# Introduction

Many North American forest tree species retreated to glacial refuges in the south during glacial periods of the late Pleistocene, and subsequently migrated northward following retreat of the ice sheets (Neale and Kremer, 2011, Zhou et al., 2014, Aitken et al., 2008). This notion has been corroborated with lot of exploration of genetic variation based models and population genetic studies (Magri et al., 2006, Pyhajarvi et al., 2007, Kremer and Goenaga, 2002). While the evidence has been overwhelming and common among various trees and plant annuals, most of these studies were limited to few loci when it came to coalescent analyses (Holliday et al., 2010, Keller et al., 2011, Pyhajarvi et al., 2007). Recent advances in coalescent theory and large scale availability of whole genome data in non-model organisms have brought major advances in our understanding of genetic variation, and what the underlying patterns of genetic variation can tell us about the organism’s history. The coalescent analyses which were limited to few loci can now use whole genome sequences to understand population history, divergence and introgression (Li and Durbin, 2011, Schiffels and Durbin, 2014, Liu and Fu, 2015).

In flowering plants, especially in the case of trees, the speciation process can be very intricate and complex (Coyne and Orr, 2004, Rieseberg and Willis, 2007). Being very different in biology from animals and microbes, most speciation and divergence models can be naïve and overly simplistic in case of trees. Trees have a very rampant gene-flow among themselves, which can extend from few to hundreds of meters (Hamrick, 2000, Savolainen et al., 2007). This kind of migration slows down the evolutionary process, and makes it very hard to decipher the signals of natural selection in genetic variation (Neale and Kremer, 2011, Savolainen et al., 2007). Moreover, range wide expansion after the clearing of glaciers have led to the recurrent bottleneck and founder affects followed by population expansion, leading to it being one of the predominant factors affecting natural variation (Excoffier et al., 2009). These colonization dynamics (population expansion and bottleneck) have been shown to have similar signals to natural selection and therefore needs to be studied before-hand, and controlled for before studying natural selection (Biswas and Akey, 2006). Having said that, trees also pose the benefit of having very less anthropogenic disturbance or domestication in its biology (Anderson et al., 2011, Neale and Kremer, 2011). This can be very beneficial, as it helps us understand the effect of neutral genetic forces like random genetic drift, and geography on their genetics in a much better way.

To advance our understanding of demographic history and colonization in case of trees, we sequenced 391 exomes of western poplar (*Populus trichocarpa*), and six genomes of three closely related species, balsam poplar (*Populus balsameifera*), Eastern cottonwood (*Populus deltoides*), and European aspen (*Populus tremula*).

The western poplar samples were chosen all along the west coast, such that they were equally distributed across the landscape. This technique of sampling was particularly useful, since it maximizes the chances to capture distinct alleles at different stages of colonization and local adaptation. We used modern coalescent methods based on linkage disequilibrium, heterozygosity and allele frequency spectrum to understand the intraspecific population history of western poplar. Combining the results from various complementary approaches, we here present the robust, approximate and integrated picture of population history of *Populus trichocarpa*. Using 1 million single nucleotide variants at 10x resolution, we investigate: 1) Population structure; 2) timing of most recent population split (tMRCA); 3) Change in effective population size (Ne) over time; 4) rate of gene-flow and 5) natural selection and local adaptation scenario in each population.

# Methods

## Sampling and sequencing

We collected leaf tissue samples from 391 *P. trichocarpa* individuals growing in the common garden located at Reynolds Homestead in Critz county, Virginia. The original locations of these samples range from 37° N to 61° N latitude (Fig. 1 and Supplementary File S1). The samples were well distributed across the whole landscape, and not just close to the rivers to prevent any sampling bias. We used sequence capture method by Agilent Technologies ((Santa Clara, CA), and designed RNA baits, at least one exon per gene model to capture most of the exome in our individuals. The results published in Zhou et al. paper (Zhou and Holliday, 2012) showed that at least 97% of the exome was captured, with some upstream regions and non-repetitive control sequences. We aligned the paired end reads to *P. trichocarpa* v3.0 reference genome using Burrows-Wheeler aligner (bwa) (Li and Durbin, 2009) and called the SNPs using combination of SAMtools (Li et al., 2009) and custom scripts. Using all these methods, we ended up with 1,058,513 SNPs with at least 10x resolution. Alignment and SNP calling details can be found in detail in Zhou et al. paper (Zhou and Holliday, 2012).

## Data Analysis

### *Admixture and PCA –* We used Admixture software (Alexander et al., 2009) and Principal Component Analysis method to explore the patterns of population structure in our dataset. We down sampled to 188774 SNPs and used Plink (Purcell et al., 2007) to select SNPs with low linkage disequilibrium (r2 < 0.1) for both of these analyses. Admixture uses the unsupervised clustering algorithm to compute number of population clusters in the sample. We tested nine ancestral clusters from K=2 through K = 10, and used the ten-fold cross-validation procedure to choose an optimal value of K for our dataset.

We used SNPRelate (Zheng et al., 2012) package in R to perform principal component analysis and used custom scripts to plot the results. Both of these analyses were performed with and without out-groups to investigate the change in clustering pattern and degree of hybridization in the *P. trichocarpa* samples with the out-groups.

### *Genome-wide heterozygosity –* We used Plink 2.0 software (Purcell et al., 2007) and custom scripts to calculate the observed heterozygosity rate for each individual in our dataset, and then used it to calculate average genome-wide heterozygosity for each population. We used the non-parametric Kruskawallis test and did the multiple comparisons between each pair of populations to find the significant differences between them

### *Introgression and gene flow –* We used two complementary approaches, TreeMix (Pickrell and Pritchard, 2012) and Estimated Effective Migration Surface (EEMS) software (Petkova et al., 2015) to estimate introgression and gene-flow in our populations and between our population and out-groups. We used 1014751 SNPs for both of these analyses and used *Populus balsamifera* as an out-group.

#### Treemix

We used Treemix software to test for patterns of historical gene-flow and admixture in our samples and used balsam poplar (*Populus balsamifera*) as an out-group. Treemix builds the maximum-likelihood tree to model evolutionary history of the population including splits and gene-flow events (Pickrell and Pritchard, 2012). We used window size of 1000 SNPs (-k option) and tested 1 to 4 migration events. Four population test (Patterson et al., 2012) and three-pop test (Reich et al., 2009) was used to confirm the gene-flow events predicted by treemix software.

#### Estimated Effective Migration Surface (EEMS)

We used EEMS software to identify the potential regions with high and low historic gene-flow. EEMS uses genetic data from the geo-referenced samples and uses “isolation by resistance” method to estimate the migration surface in the continuous landscape (Petkova et al., 2015). It can decipher complex patterns of genetic dissimilarity, which otherwise cannot be identified using PCA or admixture. We downscaled the data to 188774 SNPs in order to remove the the highly linked SNPs, and divided the region into three different parts (north, center and south), to examine the subtler variation in each subgroup. All runs were run with a different seed for at least 8 million iterations in the mcmc sampler.

### Pattern of Linkage Disequilibrium (LD)

We measured the linkage disequilibrium for all our population using the PLINK software (Purcell et al., 2007). 1014751 SNPs, which could be linked to the genetic map (kindly provided by Steve DiFaizo), were used to calculate correlation between each marker pair for our population. We calculated both correlation (rLD) and squared correlation (rLD2) in genotype frequencies as the two different measures of LD. All the genome-wide data was binned into 50 uniformly spaced recombination distance categories (0.005 – 0.25 cM) and the rLD2 was averaged for each category. This average squared correlation was compared with their respective genetic distance to look at the patterns of LD decay in each of our population.

### Effective population size (Ne) estimation

In order to explore the demographic history of our populations, we estimated the changes in effective population size over time for each subgroup. We used two different approaches, one based on linkage disequilibrium (McEvoy et al., 2011) and another one based on the SNP frequency spectrum (Stairway Plot) (Liu and Fu, 2015) to measure these changes.

#### LD-based method

We used McEvoy et al. described method (McEvoy et al., 2011), which exploits the relationship of LD with recombination distance, to calculate the effective population size in our populations. Since time (t) can be approximated to 1/2c (Hayes et al., 2003), where c is the recombination distance category, we used average rLD2 for each binned time interval (0.005-0.25cM) to calculate change in effective population size over time. The equations used to measure the effective population size are:

E(rLD2) = 1/(α+ 4Nec) − (1)

Ne = 1/[(4c) \* [(1/rLD2) - 2]] − (2)

where α = 2 and accounts for impact of mutation; c is the recombination distance category. The details of these methods can be found in McEvoy et al.l (2011)

#### Stairway plot

We also used the Stairway plot, which uses the SFS calculated from genomic data (Liu and Fu, 2015) to infer the demographic history in our ten populations. This was used as a complementary approach to confirm the results from our linkage disequilibrium based method. The stairway plot uses the SFS from the population genomic data to infer the population mutation rates (θ = 4Neμ). It uses the flexible multi-epoch demographic model and estimations of θ to estimate the changes in effective population over time. We ran this method with default parameters on the unfolded site frequency spectra calculated using *P. deltoides* as an outgroup, and used 200 bootstrap estimates to calculate the confidence intervals on our estimates. In order to prevent any skew in our SFS estimation, we removed all the singletons and low frequency variants from the file, which could be inferred both as a signature of population expansion and selective sweep.

Once done with the calculation of ancestral demography, the major bottleneck and expansion events were test using one-way anova. We also performed the Tukey’s multiple comparisons to find the significant differences between each pair of populations.

### Divergence time estimation

We used LD in two different ways to estimate divergence time in our populations. The first method (TF) looked at the relationship between FST and effective population size, and the second method used the correlation of LD for two populations to estimate the divergence time (McEvoy et al., 2011). All the estimates were done using ≤ 0.1cM distance categories, since it provided the better longer term estimate of Ne over time.

#### TF

Assuming a neutral evolutionary scenario with no migration, genetic drift will have a huge impact on genetic differentiation(FST). On the other hand, genetic drift is also affected by effective population size(Ne) and divergence time(T). According to McEvoy et al (McEvoy et al., 2011) and Nei (Nei et al., 1987), this relationship can be approximated as: FST  ≈ T/2Ne, which works out to TF ≈ 2NeFST.

We calculated the SNP-wise FST for all our populations and estimated average FST between them. Ne was calculated as harmonic mean for each population, which was averaged between two populations to calculate divergence time.

#### TLD

This method is based on the premise that linkage disequilibrium between two populations after recent split should be in perfect correlation(rpop), and that this correlation should deteriorate as the function of time. We used following formula, as described by McEvoy et al. (McEvoy et al., 2011) to estimate divergence time in our population:

ln(rpop) = c \* (-2TLD) − (3)

where, rpop is the correlation in LD pattern between the two population, c is the recombination distance category and TLD is the timing of divergence.

The files were split by chromosome, and divergence times was calculated across each chromosome to calculate confidence intervals for both our divergence estimates.

### Signatures of local adaptation

After looking at the demographic scenario in different populations of *P.trichocarpa*, we got interested in inspecting the genomic islands of differentiation, areas showing elevated and decreased level of differentiation in the genome. We used FST based single locus based approaches to inspect the past episodes of selection in our dataset.

#### Single-locus genome scans

We used LOSITAN (Antao et al., 2008) and OutFLANK (Lotterhos and Whitlock, 2014), two different FST outlier based methods to test for directional selection in our dataset. LOSITAN uses coalescent simulation based approach and uses distributions of heterozygosity to identify loci under directional selection. OutFLANK, on the other hand is more recent method, and uses likelihood based approach and trimmed FST distribution (based on heterozygosity), to identify loci under local adaptation.

For LOSITAN, we ran 1,000,000 simulations and used recommended settings of ‘Neutral mean FST’ and ‘Forcing mean FST’, which removes loci under selection while calculating mean FST. We adjusted the p-value using Bejamani and Hochberg false discovery rate method (Benjamini and Hochberg, 1995) and selected the SNPs having p-value ≤ 0.01to detect the outliers.

In OutFLANK, we removed the SNPs with minor allele frequency less than 0.05, and used the default trim fractions of 0.05, Hmin (minimum expected heterozygosity) of 0.1 and q-threshold of 0.01 to get outliers under local adaptation.

Once done with that, we used the *P. trichocarpa* v3.0 annotation file to extract the functional information about the outliers. The gene IDs and KEGG Orthology IDs (Kanehisa et al., 2016) were obtained to get functional information about the loci. Finally, we did over-representation analysis of the gene pathways (Hosack et al., 2003) to gain an in depth understanding of the biological themes in our populations.

# Results

# Genetic relationships between population

We carried out principal component analysis (PCA) to determine the structure in the genetic diversity of our populations. The first two axes explained the most variation, (PC1 = 5.2% and PC2 = 3.3%), and were driven by latitudinal differences in the populations. The Anchorage and Kodiak populations clustered separately, while rest of the populations displayed the continuous clinal variation, with no geographical separation (Fig. 2). This suggests of the geographical isolation of the Alaskan populations, and very high gene-flow (both direct and indirect) and sharing of alleles between the rest of the samples.

In order to further understand the population structure, we ran the maximum likelihood based clustering algorithm ADMIXTURE (Alexander et al., 2009) on our samples with outgroups. This method considers each sample’s genome to have originated from some hypothetical number (K) of ancestral populations, and then runs the unsupervised clustering algorithm to assign the proportion of genetic ancestry to each of these ancestral populations. We ran K = 2 through K = 13 admixture runs, and used the consistency between the runs and ADMIXTURE’s cross-validation to determine best K for our dataset. K = 8 run had lowest cross-validation error, and showed the closer relationship of Anchorage and Kodiak populations to *P. balsamifera* (balsam poplar) (Fig. 3, and Fig. S1). Other outgroups did not show any admixture with the trichocarpa samples. This was not surprising because trichocarpa is known to hybridize with balsam poplar, while other outgroups are either reproductively isolated (*P. tremula* and *P. tremuloides*) or not known to hybridize with trichocarpa in this range (*P. deltoides* and *P. fermontii*). Moreover, just like PCA, ADMIXTURE assigned Anchorage and Kodiak populations into a separate cluster, again representing geographic isolation of these populations from rest of the samples. Populations from British Columbia, Canada (BC) and Washington state (WA) were assigned to a one big cluster suggesting extensive gene-flow between these samples. Unlike the populations from BC and WA, southern samples were assigned to three distinct clusters, which represented assortative mating and some sort of isolation between these populations from rest of the samples.

Following the PCA and admixture, we used our knowledge of geography of the samples and the results of assignment tests to divide our populations into ten distinct groups. Since demographic history in trees is predominantly affected by post-glacial migration events and long distance dispersal followed by isolation (Savolainen et al., 2007, Excoffier et al., 2009), it made more sense to use both latitudinal and elevation information along with population structure results to split our populations into distinct groups. These groups were named based on their location and origin of the samples and the names are: Anchorage (AK), Kodiak (KD), Northern B.C. (NBC), Central B.C. (CBC), Coquihalla (CQ), Southern B.C. (SBC), Northern U.S. (NUS), Greenwood (GW), Oregon (OR), Sierras (CA). Please refer to supplementary information (File S1) for more detailed information about the samples and populations.

## Introgression and Gene-flow

Since the northern populations clustered closely with the balsam poplar, we examined the patterns of genomic heterozygosity in our samples to see any signs of recent hybridization. The results showed significantly higher genome-wide heterozygosity in Anchorage (HMean = 0.0913) and Kodiak populations (HMean = 0.0922) as compared to rest of the groups (Fig. 4 and S2). This was surprising because recently founded, geographically or reproductively isolated populations should be going through bottleneck and founder effects and should have lower heterozygosity, unless there was a hybridization; but the heterozygosity of these both populations was not high enough to show any signs of recent admixture. We therefore wanted to see, if these patterns were due to past introgression with the balsam poplar. We used Treemix (Pickrell and Pritchard, 2012) to build a maximum likelihood tree with our populations, using balsam poplar as an outgroup and added migration events until best fit of the tree was achieved. Looking at the results, the tree with no migration events explained 91.5% of the variation, but the residuals from the inferred tree were very high, suggestive of the presence of gene flow events. We, therefore sequentially added migration events to the tree. The new population tree allowing four migration edges explained 99.7% of the variation in the relatedness of the population, with p-values for each edge less than 2.2 X 10-308 (Please refer to Fig. S3 for the no migration graph and residuals). The new inferred graph showed the clear signs of recent admixture between balsam poplar and Kodiak and CQ populations, and showed ancestral introgression between balsam poplar and Anchorage and Kodiak and NBC populations (Fig. 5a and 5b). In order to further examine these migration events, we ran three-pop and four-pop test for treeness on our populations. The three-pop test [KD; [CBC, BAL]] and [CQ; [NUS, BAL]] failed to pass (Z = -32.6, p ≤ 1X10-30 ; Z = -31.87, p ≤ 1X10-30), confirming the gene-flow between these populations. Similarly, the four-pop test [[GW, ANCHORAGE]; [CBC, BAL]] and [[KD, NBC]; [OR, BAL]] failed to pass (Z = 35.1, p ≤ 1X10-30 ; Z = -37.3, p ≤ 1X10-30), providing the additional evidence of historical introgression between balsam poplar and Anchorage and ancestral Anchorage and Kodiak populations. Confirming these migration events, the ancestral introgression suggests of the presence of some older refugia populations in the higher latitudes, which has not been recorded or found till date.

## Effective population size and Divergence time estimation

We next examined the genome-wide linkage disequilibrium (LD) patterns, to estimate the effective population size and divergence time. For this analysis, we binned the LD estimates between SNPs to 50 evenly spaced recombination distance categories (0.005-0.25 cM) and calculate the average rLD2 for each of our populations. The results (Fig. 6) show that LD increases with increase in the geographical distance from the south. This has been previously reported (Slavov et al., 2012), and reflects of the migration of populations from the south to higher altitudes. The most extreme values of LD were found in Anchorage and Kodiak populations, with extreme values of LD over short genomic distances (0.005-0.075 cM) in Kodiak (8.835) followed by Anchorage (8.643), whereas the extreme values of LD over long genomic distances (0.075-0.25 cM) were found in Anchorage (0.591) followed by Kodiak (0.579) populations. High LD values over short genomic distances represents older population size, whereas high LD over large genomic distances represents recent colonization or extreme geographic or reproductive isolation.

In order to further investigate the demographic history of the species, we used the LD values to calculate the effective population size (Ne) in our populations. We used the McEvoy et al.l (2011) method to estimate the effective population size and calculated the harmonic mean over all recombination distance categories to get average value for each population (Refer to methods section for more details). The average Ne ranged from 172 in Anchorage to 30 in NBC populations (Fig. 7). We found a considerable amount of both temporal (genetic distance) and spatial variation (populations) in Ne (Fig. 8). Ne estimates from 100-20 KYA period showed a very large effective population size for Anchorage populations, followed by the southern populations originating sometime around 60 KYA ago. This was very surprising result, since most of the northern latitudes were covered by glaciers at that time, and most studies until now have assumed population in the north to be migrated from lower latitudes. All other populations originated between 40-20 KYA. From ~17 KYA all the recent populations stabilized over time, with Anchorage population in constant bottleneck, and Oregon populations decreasing in growth around 7 KYA. These results suggest of a cryptic refugia close to Alaska, which could be source of migration for some populations in the north, while rest of the populations could have migrated from the south with the receding of glaciers.

In order to confirm our results, we used an alternate method to estimate the demographic history in our populations. We used Stairway Plot, which uses multi-epoch model to calculate composite likelihood estimations of theta (θ) from different SNP frequency spectra to estimate Ne in our populations (Liu and Fu, 2015). The results (Fig. 9) confirm our finding showing different demographic trajectory in both Anchorage and Kodiak populations as compared to rest of the groups. Anchorage populations originated around 40 MYA ago, whereas OR populations (south) originated around 27.5 MYA ago. The results showed that Kodiak population originated from Anchorage populations, as it shared the same demographic history and originated around 5.5 MYA ago (Fig. S4-S13). In order to get the statistical significance of the difference in the origin, we compared the population sizes of each population during their first and second bottleneck. Premise behind this comparison was done based on the range expansion dynamics. The populations having common origin from the south would show similar signatures as its ancestral population, until the time it was founded, which should most likely be marked by a bottleneck effect (Excoffier et al., 2009). We standardized the Ne estimates at that particular time, and ran ANOVA. We found that there was significant statistical difference between the origins of Anchorage and Kodiak populations from rest of the groups, suggesting different trajectories of these populations (Fig. S14, S15). One thing to note is that both Stairway Plot and LD based Ne estimation has different time scales, and we believe that it’s because of the limitation of LD based method to be dependent on recombination distance categories i.e. t ≈ 1/(2c), where c is the recombination distance (Hayes et al., 2003). One the similar lines, Stairway plot doesn’t show that much temporal variation in Ne as compared to LD based Ne estimation, and that is because it uses multi-epoch demographic model, in which epochs coincide with coalescent events. It is owing to this dependency that the Stairway plot depicts Ne change in discrete time blocks (Liu and Fu, 2015).

Stairway plots also showed that Kodiak populations experienced the population decline (bottleneck), and then maintained the small population size, suggesting geographic isolation and strong founder effects in that population. Similar pattern was also seen in Anchorage populations. However, most populations in the south and mid latitudes experienced a population expansion after bottleneck, with the highest Ne in SBC and NUS populations (Fig S4-S13). This goes along the lines of abundant center hypothesis (Sagarin and Gaines, 2002), which states that the central population have higher effective population size and high migration rates as compared to peripheral populations, and has been highlighted by various studies in trees (Holliday et al., 2012, Eckert et al., 2008).

Following the estimation of demographic history, we used linkage disequilibrium estimates to measure the divergence time between each pair of our populations. LD can be used in two different ways to estimate LD. The first method uses the relationship of Ne and FST and the second one uses similarity in LD structure over time to estimate divergence time between the populations (refer to methods and McEvoy et al. for details). The results showed the signatures of isolation by distance, with the populations more geographically closer having small divergence time than the distant ones. The results of both the TF and TLD estimates are summarized in the distance matrix and neighbor-joining tree (Fig. S16-S17; File S2). One of the interesting result from the both divergence time estimates was that, the Anchorage population was more closely related to NBC, than Kodiak populations, even though Kodiak populations are more geographically closer than NBC populations. Similar results were also seen in effective population size estimates, and could be explained because of Kodiak being an island and more recently colonized as compared to NBC, which is an inland population.

## Signatures of selection and differentiation

We have showed until now that the range expansion and colonization dynamics have shaped lot of genetic variation patterns in *Populus trichocarpa*. Such kind of demographic events can produce shifts in allele frequency of the populations, and can act as a confounding factor in identifying signatures of selection (Nielsen, 2005). That said, the loci showing extra-ordinary levels of genetic differentiation can be potential targets of natural selection. We therefore, employed two different genetic differentiation (FST) based approaches, which use a robust simulation framework to remove loci affected by population subdivision and demography to identify signatures of selection. LOSITAN (Antao et al., 2008) identified 840 loci under selection, out of which only 10 loci showed balancing selection, while all other 830 loci showed positive selection. OUTFLANK (Whitlock and Lotterhos, 2015) on the other hand identified 1500 loci affected by diversifying selection. On doing the overrepresentation analysis, the results of both methods (Fig. 10, 11 and Table 1) showed over representation of Leaf Hypocotyl (LHY) gene, which is a circadian clock regulator and known to promote freezing tolerance and winter dormancy in poplar (Ibanez et al., 2010). Pseudo response regulator 5 (PRR5) was another circadian clock gene, that was over represented in the outliers, suggesting local adaptation to dormancy in the plant populations. We also found an over representation of DNA repair and nucleotide excision repair genes (ERCC-1, PRPF-40, Histone H1/5 , CNOT1and MSH-3). ERCC-1 and MSH-3 is known to the repair of UV-induced pyrimidine dimers, and correct mismatched bases caused due to UV radiation in higher plants. Moreover, previous studies in poplar have also found DNA repair related genes in their studies (Evans et al., 2014), suggesting adaptation to light radiation in the populations. Other GO terms that were highly enriched in the outliers were carbohydrate and amino acid metabolism and genes involved in the photosynthetic pathway.

# References

AITKEN, S. N., YEAMAN, S., HOLLIDAY, J. A., WANG, T. & CURTIS-MCLANE, S. 2008. Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evol Appl,* 1**,** 95-111.

ALEXANDER, D. H., NOVEMBRE, J. & LANGE, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res,* 19**,** 1655-64.

ANDERSON, J. T., WILLIS, J. H. & MITCHELL-OLDS, T. 2011. Evolutionary genetics of plant adaptation. *Trends in Genetics,* 27**,** 258-266.

ANTAO, T., LOPES, A., LOPES, R. J., BEJA-PEREIRA, A. & LUIKART, G. 2008. LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. *BMC Bioinformatics,* 9**,** 323.

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological,* 57**,** 289-300.

BISWAS, S. & AKEY, J. M. 2006. Genomic insights into positive selection. *Trends in Genetics,* 22**,** 437-446.

COYNE, J. A. & ORR, H. A. 2004. *Speciation,* Sunderland, Mass., Sinauer Associates.

ECKERT, C. G., SAMIS, K. E. & LOUGHEED, S. C. 2008. Genetic variation across species' geographical ranges: the central-marginal hypothesis and beyond. *Mol Ecol,* 17**,** 1170-88.

EVANS, L. M., SLAVOV, G. T., RODGERS-MELNICK, E., MARTIN, J., RANJAN, P., MUCHERO, W., BRUNNER, A. M., SCHACKWITZ, W., GUNTER, L., CHEN, J. G., TUSKAN, G. A. & DIFAZIO, S. P. 2014. Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. *Nat Genet,* 46**,** 1089-96.

EXCOFFIER, L., FOLL, M. & PETIT, R. J. 2009. Genetic Consequences of Range Expansions. *Annual Review of Ecology Evolution and Systematics,* 40**,** 481-501.

HAMRICK, J. L. N., J.D. 2000. Gene flow in forest trees. *Forest conservation genetics : principles and practice.* Collingwood, Vic.

Wallingford, Oxon: CSIRO ;

CABI Pub.

HAYES, B. J., VISSCHER, P. M., MCPARTLAN, H. C. & GODDARD, M. E. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res,* 13**,** 635-43.

HOLLIDAY, J. A., SUREN, H. & AITKEN, S. N. 2012. Divergent selection and heterogeneous migration rates across the range of Sitka spruce (Picea sitchensis). *Proc Biol Sci,* 279**,** 1675-83.

HOLLIDAY, J. A., YUEN, M., RITLAND, K. & AITKEN, S. N. 2010. Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests. *Molecular Ecology,* 19**,** 3857-3864.

HOSACK, D. A., DENNIS, G., SHERMAN, B. T., LANE, H. C. & LEMPICKI, R. A. 2003. Identifying biological themes within lists of genes with EASE. *Genome Biology,* 4.

IBANEZ, C., KOZAREWA, I., JOHANSSON, M., OGREN, E., ROHDE, A. & ERIKSSON, M. E. 2010. Circadian Clock Components Regulate Entry and Affect Exit of Seasonal Dormancy as Well as Winter Hardiness in Populus Trees. *Plant Physiology,* 153**,** 1823-1833.

KANEHISA, M., SATO, Y., KAWASHIMA, M., FURUMICHI, M. & TANABE, M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research,* 44**,** D457-D462.

KELLER, S. R., LEVSEN, N., INGVARSSON, P. K., OLSON, M. S. & TIFFIN, P. 2011. Local Selection Across a Latitudinal Gradient Shapes Nucleotide Diversity in Balsam Poplar, Populus balsamifera L. *Genetics,* 188**,** 941-U318.

KREMER, A. & GOENAGA, X. 2002. Special Issue: Range wide distribution of chloroplast DNA diversity and pollen deposits in European white oaks: inferences about colonisation routes and management of oak genetic resources. Preface. *Forest Ecology and Management,* 156**,** 1-3.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LI, H. & DURBIN, R. 2011. Inference of human population history from individual whole-genome sequences. *Nature,* 475**,** 493-U84.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LIU, X. M. & FU, Y. X. 2015. Exploring population size changes using SNP frequency spectra. *Nature Genetics,* 47**,** 555-U172.

LOTTERHOS, K. E. & WHITLOCK, M. C. 2014. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Mol Ecol,* 23**,** 2178-92.

MAGRI, D., VENDRAMIN, G. G., COMPS, B., DUPANLOUP, I., GEBUREK, T., GOMORY, D., LATALOWA, M., LITT, T., PAULE, L., ROURE, J. M., TANTAU, I., VAN DER KNAAP, W. O., PETIT, R. J. & DE BEAULIEU, J. L. 2006. A new scenario for the quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytol,* 171**,** 199-221.

MCEVOY, B. P., POWELL, J. E., GODDARD, M. E. & VISSCHER, P. M. 2011. Human population dispersal "Out of Africa'' estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Research,* 21**,** 821-829.

NEALE, D. B. & KREMER, A. 2011. Forest tree genomics: growing resources and applications. *Nat Rev Genet,* 12**,** 111-22.

NEI, M., NEI, M. & WILLIAM B. PROVINE COLLECTION ON EVOLUTION AND GENETICS. 1987. *Molecular evolutionary genetics,* New York, Columbia University Press.

NIELSEN, R. 2005. Molecular signatures of natural selection. *Annu Rev Genet,* 39**,** 197-218.

PATTERSON, N., MOORJANI, P., LUO, Y., MALLICK, S., ROHLAND, N., ZHAN, Y., GENSCHORECK, T., WEBSTER, T. & REICH, D. 2012. Ancient admixture in human history. *Genetics,* 192**,** 1065-93.

PETKOVA, D., NOVEMBRE, J. & STEPHENS, M. 2015. Visualizing spatial population structure with estimated effective migration surfaces. *bioRxiv*.

PICKRELL, J. K. & PRITCHARD, J. K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet,* 8**,** e1002967.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A. R., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. W., DALY, M. J. & SHAM, P. C. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics,* 81**,** 559-575.

PYHAJARVI, T., GARCIA-GIL, M. R., KNURR, T., MIKKONEN, M., WACHOWIAK, W. & SAVOLAINEN, O. 2007. Demographic history has influenced nucleotide diversity in European Pinus sylvestris populations. *Genetics,* 177**,** 1713-1724.

REICH, D., THANGARAJ, K., PATTERSON, N., PRICE, A. L. & SINGH, L. 2009. Reconstructing Indian population history. *Nature,* 461**,** 489-94.

RIESEBERG, L. H. & WILLIS, J. H. 2007. Plant speciation. *Science,* 317**,** 910-4.

SAGARIN, R. D. & GAINES, S. D. 2002. The 'abundant centre' distribution: to what extent is it a biogeographical rule? *Ecology Letters,* 5**,** 137-147.

SAVOLAINEN, O., PYHÄJÄRVI, T. & KNÜRR, T. 2007. Gene Flow and Local Adaptation in Trees. *Annual Review of Ecology, Evolution, and Systematics,* 38**,** 595-619.

SCHIFFELS, S. & DURBIN, R. 2014. Inferring human population size and separation history from multiple genome sequences. *Nat Genet,* 46**,** 919-25.

SLAVOV, G. T., DIFAZIO, S. P., MARTIN, J., SCHACKWITZ, W., MUCHERO, W., RODGERS-MELNICK, E., LIPPHARDT, M. F., PENNACCHIO, C. P., HELLSTEN, U., PENNACCHIO, L. A., GUNTER, L. E., RANJAN, P., VINING, K., POMRANING, K. R., WILHELM, L. J., PELLEGRINI, M., MOCKLER, T. C., FREITAG, M., GERALDES, A., EL-KASSABY, Y. A., MANSFIELD, S. D., CRONK, Q. C., DOUGLAS, C. J., STRAUSS, S. H., ROKHSAR, D. & TUSKAN, G. A. 2012. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree Populus trichocarpa. *New Phytol,* 196**,** 713-25.

WHITLOCK, M. C. & LOTTERHOS, K. E. 2015. Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of F-ST. *American Naturalist,* 186**,** S24-S36.

ZHENG, X., LEVINE, D., SHEN, J., GOGARTEN, S. M., LAURIE, C. & WEIR, B. S. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics,* 28**,** 3326-8.

ZHOU, L., BAWA, R. & HOLLIDAY, J. A. 2014. Exome resequencing reveals signatures of demographic and adaptive processes across the genome and range of black cottonwood Populus trichocarpa. *Molecular Ecology,* 23**,** 2486-2499.

ZHOU, L. C. & HOLLIDAY, J. A. 2012. Targeted enrichment of the black cottonwood (Populus trichocarpa) gene space using sequence capture. *Bmc Genomics,* 13.