics: whole-genome analysis of polymorphism and divergence in
774 *Drosophila simulans*. *PLoS Biol.* 5:e310.

775 Branca, A., T.D. Paape, P. Zhou, R. Briskine, A.D. Farmer *et al.*, 2011 Whole-
776 genome nucleotide diversity, recombination, and linkage disequilibrium in the model
777 legume *Medicago truncatula*. *Proc. Natl. Acad. Sci. USA* 108: E864-E870.

778 Callahan, C.M., C.A. Rowe, R.J. Ryel, J.D. Shaw, M.D. Madritch *et al.*, 2013
779 Continental - scale assessment of genetic diversity and population structure in
780 quaking aspen (*Populus tremuloides*). *J. Biogeogr.* 40:1780-1791.

781 Campos, J.L., D.L. Halligan, P.R. Haddrill, and B. Charlesworth, 2014 The relation
782 between recombination rate and patterns of molecular evolution and variation in
783 *Drosophila melanogaster*. *Mol. Biol. Evol.* 31:1010-1028.

784 Carneiro, M., F.W. Albert, J. Melo-Ferreira, N. Galtier, P. Gayral *et al.*, 2012
785 Evidence for widespread positive and purifying selection across the European rabbit
786 (*Oryctolagus cuniculus*) genome. *Mol. Biol. Evol.* 29: 1837-1849.

787 Charlesworth, B., and J.L. Campos, 2014 The relations between recombination rate
788 and patterns of molecular variation and evolution in *Drosophila*. *Annu. Rev. Genet.*
789 48:383-403.

790 Charlesworth, B., M. Morgan, and D. Charlesworth, 1993 The effect of deleterious
791 mutations on neutral molecular variation. *Genetics* 134:1289-1303.

792 Corbett-Detig, R.B., D.L. Hartl, and T.B. Sackton, 2015 Natural selection constrains
793 neutral diversity across a wide range of species. *PLoS Biol.* 13:e1002112.

794 Cutter, A.D., and J.Y. Choi, 2010 Natural selection shapes nucleotide
795 polymorphism across the genome of the nematode *Caenorhabditis briggsae*. *Genome*
796 *Res.* 20:1103-1111.

797 Cutter, A.D., R. Jovelín, and A. Dey, 2013 Molecular hyperdiversity and evolution in
798 very large populations. *Mol. Ecol.* 22:2074-2095.

799 Cutter, A.D., and B.A. Payseur, 2003 Selection at linked sites in the partial selfer
800 *Caenorhabditis elegans*. *Mol. Biol. Evol.* 20:665-673.

801 Cutter, A.D., and B.A. Payseur, 2013 Genomic signatures of selection at linked sites:
802 unifying the disparity among species. *Nat. Rev. Genet.* 14:262-274.

803 De Carvalho, D., P.K. Ingvarsson, J. Joseph, L. Suter, C. Sedivy *et al.*, 2010
804 Admixture facilitates adaptation from standing variation in the European aspen
805 (*Populus tremula* L.), a widespread forest tree. *Mol. Ecol.* 19:1638-1650.

806 DePristo, M.A., E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire *et al.*, 2011 A
807 framework for variation discovery and genotyping using next-generation DNA
808 sequencing data. *Nature Genet.* 43:491-498.

809 Dickmann, D.I., and J. Kuzovkina, 2014 Poplars and willows of the world, with
810 emphasis on silviculturally important species, pp.8-91 in *Poplars and Willows; trees*
811 *for society and the environment*, edited by J. D. Isebrands and J. Richardson. The
812 Food and Agriculture Organization of the United Nations (FAO) and CAB
813 International (CABI). Rome.

814 Eckenwalder, J.E., 1996 Systematics and evolution of *Populus*, pp. 7-32 in *Biology of*
815 *Populus and its Implications for Management and Conservation* (Part I), edited by R.
816 F. Stettler, H. D. Bradshaw, P. E. Heilman, T. M. Hinckley. NRC Research Press.
817 Ottawa.

818 Ellegren, H., 2014 Genome sequencing and population genomics in non-model
819 organisms. *Trends Ecol. Evol.* 29: 51-63.

820 Evans, L.M., G.T. Slavov, E. Rodgers-Melnick, J. Martin, P. Ranjan *et al.*, 2014
821 Population genomics of *Populus trichocarpa* identifies signatures of selection and
822 adaptive trait associations. *Nature Genet.* 46: 1089–1096.

823 Eyre-Walker, A., and P.D. Keightley, 2009 Estimating the rate of adaptive molecular
824 evolution in the presence of slightly deleterious mutations and population size change.
825 *Mol. Biol. Evol.* 26: 2097-2108.

826 Fay, J.C., G.J. Wyckoff, and C.-I. Wu, 2001 Positive and negative selection on the
827 human genome. *Genetics* 158: 1227-1234.

828 Flowers, J.M., J. Molina, S. Rubinstein, P. Huang, B.A. Schaal *et al.*, 2012 Natural
829 selection in gene-dense regions shapes the genomic pattern of polymorphism in wild
830 and domesticated rice. *Mol. Biol. Evol.* 29:675-687.

831 Fumagalli, M., F.G. Vieira, T. Linderöth, and R. Nielsen, 2014 ngsTools: methods for
832 population genetics analyses from next-generation sequencing data. *Bioinformatics*
833 30:1486-1487.

834 Gaut, B.S., S.I. Wright, C. Rizzon, J. Dvorak, and L.K. Anderson, 2007
835 Recombination: an underappreciated factor in the evolution of plant genomes. *Nat.*
836 *Rev. Genet.* 8: 77-84.

837 González - Martínez, S.C., K.V. Krutovsky, and D.B. Neale, 2006 Forest - tree
838 population genomics and adaptive evolution. *New Phytol.* 170: 227-238.

839 Gossmann, T.I., B.-H. Song, A.J. Windsor, T. Mitchell-Olds, C.J. Dixon *et al.*, 2010
840 Genome wide analyses reveal little evidence for adaptive evolution in many plant
841 species. *Mol. Biol. Evol.* 27: 1822-1832.

842 Haddrill, P.R., D.L. Halligan, D. Tomaras, and B. Charlesworth, 2007 Reduced
843 efficacy of selection in regions of the *Drosophila* genome that lack crossing over.
844 *Genome Biol.* 8:R18.

845 Halligan, D.L., F. Oliver, A. Eyre-Walker, B. Harr, and P.D. Keightley, 2010
846 Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genet.* 6:
847 e1000825.

848 Hamzeh, M., and S. Dayanandan, 2004 Phylogeny of *Populus* (Salicaceae) based on
849 nucleotide sequences of chloroplast trnT-trnF region and nuclear rDNA. *Am. J. Bot.*
850 91:1398-1408.

851 Hellmann, I., K. Prüfer, H. Ji, M.C. Zody, S. Pääbo *et al.*, 2005 Why do human
852 diversity levels vary at a megabase scale? *Genome Res.* 15:1222-1231.

853 Hellsten, U., K.M. Wright, J. Jenkins, S. Shu, Y. Yuan *et al.*, 2013 Fine-scale
854 variation in meiotic recombination in *Mimulus* inferred from population shotgun
855 sequencing. *Proc. Natl. Acad. Sci. USA* 110:19478-19482.

856 Hill, W.G., and A. Robertson, 1966 The effect of linkage on limits to artificial
857 selection. *Genetical Res.* 8:269-294.

858 Hough, J., R.J. Williamson, and S.I. Wright, 2013 Patterns of selection in plant
859 genomes. *Annu. Rev. Ecol. Evol. Syst.* 44: 31-49.

860 Hudson, R.R., M. Kreitman, and M. Aguadé, 1987 A test of neutral molecular
861 evolution based on nucleotide data. *Genetics* 116: 153-159.

862 Hufford, M.B., X. Xu, J. Van Heerwaarden, T. Pyhäjärvi, J.-M. Chia *et al.*, 2012
863 Comparative population genomics of maize domestication and improvement. *Nature*
864 *Genet.* 44:808-811.

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

867 Ingvarsson, P.K., 2010 Natural selection on synonymous and nonsynonymous
868 mutations shapes patterns of polymorphism in *Populus tremula*. *Mol. Biol. Evol.*
869 27:650-660.

870 Jansson, S., R.P. Bhalerao, and A.T. Groover, 2010 *Genetics and genomics of*
871 *Populus*: Springer.

872 Jansson, S., and C.J. Douglas, 2007 *Populus*: a model system for plant biology. *Annu.*
873 *Rev. Plant Biol.* 58:435-458.

874 Jukes, T.H., and C.R. Cantor, 1969 Evolution of protein molecules. pp. 21-132 in
875 *Mammalian protein metabolism*, edited by H.N. Munro. Academic Press. New York.

876 Keightley, P.D., and A. Eyre-Walker, 2007 Joint inference of the distribution of
877 fitness effects of deleterious mutations and population demography based on
878 nucleotide polymorphism frequencies. *Genetics* 177: 2251-2261.

879 Kim, S.H., and V.Y. Soojin, 2007 Understanding relationship between sequence and
880 functional evolution in yeast proteins. *Genetica* 131: 151-156.

881 Kim, S.Y., K.E. Lohmueller, A. Albrechtsen, Y. Li, T. Korneliussen *et al.*, 2011
882 Estimation of allele frequency and association mapping using next-generation
883 sequencing data. *BMC bioinformatics* 12:231.

884 Kimura, M., 1983 *The neutral theory of molecular evolution*: Cambridge University
885 Press.

886 Korneliussen, T.S., A. Albrechtsen, and R. Nielsen, 2014 ANGSD: analysis of next
887 generation sequencing data. *BMC bioinformatics* 15:356.

888 Kulathinal, R.J., S.M. Bennett, C.L. Fitzpatrick, and M.A. Noor, 2008 Fine-scale
889 mapping of recombination rate in *Drosophila* refines its correlation to diversity and
890 divergence. *Proc. Natl. Acad. Sci. USA* 105:10051-10056.

891 Larracuente, A.M., T.B. Sackton, A.J. Greenberg, A. Wong, N.D. Singh *et al.*, 2008
892 Evolution of protein-coding genes in *Drosophila*. *Trends Genet.* 24:114-123.

893 Lawrie, D.S., and D.A. Petrov, 2014 Comparative population genomics: power and
894 principles for the inference of functionality. *Trends Genet.* 30 (4):133-139.

895 Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with
896 BWA-MEM. Preprint. Available: arXiv:1303.3997.

897 Li, H., and R. Durbin, 2011 Inference of human population history from individual
898 whole-genome sequences. *Nature* 475:493-496.

899 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence
900 alignment/map format and SAMtools. *Bioinformatics* 25:2078-2079.

901 Liu, S., E.D. Lorenzen, M. Fumagalli, B. Li, K. Harris *et al.*, 2014 Population
902 genomics reveal recent speciation and rapid evolutionary adaptation in polar bears.
903 *Cell* 157:785-794.

904 Locke, D.P., L.W. Hillier, W.C. Warren, K.C. Worley, L.V. Nazareth *et al.*, 2011
905 Comparative and demographic analysis of orang-utan genomes. *Nature* 469:529-533.

906 Lohmueller, K.E., A. Albrechtsen, Y. Li, S.Y. Kim, T. Korneliussen *et al.*, 2011
907 Natural selection affects multiple aspects of genetic variation at putatively neutral
908 sites across the human genome. *PLoS Genet.* 7:e1002326.

909 Lohse, M., A. Bolger, A. Nagel, A.R. Fernie, J.E. Lunn *et al.*, 2012 RobiNA: a user-
910 friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic
911 Acids Res.* 40:W622-W627.

912 Luikart, G., P.R. England, D. Tallmon, S. Jordan, and P. Taberlet, 2003 The power
913 and promise of population genomics: from genotyping to genome typing. *Nat. Rev.
914 Genet.* 4:981-994.

915 Lynch, M., 2015 Genetics: Feedforward loop for diversity. *Nature* 523:414-416.

916 Mackay, T.F., S. Richards, E.A. Stone, A. Barbadilla, J.F. Ayroles *et al.*, 2012 The
917 *Drosophila melanogaster* genetic reference panel. *Nature* 482:173-178.

918 Macpherson, J.M., G. Sella, J.C. Davis, and D.A. Petrov, 2007 Genomewide spatial
919 correspondence between nonsynonymous divergence and neutral polymorphism
920 reveals extensive adaptation in *Drosophila*. *Genetics* 177:2083-2099.

921 McGaugh, S.E., C.S. Heil, B. Manzano-Winkler, L. Loewe, S. Goldstein *et al.*, 2012
 922 Recombination modulates how selection affects linked sites in *Drosophila*. PLoS
 923 Biol. 10:e1001422.

924 McVean, G.A., S.R. Myers, S. Hunt, P. Deloukas, D.R. Bentley *et al.*, 2004 The fine-
 925 scale structure of recombination rate variation in the human genome. Science
 926 304:581-584.

927 Neale, D.B., and A. Kremer, 2011 Forest tree genomics: growing resources and
 928 applications. Nat. Rev. Genet. 12:111-122.

929 Nevado, B., S. Ramos - Onsins, and M. Perez - Enciso, 2014 Resequencing studies
 930 of nonmodel organisms using closely related reference genomes: optimal
 931 experimental designs and bioinformatics approaches for population genomics. Mol.
 932 Ecol. 23:1764-1779.

933 Nielsen, R., T. Korneliussen, A. Albrechtsen, Y. Li, and J. Wang, 2011 SNP calling,
 934 genotype calling, and sample allele frequency estimation from New-Generation
 935 Sequencing data. PloS One 7:e37558.

936 Nordborg, M., T.T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian *et al.*, 2005 The pattern of
 937 polymorphism in *Arabidopsis thaliana*. PLoS Biol. 3:1289.

938 Payseur, B.A., and M.W. Nachman, 2002 Gene density and human nucleotide
 939 polymorphism. Mol. Biol. Evol. 19:336-340.

940 Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira *et al.*, 2007 PLINK:
 941 a tool set for whole-genome association and population-based linkage analyses. Am.
 942 J. Hum. Genet. 81:559-575.

943 R Development Core Team, 2014 R: A language and environment for statistical
944 computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-
945 900051-07-0.

946 Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt *et al.*,
947 2001 Structure of linkage disequilibrium and phenotypic associations in the maize
948 genome. Proc. Natl. Acad. Sci. USA 98:11479-11484.

949 Sella, G., D.A. Petrov, M. Przeworski, and P. Andolfatto, 2009 Pervasive natural
950 selection in the Drosophila genome. PLoS Genet. 5: e1000495.

951 Silva - Junior, O.B., and D. Grattapaglia, 2015 Genome - wide patterns of
952 recombination, linkage disequilibrium and nucleotide diversity from pooled
953 resequencing and single nucleotide polymorphism genotyping unlock the evolutionary
954 history of *Eucalyptus grandis*. New Phytol. doi: 10.1111/nph.13505.

955 Slotte, T., 2014 The impact of linked selection on plant genomic variation. Brief.
956 Funct. Genomics 13:268-275.

957 Slotte, T., J.P. Foxe, K.M. Hazzouri, and S.I. Wright, 2010 Genome-wide evidence
958 for efficient positive and purifying selection in *Capsella grandiflora*, a plant species
959 with a large effective population size. Mol. Biol. Evol. 27: 1813-1821.

960 Strasburg, J.L., N.C. Kane, A.R. Raduski, A. Bonin, R. Michelmore *et al.*, 2011
961 Effective population size is positively correlated with levels of adaptive divergence
962 among annual sunflowers. Mol. Biol. Evol. 28: 1569-1580.

963 Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by
964 DNA polymorphism. Genetics 123:585-595.

965 Tarailo - Graovac, M., and N. Chen, 2009 Using RepeatMasker to identify repetitive
966 elements in genomic sequences. Curr. Protoc. in Bioinformatics 25:4.10. 11-14.10.
967 14.

968 Tenesa, A., P. Navarro, B.J. Hayes, D.L. Duffy, G.M. Clarke *et al.*, 2007 Recent
969 human effective population size estimated from linkage disequilibrium. Genome Res.
970 17:520-526.

971 Tuskan, G.A., S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev *et al.*, 2006 The
972 genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596-
973 1604.

974 Wang, J., D. Scofield, N.R. Street, and P.K. Ingvarsson, 2015 Variant calling using
975 NGS data in European aspen (*Populus tremula*). pp.43-61 in *Advances in the*
976 *Understanding of Biological Sciences Using Next Generation Sequencing (NGS)*
977 *Approaches*, edited by G. Sablok, S. Kumar, S. Ueno, J. Kuo, C. Varotto. Springer.

978 Wang, Z., S. Du, S. Dayanandan, D. Wang, Y. Zeng *et al.*, 2014 Phylogeny
979 Reconstruction and Hybrid Analysis of *Populus* (Salicaceae) Based on Nucleotide
980 Sequences of Multiple Single-Copy Nuclear Genes and Plastid Fragments. PloS One
981 9:e103645.

982 Watterson, G., 1975 On the number of segregating sites in genetical models without
983 recombination. Theor. Pop. Biol. 7:256-276.

984 Williamson, R.J., E.B. Josephs, A.E. Platts, K.M. Hazzouri, A. Haudry *et al.*, 2014
985 Evidence for Widespread Positive and Negative Selection in Coding and Conserved
986 Noncoding Regions of *Capsella grandiflora*. PLoS Genet. 10: e1004622.

987 Yang, S., L. Wang, J. Huang, X. Zhang, Y. Yuan *et al.*, 2015 Parent-progeny
988 sequencing indicates higher mutation rates in heterozygotes. Nature 523: 463-467.

989 Zhou, L., R. Bawa, and J. Holliday, 2014 Exome resequencing reveals signatures of
 990 demographic and adaptive processes across the genome and range of black
 991 cottonwood (*Populus trichocarpa*). Mol. Ecol. 23:2486-2499.

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

Tables

Table 1. Summary of the correlation coefficients (Spearman's rank correlation coefficient) between levels of neutral polymorphism (Θ), divergence (d) and recombination rate (ρ) in genic and intergenic regions among all three *Populus* species.

Dataset	Species	ρ vs. $\Theta_{4\text{-fold}}$		ρ vs. $d_{4\text{-fold}}$	ρ vs. $\Theta_{\text{Intergenic}}$		ρ vs. $d_{\text{Intergenic}}$
		Pairwise	Partial ^a		Pairwise	Partial ^b	
100Kbp	<i>P. tremula</i>	0.339***	0.309***	0.043	0.062**	0.142***	-0.077**
	<i>P. tremuloides</i>	0.310***	0.284***	0.061**	-0.037	0.100**	-0.029
	<i>P. trichocarpa</i>	0.011	-0.024	0.053*	-0.080**	-0.002	-0.015
1Mbp	<i>P. tremula</i>	0.647***	0.573***	-0.070	0.201**	0.348**	-0.209**
	<i>P. tremuloides</i>	0.400**	0.363**	-0.033	0.032	0.320**	-0.127*
	<i>P. trichocarpa</i>	0.227**	0.151*	-0.027	-0.072	0.165*	-0.120*

^aPartial correlation controls for GC content, gene density, divergence of 4-fold synonymous sites between aspen and *P. trichocarpa*, and coverage (the number of 4-fold synonymous bases covered by sequencing data).

^bPartial correlation controls for GC content, gene density, divergence of intergenic sites between aspen and *P. trichocarpa*, and coverage (the number of intergenic bases covered by sequencing data).

* $P < 0.05$

** $P < 0.001$

*** $P < 2.2 \times 10^{-16}$

Table 2. Summary of the correlation coefficients (Spearman's rank correlation coefficient) between recombination rate (ρ) and the ratio of non-synonymous to synonymous polymorphism ($\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$) and divergence ($d_{0\text{-fold}}/d_{4\text{-fold}}$).

Dataset	Species	ρ vs. $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$		ρ vs. $d_{0\text{-fold}}/d_{4\text{-fold}}$	
		Pairwise	Partial ^a	Pairwise	Partial ^a
100Kbp	<i>P. tremula</i>	-0.057*	-0.075**	-0.012	-0.005
	<i>P. tremuloides</i>	-0.118**	-0.122**	-0.003	-0.002
	<i>P. trichocarpa</i>	-0.004	-0.002	-0.026	-0.020
1Mbp	<i>P. tremula</i>	-0.063	-0.045	-0.007	0.017
	<i>P. tremuloides</i>	-0.142*	-0.092	0.014	0.020
	<i>P. trichocarpa</i>	0.035	-0.002	0.030	0.036

^aPartial correlation controls for GC content, gene density, and the number of 4-fold synonymous and 0-fold non-synonymous bases covered by sequencing data.

* $P < 0.05$

** $P < 0.001$

Table 3. Summary of the correlation coefficients (Spearman's rank correlation coefficient) between gene density and population recombination rate (ρ), neutral polymorphism in genic ($\Theta_{4\text{-fold}}$) and intergenic regions ($\Theta_{\text{intergenic}}$) over 1 Mbp non-overlapping windows in three *Populus* species.

Species	Correlation type	Gene density vs. ρ^a		Gene density vs. $\Theta_{4\text{-fold}}^b$		Gene density vs. $\Theta_{\text{intergenic}}^c$	
		low	high	low	high	low	high
<i>P. tremula</i>	Pairwise	0.674**	-0.112	0.601**	-0.180*	0.431**	-0.605***
	Partial	0.516**	0.263*	0.191*	0.110	0.263*	-0.438**
<i>P. tremuloides</i>	Pairwise	0.527**	0.006	0.576**	-0.077	0.419**	-0.600***
	Partial	0.315**	0.048	0.407**	0.280**	0.363**	-0.444**
<i>P. trichocarpa</i>	Pairwise	0.609**	0.168*	0.417**	-0.033	0.529**	-0.513***
	Partial	0.477**	0.193*	0.242*	0.263**	0.432**	-0.273**

^aPartial correlation controls for GC content and the number of bases covered by the data

^bPartial correlation controls for GC content, population recombination rate, divergence of 4-fold synonymous sites between aspen and *P. trichocarpa*, and coverage (the number of 4-fold synonymous bases covered by sequencing data).

^cPartial correlation controls for GC content, population recombination rate, divergence of intergenic sites between aspen and *P. trichocarpa*, and coverage (the number of intergenic bases covered by sequencing data).

* $P < 0.05$

** $P < 0.001$

*** $P < 2.2 \times 10^{-16}$

Table 4. Summary of the correlation coefficients (Spearman's rank correlation coefficient) between levels of synonymous diversity ($\Theta_{4\text{-fold}}$) and non-synonymous divergence ($d_{0\text{-fold}}$) at different physical scales in three *Populus* species.

Dataset	Species	$d_{0\text{-fold}}$ vs. $\Theta_{4\text{-fold}}$	
		Pairwise	Partial
100 Kbp ^a	<i>P. tremula</i>	-0.029	-0.032
	<i>P. tremuloides</i>	-0.021	-0.025
	<i>P. trichocarpa</i>	-0.053*	-0.051*
1 Mbp ^a	<i>P. tremula</i>	-0.049	0.043
	<i>P. tremuloides</i>	-0.069	-0.008
	<i>P. trichocarpa</i>	-0.086	-0.006
Single-genes ^b	<i>P. tremula</i>	-0.087***	-0.185***
	<i>P. tremuloides</i>	-0.087***	-0.192***
	<i>P. trichocarpa</i>	-0.148***	-0.218***

^aPartial means partial correlation controls for GC content, gene density, population recombination rate, divergence of 4-fold synonymous sites between aspen and *P. trichocarpa*, the number of 4-fold synonymous bases and 0-fold non-synonymous bases covered by sequencing data.

^bPartial means correlation between $d_{0\text{-fold}}$ and $\Theta_{4\text{-fold}}/d_{4\text{-fold}}$.

* $P < 0.05$

** $P < 0.001$

*** $P < 2.2 \times 10^{-16}$

1080 **Figure legends**

1081

1082 **Figure 1. Genome-wide patterns of polymorphism among three *Populus* species.**

1083 Nucleotide diversity (Θ_π) was calculated over 100 Kbp non-overlapping windows in
1084 *P. tremula* (orange line), *P. tremuloides* (blue line) and *P. trichocarpa* (green line)
1085 along the 19 chromosomes.

1086

1087 **Figure 2. Distribution and correlations of (a) polymorphism (Θ_π), (b) Tajima's D**

1088 **and (c) population-scaled recombination rate (ρ) between pairwise comparisons**

1089 **of *P. tremula*, *P. tremuloides* and *P. trichocarpa* over 100 Kbp non-overlapping**

1090 **windows.** The red to yellow to blue gradient indicates decreased density of observed

1091 events at a given location in the graph. Spearman's rank correlation coefficient (ρ)

1092 and the *P*-value are shown in each subplot. (*** $P < 2.2 \times 10^{-16}$, ** $P < 0.001$). The dotted

1093 grey line in each subplot indicates simple linear regression line with intercept being

1094 zero and slope being one.

1095

1096 **Figure 3. Estimates of purifying and positive selection at 0-fold non-synonymous**

1097 **sites in three *Populus* species.** (a) The distribution of fitness effects of new amino

1098 acid mutations (DFE), (b) the proportion of adaptive substitution (α), and (c) the rate

1099 of adaptive non-synonymous to synonymous substitutions (ω) in *P. tremula* (orange

1100 bar), *P. tremuloides* (blue bar) and *P. trichocarpa* (green bar). Error bars represent 95%

1101 bootstrap confidence intervals.

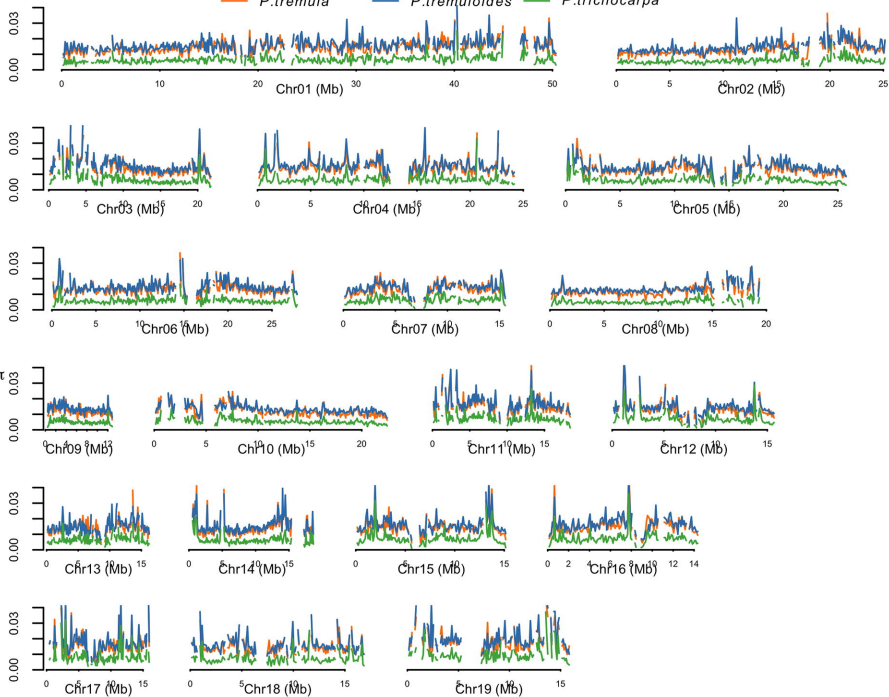
1102

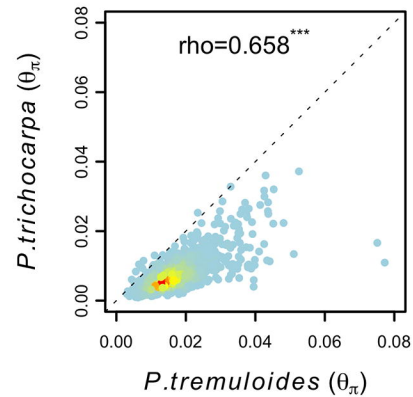
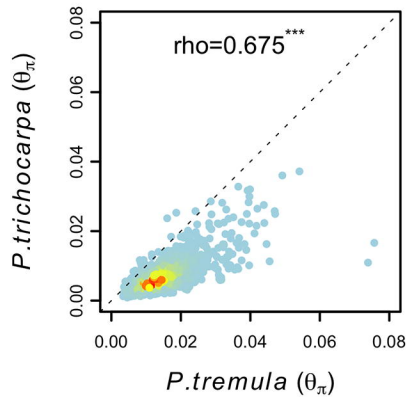
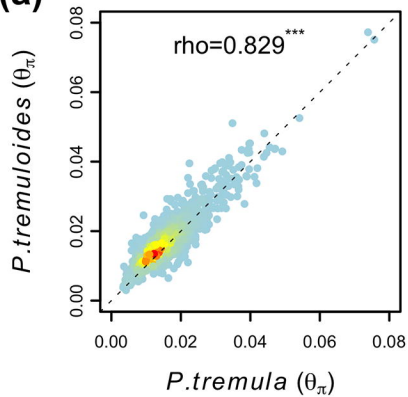
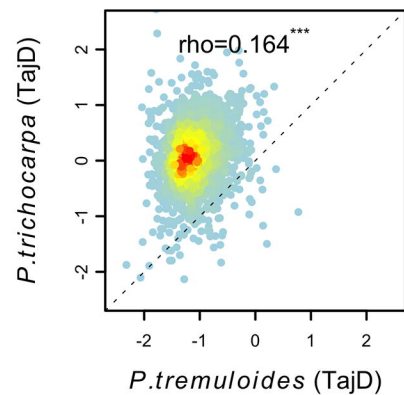
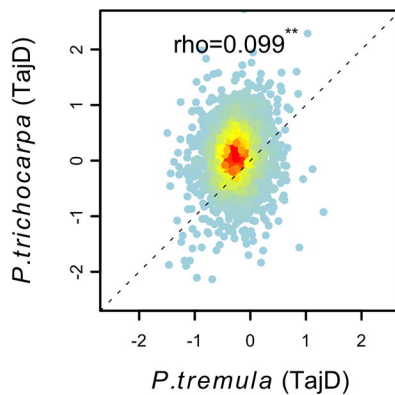
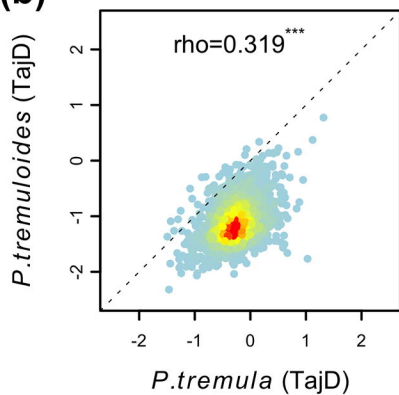
1103 **Figure 4. Correlations of estimates between neutral genetic diversity ($\Theta_{4\text{-fold}}$) (left**
 1104 **panel), neutral genetic divergence ($d_{4\text{-fold}}$) (right panel) and population-scaled**
 1105 **recombination rates (ρ) over 1Mbp non-overlapping windows. Linear regression**
 1106 **lines are colored according to species: (a) *P. tremula* (orange line), (b) *P. tremuloides***
 1107 **(blue line) and (c) *P. trichocarpa* (green line).**

1108
 1109 **Figure 5. Correlations of estimates between (a) population-scaled recombination**
 1110 **rates (ρ), (b) genic genetic diversity ($\Theta_{4\text{-fold}}$), (c) intergenic genetic diversity**
 1111 **($\Theta_{\text{Intergenic}}$) and gene density over 1 Mbp non-overlapping windows in *P. tremula***
 1112 **(left panel), *P. tremuloides* (middle panel) and *P. trichocarpa* (right panel). Grey**
 1113 **points represent the statistics computed over 1Mbp non-overlapping windows.**
 1114 **Colored lines denote the lowess curves fit to the analyzed two variables in each**
 1115 **species.**

1116

— *P. tremula* — *P. tremuloides* — *P. trichocarpa*



(a)**(b)****(c)**