Comparative population genomics of three related *Populus* species

Jing Wang[1], Nathaniel R. Street[2], Douglas G. Scofield[1,3,4], Pär K. Ingvarsson[1*]

[1] Department of Ecology and Environmental Science, Umeå University, Umeå, Sweden

[2] Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, Umeå, Sweden

[3] Department of Ecology and Genetics: Evolutionary Biology, Uppsala University, Uppsala, Sweden

[4] Uppsala Multidisciplinary Center for Advanced Computational Science, Uppsala University, Uppsala, Sweden

* Corresponding author

Email: par.ingvarsson@umu.se

Abstract

A central aim of evolutionary genomics is to identify the relative roles that various evolutionary forces have played in generating and shaping genetic variation within and among species. Here we used whole-genome re-sequencing data from three related *Populus* species to characterize and compare genome-wide patterns of nucleotide polymorphism, site frequency spectrum, population-scaled recombination rate and linkage disequilibrium. Our analyses revealed that *P. tremuloides* has the highest level of genome-wide variation, skewed allele frequencies and population-scaled recombination rates, whereas *P. trichocarpa* harbors the lowest. Consistent with this, linkage disequilibrium decay was fastest in *P. tremuloides* and slowest in *P. trichocarpa*. Pervasive natural selection has been proven to be the primary force creating significant positive correlations between neutral polymorphism and recombination rate in all three species. Disparate effective population sizes and recombination rates among species, on the other hand, drive the distinct magnitudes and signatures of linked selection and consequent heterogeneous patterns of genomic variation among them. We find that purifying selection against slightly deleterious non-synonymous mutations is more effective in regions experiencing high recombination, which may provide one explanation for a partially positive association between recombination rate and gene density in these species. Moreover, distinct signatures of linked selection dependent on gene density are found between genic and intergenic regions within each species. To our knowledge, the present work is the first comparative population genomic study among forest tree species and represents an important step toward dissecting how the interactions of various evolutionary forces have shaped genomic variation within and among these ecologically and economically important tree species.

Author Summary

A fundamental goal of population genetics is to understand how various evolutionary forces shape the heterogeneity of genomic variation within and among species. Here, we characterize and compare genome-wide patterns of nucleotide diversity, site frequency spectra, population-scaled recombination rates and linkage disequilibrium among three related *Populus* species: *Populus tremula*, *P. tremuloides* and *P. trichocarpa*. Pervasive natural selection, mediated by the local recombination environments, is supposed to be the primary force shaping heterogeneous patterns of neutral polymorphism throughout the genome. The disparate magnitudes and signatures of linked selection among the three species, however, likely result from either different effective population sizes and/or differences in recombination rates among them. Moreover, we find distinct patterns of selection between genic and intergenic regions in all three species, indicating these two types of sites may have undergone independent evolutionary responses to selection in *Populus*. To our knowledge, the present work provides the first phylogenetic comparative study of genome-wide patterns of variation between closely related forest tree species. This information will also improve our ability to understand how various evolutionary forces have interacted to influence genome evolution among related species.

Introduction

A major goal in evolutionary genetics is to understand how genomic variation is established, maintained and diverge within and between species [1, 2]. Various evolutionary forces are known to have substantial impacts in shaping genetic variation and linkage disequilibrium throughout the genome [3]. Under the neutral theory, genetic variation is the manifestation of the balance between mutation and genetic drift [4]. Demographic fluctuations, such as population expansion and/or bottlenecks, can cause patterns of genome-wide variation to deviate from standard neutral model in various ways [5]. Natural selection, via positive selection favoring beneficial mutations (genetic hitchhiking) and/or negative selection against deleterious mutations (background selection), plays an important role in sculpting the landscape of polymorphism across the genome [2, 6-8]. The signature and magnitude of apparent selection at linked sites depends heavily on the local environment of recombination [9, 10]. Linked selection is expected to remove more neutral polymorphism in low-recombination regions compared to high-recombination regions [10-12]. In addition to indirectly affecting genetic variation via linked selection, the rate of recombination can also shape the landscape of genomic features, such as base composition and gene density [13, 14]. However, there remains much to be learned about how these various evolutionary forces have shaped the heterogeneous patterns of genomic polymorphism within and between species [2, 6, 15]. With the advance of next-generation sequencing technology, sufficient genome-wide data among multiple related species are becoming available [16, 17]. Phylogenetic comparative approaches using these data will place us in a stronger position to understand the relative importance of mutation, genetic drift, natural selection and recombination in determining patterns of genome evolution [18, 19].

Thus far, genome-wide comparative studies have largely dealt with experimental model species, mammals, and cultivated plants of either agricultural or horticultural interest [19-21]. Forest trees, as a group, are characterized by extensive geographical distributions and are of high ecological and economic value [22]. Most forest trees have largely persisted in an undomesticated state and, until quite recently, without anthropogenic influence [22]. Accordingly, in contrast to crop and livestock lineages that have been through strong domestication bottlenecks, most extant populations of forest trees harbor a wealth of genetic variation and they may be excellent model systems for dissecting the dominant evolutionary forces that sculpt patterns of variation throughout the genome [22, 23]. Among forest tree species, the genus *Populus* represents a particularly attractive choice because of its wide geographic distribution, important ecological role in a wide variety of habitats, multiple economic uses in wood and energy products, and relatively small genome size [24, 25]. Here, we studied three *Populus* species which differ in their morphology, geographic distribution, population size and phylogenetic relationship (S1 Fig) [26, 27]. *P. tremula* and *P. tremuloides* (collectively 'aspens') have wide native ranges across Eurasia and North America respectively, and are closely related, belonging to the same section of the genus *Populus* (section *Populus*) [27]. In contrast, *P. trichocarpa* belongs to a different section of the genus (section *Tacamahaca*) that is reproductively isolated from members of the section *Populus* [27]. The distribution of *P. trichocarpa* is restricted to western North America and it's range is considerably smaller than the two aspen species [28]. Importantly, *P. trichocarpa* also represents the first tree species to have its genome sequenced [29] 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 [26, 30].

In this study, we used novel and existing Illumina short read (2 × 100 bp) datasets to characterize, compare and contrast genome-wide patterns of nucleotide diversity, recombination rate, and linkage disequilibrium, and to infer contextual patterns of selection throughout the genomes of all three species.

Results

We generated whole-genome sequencing data of 24 genomes of *P. tremula* and 22 genomes of *P. tremuloides* (S1 Table) with all samples sequenced to relatively high depth (24.2×-69.2×; S2 Table). We also downloaded 24 genomes from already published data of *P. trichocarpa* [31]. After adapter removal and quality trimming, 949.2 Gb of high quality sequence data remained (S2 Table, S2 Fig). Reads from the three species were mapped to the *P. trichocarpa* reference genome [29] using BWA-MEM [32], with the mean mapping rate being 89.8% for individuals of *P. tremula*, 91.1% for individuals of *P. tremuloides*, and 95.2% for individuals of *P. trichcoarpa* (S2 Table). On average, the genome-wide coverage of uniquely mapped reads was more than 20× for each species (S2 Table). After excluding sites with extreme coverage, low mapping quality, or those overlapping with annotated repetitive elements separately in each species (see Materials and Methods), 42.8% of collinear genomic sequences remained for downstream analyses. Among all retained genomic regions, 54.9% were located within gene boundaries, which covers 70.1% of all genic regions predicted from *P. trichocarpa* assembly, and the remainder (45.1%) was located in intergenic regions.

Negligible population substructure in samples of all three species

Given the great dispersal capabilities of pollen and seed in *Populus* [24, 25], population genetic structure appears to be generally weak in most *Populus* species [31, 33]. In order to ascertain the population structure within and between species, we used a model-based clustering algorithm, implemented in ADMIXTURE [34], to cluster sampled individuals using only 4-fold synonymous single nucleotide polymorphisms (SNPs) with minor allele frequency greater than 10%. When we analyzed population structure between species, we found the model exhibits the lowest cross-validation error when K=3 (S3b Fig), which clearly subdivides the three species into three distinct clusters (S3a Fig). When we analyzed local population structure within each species, K=1 minimized the cross-validation error in all species, implying that extremely weak population structure in the samples of the three *Populus* species (S3c-e Fig). Therefore, intra-species population structure likely play a negligible role in our comparative population genetic analyses among the three species.

Patterns of divergence among the three *Populus* species

We measured pairwise nucleotide divergence (d_{xy}) among pairs of the three species across the genome in non-overlapping 100 kilobase pairs (Kbp) windows. Between either of the two aspen species and *P. trichocarpa*, d_{xy} was significantly higher than between the two aspen species (Wilcoxon rank sum test, P-value<0.001) (S4 Fig). We found extremely consistent patterns of divergence between the two aspen species and *P. trichocarpa* (Spearman's ρ = 0.994, P-value<0.001) (S5a Fig), reflecting the historical divergence of the common ancestor of two aspen species from *P.*

trichocarpa and the relatively recent divergence of the two aspen species. In addition, we found that the divergence was significantly correlated between the evolutionarily independent lineages (*P. tremula*-*P. tremuloides)* vs. (aspens-*P. trichocarpa*) (S5b,c Fig), suggesting that patterns of genome-wide variation of mutation rates and/or selective constraints are relatively conserved across these three *Populus* species.

Polymorphism varies, but is highly correlated, between species

Fig 1 shows genome-wide estimates of nucleotide diversity among all three species over non-overlapping 100 Kbp windows. We also performed the analyses using 1 megabase pair (Mbp) windows, with the results being nearly identical (S6a Fig). We found that both aspen species harbor substantial levels of nucleotide diversity (Θ_{Π}=0.0133 in *P. tremula*; Θ_{Π}=0.0144 in *P. tremuloides*), approximately two-fold higher than the diversity in *P. trichocarpa* (Θ_{Π}=0.0059) (Table 1; Fig 1; S6a Fig). The overall nucleotide diversity we observe in *P. trichocarpa* was a slightly higher than the value reported in [31]. This is likely due to differences in the methods used between the two studies. In this study, we utilized the full information of the filtered data and estimated the population genetic statistics directly from genotype likelihoods, which takes statistical uncertainty of SNP and genotype calling into account and should give more accurate estimates [35, 36]. In accordance with the highly consistent genome-wide distribution of Θ_{Π} among the three *Populus* species (Fig 1), we observed a significantly positive correlation of Θ_{Π} between each pair of species across the whole genome (Fig 2a). Such strong correlations of polymorphism suggest that mutation rates and/or selective constraints are highly conserved among the species despite the clades represented by genetic sections *Populus* and *Tacamahaca* having diverged for ~4.5 million years [37]. Not surprisingly, we found higher correlation of Θ_{Π} between *P. tremula* and *P. tremuloides* (Spearman's ρ=0.829, *P*-value<0.001, Fig 2a), which both belong to section *Populus*, compared to the correlation between the two aspen species and *P. trichocarpa*, which is likely due to the higher levels of shared ancestral polymorphism between the aspens [26].

Table 1. Diversity statistics (median and central 95% range) for various genomic contexts over 100 Kbp non-overlapping windows across genome

	Filtered bases (Mbp)	***P. tremula***		***P. tremuloides***		***P. trichocarpa***	
		Θ_{Π}	**Tajima's D**	**Θ_{Π}**	**Tajima's D**	**Θ_{Π}**	**Tajima's D**
Total	136.47	0.0133(0.0076-0.0236)	-0.2723(-0.7727-0.2941)	0.0144(0.0091-0.0247)	-1.1688(-1.6003--0.4899)	0.0059(0.0031-0.0125)	0.0643(-0.8266-0.9496)
0-fold[a]	16.52	0.0035***(0.0011-0.0085)	-1.0913***(-2.2240-0.0128)	0.0044***(0.0018-0.0091)	-2.1717***(-2.7932--1.2275)	0.0013***(0.0003-0.0043)	-0.4090***(-1.7625-1.4391)
4-fold	3.40	0.0108(0.0035-0.0207)	-0.2220(-1.4676-0.9667)	0.0120(0.0044-0.0214)	-1.3689(-2.2458--0.2602)	0.0040(0.0008-0.0104)	0.0084(-1.5663-1.7753)
Introns[a]	31.89	0.0096***(0.0038-0.0182)	-0.2669**(-1.3490-0.7526)	0.0106***(0.0046-0.0184)	-1.4286**(-2.2283--0.4173)	0.0038**(0.0013-0.0094)	-0.0245*(-1.5429-1.6489)
UTR 5'[a]	4.02	0.0091***(0.0033-0.0192)	-0.5642***(-1.7370-0.6994)	0.0104***(0.0041-0.0197)	-1.5829***(-2.3688--0.4202)	0.0038**(0.0007-0.0102)	-0.1040**(-1.6203-1.6915)
UTR 3'[a]	7.19	0.0108(0.0039-0.0190)	-0.3081**(-1.4577-0.7662)	0.0121(0.0050-0.0204)	-1.3842*(-2.2319--0.3766)	0.0043(0.0012-0.0101)	-0.0033(-1.5265-1.6444)
Intergenic[a]	73.46	0.0184(0.0110-0.0313)	-0.3062**(-0.8360-0.4298)	0.0198(0.0130-0.0326)	-1.1843(-1.6698--0.3969)	0.0088(0.0044-0.0175)	0.1042(-0.9211-1.1131)

[a] One-sided paired Mann-Whitney U test in comparison to the 4-fold synonymous

* *P*<0.05

** *P*<0.001

*** $P<2.2\times10^{-16}$

Fig 1. Genome-wide patterns of polymorphism among three *Populus* species. Nucleotide diversity (Θ_{π}) was calculated over 100 Kbp non-overlapping windows in

P. tremula (orange line), *P. tremuloides* (blue line) and *P. trichocarpa* (green line) along the 19 chromosomes.

Fig 2. Correlations of polymorphism, Tajima's D and population recombination rate between species. Distributions and correlations of (a) pairwise nucleotide diversity (Θ_π), (b) Tajima's D, (c) population-scaled recombination rate (ρ) between pairwise comparisons of *P. tremula*, *P. tremuloides* and *P. trichocarpa*, over 100 Kbp non-overlapping windows. The red to yellow to blue gradient indicates decreased density of observed events at a given location in the graph. Spearman's rank correlation coefficient (*ρ*) and the *P*-value are shown in each subplot. (*** $P<2.2\times10^{-16}$, ***P*<0.001). The dotted grey line in each subplot indicates simple linear regression line with intercept being zero and slope being one.

Along all chromosomes, the distribution of polymorphisms was more variable (average coefficient of variation (CV) of Θ_Π among three species=0.3362) than was divergence (average CV of d_{xy} =0.1670) (Fig 1; S4 Fig). As the *Populus* karyotype has not been established and thus the locations of centromeres and telomeres remains unknown, we can only speculate that the genomic regions with long measurement gaps that failed to pass our quality requirements may represent repetitive regions of chromosomes located near centromeres, and with distal chromosomal regions being the approximate locations of telomeres. Fig 1 shows that diversity generally declines near the supposed locations of centromeres and telomeres in all three *Populus* species. Divergence, however, did not show similar patterns of decline in such regions (S4 Fig), potentially indicating that the reduced polymorphism in these regions is most likely due to a greater influence of selection at linked sites because of reduced recombination in these regions (see below) rather than reduced neutral mutation rates [2].

Tajima's D varies, and is weakly correlated, between species

Genome-wide allele frequency distributions can also help elucidate the relative contributions of different evolutionary dynamics in charactering patterns of polymorphism. We compared the site frequency spectrum among the three species based on the Tajima's D statistic [38], which is the standardized difference between the average pairwise sequence diversity (Θ_{Π}) and the number of segregating sites (Θ_{W}). Under the standard neutral model, the expected value of Tajima's D is roughly equal to 0 ($\Theta_{\Pi} = \Theta_{W}$). Negative Tajima's D ($\Theta_{\Pi} < \Theta_{W}$; an excess of rare alleles) usually results from purifying selection, selective sweeps, or population expansion, whereas positive Tajima's D ($\Theta_{\Pi} > \Theta_{W}$; an excess of common alleles) indicates either balancing selection or a decrease in population size. We found dramatically different patterns in the genome-wide distribution of Tajima's *D* among the three species (Fig 3). The genome-wide average of Tajima's D was slightly positive in *P. trichocarpa*, whereas *P. tremula* had negative genome-wide averages of Tajima's D (Table1; Fig 3; S6c Fig). Compared to *P. trichocarpa* (average Tajima's D=0.064) and *P. tremula* (average Tajima's D=-0.272), *P. tremuloides* (average Tajima's D=-1.169) showed substantially more negative values of Tajima's D along all chromosomes (Fig 3; S6c Fig), reflecting a large excess of low-frequency polymorphisms across the genome. As natural selection is usually expected to act on a relatively small number of genomic regions, the marked genome-wide negative Tajima's D is most likely to be explained by a recent substantial expansion in population size in *P. tremuloides*. The weakly negative Tajima's D in *P. tremula* could also reflect an increase in population size although not as great as that experienced by *P. tremuloides*. The slightly positive Tajima's D in *P. trichocarpa*, however, implies that it may have experienced a recent population contraction as was also suggested by [39]. In contrast to the significantly positive correlations of nucleotide diversity among the three species, the much weaker correlations seen for Tajima's D (Fig 2b) could be ascribed to either the different demographic histories of these species, or result from different targets of divergent selection due to different environmental conditions experienced by the species since their divergence [37].

Fig 3. Genome-wide patterns of allele frequency distribution among three *Populus* species. Tajima's D was calculated over 100 Kbp non-overlapping windows in *P. tremula* (orange line), *P. tremuloides* (blue line) and *P. trichocarpa* (green line) along the 19 chromosomes.

Little confounding effects of population structure, biased sampling schemes and hybridization

The observed variation of intra-species population genetic patterns could also be caused by other factors, such as population sub-structure, biased sampling schemes and/or hybridization [40, 41]. We found no obvious population sub-structure in samples among all three species (S3 Fig) and thus expect that any effect of population structure is negligible in structuring polymorphisms in these three species. Biased sampling schemes could lead to biased estimates of genome-wide diversity and allele frequency spectrum as the samples of *P. tremula* and *P. trichocarpa* were all collected from continuous local populations whereas those of *P. tremuloides* were collected from two discrete populations (S1 Fig). The lack of population substructure suggests that bias is unlikely, but to be sure we tested this by calculating Θ_Π and Tajima's D separately for the two local *P. tremuloides* populations (Alberta and Wisconsin). We observed remarkably similar values of Θ_Π in both local population samples and the pooled samples (S7a Fig), confirming that structured sampling in this species does not affect our results. Values of Tajima's D were slightly skewed toward more negative for the pooled samples compared with single sampling localities (S7b Fig), likely reflecting low sharing of low-frequency polymorphisms between these localities which is consistent with widespread population expansion. This seems unlikely to influence the comparison of the overall patterns of genetic variation among the species. An additional possibility is that the excess of rare alleles we observed in *P. tremuloides* could be derived from one or few "outlier" individuals that are misidentified or are recent inter-specific hybrids. To assess this possibility we calculated the number of singletons contributed by each individual in the dataset. We found an overall higher number of singletons in individuals of *P. tremuloides* relative to the other two species, which was expected from patterns of Tajima's D, but there were no outlier individuals in *P. tremuloides* that contribute disproportionally large numbers of singletons (S8 Fig). Together these results indicate that the genome-wide excess of rare variants we observed in *P. tremuloides* is a species-wide pattern rather than being population or individual specific.

Patterns of polymorphism and divergence vary by genomic contexts

We compared patterns of nucleotide diversity and divergence across different genomic contexts and in all comparisons levels of nucleotide diversity and divergence were highest for intergenic sites, followed by 4-fold synonymous sites, 3'UTRs, 5'UTRs, introns and were lowest at 0-fold non-synonymous sites (Table 1; S3 Table;

S9 Fig; S11 Fig). The extremely high levels of diversity and divergence in intergenic regions could arise due artifacts of mapping errors in repetitive sequences [42]. However, we applied the same strict filtering steps in both genic and intergenic regions, making this error bias less likely. Therefore, the markedly higher levels of diversity and divergence in intergenic regions probably result from a higher mutation rate, a relaxed selective constraint or both [2]. If we assume that the mutation rate of intergenic regions does not differ from that in genic regions, we could infer that there is strong selective constraint on all genic features throughout the genomes. Nevertheless, the relative contribution of alternative factors to the higher divergence rate in intergenic regions requires further investigation.

Within genic regions, 3’ UTRs showed only slightly lower levels of divergence and similar levels of diversity and allele frequency distribution compared to 4-fold synonymous sites (Table 1; S3 Table). This suggests that the large majority of sites in 3’ UTRs are effectively neutral or are subject to purifying selection to an extent comparable to 4-fold synonymous sites. We found a slight, but significant, reduction in diversity, Tajima’s D and divergence in introns and 5’ UTRs, consistent with the notion that introns and 5’UTRs have undergone stronger selective constraint than 4-fold synonymous sites (Table 1; S3 Table). Finally, both diversity and divergence at 0-fold non-synonymous sites was nearly three times lower than 4-fold synonymous sites. In accordance with this, we found significantly lower Tajima’s D at 0-fold non-synonymous sites compared to 4-fold synonymous sites (P<0.001, Mann-Whitney U test) (Table 1), indicating that a large majority of amino acid substitutions are under strong purifying selection [43].

Linkage Disequilibrium (LD) and Recombination

Populus species are predominantly outcrossing and thus the expectation is thus that LD decays rapidly and that the rates of scaled recombination are high [44]. However, a recent genome-wide analysis in *P. trichocarpa* has revealed more extensive LD across the genome that was expected base on earlier studies [45]. We found that the average LD (r^2) between pairs of SNPs fell to lower than 0.2 within approximately 6-7 Kbp in *P. trichocarpa* (Fig 4), which is consistent with values previously reported in this species [45]. In *P. tremula*, mean r^2 dropped below 0.2 within about 5 Kbp, which is substantially greater than reported in earlier studies that were based on a small number of candidate gene fragments [44]. Finally, LD decayed considerably more rapidly in *P. tremuloides* compared to the other two species, with mean r^2 dropping below 0.2 within ~2-3 Kbp (Fig 4).

Fig 4. Decay of linkage disequilibrium (LD). The decay of LD (estimated as r^2) with physical distance in *P. tremula* (orange line), *P. tremuloides* (blue line) and *P. trichocarpa* (green line).

We also estimated population-scaled recombination rates (ρ) in each species. There was considerable large-scale variation in recombination rates throughout the genomes of all three species, with ρ in *P. tremuloides* consistently being higher than in the other two species (Fig 5). In accordance with the genome-wide patterns of diversity, we also found patterns of decreasing ρ near the putative locations of centromeres and telomeres in all three species (Fig 5). When we measured the average r^2 over 100 Kbp non-overlapping windows across the genome, we found population recombination rates were significantly correlated with the extent of LD (mean pairwise r^2) in all species (S12 Fig). The mean ρ computed from 100 Kbp windows in

P. tremuloides was 8.42 Kbp^{-1} (standard deviation of 4.71 Kbp^{-1}), and the mean ρ in *P. tremula* was 3.23 Kbp^{-1} (standard deviation: 1.66 Kbp^{-1}). The genome-wide average ρ in *P. trichocarpa* was 2.19 Kbp^{-1} (standard deviation: 1.11 Kbp^{-1}), which is consistent with the previously reported ρ value estimated from exome re-sequencing data [39]. Concordant ρ values for all three species were also observed in 1Mbp windows (S6d Fig). In comparison to the extremely high correlation of diversity and low correlation of allele frequency spectrum among the three *Populus* species (Fig 2a,b), we found an intermediate correlation in recombination rates between species, suggesting that the overall recombination environment is only partially conserved among the three species (Fig 2c).

Fig 5. Genome-wide patterns of population-scaled recombination rate among three *Populus* species. Population-scaled recombination rate (ρ) was averaged over 100 Kbp non-overlapping windows in *P. tremula* (orange line), *P. tremuloides* (blue line) and *P. trichocarpa* (green line) along the 19 chromosomes.

For populations under drift-mutation-recombination equilibrium, $\rho = 4N_ec$ (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 (S13 Fig). The mean c/μ in *P. tremuloides* and *P. trichocarpa* was 0.39 and 0.38 respectively, indicating that mutations occur approximately two to three times more frequently than recombination events. On the other hand, the average value of c/μ in *P. tremula* was 0.22, implying that recombination is less important than mutation in generating diversity in *P. tremula* compared to the other two *Populus* species.

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

If natural selection is pervasive across the genome, positive correlations between levels of neutral polymorphisms and recombination rates are expected since demography alone is unlikely to generate these patterns [8]. If selection is the primary force driving the association of neutral polymorphism and recombination rate, the association should be stronger in genic regions of the genome than in intergenic regions since genes are more likely to be targets of selection. In order to examine these correlations, we first assumed that 4-fold synonymous sites in genic regions represent selectively neutral sites, as every possible mutation in 4-fold degenerate sites is synonymous. In the following we refer to the pairwise nucleotide diversity at 4-fold synonymous sites ($\theta_{4\text{-fold}}$) as "neutral polymorphism". We then measured the 4-fold synonymous substitution rate ($d_{4\text{-fold}}$) between either of the two aspen species and *P. trichocarpa* and used this to represent "neutral divergence", which was further taken as a proxy for the neutral mutation rate [46]. As many other genomic features may also influence the variation of neutral polymorphism, we also tabulated GC content, gene density and the number of neutral bases covered by sequencing data for all three species. All measurements were carried out in non-overlapping windows that were either 100 Kbp or 1Mbp in size.

We found significantly positive correlations between the level of neutral polymorphism ($\theta_{4\text{-fold}}$) and population recombination rate for the two aspen species (Table 2), with correlations being stronger in *P. tremula* compared to *P. tremuloides*.

In *P. trichocarpa*, however, we found either no or weak correlation between diversity and recombination (Table 2). Compared to 100 Kbp windows, the correlations were stronger in 1Mbp windows among all species, which most likely results from the higher signal-to-noise ratio provided by larger genomic regions (Table 2). In the remainder of this paper we therefore focus our analyses primarily on data generated using 1Mbp window size. We performed simple linear regression analysis between recombination rate and diversity, and 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 (Fig 6).

Table 2. Summary of the correlation coefficients (Spearman's ρ) between levels of neutral polymorphism, divergence and recombination rate in 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	*P. tremula*	0.339^{***}	0.309^{***}	0.043	0.062^{**}	0.142^{***}	-0.077^{**}
	P. tremuloides	0.310^{***}	0.284^{***}	0.061^{**}	-0.037	0.100^{**}	-0.029
	P. trichocarpa	0.011	-0.024	0.053^{*}	-0.080^{**}	-0.002	-0.015
1Mbp	*P. tremula*	0.647^{***}	0.573^{***}	-0.070	0.201^{**}	0.348^{**}	-0.209^{**}
	P. tremuloides	0.400^{**}	0.363^{**}	-0.033	0.032	0.320^{**}	-0.127^{*}
	P. trichocarpa	0.227^{**}	0.151^{*}	-0.027	-0.072	0.165^{*}	-0.120^{*}

[a] Partial 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).

[b] Partial 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

*** $P<2.2\times10^{-16}$

Fig 6. Correlations between estimates of neutral genetic diversity and divergence with population recombination rates over 1Mbp non-overlapping windows. Correlations between estimates of 4-fold synonymous diversity ($\Theta_{4\text{-fold}}$) (left panel) and divergence ($d_{4\text{-fold}}$) (right panel) with population-scaled recombination rate (ρ) over 1Mbp non-overlapping windows. Linear regression lines are colored according to species: (a) *P. tremula* (orange line), (b) *P. tremuloides* (blue line) and (c) *P. trichocarpa* (green line).

If the relationship between diversity and recombination rate was merely caused by the mutagenic effect of recombination, similar correlations should also be observed between divergence and recombination rate. However, no such correlations were observed in any of the three species (Table 2; Fig 6). The association between recombination rate and nucleotide diversity, and not with divergence, is thus most likely caused by the effects of linked natural selection, where the elimination of linked polymorphisms caused by selection is disproportionally stronger in low-recombination genomic regions relative to regions of high recombination [8-10, 47]. Moreover, among all three species, the correlations between neutral polymorphism and recombination rate remained significant even after we performed partial correlation analyses to control for several possible confounding factors such as GC content, gene density, divergence at neutral sites, and the number of neutral bases covered by sequencing data (Table 2).

In accordance with the view that genes represent the most likely targets of natural selection, the correlations between intergenic diversity and recombination rate were substantially weaker than those correlations in genic regions (Table 2). Only 7.3% of intergenic genetic variation in *P. tremula* could be explained by recombination, whereas the impact of recombination rate on intergenic diversity in *P. tremuloides* and *P. trichocarpa* was <1% and could be considered negligible (Table 2; S14 Fig). In addition, we found slightly negative correlation between the divergence and recombination in intergenic regions (Table 2; S14 Fig). This pattern is likely to be explained by Hill-Robertson interference where weakly deleterious intergenic mutations would reach fixation due to ineffective purifying selection in regions of low recombination [48]. Further investigation is required to support this assertion. Notably, after controlling for GC content, gene density, divergence and the number of covered intergenic bases using partial correlation analyses, the correlations between intergenic diversity and recombination rate become significant in all species, but remained relatively weak compared to the values for genic regions in the two aspen species (Table 2).

Inconsistent effect of gene density on patterns of polymorphism in genic vs. intergenic regions

Genome-wide signatures of linked selection are not only influenced by the local environments of recombination rate, but also are sensitive to the density of functionally important sites within specific genomic regions [14]. Genomic regions with a high density of genes are therefore expected to have undergone stronger effects of linked selection and should therefore exhibit lower levels of neutral polymorphism [1, 14]. However, a positive or negative co-variation of gene density and recombination rate would either act to obscure or strengthen the genome-wide signatures of linked selection, respectively [7, 14, 49]. We measured gene density as the number of protein-coding genes in each 1Mbp window, which was unsurprisingly also found to be highly correlated with the proportion of coding bases in each window (S15 Fig). In all three *Populus* species, we found significantly positive correlation between population recombination rate and gene density (Fig 7a). However, rather than being linear, the relationships between recombination rate and gene density were found to be curvilinear in all three species, with a significant positive correlation observed only in regions of lower gene density (gene number smaller than ~85 within each 1Mbp window) (Table 3). In clear contrast, in high gene density regions (gene number greater than ~85 within each 1Mbp window) we observed no correlations between recombination rate and gene density in any of the two aspen species, and only weak correlation in *P. trichocarpa* (Table 3; Fig 7a). These correlation patterns persisted after controlling for the GC content and the number of bases covered by sequencing data in each window (Table 3).

Fig 7. Correlations between estimates of population recombination rates, genic and intergenic genetic diversity with gene density over 1 Mbp non-overlapping windows. (a) Correlations between gene density and population-scaled recombination rate (ρ) in *P. tremula* (left panel), *P. tremuloides* (middle panel) and *P. trichocarpa* (right panel). (b) Correlations between gene density and neutral genetic diversity ($\Theta_{4\text{-fold}}$) in *P. tremula*, *P. tremuloides* and *P. trichocarpa*. (c) Correlations between gene density and intergenic genetic diversity ($\Theta_{\text{Intergenic}}$) in *P. tremula*, *P. tremuloides* and *P. trichocarpa*. Grey points represent the statistics computed over 1Mbp non-overlapping windows. Colored lines denote the lowess curves fit to the two variables in each species.

Table 3. Summary of the correlation coefficients (Spearman's ρ) between gene density and population recombination rate, neutral polymorphism in genic and intergenic regions 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**
P. tremula	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^{**}
P.tremuloides	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^{**}
P.trichocarpa	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^{**}

[a] Partial correlation controls for GC content and the number of bases covered by the data

[b] Partial 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).

[c] Partial 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).

* $P<0.05$

** $P<0.001$

*** $P<2.2\times10^{-16}$

We then examined the correlation between neutral polymorphism and gene density. Compared to the prediction of lower diversity in regions with higher functional density [50], we found that the correlation pattern between genic diversity and gene density was highly consistent with the pattern found in recombination rate, where significantly positive correlations were found in regions of lower gene density and either no correlation or weak negative correlations were found in regions of higher gene density (Table 3; Fig 7b). After controlling for potential confounding variables such as GC content, recombination rate, neutral divergence, and the number of covered sites in each window, weaker but significant positive correlations between neutral diversity and gene density remained in all three species in regions of low gene density (Table 3). Positive associations between neutral diversity and gene density were also found in high gene-density regions (Table 3).

Compared with genic regions, different correlation patterns between intergenic diversity and gene density were found in all three species (Fig 7c). In accordance with genic regions, we found significantly positive correlation between intergenic diversity and gene density in regions of lower gene density. However, in regions of higher gene density, strongly negative correlations between intergenic diversity and gene density were observed in all three species (Table 3; Fig 7c). Due to the lack of correlation between intergenic divergence and gene density (S16 Fig), our findings suggest that the levels of intergenic polymorphism are also largely affected by natural selection, with the intensity of selection increasing with an increase of gene density. These correlations remained significant even after controlling for possible confounding variables (Table 3).

Lack of correlation between synonymous diversity and non-synonymous divergence

A distinctive signature of recurrent selective sweeps is the local reduction of linked neutral polymorphism due to frequent adaptive substitutions [51]. Given amino acid substitutions compose a substantial number of adaptive substitutions, negative correlation between neutral polymorphism and non-synonymous divergence can be particularly informative of the prevalence of selective sweeps [52]. However, in all three species, we found either no or weak negative correlations between neutral polymorphism ($\theta_{4\text{-fold}}$) and the rate of non-synonymous substitutions ($d_{0\text{-fold}}$) in both 100 Kbp and 1 Mbp windows (S4 Table). The correlational patterns did not change after we controlled for GC content, recombination rate, gene density, neutral divergence rate, and the number of 4-fold synonymous and 0-fold non-synonymous sites covered by the data (S4 Table). This result contrasts with our previous study reported from a small number of candidate genes, where we found a significant negative correlation between polymorphism at synonymous sites and amino acid divergence in *P. tremula* [53]. One possible explanation for the different patterns between these two studies is that they are based on different scales of measurement, from single genes to 100 Kbp and 1 Mbp windows [52]. Accordingly, additional future analyses are still needed to examine the relationship between the synonymous polymorphism and the rate of amino acid evolution on a genic scale.

The effect of recombination on the efficacy of natural selection

We next characterized the ratio of non-synonymous to synonymous polymorphism ($\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$) and divergence ($d_{0\text{-fold}}/d_{4\text{-fold}}$) for each of the three *Populus* species in order to assess whether there is a relationship between the efficacy of natural selection and the rate of recombination (Table 4). Once GC content, gene density and number of 4-fold synonymous and 0-fold non-synonymous sites were taken into account, we found no correlation between recombination rate and $d_{0\text{-fold}}/d_{4\text{-fold}}$ in all three species (Table 4). We did not observe any significant correlations between recombination rate and $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ in the 1 Mbp windows after controlling for the various confounding factors (Table 4). However, for 100 Kbp windows, we found significantly negative correlations between recombination rate and $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ in *P. tremula* and *P.*

tremuloides, but not in *P. trichocarpa*. The relative overabundance of non-synonymous polymorphism in regions of low recombination most likely suggests that the effective elimination of weakly deleterious non-synonymous mutations was reduced in low recombination regions in the two aspen species [12]. The lack of such correlation in *P. trichocarpa* may reflect its lower effective population size and accordingly weaker efficacy of selection across the genome [54]. In addition, since no correlation between $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ and recombination rate was observed at a broad scale (1 Mbp) in any of the three species, it is likely that interference between weakly selected mutations is more easier to be detected at fine scales [54], although this requires further investigation.

Table 4. Summary of the correlation coefficients (Spearman's ρ) between recombination rate 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	*P. tremula*	-0.057*	-0.075**	-0.012	-0.005
	P. tremuloides	-0.118**	-0.122**	-0.003	-0.002
	P. trichocarpa	-0.004	-0.002	-0.026	-0.020
1Mbp	*P. tremula*	-0.063	-0.045	-0.007	0.017
	P. tremuloides	-0.142*	-0.092	0.014	0.020
	P. trichocarpa	0.035	-0.002	0.030	0.036

[a] Partial 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

Discussion

We have characterized and compared genome-wide nucleotide polymorphism, site frequency spectra, linkage disequilibrium (LD), and population-scaled recombination rates among three related *Populus* species. Widespread variation in nucleotide diversity is found throughout the genomes of all three species and we found significant genome-wide correlations of diversity among the three *Populus* species. This likely results from shared selective constraints and/or patterns of conserved variation in mutation rate between these related species [55, 56]. Compared to *P. trichocarpa*, levels of diversity in *P. tremula* and *P. tremuloides* are more than two-fold higher throughout the genome. The higher diversity we find in both aspen species is likely due to their larger effective population sizes (N_e) because consistent patterns of interspecific sequence divergence between independent evolutionary lineages (S5 Fig) indicate that mutation rates are likely to be conserved among the three species [4]. Larger effective population sizes in the two aspen species are also in agreement with their larger current census population size and substantially more extensive geographic ranges [24]. Assuming that mutation rates do not differ dramatically among the three species, we could infer that the effective population size in the two aspen species are more than twice as large as in *P. trichocarpa* [4]. However, the relative importance of mutation rate variation in determining diversity levels across related species obviously deservs to be studies further, particularly in light of very recent results indicating that high levels of heterozygosity, as are observed in these species, can increase local and genome-wide mutation rates [57].

Compared to the consistent patterns of diversity among species, the much weaker correlations observed for allele frequency spectrum (Tajima's D) could either be ascribed to divergent selective targets to different environments since their divergence, or different demographic histories experienced by the three species during the Quaternary ice ages [39, 44, 58]. In particular, the genome-wide excess of rare frequency alleles in *P. tremuloides* is most likely explained by a recent substantial population expansion that was specific to this species. Many other factors, such as population structure, an unbalanced sampling scheme, and hybridization, can influence estimates of genomic variation and may therefore contribute to the different patterns observed among species [40, 41, 59], but we were able to exclude each of these. First, in accordance with mating characteristics of the genus where the seed and pollen are both wind-dispersed, we found little evidence of population structure in any of the three *Populus* species in this study. Second, despite a potentially biased sampling scheme in *P. tremuloides*, where the samples were collected from two geographically distinct populations, when we analyzed the genome-wide patterns of polymorphism separately we found the same patterns as those obtained when analyzing the populations jointly. Third, with regard to the influence of hybridization, it should be noted that there are no other species of *Populus* that occur naturally in the regions from where the *P. tremula* samples were collected [60]. For *P. tremuloides*, naturally occurring hybridization is only known to occur at very low levels with *P. grandidentata* [61]. These two species occur sympatrically in central and eastern North America, so in our study any possible hybridization in *P. tremuloides* would be limited to samples from the Wisconsin population. Although hybridization with other nearby *Populus* species is more frequent in *P. trichocarpa* [41], in this study we only used individuals of *P. trichocarpa* that have previously been shown having no evidence of admixture with other species [31].

The recombination rate and the extent of linkage disequilibrium (LD) are key factors influencing the feasibility and power of genome-wide association studies [62, 63]. The three *Populus* species exhibit different patterns in the decay of LD (r^2) with physical distance, with LD decaying fastest in *P. tremuloides* and slowest in *P. trichocarpa*. This reflects the rank order of their population-scaled recombination rate (ρ=4N_ec), for which *P. tremuloides* is the highest (8.42 Kbp^{-1}), followed by *P. tremula* (3.23 Kbp^{-1}), and *P. trichocarpa* is the lowest (2.19 Kbp^{-1}). It is important to note that differences in ρ among the species cannot simply reflect differences in species' N_e, because recombination rate correlations in 100 Kbp windows show that ρ is only partially (not highly) conserved among these species (Fig 2c) [64]. This suggests that even with conserved gene function and synteny, associations might be more easily discovered in one *Populus* species than another.

The genome-wide ratio of recombination to mutation rate (ρ/θ_W or c/μ) was similar between *P. tremuloides* (0.39) and *P. trichocarpa* (0.38), but substantially smaller in *P. tremula* (0.22). If mutation rate is indeed unchanged between species, the lower estimate of c/μ in *P. tremula* indicates a considerably smaller recombination rate relative to the other species. Nevertheless, these c/μ estimates are of the same order of magnitude as recent genome-wide estimates of other plant species, such as *Medicago truncatula* (0.29) [15], *Mimulus guttatus* (0.8) [65] and the tree *Eucalyptus grandis* (0.65) [66]. However, the discrepant results obtained from patterns of polymorphism and recombination between *P. tremula* and *P. tremuloides* are likely due to differences in the effective population sizes influencing patterns of nucleotide diversity and linkage disequilibrium [67]. These processes operate over different time-scales and are therefore subject to temporal variation in the effective population size [67, 68]. The recent population size expansion that we infer to have taken place in *P. tremuloides* can thus also explain why its recombination rate is higher than *P. tremula*, even if they share similar levels of genome-wide polymorphism.

In addition to the historical patterns of mutation, recombination and demographic processes, patterns of genomic variation also contain much information about natural selection [54]. In all three species, as expected we find 0-fold non-synonymous sites exhibit significantly lower levels of polymorphism and divergence compared to 4-fold synonymous sites. The 0-fold non-synonymous sites are likely experiencing strong selective constraint, consistent with their excess of ultra-rare variants as indicated by Tajima's D [69]. In addition, introns and 5' UTR sites are also likely to be under some degree of selective constraint, although much weaker than non-synonymous sites. The 3' UTR sites seem to be either neutral or under comparable extent of selective constraint as 4-fold synonymous sites [70]. In contrast to all genic categories, we find there are substantially higher levels of polymorphism and divergence in intergenic regions throughout the genome, reflecting either higher mutation rates, relaxed selective constraint or both in these regions [2].

Apart from strong selective constraints on protein-coding genes, multiple lines of evidence indicate that genomic patterns of polymorphism have been primarily shaped by widespread natural selection in all three *Populus* species. First, we find significantly positive correlations between neutral polymorphism and population-scaled recombination rate in both genic and intergenic regions, even after controlling for the confounding variables such as GC content, gene density, mutation rate and number of covered sites by the data. Such patterns could be explained by both background selection and recurrent selective sweeps, where perturbations of linked selection on neutral genetic variation are more drastic and extensive in regions of low recombination compared to high recombination regions [8, 9, 71]. An alternative explanation to natural selection would be that recombination itself has a mutagenic effect [47]. In this case, the neutral theory predicts that we would also detect a correlation between nucleotide divergence and recombination rate [10, 47] but this relationship was not observed for any of the three species. Thus our findings support the notion that ubiquitous linked selection, as selective sweeps of adaptive alleles and/or background selection against deleterious alleles, is the dominant force shaping the observed associations between recombination and neutral polymorphism in all three species [72]. In addition, the extent of such associations can also reflect the magnitude of the impact of linked selection on genomes [54, 73]. Here, we tried to decipher the factors that may contribute to inconsistent signatures and magnitudes of linked selection across the three species. First of all, the genome-wide effects of linked selection ought to be influenced by effective population size (N_e) across species, where the impact of selection at linked sites should be more severe in larger populations [73, 74]. As a result, the substantially stronger signatures of linked selection in *P. tremula* and *P. tremuloides* are most likely due to their larger N_e compared to *P. trichocarpa*. Furthermore, just as the impact of natural selection at linked sites depends on the local environment of recombination, we expect that the disparate patterns of linked selection among species is also likely to be caused by the various recombination rates across genomes [54, 71]. In particular, compared with *P. tremuloides*, the stronger signature of linked selection in *P. tremula* is supposed to be primarily driven by its lower average levels of recombination across the genome. More broadly, the different magnitude of linked selection may provide one of the major explanations for the disparate patterns of genomic variation across related species [73].

In addition to the association between recombination and neutral polymorphism, we find slightly negative correlations between recombination rate and the ratio of non-synonymous- to synonymous- polymorphism, but not divergence, in

P. tremula and *P. tremuloides* after controlling for the confounding variables. This pattern indicates a potential reduced efficacy of purifying selection at eliminating weakly deleterious non-synonymous mutations in low recombination regions [7, 48]. As a consequence, such Hill-Robertson interference (HRI) may help to understand patterns of partially positive correlations between gene density and recombination rate among all three species [13]. Given the relaxed efficacy of purifying selection in regions of low recombination where weakly deleterious mutations are more likely to accumulate at a high rate, important functional elements should thus not cluster in these regions, as has already been shown in several other plant species [14, 15, 75] Consistent with this prediction [76], we find positive association between gene density and recombination rate in regions that experience low rates of recombination. In high-recombination regions where selection is more effective at eliminating slightly deleterious mutations, the association between gene density and recombination become much weaker in all three species. However, it remains unclear whether it is the effects of recombination gradients that drive the functional organization of genomes in response to selection, or it is the gradients of functional genomic elements that modify the evolution of recombination rates in *Populus*.

By examining the relationship of neutral polymorphism, recombination rate and gene density, we find that levels of neutral polymorphism in genic regions are primarily dominated by local rates of recombination, regardless of the density of functional genes nearby. This suggests that widespread selection might have uniformly shaped the patterns of neutral polymorphism in genic regions across the genome, with variation of genetic diversity primarily relying on the variation of local recombination rates [7, 49, 71]. However, there is a more complex pattern in intergenic regions where levels of intergenic polymorphism are mainly dominated by recombination rates in regions of lower gene density, while in regions of higher gene density, levels of intergenic diversity are primarily shaped by the density of genes nearby. Patterns of polymorphism vary, on both quantitative and qualitative scales, between genic and intergenic sequences, with the latter exhibiting substantially higher diversity, divergence and more non-uniformly distributed selective effects compared with the former [54]. In addition, 84.2% of the intergenic sites included in this study are located within 5-Kbp upstream/downstream regions of functional genes. This suggests that many of these intergenic regions may have important functions in gene regulation, in accordance with the widespread signatures of linked selection as we found in these regions [77]. In these cases, we could argue that the differences in neutral mutation rate alone are not sufficient to explain the distinct patterns of genetic variation between genic and intergenic sites. Various rates, distributions, and selective coefficients for either adaptive or deleterious mutations, however, may at least in part drive the distinct patterns of polymorphism and divergence between these different genomic environments.

In conclusion, we have examined and compared the relative roles of mutation, population history, recombination and natural selection in forging the landscape heterogeneity of genomic variation within and among three related *Populus* species. We find substantially different magnitudes and signatures of linked selection among species, with selection effects being strongest in *P. tremula* and weakest in *P. trichocarpa*. Various effective population sizes and genome-wide recombination rates are likely to be the primary factors causing the disparate genome-wide signatures of linked selection among species. By analyzing the ratio of non-synonymous- to synonymous- polymorphism along recombination gradients, we find that purifying selection at purging slightly deleterious non-synonymous mutations is more effective in regions experiencing high recombination. Such selective interaction between recombination and selection may provide one of the explanations for the co-varying patterns of gene density and recombination in the *Populus* species, where functional genes are more likely to cluster in high-recombination regions. Finally, we find distinct genomic signatures of selection between genic and intergenic regions. The recombination rate-dependent effect of selection dominates levels of polymorphism at genic sites, while patterns of linked selection at intergenic sites are shaped by interactions between recombination and local gene density. Thus, our study provides a promising avenue to dissect how interactions of various evolutionary forces are driving the evolution of genomes for even closely related species.

Materials and Methods

Samples and sequencing

Leaf samples were separately collected from 24 genotypes of *P. tremula* and 24 genotypes of *P. tremuloides* (S1 Table). 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, and 2×100-bp paired-end reads were generated for all genotypes. As two samples of *P. tremuloides* failed to obtain the expected coverage, all analyses are based on data from 24 *P. tremula* genotypes and 22 *P. tremuloides* genotypes. All newly generated Illumina reads from this study have been submitted to the Short Read Archive at NCBI under accession IDs ranging from XXXXXX-XXXXXX. We obtained publicly available short read Illumina data for 24 *P. trichocarpa* individuals from NCBI SRA (S1 Table).

Individuals were selected to have a similar read depth as the samples of the two aspen species. The accession numbers of *P. trichcoarpa* samples can be found in [31]. These data are paired-end 100bp reads generated on the Illumina HiSeq2000 platform.

Raw read filtering, read alignment and post-processing alignment

Prior to read alignment, we used Trimmomatic [78] to remove adapter sequences from reads. Since the quality of reads always drops towards the end of reads, we used Trimmomatic to cut bases off the start and end of each read when the quality values dropped below 20. If the length of the processed reads was reduced to below 36 bases after trimming, reads were completely discarded. FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to check and compare the per base sequence quality of the raw sequence data and the filtered data. After quality control, all paired-end and orphaned single-end reads of each sample were mapped to the *P. trichocarpa* version 3 (v3.0) genome [29] using the BWA-MEM algorithm with default parameters in bwa-0.7.10 [32].

Several post-processing steps of alignments were performed in order to minimize the number of artifacts in downstream analysis: First, indel realignment was performed as sequence reads are often mapped with mismatching bases in regions with insertions and deletions (indels). The RealignerTargetCreator in GATK (The Genome Analysis Toolkit) [79] was first used to find suspicious-looking intervals which 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 or high read depths. After investigating the empirical distribution of read coverage, we filtered out sites with a total coverage less than 100X or greater than 1200X 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 for each species, we removed those sites if there were more than 20 mapped reads with mapping score equal to zero across all individuals. 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 [80]. In the end, the subset of sites that passed all these filtering criteria in the three *Populus* species were used in all following analyses.

SNP and genotype calling

We implemented two complementary bioinformatics approaches in downstream analyses:

(i) Population genetic inferences that rely on the site frequency spectrum (SFS). Recently, many studies pointed out the bias introduced in population genetic estimates by inaccurate genotype calls from NGS data [35, 81]. Either the single-sample genotype calling (calling genotypes for each individual separately and then merging them later) or the multi-sample genotype calling (jointly calling genotypes for all individuals) can result in a bias in the estimation of SFS, as the former method usually leads to overestimation of rare variants, whereas the latter often leads to the opposite [35]. Therefore, all the population genetic statistics that based on the SFS in this study were estimated directly and jointly from filtered sites and individuals without calling genotypes, as implemented in the software package Analysis of Next-Generation Sequencing Data (ANGSD v0.602) [82].

(ii) Analyses based on accurate SNP and genotype calls. We performed SNP calling with HaplotypeCaller of the GATK v3.2.2 [79], which called SNPs and indels simultaneously via local re-assembly of haplotypes for each individual and created single-sample gVCFs. GenotypeGVCFs in GATK was then used to merge multi-sample records together, correct genotype likelihoods, and re-genotype the newly merged record and perform re-annotation. Several filtering steps were then used to reduce the number of false positive SNPs and retain high-quality SNPs: (1) we removed all SNPs that were overlapped with sites excluded by the previous filtering criteria. (2) only biallelic SNPs with a distance of more than 5bp away from indels were retained for further analysis. (3) Genotypes were accepted for each SNP and each individual only if the genotype quality score (GQ) was ≥10, otherwise that specific genotype was treated as missing data. (4) SNPs with missing rate higher than 20% were removed from downstream analysis. (5) SNPs that showed significant deviation from Hardy-Weinberg Equilibrium (P<0.001) were removed from further downstream analysis.

Population structure

We used only 4-fold synonymous SNPs with minor allele frequency >0.1 to perform population structure analyses with ADMIXTURE [34]. We ran ADMIXTURE separately on all the sampled individuals among species and on the samples within each species, varying the number of genetic clusters *K* from 1 to 6. The most likely number of genetic cluster was selected by minimizing the cross-validation error in ADMIXTURE.

Diversity and divergence - related summary statistics

For nucleotide diversity and divergence estimates, only the reads with mapping quality above 30 and the bases with quality score higher than 20 were used in all of the following analyses with ANGSD [82] and ngsTools [83]. To infer the global SFS, we firstly used the -doSaf implementation in ANGSD to calculate the site allele frequency likelihood based on the SAMTools genotype likelihood model [84]. Then, we used the –realSFS implementation in ANGSD to obtain an optimized folded global SFS using Expectation Maximization (EM) algorithm for each species. Based on the global SFS, we used the –doThetas function in ANGSD to estimate the per-site nucleotide diversity from posterior probability of allele frequency based on a maximum likelihood approach [36]. Two standard estimates of nucleotide diversity, the average pairwise nucleotide diversity (Θ_π) [38] and the proportion of segregating sites (Θ_W) [85], and one neutrality statistic test Tajima's D [38] were then summarized along all 19 chromosomes using non-overlapping sliding windows of 100 Kbp and 1 Mbp. Windows with less than 10% covered sites left after previous quality filtering steps were excluded. Accordingly, 3340 100-Kbp and 343 1-Mbp windows, with an average of 50,538 and 455,910 covered bases per window, were respectively included for downstream analyses. Based on posterior probabilities of sample allele frequencies at each site, we further used the ngsTools [83] to calculate pairwise nucleotide divergence, d_{xy}, between pairs of species over all non-overlapping 100-Kbp and 1-Mbp windows.

All these statistics were also calculated for each type of functional element (0-fold non-synonymous, 4-fold synonymous, intron, 3’ UTRs, 5’ UTR, and intergenic sites) over the non-overlapping 100-Kbp and 1-Mbp windows in all three *Populus* species. The category of gene models we used followed the gene annotation of *P. trichocarpa* version 3.0 [29]. For protein-coding genes, we only included genes with at least 90% covered sites left from previous filtering steps to ensure that the three species have same gene structures. We also excluded genes overlapping with other genes. For the remaining genes, we selected the transcript with the highest content of protein-coding sites. For regions overlapped by different transcripts in each gene, we classified each site according to the following hierarchy (from highest to lowest): Coding regions (CDS), 3’UTR, 5’UTR, Intron. Thus, if a site resides in a 3’UTR in one transcript and CDS for another, the site was classified as CDS. In the end, a respective of 16.52, 3.4, 7.19, 4.02, 31.89, 73.46 megabases (Mbp) were partitioned into 0-fold non-synonymous (where all DNA sequence changes lead to protein sequence changes), 4-fold synonymous (where all DNA sequence changes lead to the same protein sequences), 3’UTR, 5’UTR, intron, and intergenic categories. Windows were not used if there were less than 100 sites left for any of the functional elements.

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

A total of 1,409,377 SNPs, 1,263,661 SNPs and 710,332 SNPs with minor allele frequency higher than 10% were used for the analysis of LD and ρ in *P. tremula*, *P. tremuloides* and *P. trichocarpa*, respectively. To estimate and compare the rate of LD decay in the three *Populus* species, we firstly used PLINK 1.9 [86] to randomly thin the number of SNPs to 100,000 in each species. Then we calculated the squared correlation coefficients (r^2) between all pairs of SNPs within 50 Kbp windows using PLINK 1.9. The decay of LD against physical distance was estimated using nonlinear regression of pairwise r^2 vs. the physical distance between sites in base pairs [87]. Furthermore, we estimated the population-scaled recombination rate ρ using the Interval program of LDhat 2.2 [88] with 1,000,000 MCMC iterations sampling every 2,000 iterations and a block penalty parameter of five. The first 100,000 iterations of the MCMC iterations were discarded as a burn-in. We then calculated the scaled value of ρ in each 100-Kbp and 1-Mbp window as the average across SNPs in that window. In order to evaluate the extent of correlation between the estimated ρ and the pattern of LD, we also calculated the scaled r^2 by averaging r^2 over all pairwise SNPs in each 100 Kbp and 1 Mbp window. Only windows with more than 10,000 (in 100 Kbp windows) and 100,000 bases (in 1 Mbp windows) and 100 SNPs left after previous filtering steps were used for the estimation of ρ and r^2.

Genomic correlates of diversity

Within each non-overlapping 100 Kbp or 1 Mbp window, levels of neutral polymorphism in genic and intergenic regions were tabulated as the pairwise nucleotide diversity (Θ_π) at 4-fold synonymous and intergenic sites respectively. In order to examine the factors influencing levels of neutral polymorphism in all three *Populus* species, we further tabulated several genomic features within each window. First, we summarized population-scaled recombination rate (ρ) as described above for each species. Second, we tabulated GC content as the fraction of bases where the reference sequence (*P. trichocarpa* v3.0) was a G or a C. Third, we measured the gene density as the number of functional genes within each window according to the gene annotation of *P. trichocarpa* version 3.0. Fourth, we accounted for the variation of mutation rate by calculating the number of fixed differences between aspen and *P. trichocarpa* per neutral site (either 4-fold synonymous site or intergenic site) within each window. The reason why we used divergence between aspen and *P. trichocarpa* to measure mutation rate is because they are distantly related [26], and thus the estimate of divergence are unlikely to be influenced by shared ancestral polymorphisms between species. Fifth, we tabulated the number of covered bases in each window as those met the filtering criteria described above.

We used Spearman's rank-order correlation tests to examine pairwise correlations between the variables as described above. In order to account for the autocorrelation between many of these variables, we calculated partial correlations between the interested variables [89], which simultaneously remove the confounding effects of other variables. All statistical tests were performed using R version 3.2.0 unless stated otherwise.

Reference

1. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, et al. The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. 2005;3:1289.

2. Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, et al. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol. 2007;5:e310.

3. Hellmann I, Prüfer K, Ji H, Zody MC, Pääbo S, Ptak SE. Why do human diversity levels vary at a megabase scale? Genome Res. 2005;15:1222-1231.

4. Kimura M. The neutral theory of molecular evolution: Cambridge University Press; 1984.

5. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493-496.

6. Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. The *Drosophila melanogaster* genetic reference panel. Nature. 2012;482:173-178.

7. Cutter AD, Choi JY. Natural selection shapes nucleotide polymorphism across the genome of the nematode *Caenorhabditis briggsae*. Genome Res. 2010;20:1103-1111.

8. Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature. 1992;356:519 - 520.

9. McGaugh SE, Heil CS, Manzano-Winkler B, Loewe L, Goldstein S, Himmel TL, et al. Recombination modulates how selection affects linked sites in *Drosophila*. PLoS Biol. 2012;10:e1001422.

10. Kulathinal RJ, Bennett SM, Fitzpatrick CL, Noor MA. Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. Proc Natl Acad Sci U S A. 2008;105:10051-10056.

11. Campos JL, Halligan DL, Haddrill PR, Charlesworth B. The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. Mol Biol Evol. 2014;31:1010-1028.

12. Charlesworth B, Campos JL. The relations between recombination rate and patterns of molecular variation and evolution in *Drosophila*. Annu Rev Genet. 2014;48:383-403.

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

14. Flowers JM, Molina J, Rubinstein S, Huang P, Schaal BA, Purugganan MD. Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. Mol Biol Evol. 2012;29:675-687.

15. Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J, et al. Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. Proc Natl Acad Sci U S A. 2011;108:E864-E870.

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

17. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol. 2014;29:51-63.

18. Lawrie DS, Petrov DA. Comparative population genomics: power and principles for the inference of functionality. Trends Genet. 2014;30:133-139.

19. Hufford MB, Xu X, Van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, et al. Comparative population genomics of maize domestication and improvement. Nature Genet. 2012;44:808-811.

20. Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, et al. Comparative and demographic analysis of orang-utan genomes. Nature. 2011;469:529-533.

21. Liu S, Lorenzen ED, Fumagalli M, Li B, Harris K, Xiong Z, et al. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell. 2014;157:785-794.

22. Neale DB, Kremer A. Forest tree genomics: growing resources and applications. Nat Rev Genet. 2011;12:111-122.

23. González‐Martínez SC, Krutovsky KV, Neale DB. Forest‐tree population genomics and adaptive evolution. New Phytol. 2006;170:227-238.

24. Eckenwalder JE. Systematics and evolution of *Populus*. In: Stettler RF, Bradshaw HD, Heilman PE, Hinckley TM, editors. Biology of *Populus* and its Implications for Management and Conservation. Ottawa: NRC Research Press; 1996.pp.7-32.

25. Jansson S, Douglas CJ. *Populus*: a model system for plant biology. Annu Rev Plant Biol. 2007;58:435-458.

26. Wang Z, Du S, Dayanandan S, Wang D, Zeng Y, Zhang J. Phylogeny Reconstruction and Hybrid Analysis of *Populus* (Salicaceae) Based on Nucleotide Sequences of Multiple Single-Copy Nuclear Genes and Plastid Fragments. Plos One. 2014;9:e103645.

27. Jansson S, Bhalerao RP, Groover AT. Genetics and genomics of *Populus*: Springer; 2010.

28. Dickmann DI, Kuzovkina J. Poplars and willows of the world, with emphasis on silviculturally important species. In: Isebrands JG, Richardson J, editors. Poplars and Willows: trees for society and the environment. Rome: The Food and Agriculture Organization of the United Nations (FAO) and CAB International (CABI); 2014:8-91.

29. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science. 2006;313:1596-604.

30. Hamzeh M, Dayanandan S. Phylogeny of *Populus* (Salicaceae) based on nucleotide sequences of chloroplast trnT-trnF region and nuclear rDNA. Am J Bot. 2004;91:1398-408.

31. Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, et al. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. Nature Genet. 2014;46:1089–1096.

32. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM; 2013. Preprint. Available: arXiv:13033997.

33. Hall D, Luquez V, Garcia VM, St Onge KR, Jansson S, Ingvarsson PK. Adaptive population differentiation in phenology across a latitudinal gradient in European aspen (*Populus tremula*, L.): a comparison of neutral markers, candidate genes and phenotypic traits. Evolution. 2007;61:2849-2860.

34. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655-1664.

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

36. Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, Tian G, et al. Estimation of allele frequency and association mapping using next-generation sequencing data. BMC bioinformatics. 2011;12:231.

37. Ismail M, Soolanayakanahally RY, Ingvarsson PK, Guy RD, Jansson S, Silim SN, et al. Comparative nucleotide diversity across North American and European *Populus* species. J Mol Evol. 2012;74:257-272.

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

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

40. Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. Genetics. 2010;186:983-995.

41. Huang DI, Hefer CA, Kolosova N, Douglas CJ, Cronk QC. Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, *Populus balsamifera* and *P. trichocarpa* (Salicaceae). New Phytol. 2014;204:693-703.

42. Wang J, Scofield D, Street NR, Ingvarsson PK. Variant calling using NGS data in European aspen (*Populus tremula*). In: Sablok G, Kumar S, Ueno S, Kuo J, Varotto C, editors. Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches. Springer; 2015. pp.43-61.

43. Begun DJ. Population genetics of silent and replacement variation in *Drosophila simulans* and *D. melanogaster*: X/autosome differences? Mol Biol Evol. 1996;13:1405-1407.

44. Ingvarsson PK. Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. Genetics. 2008;180:329-340.

45. Slavov GT, DiFazio SP, Martin J, Schackwitz W, Muchero W, Rodgers - Melnick E, et al. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. New Phytol. 2012;196:713-725.

46. Comeron JM, Kreitman M. The correlation between synonymous and nonsynonymous substitutions in *Drosophila*: mutation, selection or relaxed constraints? Genetics. 1998;150:767-775.

47. Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M. A neutral explanation for the correlation of diversity with recombination rates in humans. Am J Hum Genet. 2003;72:1527-1735.

48. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genetical Res. 1966;8:269-294.

49. Cutter AD, Payseur BA. Selection at linked sites in the partial selfer *Caenorhabditis elegans*. Mol Biol Evol. 2003;20:665-673.

50. Payseur BA, Nachman MW. Gene density and human nucleotide polymorphism. Mol Biol Evol. 2002;19:336-340.

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

52. Macpherson JM, Sella G, Davis JC, Petrov DA. Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. Genetics. 2007;177:2083-2099.

53. Ingvarsson PK. Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*. Mol Biol Evol. 2010;27:650-660.

54. Cutter AD, Payseur BA. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat Rev Genet. 2013;14:262-274.

55. Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. Genetics. 1987;116:153-159.

56. Charlesworth B, Morgan M, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134:1289-1303.

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

58. Callahan CM, Rowe CA, Ryel RJ, Shaw JD, Madritch MD, Mock KE. Continental - scale assessment of genetic diversity and population structure in quaking aspen (*Populus tremuloides*). J Biogeogr. 2013;40:1780-1791.

59. Städler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. Genetics. 2009;182:205-216.

60. Lexer C, Fay M, Joseph J, Nica MS, Heinze B. Barrier to gene flow between two ecologically divergent *Populus* species, *P. alba* (white poplar) and *P. tremula* (European aspen): the role of ecology and life history in gene introgression. Mol Ecol. 2005;14:1045-1057.

61. Pregitzer KS, Barnes BV. Flowering phenology of *Populus tremuloides* and *P. grandidentata* and the potential for hybridization. Can J Forest Res. 1980;10:218-223.

62. Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, et al. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nature Genet. 2007;39:1151-1155.

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

64. Smukowski C, Noor M. Recombination rate variation in closely related species. Heredity. 2011;107:496-508.

65. Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, Wessler SR, et al. Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. Proc Natl Acad Sci U S A. 2013;110:19478-19482.

66. Silva ‐ Junior OB, Grattapaglia D. Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of *Eucalyptus grandis*. New Phytol. 2015; doi: 10.1111/nph.13505.

67. Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, et al. Recent human effective population size estimated from linkage disequilibrium. Genome Res. 2007;17:520-526.

68. Cutter AD, Jovelin R, Dey A. Molecular hyperdiversity and evolution in very large populations. Mol Ecol. 2013;22:2074-2095.

69. Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, et al. Evolution of protein-coding genes in *Drosophila*. Trends Genet. 2008;24:114-123.

70. Andolfatto P. Adaptive evolution of non-coding DNA in *Drosophila*. Nature. 2005;437:1149-1152.

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

72. Hahn MW. Toward a selection theory of molecular evolution. Evolution. 2008;62:255-265.

73. Corbett-Detig RB, Hartl DL, Sackton TB. Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 2015;13:e1002112.

74. Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, et al. Revisiting an old riddle: what determines genetic diversity levels within species. PLoS Biol. 2012;10:e1001388.

75. Anderson LK, Lai A, Stack SM, Rizzon C, Gaut BS. Uneven distribution of expressed sequence tag loci on maize pachytene chromosomes. Genome Res. 2006;16:115-122.

76. Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. Genome Biol. 2007;8:R18.

77. Haygood R, Fedrigo O, Hanson B, Yokoyama K-D, Wray GA. Promoter regions of many neural-and nutrition-related genes have experienced positive selection during human evolution. Nature Genet. 2007;39:1140-1144.

78. Lohse M, Bolger A, Nagel A, Fernie AR, Lunn JE, Stitt M, et al. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. 2012;40:W622-W627.

79. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genet. 2011;43:491-498.

80. Tarailo - Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc in Bioinformatics. 2009; 25:4.10.1-4.10.14.

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

82. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC bioinformatics. 2014;15:356.

83. Fumagalli M, Vieira FG, Linderoth T, Nielsen R. ngsTools: methods for population genetics analyses from next-generation sequencing data. Bioinformatics. 2014;30:1486-1487.

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

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

86. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559-575.

87. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, et al. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci U S A. 2001;98:11479-11484.

88. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004;304:581-584.

89. Kim S-H, Soojin VY. Understanding relationship between sequence and functional evolution in yeast proteins. Genetica. 2007;131:151-156.

Supporting Information

S1 Fig. Sampling localities (details in S1 Table, black star symbols) and distributions of *P. tremula* (orange areas), *P. tremuloides* (blue areas) and *P. trichocarpa* (green areas).

S2 Fig. Comparison of per-base sequence quality between raw and filtered sequence data. Per-base sequence quality comparison between raw paired-end sequence data (forward reads: top left and reverse reads: top right), and filtered sequence data with both forward (bottom left) and reverse (bottom middle) reads left or only single-end (bottom right) reads left. The x-axis of the BoxWhisker plot shows the position in read, and the y-axis shows the quality scores. The higher the score the better the base call. The background of the plot divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The central red line is the median quality value, the yellow box represents the inter-quartile of quality, the upper and lower whiskers represent the 10% and 90% points, the blue line represents the mean quality. (a) Sample SwAsp009 of *Populus tremula*. (b) Sample Alb16-1 of *P. tremuloides*. (c) Sample GW-9772 (accession number in SRA: SRR1571518) of *P. trichocarpa*.

S3 Fig. Population structure within and between species. (a) Genetic structure of three *Populus* inferred using ADMIXTURE when it identifies three genetic clusters in the dataset. (b) The cross-validation error when *K* varies from 1 to 6 across the three species. (c,d,e) The cross-validation error when *K* varies from 1 to 6 separately in samples of *P. tremula*, *P. tremuloides*, *P. trichocarpa*.

S4 Fig. Genome-wide patterns of divergence among three *Populus* species. Mean pairwise divergence (d_{xy}) between pairs of three *Populus* species was calculated over 100 Kbp non-overlapping windows along the 19 chromosomes. *P. tremula-P.tremuloies*: red, *P. tremula-P. trichocarpa*: light purple, *P. tremuloides-P.trichocarpa*: purple.

S5 Fig. Correlations of divergence between independent pairs of the three *Populus* species. Spearman's correlations of pairwise nucleotide divergence (d_{xy}) between $d_{xy(P.\ tremula\text{-}P.\ trichocarpa)}$ and $d_{xy(P.\ tremulodies\text{-}P.\ trichocarpa)}$ (a); between $d_{xy(P.\ tremula\text{-}P.\ tremuloides)}$ and $d_{xy(P.\ tremula\text{-}P.\ trichocarpa)}$ (b); and between $d_{xy(P.\ tremula\text{-}P.\ tremuloides)}$ and $d_{xy(P.\ tremuloides\text{-}P.\ trichocarpa)}$ (c). All datasets are based on 100 Kbp non-overlapping windows across the genome.

S6 Fig. The distributions of estimates of (a) pairwise sequence diversity (Θ_Π), (b) the number of segregating sites (Θ_W), (c) Tajima's D and (d) population-scaled recombination rate (ρ) over 1Mbp non-overlapping windows in *P. tremula* (orange), *P. tremuloides* (blue) and *P. trichocarpa* (green).

S7 Fig. Estimates of (a) pairwise sequence diversity (Θ_Π) and (b) Tajima's D in samples of Alberta (light blue), Wisconsin (light green) and all samples of *P. tremuloides* (blue) over 100 Kbp non-overlapping windows.

S8 Fig. Number of singletons in samples of (a) *P. tremula*, (b) *P. tremuloides*, and (c) *P. trichocarpa*.

S9 Fig. The distributions of estimates of pairwise sequence diversity (Θ_Π) in *P. tremula* (orange), *P. tremuloides* (blue) and *P. trichocarpa* (green) over 1 Mbp non-overlapping windows in different site categories.

S10 Fig. The distributions of estimates of Tajima's D in *P. tremula* (orange), *P. tremuloides* (blue) and *P. trichocarpa* (green) over 1 Mbp non-overlapping windows in different site categories.

S11 Fig. The distributions of estimates of nucleotide divergence (d_{xy}) between pairs of the three *Populus* species over 1 Mbp non-overlapping windows in different site categories.

S12 Fig. Relationship between population-scaled recombination rate and linkage disequilibrium. Scatter plots display correlations between population-scaled population rates (ρ) and linkage disequilibrium (r^2) over 100 Kbp non-overlapping windows in (a) *P. tremula*, (b) *P. tremuloides*, and (c) *P. trichocarpa*. The red to yellow to blue gradient indicates decreased density of observed events at a give location in the graph.

S13 Fig. Distributions of the ratio of population-scaled recombination rate to nucleotide diversity (ρ/θ_W) over 100 Kbp non-overlapping windows in *P. tremula* (orange line), *P. tremuloides* (blue line) and *P. trichocarpa* (green line).

S14 Fig. Correlations between estimates of intergenic genetic diversity ($\Theta_{Intergenic}$) (left panel) and divergence ($d_{Intergenic}$) (right panel) with population-scaled recombination rate (ρ) over 1 Mbp non-overlapping windows. Linear regression lines are colored according to species: (a) *P. tremula* (orange line), (b) *P. tremuloides* (blue line) and (c) *P. trichocarpa* (green line).

S15 Fig. Relationship between gene number and the proportion of coding bases within 1 Mbp non-overlapping windows.

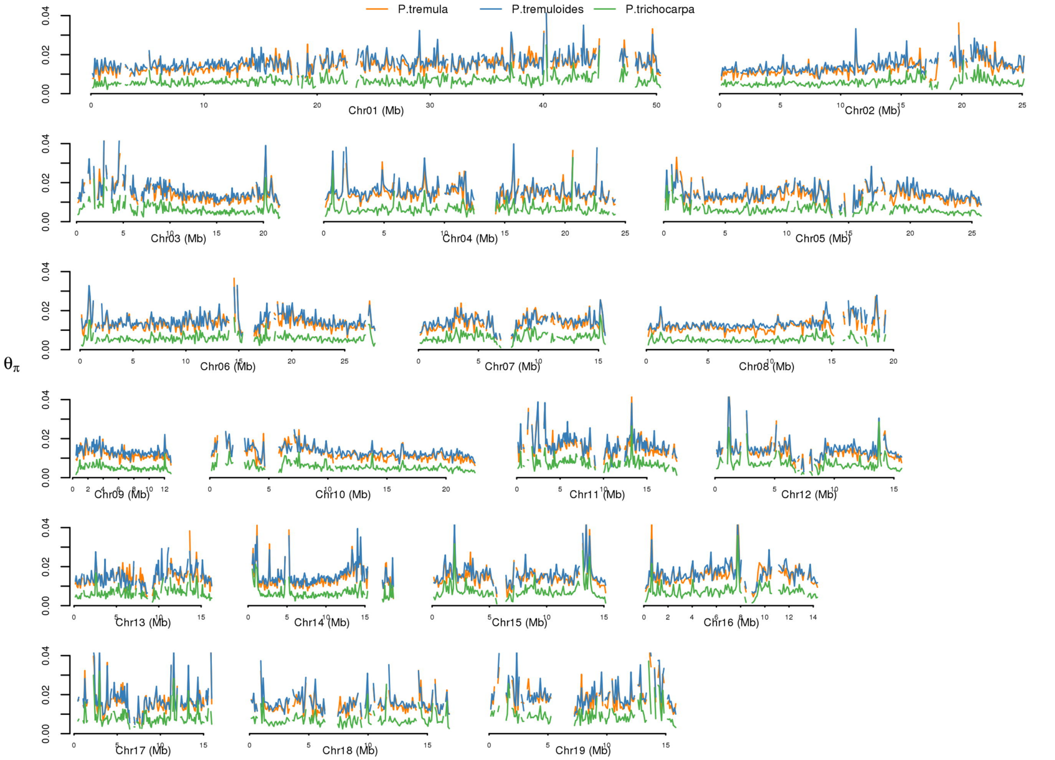
S16 Fig. Correlations between estimates of genic and intergenic genetic divergence with gene density over 1 Mbp non-overlapping windows. Correlations between estimates of genetic divergence at 4-fold synonymous sites ($d_{4\text{-fold}}$) (left panel) and intergenic sites ($d_{Intergenic}$) (right panel) with gene density over 1Mbp non-overlapping windows. Linear regression lines are colored according to species: (a) *P. tremula* (orange line), (b) *P. tremuloides* (blue line) and (c) *P. trichocarpa* (green line).

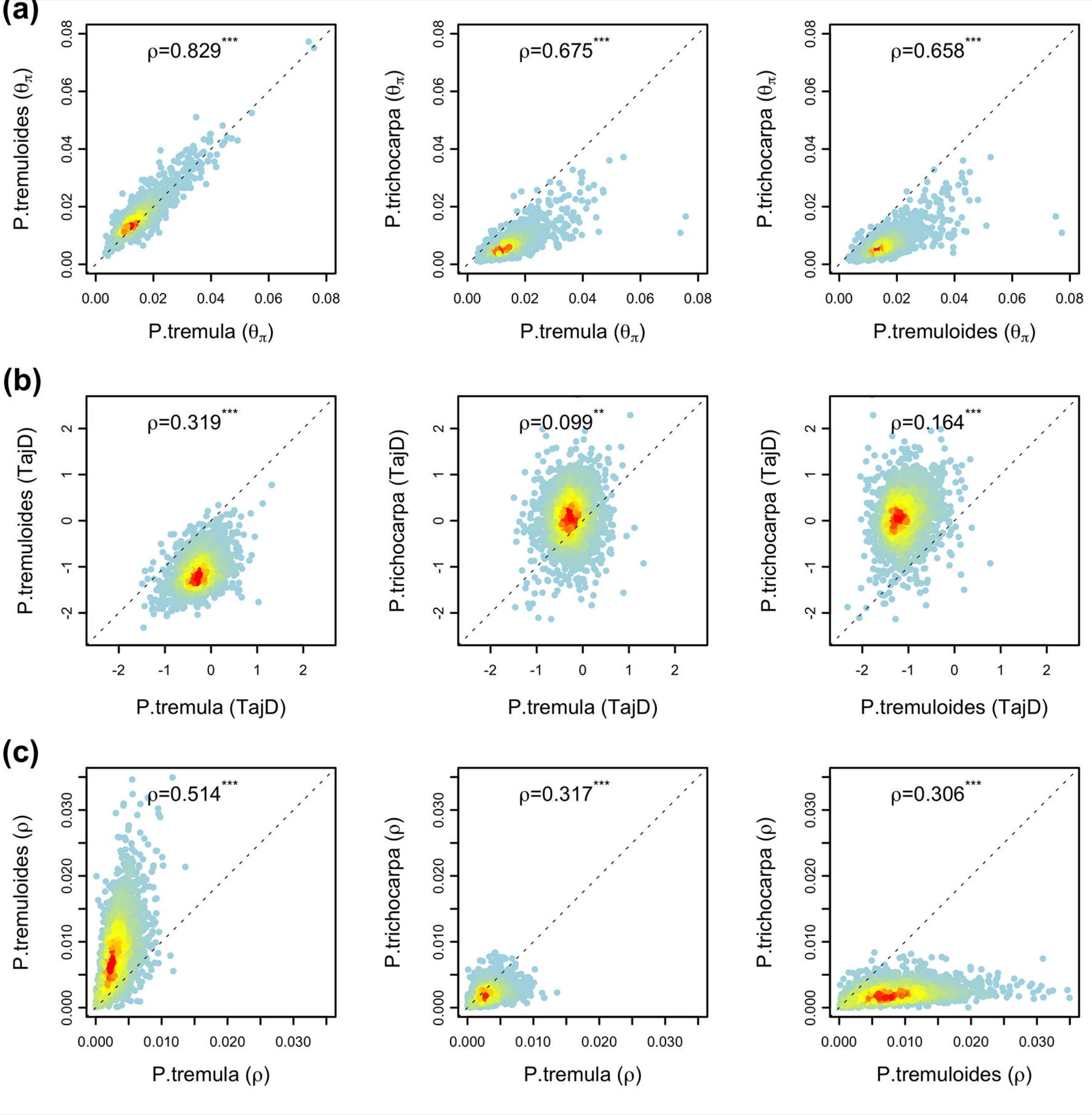
S1 Table. Samples used in this study.

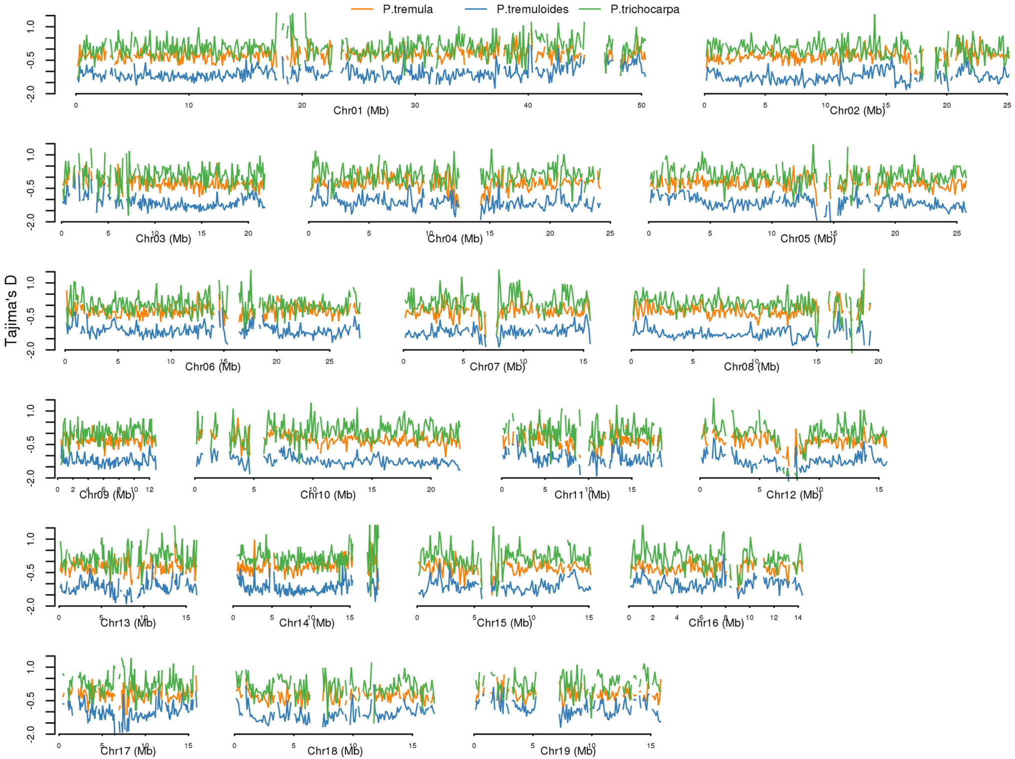
S2 Table. Summary statistics of Illumina re-sequencing data per sample.

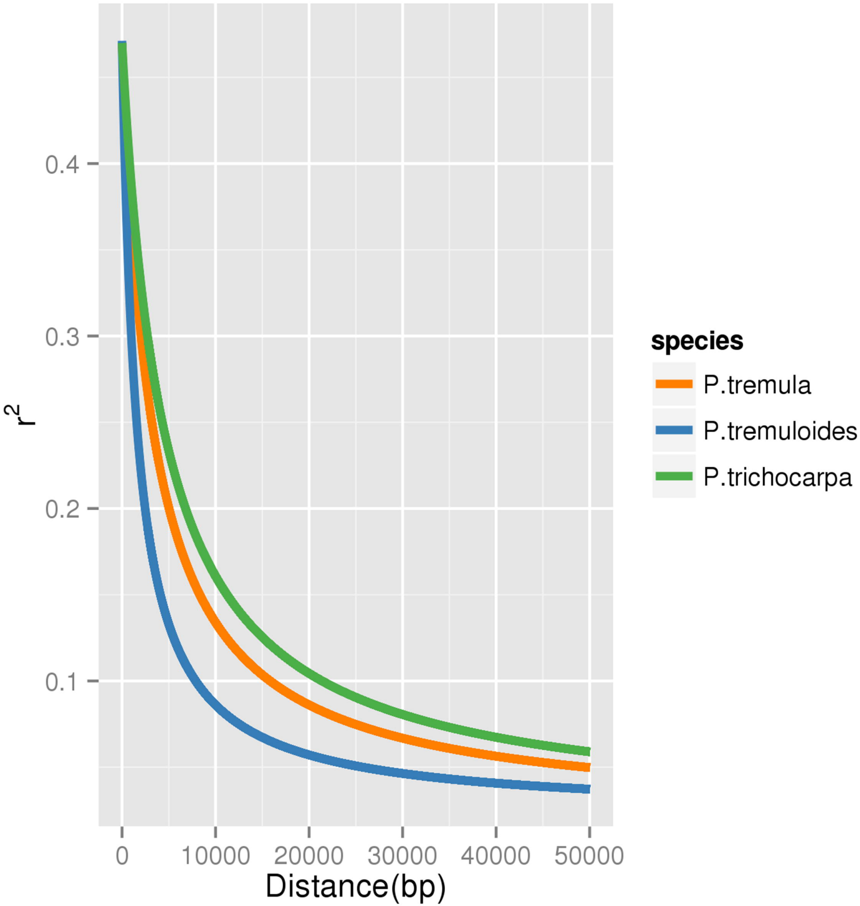
S3 Table. Pairwise divergence (d_{xy}) (median and central 95% range) between *P. tremula, P. tremuloides* and P. *trichocarpa* for various genomic contexts over 100 Kbp non-overlapping windows across genomes.

S4 Table. Summary of the correlation coefficients (Spearman's ρ) between levels of synonymous diversity and non-synonymous divergence.

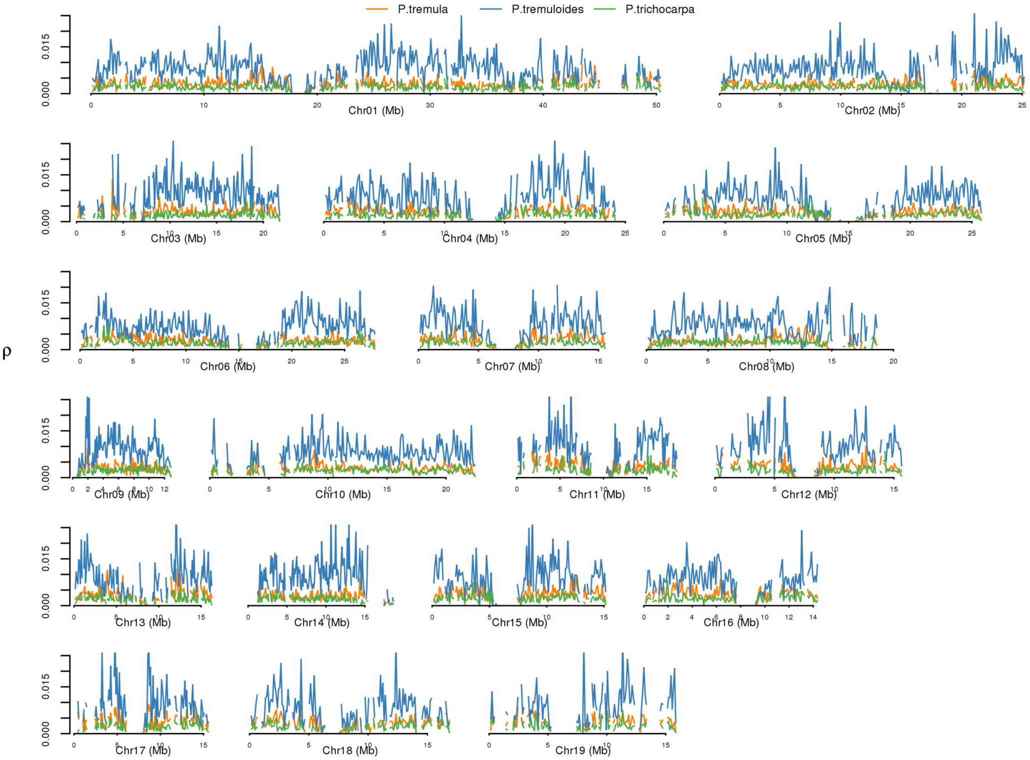

P.tremula
P.tremuloides
P.trichocarpa
θ_π
0.00
0.02
0.04
Chr01 (Mb)
0 10 20 30 40 50
Chr02 (Mb)
0 5 10 15 20 25
Chr03 (Mb)
0 5 10 15 20
Chr04 (Mb)
0 5 10 15 20 25
Chr05 (Mb)
0 5 10 15 20 25
Chr06 (Mb)
0 5 10 15 20 25
Chr07 (Mb)
0 5 10 15
Chr08 (Mb)
0 5 10 15 20
Chr09 (Mb)
0 2 4 6 8 10 12
Chr10 (Mb)
0 5 10 15 20
Chr11 (Mb)
0 5 10 15
Chr12 (Mb)
0 5 10 15
Chr13 (Mb)
0 5 10 15
Chr14 (Mb)
0 5 10 15
Chr15 (Mb)
0 5 10 15
Chr16 (Mb)
0 2 4 6 8 10 12 14
Chr17 (Mb)
0 5 10 15
Chr18 (Mb)
0 5 10 15
Chr19 (Mb)
0 5 10 15

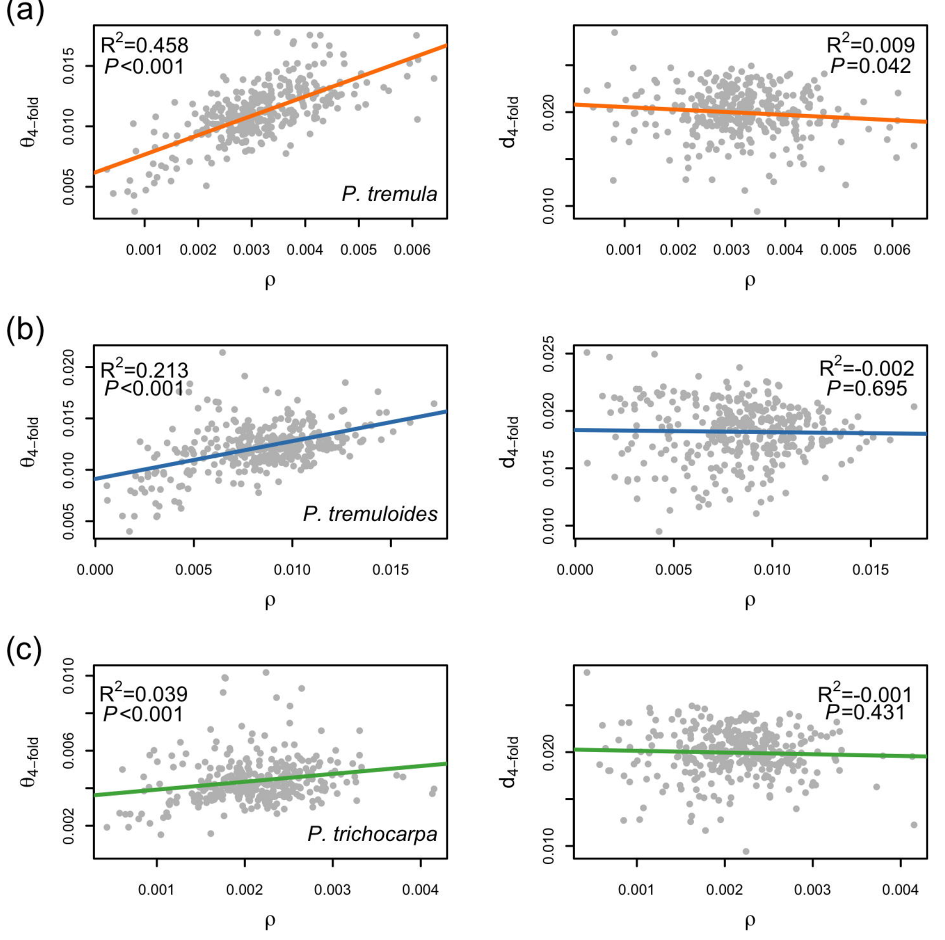

(a)
ρ=0.829***
ρ=0.675***
ρ=0.658***
P.tremuloides (θπ)
P.trichocarpa (θπ)
P.trichocarpa (θπ)
P.tremula (θπ)
P.tremula (θπ)
P.tremuloides (θπ)
0.00
0.02
0.04
0.06
0.08
(b)
ρ=0.319***
ρ=0.099**
ρ=0.164***
P.tremuloides (TajD)
P.trichocarpa (TajD)
P.trichocarpa (TajD)
P.tremula (TajD)
P.tremula (TajD)
P.tremuloides (TajD)
-2
-1
0
1
2
(c)
ρ=0.514***
ρ=0.317***
ρ=0.306***
P.tremuloides (ρ)
P.trichocarpa (ρ)
P.trichocarpa (ρ)
P.tremula (ρ)
P.tremula (ρ)
P.tremuloides (ρ)
0.000
0.010
0.020
0.030

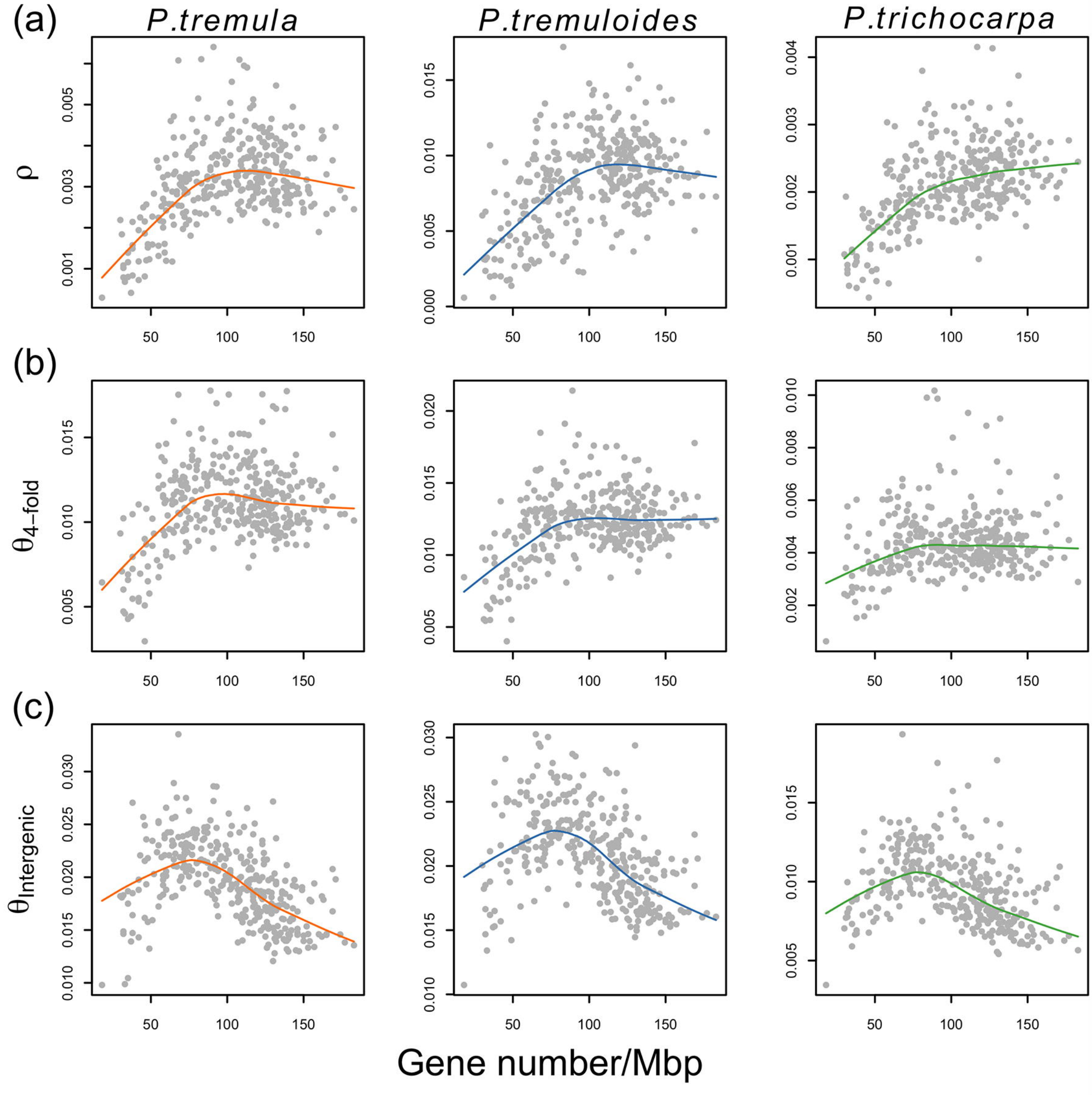

P.tremula
P.tremuloides
P.trichocarpa
Tajima's D
1.0
-0.5
-2.0
Chr01 (Mb)
0 10 20 30 40 50
Chr02 (Mb)
0 5 10 15 20 25
Chr03 (Mb)
0 5 10 15 20
Chr04 (Mb)
0 5 10 15 20 25
Chr05 (Mb)
0 5 10 15 20 25
Chr06 (Mb)
0 5 10 15 20 25
Chr07 (Mb)
0 5 10 15
Chr08 (Mb)
0 5 10 15 20
Chr09 (Mb)
0 2 4 6 8 10 12
Chr10 (Mb)
0 5 10 15 20
Chr11 (Mb)
0 5 10 15
Chr12 (Mb)
0 5 10 15
Chr13 (Mb)
0 5 10 15
Chr14 (Mb)
0 5 10 15
Chr15 (Mb)
0 5 10 15
Chr16 (Mb)
0 2 4 6 8 10 12 14
Chr17 (Mb)
0 5 10 15
Chr18 (Mb)
0 5 10 15
Chr19 (Mb)
0 5 10 15



r^2
0.4
0.3
0.2
0.1
0
10000
20000
30000
40000
50000
Distance(bp)
species
P.tremula
P.tremuloides
P.trichocarpa


P.tremula
P.tremuloides
P.trichocarpa
ρ
0.000
0.015
0 10 20 30 40 50
Chr01 (Mb)
0 5 10 15 20 25
Chr02 (Mb)
0 5 10 15 20
Chr03 (Mb)
0 5 10 15 20 25
Chr04 (Mb)
0 5 10 15 20 25
Chr05 (Mb)
0 5 10 15 20 25
Chr06 (Mb)
0 5 10 15
Chr07 (Mb)
0 5 10 15 20
Chr08 (Mb)
0 2 4 6 8 10 12
Chr09 (Mb)
0 5 10 15 20
Chr10 (Mb)
0 5 10 15
Chr11 (Mb)
0 5 10 15
Chr12 (Mb)
0 5 10 15
Chr13 (Mb)
0 5 10 15
Chr14 (Mb)
0 5 10 15
Chr15 (Mb)
0 2 4 6 8 10 12 14
Chr16 (Mb)
0 5 10 15
Chr17 (Mb)
0 5 10 15
Chr18 (Mb)
0 5 10 15
Chr19 (Mb)


(a)
R²=0.458
P<0.001
P. tremula
θ4−fold
0.005
0.010
0.015
0.001
0.002
0.003
0.004
0.005
0.006
ρ
R²=0.009
P=0.042
d4−fold
0.010
0.020
0.001
0.002
0.003
0.004
0.005
0.006
ρ
(b)
R²=0.213
P<0.001
P. tremuloides
θ4−fold
0.005
0.010
0.015
0.020
0.000
0.005
0.010
0.015
ρ
R²=-0.002
P=0.695
d4−fold
0.010
0.015
0.020
0.025
0.000
0.005
0.010
0.015
ρ
(c)
R²=0.039
P<0.001
P. trichocarpa
θ4−fold
0.002
0.006
0.010
0.001
0.002
0.003
0.004
ρ
R²=-0.001
P=0.431
d4−fold
0.010
0.020
0.001
0.002
0.003
0.004
ρ


(a)
(b)
(c)
P.tremula
P.tremuloides
P.trichocarpa
ρ
$\theta_{4\text{-fold}}$
$\theta_{\text{Intergenic}}$
0.001
0.003
0.005
0.000
0.005
0.010
0.015
0.001
0.002
0.003
0.004
0.005
0.010
0.015
0.005
0.010
0.015
0.020
0.002
0.004
0.006
0.008
0.010
0.010
0.015
0.020
0.025
0.030
0.010
0.015
0.020
0.025
0.030
0.005
0.010
0.015
50
100
150
Gene number/Mbp