Natural variation at *FLOWERING LOCUS T2* mediates local adaptation in a key life history trait in European aspen

Jing Wang[1,8], Jihua Ding[2], Biyue Tan[1,3], Kathryn M. Robinson[4], Ingrid H. Michelson[4], Anna Johansson[5], Björn Nystedt[5], Douglas G. Scofield[1,6,7], Ove Nilsson[3], Stefan Jansson[4], Nathaniel R. Street[4], Pär K. Ingvarsson[1,9]

[1]Umeå Plant Science Centre, Department of Ecology and Environmental Science, Umeå University, SE-90187, Umeå, Sweden

[2]Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 83 Umeå, Sweden

[3]Stora Enso Biomaterials, SE-13104, Nacka, Sweden

[4]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, SE-90187, Umeå, Sweden

[5]Wallenberg Advanced Bioinformatics Infrastructure, Science for Life Laboratory, Uppsala University, Uppsala, Sweden

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

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

[8]Present address: Centre for Integrative Genetics, Department of Animal and Aquacultural Sciences, Faculty of Life Sciences, Norwegian University of Life Science, PO Box 5003, Ås, Norway

[9]Present address: Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Sciences, PO Box 7080, SE-750 07 Uppsala, Sweden

Number of words (excluding introductory paragraph): 1420, Number of tables: 0, Number of figures: 3

The timing of growth and dormancy represent critical life-history trade-offs in perennial plants and often show strong local adaptation. Despite their importance, the genetic architecture of phenological traits remains poorly understood. Here we identify a ~700 Kbp region mediating local adaptation in the timing of bud set in *Populus tremula* across a latitudinal gradient (~56-66°N) in Sweden, where extensive gene flow has removed almost all traces of population structure. The strongest genomic associations with bud set are centered on a *P. tremula* homolog of *FLOWERING LOCUS T*2 (*PtFT*2) and explain 32-52% of the variation in bud set across sites and years. This region also shows multiple signs of a recent selective sweep restricted to the northernmost populations. Under field and greenhouse conditions, variation in bud set among accessions originating from different latitudes is strongly associated with variation in *PtFT*2 gene expression. Finally, transgenic down-regulation of *PtFT*2 yields a phenotype that, under field conditions, closely mimics variation observed between phenotypic extremes in natural populations. Our results thus provide evidence of one genomic region in *P. tremula*, centered on *PtFT2*, with major effects on adaptation to shorter growing seasons and colder climates following post-glacial colonization.

The most important environmental cue regulating the cessation of growth, the initiation of dormancy and bud set in perennial plants growing at northern latitudes is a shortening of the photoperiod [1]. We have integrated whole genome re-sequencing with field and greenhouse experiments to characterize the genetic architecture of adaptive variation affecting bud set in the obligately outcrossing, woody perennial plant European aspen (*Populus tremula*). We performed whole-genome re-sequencing on 94 unrelated *P. tremula* trees that display a strong latitudinal cline in bud set (Fig. 1A,B; Supplementary Table 1) [2-4]. After stringent variant calling and filtering (see "Online methods"), we identified a total of 4,425,109 high-quality single nucleotide polymorphisms (SNPs) with a minor allele frequency (MAF) greater than 5%. We found very weak population structure across the entire range using principal component analysis (PCA) [5], with a single significant axis separating individuals according to latitude (r=0.889, P-value <0.001) but explaining only 1.3% of the total genetic variance (Fig. 1C; Supplementary Table 2). Consistent with this, a Mantel test also showed a weak pattern of isolation by distance (r=0.210; P-value=0.047, Supplementary Fig. 1). Although recent admixture of divergent postglacial lineages following the Last Glacial Maximum (LGM) is capable of generating a genome-wide pattern of clinal variation [6], the extremely low genetic differentiation among populations (mean F_{ST}=0.0021; Supplementary Fig. 2) suggests that extensive gene flow within *P. tremula* has almost eradicated this signal.

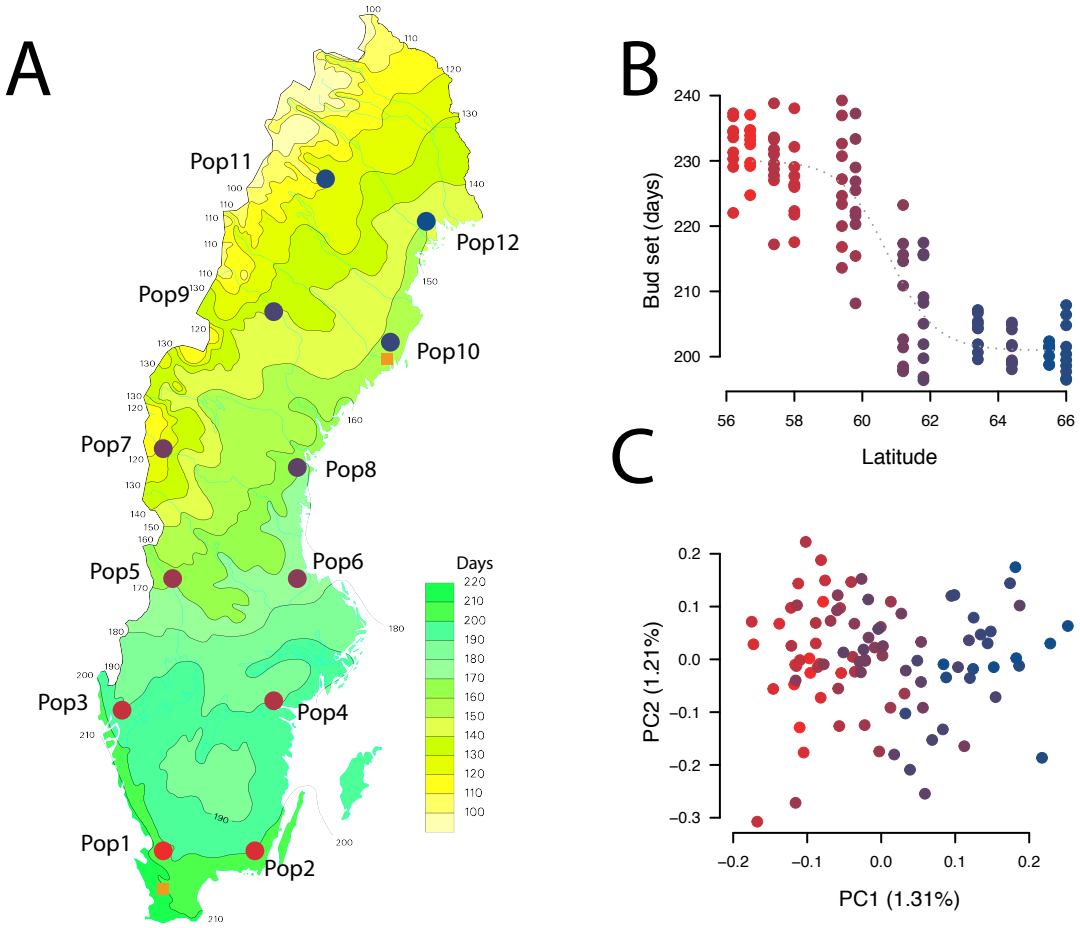


Figure 1 A) Location of the twelve original sample sites of the SwAsp collection (circles) and the location of the two common garden sites (squares). The original collection sites span a latitudinal gradient of c. 10 latitude degrees across Sweden. B) Breeding values for date of bud set for the 94 individuals included in the study across the two common gardens and three years (2005, 2006 and 2007). C) Population structure in the SwAsp collection based a principle component analysis of 217,489 SNPs that were pruned to remove SNPs in high linkage disequilibrium (SNPs included all have r^2<0.2). Although two axes are shown, only the first axis is significant (P=3.65×10^{-12}, Tacey-Widom test, 1.31% variance explained).

We used three complementary approaches to identify candidate SNPs involved in local adaptation (See "Online Methods"). We identified SNPs that were most strongly associated with observed population structure using PCAdapt [7], SNPs showing strong associations with environmental variables based on a latent factor mixed-effect model (LFMM) [8] and finally we

77 performed genome-wide association mapping (GWAS) on the timing of bud set, our target
78 adaptive trait, using GEMMA (Fig. 2A, [9]). SNPs identified as significant (false discovery
79 rate<0.05) by the three methods showed a large degree of overlap (Supplementary Fig. 3) and
80 for subsequent analyses we consider SNPs that were identified as significant by at least two of
81 the three methods to be involved in local adaptation. 99.2% of the 910 SNPs identified by all
82 three methods and 89.1% of the additional 705 SNPs identified by two methods were located
83 in a single region spanning c. 700 Kbp on chromosome 10 (Fig. 2A,B; Supplementary Fig. S4
84 and Supplementary Table 3). SNPs associated with local adaptation displayed strong clinal
85 patterns in allele frequencies with latitude, in stark contrast to 10,000 SNPs randomly selected
86 from across the genome that displayed no or negligible differences among populations
87 (Supplementary Fig. 5). The 700 Kbp region on chromosome 10 encompasses 92 genes and
88 the most strongly associated variants for all three tests were located in a region containing two
89 *P. tremula* homologs of *FLOWERING LOCUS T2* (*PtFT2*; Potra001246g10694 and an
90 unannotated copy located c. 20 Kbp upstream of *PtFT*2) (Fig. 2B; Supplementary Figs. 6).
91 *PtFT2* is known to be involved in controlling seasonal phenology in perennial plants [10] and
92 has previously been implicated in regulating short-day induced growth cessation, bud set and
93 dormancy induction in *Populus* [11-13]. Although both copies of *PtFT*2 appear to be expressed
94 (Supplementary Fig. 6), the SNP showing the strongest signal of local adaptation across all
95 three methods (Potra001246:25256) was located in the third intron of the annotated copy of
96 *PtFT2* (Potra001246g10694, Supplementary Figs. 6C). This SNP explain 32-52% of the
97 observed phenotypic variation in the timing of bud set across years and sites (Supplementary
98 Table 4).
99

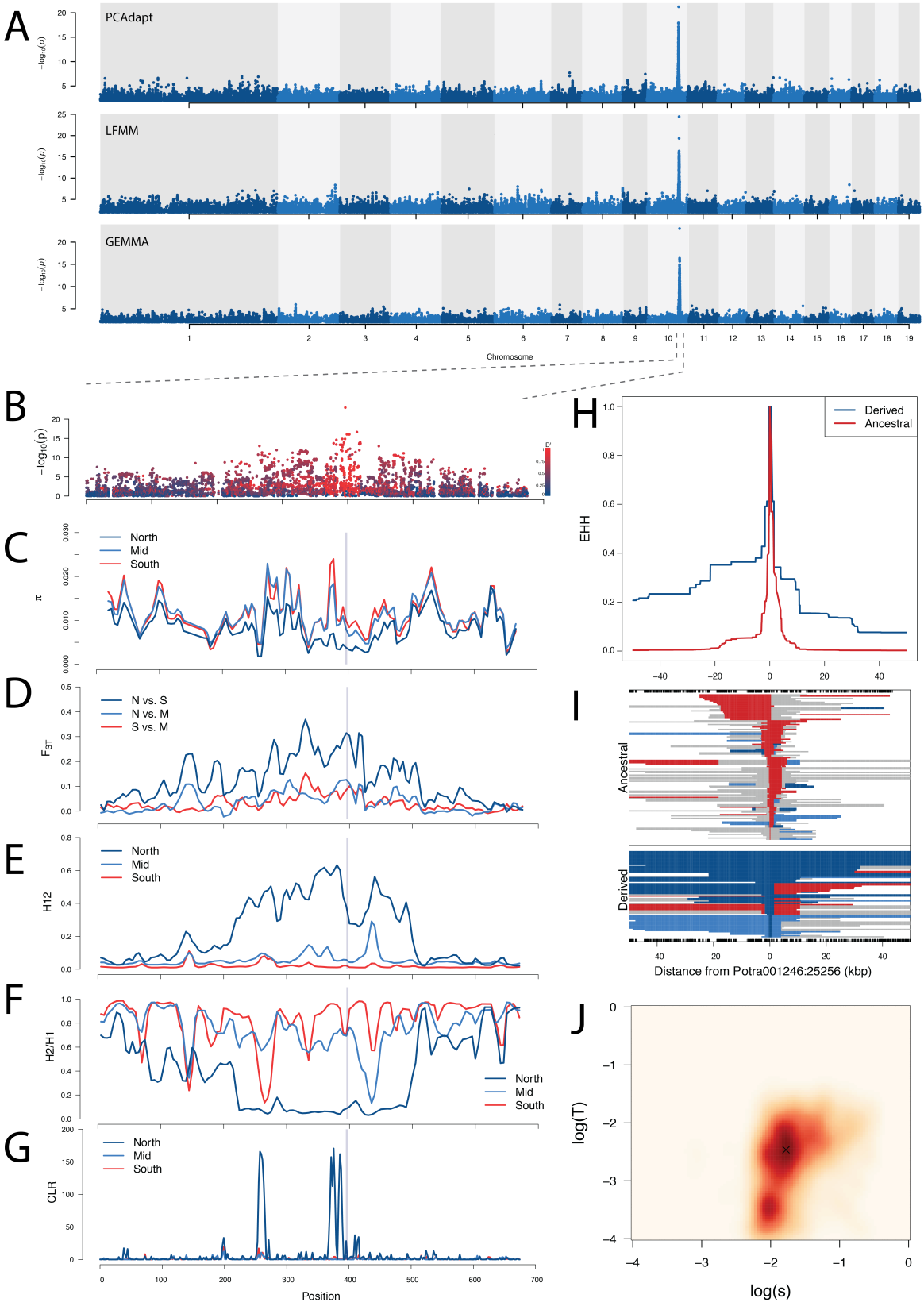


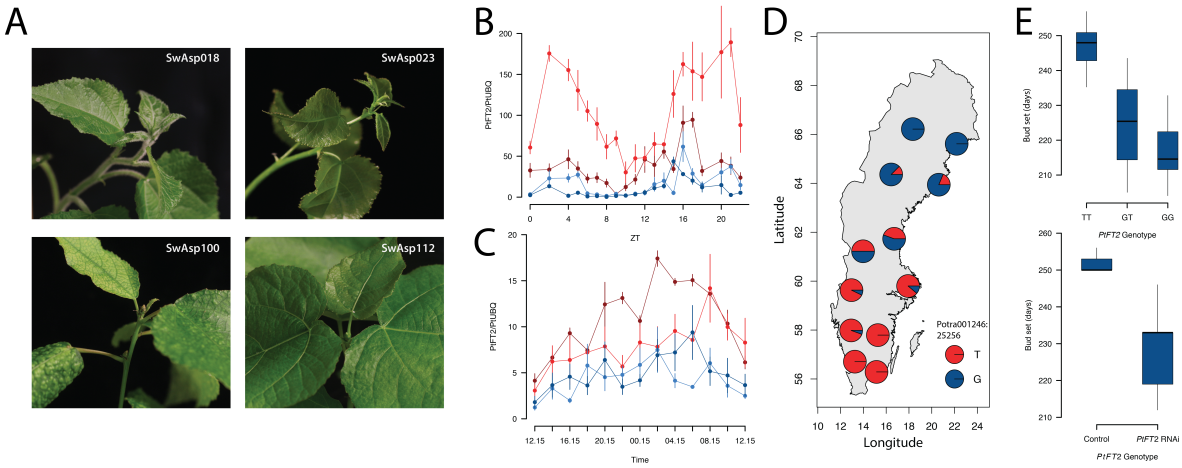
Figure 2 A) Manhattan plots for SNPs associated with local adaptation using three approaches, PCAdapt, LFMM and GEMMA. Negative $\log_{10}$-transformed *P* values from

genome-wide scans are plotted against genome position for each of 19 chromosomes (depicted in alternating colours). B) magnification of the GWAS results for the region on Chr10. Individual data points are coloured according to LD with the most strongly associated SNP (Potra001246:25256). C-G) A magnified view of the candidate region on chromosome 10 showing values for different statistics sensitive to selective sweeps. Statistics were calculated separately for individuals from southern (population 1-6), middle (populations 7-8) and northern (populations 9-12) Sweden. C) Nucleotide diversity, π D) Genetic differentiation, F_{ST}, E) H12, F) H2/H1, G) composite likelihood ratio (CLR) test for the presence of a selective sweep. In C-G, the grey bar marks the location of the *PtFT*2 locus. and H) the decay of extended haplotype homozygosity (EHH) of the derived (blue) and ancestral (red) alleles for the SNP Potra001246:25256. I) The extent of the three most common haplotypes at Potra001246:25256. Rare recombinant haplotypes were pooled and are displayed in grey. J) Joint inference of allele age and selection coefficient for the region surrounding *PtFT*2.

In order to further understand the evolution of functional differences between northern and southern *PtFT2* alleles, we analysed sequence diversity in the region surrounding *PtFT*2 and identified multiple signs of a strong and recent selective sweep that has been restricted to the northern-most populations. We observed a strong reduction in genetic diversity in the northern populations and high genetic differentiation between southern and northern populations around the *PtFT2* locus (Fig. 2C,D; Supplementary Fig. 7 and Supplementary Table 5). Furthermore, height H12 but low H2/H1 statistics [14] provide a clear indication of a single adaptive haplotype that has risen to high frequency in northern populations only (Fig. 2E,F; Supplementary Figs. 7, 8, and 9). The occurrence of a recent geographically restricted selective sweep is further supported by a composite-likelihood based (CLR) test that documents a distorted site frequency spectrum with an excess of rare and high frequency derived variants near the *PtFT2* gene in northern populations only (Fig. 2G; Supplementary Table 5). Two haplotype homozygosity-based tests, iHS [15]and nSL [16], also reveal a high concentration of significant SNPs with signals of positive selection within this region (Supplementary Fig. 10). Using the most strongly associated SNP (Potra001246:25256) as a proxy we observed that haplotypes carrying the derived allele have elevated extended haplotype homozygosity (EHH, [17]) relative to the ancestral allele (Fig. 2H,I) and the high EHH allele is largely restricted to high-latitude populations (Fig. 3D). We assumed that the

136 allele present in northern populations has been driven to fixation by a hard selective sweep
137 and used an Approximate Bayesian Computation (ABC) method [18] to jointly infer the age and
138 strength of selection acting on the northern allele. The results (Fig. 2J) point to a recent origin
139 of the northern allele (T=18952 years, 95% credible interval 719 - 114122 years) and that
140 selection during the sweep has been relatively strong (s=0.016, 95% credible interval 0.006 -
141 0.192).

142



143

144 **Figure 3** A) Bud set phenotype under 19hr day-length conditions. Two southern clones
145 (SwAsp 18,Ronneby, latitude 56.2 ˚N; SwAsp 23, Vårgårda, latitudes 58 ˚N) and two
146 northern clones (SwAsp 100, Umeå , latitude 63.9 ˚N; SwAsp 112, Luleå , latitudes 65.7
147 ˚N) were chosen to be analyzed. Trees were grown under 23hr day-length for one
148 month and then shifted to 19hr day-length, Photos were taken one month after the shift
149 to 19hr day-length. B) Dynamic expression analysis of *PtFT2* in two southern clones
150 (red) and two northern clones (blue) from the greenhouse experiment (same clone
151 number as in Fig 3A). Samples for RT-PCR were taken two weeks after the trees were
152 shifted to 19hr day-length. C) Dynamic expression analysis of *PtFT2* in two southern
153 clones (red) and two northern clones (blue) from common garden experiment. Samples
154 were collected from the in Sävar common garden in early July 2014. D) Allele
155 frequencies of the most strongly associated SNP Potra001246:25256 for the 12 original
156 populations of the SwAsp collection. E) The top is genotypic means of the timing of bud
157 set for the three genotypes at the SNP Potra001246:25256, which displayed the
158 strongest signal of local adaptation identified by all three methods as shown in Fig. 2A.
159 The effect displayed is the mean time to bud set of genotypes after correcting for the
160 effects of common garden site, year and block. The bottom is the average timing of bud

161 set for wild type control lines and transgenic *PtFT2* lines.
162

163 While the most highly associated SNP (Potra001246:25256) is a strong candidate for
164 representing the causal variant regulating the timing of growth cessation, many other SNPs
165 surrounding the *PtFT*2 locus that are not in tight linkage disequilibrium with
166 Potra001246:25256 also showed strong signals of local adaptation (Supplementary Fig. 11).
167 One possible role for the functional variation mediated by these SNPs surrounding *PtFT2* is
168 regulatory, by altering levels or timing of *PtFT*2 expression across the latitudinal gradient. To
169 further assess this possibility, we performed greenhouse experiments and found that two
170 northern genotypes showed rapid growth cessation and bud set following a shift from long
171 (23hr day length) to short day (19hr day length) conditions whereas two southern genotypes
172 continued active growth under the same conditions (Fig. 3A). Analyses of *PtFT*2 gene
173 expression in these genotypes show a strong down-regulation of *PtFT*2 in the northern
174 genotypes in conjunction with growth cessation and bud set (Fig. 3B; Supplementary Table
175 6). Notably, northern genotypes showed lower expression of *PtFT*2 under field conditions at a
176 time point when all genotypes were actively growing (Fig. 3C). Furthermore, when *PtFT*2
177 expression was reduced to about 20% of wild type levels through RNAi, bud set under field
178 conditions was accelerated by c. 23 days compared to the wild type, mimicking the phenotype
179 differences we observed among the most extreme wild-collected genotypes from opposite
180 ends of the latitudinal cline (Fig. 3E). Specifically trees carrying the derived G allele in
181 homozygous form for the most strongly associated SNP in *PtFT2* (Potra001246:25256) set
182 bud on average 33 days earlier than those homozygous for the ancestral T allele, with the
183 derived G allele showing partial dominance (Fig. 3D,E).

184 When considered together with the very weak population structure in *P. tremula* across
185 Sweden, our findings provide empirical evidence supporting recent theoretical predictions
186 that local adaptation in the face of high levels of gene flow tends to favor a genomic
187 architecture of one or a few tightly clustered loci of large-effect as opposed to the more
188 traditional architecture of many loci of small effect expected for most complex traits [19]. This
189 is likely because a single large-effect or a group of tightly linked loci of smaller effects are
190 better able to resist the homogenizing effect of gene flow than independent small-effect loci
191 do [20]. While small-effect loci likely also contribute to local adaptation, variation at such loci is
192 expected to be transient and/or difficult to distinguish from neutral background variation,

193 especially when recent demographic history coincides with local adaptation, as is the case in
194 the present study (Supplementary Fig. S12) [21,22].

195 In summary, we identified a recent and strong selective sweep at *PtFT*2, a *P. tremula*
196 homolog of *FT*, which is regionally restricted to the northernmost populations and likely
197 occurred in response to the substantially shorter growing seasons that are characteristic of
198 northern Scandinavia. While *FT* is mainly involved in controlling flowering in plants,
199 functional diversification of *FT* has occurred repeatedly in many plants [23]. The clustering of
200 adaptive variation surrounding *PtFT2* together with its position as a key integrator of diverse
201 environmental signals makes it a likely target loci of selection for rapid adaptation to
202 changing environmental conditions in the face of high gene flow.

203

204 ## Acknowledgements

205 We thank Carin Olofsson for extracting DNA for all samples used in this study.
206 STRÅNG data are obtained from the Swedish Meteorological and Hydrological Institute
207 (SMHI), which were produced with support from the Swedish Radiation Protection Authority
208 and the Swedish Environmental Agency. The research was funded through grants from
209 Vetenskapsrådet, Knut and Alice Wallenbergs stiftelse and a Young Researcher Award from
210 Umeå University to PKI. JW was supported by a scholarship from the Chinese Scholarship
211 Council. BT is supported by the UPSC "Industrial graduate school in forest genetics,
212 biotechnology and breeding". NRS is supported by the Trees and Crops for the Future (TC4F)
213 project. The authors also would like to acknowledge support from Science for Life
214 Laboratory and the National Genomics Infrastructure (NGI) for providing assistance with
215 massive parallel sequencing. All analyses were performed on resources provided by the
216 Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center
217 for Advanced Computational Science (UPPMAX) under the projects b2010014 and
218 b2011141.

219 ## Author contributions

220 JW, ON, SJ, NS and PKI conceived of and designed the experiments. JW, BT, AJ, BN,
221 DGS, NS, PKI carried out all population genetic analyses. JD performed greenhouse and RT-

PCR experiments, KMR and IHM collected data from common JW and PKI wrote the paper. All other authors commented on the manuscript.

Literature Cited

1. Singh, R. K., Svystun, T., AlDahmash, B., Jönsson, A. M. & Bhalerao, R. P. Photoperiod- and temperature-mediated control of phenology in trees - a molecular perspective. *The New phytologist* **213,** 511–524 (2017).

2. Hall, D. *et al.* Adaptive population differentiation in phenology across a latitudinal gradient in European aspen (*Populus tremul*a, L.): a comparison of neutral markers, candidate genes and phenotypic traits. *Evolution* **61,** 2849–2860 (2007).

3. Ma, X.-F., Hall, D., Onge, K. R. S., Jansson, S. & Ingvarsson, P. K. Genetic differentiation, clinal variation and phenotypic associations with growth cessation across the *Populus tremula* photoperiodic pathway. *Genetics* **186,** 1033–1044 (2010).

4. Luquez, V. *et al.* Natural phenological variation in aspen (*Populus tremula*): the SwAsp collection. *Tree Genetics & Genomes* **4,** 279–292 (2008).

5. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genetics* **2,** e190 (2006).

6. De Carvalho, D. *et al.* Admixture facilitates adaptation from standing variation in the European aspen (*Populus tremula* L.), a widespread forest tree. *Mol Ecol* **19,** 1638–1650 (2010).

7. Duforet-Frebourg, N., Bazin, É. & Blum, M. G. B. Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Mol Biol Evol* **31,** 2483–2495 (2014).

8. Frichot, É., Schoville, S. D., Bouchard, G. & François, O. Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol Biol Evol* **30,** 1687–1699 (2013).

9. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44,** 821–824 (2012).

10. Ding, J. & Nilsson, O. Molecular regulation of phenology in trees-because the seasons they are a-changin'. *Curr. Opin. Plant Biol.* **29,** 73–79 (2016).

11. Böhlenius, H. *et al. CO*/*FT* regulatory module controls timing of flowering and seasonal growth cessation in trees. *Science* **312,** 1040–1043 (2006).

12. Hsu, C.-Y. *et al. FLOWERING LOCUS T* duplication coordinates reproductive and vegetative growth in perennial poplar. *Proc Natl Acad Sci U S A.* **108,** 10756–10761 (2011).

13. Evans, L. M. *et al.* Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nat Genet* **46,** 1089–1096 (2014).

14. Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. *PLoS Genetics* **11,** e1005004 (2015).

15. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive

selection in the human genome. *PLoS Biol.* **4,** e72 (2006).

16. Ferrer-Admetlla, A., Liang, M., Korneliussen, T. & Nielsen, R. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol Biol Evol* **31,** 1275–1291 (2014).

17. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).

18. Ormond, L., Foll, M., Ewing, G. B., Pfeifer, S. P. & Jensen, J. D. Inferring the age of a fixed beneficial allele. *Mol Ecol* **25,** 157–169 (2016).

19. Yeaman, S., Aeschbacher, S. & Bürger, R. The evolution of genomic islands by increased establishment probability of linked alleles. *Mol Ecol* **25,** 2542–2558 (2016).

20. Yeaman, S. & Whitlock, M. C. The genetic architecture of adaptation under migration-selection balance. *Evolution* **65,** 1897–1911 (2011).

21. Yeaman, S. Local Adaptation by alleles of small effect. *Am Nat* **186,** S74–89 (2015).

22. Bergland, A. O., Tobler, R., González, J., Schmidt, P. & Petrov, D. Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila melanogaster*. *Mol Ecol* **25,** 1157–1174 (2016).

23. Pin, P. A. & Nilsson, O. The multifaceted roles of *FLOWERING LOCUS T* in plant development. *Plant Cell Environ* **35,** 1742–1755 (2012).

Online Materials and Methods

1. Sample collection and sequencing

We collected material from all available trees in the Swedish Aspen (SwAsp), which consists of 116 individuals collected from 12 different locations spanning the distribution range in Sweden [1] (Fig. 1A). Leaf material was sampled from one clonal replicate of each individual growing at a common garden experiment located in Sävar, northern Sweden. Total genomic DNA for each individual was extracted from frozen leaf tissue using the DNeasy plant mini prep kit (QIAGEN, Valencia, USA). Paired-end sequencing libraries with an average insert size of 650 bp were constructed for all samples according to the Illumina manufacturer's instructions. Whole genome sequencing and base calling were performed on the Illumina HiSeq 2000 platform for all individuals to a mean, per-sample depth of approximately 30× at the Science for Life Laboratory, Stockholm, Sweden.

2. Polymorphism detection

2.1 Sequence quality checking, read mapping and post-mapping filtering

A total of 103 SwAsp individuals were successfully sequenced. Prior to read mapping, we used Trimmomatic v0.30 [2] to identify reads with adapter contamination and to trim adapter sequences from reads. After checking the quality of the raw sequencing data using FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/), the quality of sequencing reads was found to drop towards the ends of reads (Supplementary Fig. 13). We therefore used Trimmomatic v0.30 to trim bases from both ends of the reads if their qualities were lower than 20. Reads shorter than 36 bases after trimming were discarded completely.

After quality control, all high-quality reads were mapped to a *de novo* assembly of the *P. tremula* genome (available at http://popgenie.org; [3]) using the BWA-MEM algorithm with default parameters using bwa-0.7.10 [4]. We used MarkDuplicates methods from the Picard packages (http://picard.sourceforge.net) to correct for the artifacts of PCR duplication by only keeping one read or read-pair with the highest summed base quality among those of identical external coordinates and/or same insert lengths. Alignments of all paired-end and single-end reads for each

sample were then merged using SAMtools 0.1.19 [5]. Sequencing reads in the vicinity of insertions and deletions (indels) were globally realigned using the RealignerTargetCreator and IndelRealigner in the Genome Analysis Toolkit (GATK v3.2.2) [6]. To minimize the influence of mapping bias, we further discarded the following site types: (i) sites with extremely low (<400× across all samples, i.e. less than an average of 4× per sample) or extremely high coverage (>4500×, or approximately twice the mean depth at variant sites) across all samples after investigating the coverage distribution empirically; (ii) sites with a high number of reads (>200×, that is on average more than two reads per sample) with mapping score equaling zero; (iii) sites located within repetitive sequences as identified using RepeatMasker [7]; (iv) sites that were in genomic scaffolds with a length shorter than 2 Kbp.

2.2 SNP and genotype calling

SNP calling in each sample was performed using the GATK HaplotypeCaller, and GenotypeGVCFs were then used to perform the multi-sample joint aggregation, re-genotyping and re-annotation of the newly merged records among all samples. We performed several filtering steps to minimize SNP calling bias and to retain only high-quality SNPs: (i) Remove SNPs at sites not passing all previous filtering criteria; (ii) Retain only bi-allelic SNPs with a distance of more than 5 bp away from any indels; (iii) Remove SNPs for which the available information derived from <70% of the sampled individuals after treating genotypes with quality score (GQ) lower than 10 as missing; (iv) Remove SNPs with an excess of heterozygotes and deviates from Hardy-Weinberg Equilibrium test (*P*-value <1e-8). After all steps of filtering, a total of 4,425,109 SNPs with minor allele frequency higher than 5% were left for downstream analysis. Finally, the effect of each SNP was annotated using SnpEff version 3.6 [8] based on gene models from the *P. tremula* reference genome (available at http://popgenie.org; [3]), and the most deleterious effect was selected if multiple effects occurred for the same SNP using a custom Perl script.

3. Relatedness, population structure and isolation-by-distance

To identify closely related individuals and to infer population structure among the sampled individuals, we discarded SNPs with missing rate >10%, MAF < 5% and that

failed the Hardy-Weinberg equilibrium test (P <1×10^{-6}) after all filtering steps as shown above. We also generated LD-trimmed SNP sets by removing one SNP from each pair of SNPs when the correlation coefficients (r^2) between SNPs exceed 0.2 in blocks of 50 SNPs using PLINK v1.9 [9]. This yielded 217,489 independent SNPs that were retained for downstream analyses of population structure. First, we used PLINK v1.9 to estimate identity-by-state (IBS) scores among pairs of all individuals. Nine individuals were excluded from further analyses due to their high pairwise genetic similarity with another sampled individual (IBS>0.8), leaving a total of 94 'unrelated' individuals for all subsequent analyses (Supplementary Fig. 14). Then, we used the smartpca program in EIGENSOFT v5.0 [10] to perform the principal component analysis (PCA) on the reduced set of genome-wide independent SNPs. A Tracey-Widom test, implemented in the program twstats in EIGENSOFT v5.0, was used to determine the significance level of the eigenvectors. Finally, isolation by distance (IBD) analysis was computed based on the pairwise comparison of the genetic and geographic distances between populations. We calculated the population differentiation coefficient (F_{ST}) [11] for each pair of the twelve populations using VCFtools v0.1.12b [12]. The relationship between genetic distance measured as F_{ST}/(1-F_{ST}) and geographic distance (km) was evaluated using Mantel tests in the R package "vegan" [13], and the significance of the correlation was estimated based on 9999 permutations.

4. Screening for SNPs associated with local adaptation

We used three conceptually different approaches to test for genome-wide signatures of local adaptation.

4.1 Population differentiation

First, we detected candidate SNPs involved in local adaptation using the principal component analysis as implemented in PCAdapt [14]. PCAdapt examines the correlations (measured as the squared loadings ρ^2_{jk}, which is the squared correlation between the jth SNP and the kth principal component) between genetic variants and specific principal components (PCs) without any prior definition of populations. As only the first principal component was significant from the PC analysis (Supplementary Table 2), we only estimated the squared loadings ρ^2_{j1} with PC1 to identify SNPs involved in local adaptation. Our results showed that most outlier SNPs that were highly correlated with the first population structure PC also had high F_{ST}

values between populations (Supplementary Fig. 15). Assuming a chi-square distribution for the squared loadings ρ^2_{j1}, as suggested by [14], we used PCAdapt to compute *P*-values for all SNPs, and then calculated the false discovery rate (FDR) using the method of [15] to generate a list of candidate SNPs showing significant associations to population structure. Only SNPs with FDR < 5% were retained as those significantly involved in local adaptation.

4.2 Environmental associations

We tested for the presence of candidate SNPs that exhibited high correlations with environmental gradients. To do this, a total of 39 environmental variables were analysed. Precipitation and temperature values were retrieved from WorldClim version 1 [16]. Sunshine hours, photosynthetically active radiation and UV radiation were obtained using the STRÅNG data model at the Swedish Meteorological and Hydrological Institute (SMHI) (http://strang.smhi.se). Values were collected from the years 2002-2012 for the original sample coordinates of each SwAsp individual and the average values over years were then calculated. The environmental variables include latitude, longitude, altitude, the number of days with temperatures higher than 5 ºC, UV irradiance, the photosynthetic photon flux density (PPFD), sunshine duration, monthly and annual average precipitation and temperature. Due to the high degree of correlation among these environmental variables (Supplementary Fig.16A), we performed a PCA on these variables using the 'prcomp' function in R to identify PCs that best summarized the range of environmental variation. The first environmental PC, which explained > 60% of the total variance (Supplementary Fig.16B,C) and had the strongest loadings for the length of growing season (Supplementary Fig.16D), was kept to represent our target environmental variable for further analyses. We then used a latent factor mixed-effect model (LFMM) implemented in the package LEA in R [17] to investigate associations between SNPs and the first environmental PC while simultaneously accounting for population structure by introducing unobserved latent factors into the model [18]. Due to the weak population structure found in the SwAsp collection (see Results), we ran the LEA function *lfmm* with the number of latent factors (*K*) ranging from one to three, using 5000 iterations as burn-in followed by 10,000 iterations to compute LFMM parameters for all SNPs. This was performed five times for each value of *K*, and we

observed identical results both across different values of *K* and across independent runs within each value of *K* (data not shown). We only showed the results using *K*=2 to account for the background population structure. LFMM outliers were detected as those SNPs with FDR < 0.05 after using the method of [15] to account for multiple testing.

4.3 Genome-wide association mapping

We obtained previously published measurements of the timing of bud set, which is a highly heritable trait that shows strong adaptive differentiation along the latitudinal gradient [19,20]. To measure phenotypic traits, all SwAsp individuals have previously been clonally replicated (four ramets per individual) and planted at two common garden sites in 2004 (Sävar, 63ºN, and Ekebo, 56ºN) (Fig. 1A). The common garden setup is described in detail in [1]. The timing of bud set was scored twice weekly starting from mid-July and continuing until all trees had set terminal buds. Bud set measurements were scored in three consecutive years, from 2005 to 2007, in both common gardens [19] (Figure 1A). A severe drought in Sävar caused most of the trees to set bud prematurely in 2006, and we therefore excluded data from Sävar in 2006 in all downstream analyses (see [20] for further discussion). We combined data on bud set from the two common garden sites and years by predicting genetic values with best linear unbiased prediction (BLUP) for all individuals. The ASReml [21] was used to fit Equation 1 to the data for calculating BLUP using restricted maximum-likelihood techniques.

$$z_{ijklm}=\mu + s_i + b_{j(i)} + y_{k(i)} + \beta_l + \varepsilon_{ijklm} \qquad (1)$$

where z_{ijklm} is the phenotype of the *m*th individual in the *j*th block in the *k*th year of the *l*th clone from the *i*th site. In Equation 1, μ denotes the grand mean and ε_{ijklm} is the residual term. The clone (β_l, BLUP) and residual term (ε_{ijklm}) were modeled as random effects, whereas the site (s_i), site/block ($b_{j(i)}$) and site/year ($y_{k(i)}$) were treated as fixed effects. The genetic value of each individual was then used as the dependent trait in an univariate linear mixed model for SNP-trait association analyses performed with GEMMA [22]. This method takes relatedness among samples into account through the use of a kinship matrix. The mixed model approach implemented in GEMMA has

been shown to outperform methods that try to correct for population structure by including it as a fixed effect in the GWAS analyses [23]. Given the extremely weak population structure we observe in our GWAS population (see Results) we did not pursue any further corrections for population structure in the association analyses as this likely would severely reduce our power to detect significant associations. As described previously, we used a FDR < 5% [15] to control for the multiple testing across the 4,425,109 SNPs.

5. Testing for positive selection on the region surrounding the *PtFT2* locus

5.1 Genotype imputation

For some haplotype-based selection tests, imputed and phased data sets were needed. We therefore used BEAGLE v4.1 [24] to perform imputation and haplotype phasing on genotypes of 94 individuals with default parameters. Before performing genotype imputation, we first used Chromosemble from the Satsuma packages [25] to order and orient the scaffolds of the *P. tremula* assembly to 19 pseudo-chromosomes according to synteny with the *P. trichocarpa* genome. We then performed pairwise genome alignment between scaffolds of *P. tremula* and the 19 pseudochromosomes using the BLAST algorithm (*E*-value cutoff of 1e-50), and finally, more than 99% of the SNPs (4,397,537 out of 4,425,109) were anchored on the 19 pseudochromosomes.

To test for the accuracy of imputation, and its relationship with the MAF cutoff and the missing rate of genotypes in our dataset, we selected 346,821 SNPs with a rate of missing genotypes lower than 10% from the pseudo-chromosome 2 (~32.6 Mb) for the simulation analysis. We randomly masked out varying proportions (5-50%) of SNPs, which were treated as missing. BEAGLE v 4.1 was then used to impute genotypes at the masked positions. We found high imputation accuracy (>0.97) across a wide range of MAF when rates of missing genotypes were less than 30% (Supplementary Fig.17), suggesting imputation and phasing by BEAGLE should not bias the accuracy of our results. We therefore phased and imputed genotypes of the SNPs anchored on pseudo-chromosomes using BEAGLE v 4.1.

5.2 Estimation of ancestral states for all SNPs

Since the ancestral states of SNPs are usually used for selection detection, for each SNP, we classified alleles as either ancestral or derived on the basis of comparisons with two outgroup species: *P. tremuloides* and *P. trichocarpa*. We obtained publicly available short read Illumina data for one *P. tremuloides* (SRA ID: SRR2749867) and one *P. trichocarpa* (SRA ID: SRR1571343) individual from the NCBI Sequence Read Archive (SRA) [26]. We individually aligned the reads from these two samples to the *de novo P. tremula* assembly (Potra v1.1, available at PopGenIE.org) and used UnifiedGenotyper in GATK to call SNPs at all sites (--output_mode EMIT_ALL_SITES). For each SNP, two procedures were performed to define their ancestral states: (1) because *P. trichocarpa* is more distantly related to *P. tremula* compared to *P. tremuloides* [27] and from our previous study there were less than 1% polymorphic sites shared between *P. tremula* and *P. trichocarpa* [26], we inferred the ancestral state as the *P. trichocarpa* allele at sites where the *P. trichocarpa* individual was homozygous and matched one of the *P. tremula* alleles; Otherwise, (2) we inferred the ancestral state as the *P. tremuloides* allele at sites where the *P. tremuloides* individual was homozygous and matched one of the *P. tremula* alleles. If the above two requirements were not met, the ancestral state was defined as missing. In total, we obtained information of ancestral states for 96.3% of all SNPs.

5.3 Refine the anchoring and orientation of SNPs associated with local adaptation to a single region on chromosome 10

As we found that a large majority of significant SNPs (>90%) detected by at least two of the three methods (PCAdapt, LFMM, and GEMMA) were clustered in a single genomic region on pseudo-chromosome 10, we performed several further steps to refine the anchoring and orientation of these SNPs within this region. First, we used ColaAlignSatsuma from the Satsuma packages [25] to align the genomes of *P. tremula* and *P. trichocarpa* using default settings. The output was then converted and filtered into GBrowse synteny compatible format that was available at http://popgenie.org [3]. Based on the alignment of the two genomes, 15 scaffolds from the *P. tremula* assembly that contain SNPs inferred to be associated with local adaptation were completely or partially mapped to a single region on chromosome 10 of *P. trichocarpa* genome (Supplementary Table 3). We then retained only seven scaffolds that were completely mapped to the region and with length longer than 10

Kbp. The seven scaffolds contained more than 95% (1465 out of 1528) of the total number of significant SNPs in the single region of chromosome 10. Lastly, according to the alignment results between the genome of *P. tremula* and *P. trichocarpa*, we re-ordered and re-oriented the seven scaffolds to a ~700 Kbp region for all downstream selection tests (Supplementary Fig. 4).

5.4 Linkage disequilibrium

A possible explanation for the occurrence of the clustering of SNPs involved in local adaptation in the ~700 Kbp region is the presence of genomic rearrangements, such as chromosomal inversions, that suppress recombination and create extensive linkage disequilibrium (LD) spanning the region within the rearrangement [28]. We calculated correlations (r^2) between all pairwise common SNPs (MAF>5%, 9149 SNPs) in the ~700 Kbp region using PLINK 1.9 [9]. To explore and compare patterns of LD between the ~700 Kbp region on chromosome 10 and genome-wide levels, we first used PLINK 1.9 to randomly thin the number of common SNPs across the genome to 200,000, and then calculated the squared correlation coefficients (r^2) between all pairs of SNPs that were within a distance of 100 Kbp. 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 [29]. In contrast to the expectation of the presence of an inversion, we did not observe blocks of elevated LD across the region on chromosome 10 (Supplementary Fig. 18A,B). Although LD is slightly higher in this region compared to genome-wide averages, LD still decays rapidly and falls to background levels within a few thousand bases (Supplementary Fig. 18A). This suggests that frequent recombination has occurred in this region and that the clustering of SNPs involved in local adaptation more likely arose from selective sweeps [30]. Nonetheless, owing to the limited ability to detect inversions using short-insert paired reads, future characterization of structural variation across the genome is clearly required to determine whether, and how, genomic rearrangements are involved in mediating adaptation in *Populus*.

5.5 Positive selection detection

We measured two haplotype-based tests, integrated haplotype score (iHS) [31] and the number of segregating sites by length (nS_L) [32], to test for possible positive selection. These statistics were calculated for all SNPs with MAF higher than 0.05

and with information on ancestral state across the genome using the software `selscan` v1.1.0a [33] with its assumed default parameters. The iHS and the nS_L values were then normalized in frequency bins across the whole genome (we used 100 bins). To test for whether there is significant concentration of selection signals on the region surrounding the *PtFT2*, we divided the 19 pseudo-chromosomes (without the seven scaffolds around the *PtFT2* locus) into non-overlapping windows of 700 Kbp and calculated the proportion of SNPs with |iHS| > 2 or with $[nS_L|>2$ in each window. Statistical significance was assessed using the ranking of genome-wide windows, with windows having fewer than 100 SNPs being excluded.

5.6 Population-specific selective sweeps

Several standard methods were further applied to search for signs of selective sweeps in different groups of populations: (i) pairwise nucleotide diversity (π) [34], which is expected to have a local reduction following a selective sweep, was calculated using a sliding window approach with window size of 10 Kbp and moving step of 5 Kbp using the software package - Analysis of Next-Generation Sequencing Data (ANGSD v0.602)[35] separately for South (pop 1-6), Mid (pop 7-8) and North (pop 9-12) populations. Only the reads with mapping quality > 30 and the bases with quality score >20 were used in the analysis. Windows with <10% of covered sites remaining from the previous filtering steps (section 2.1) were excluded; (ii) Weir and Cockerham's F_{ST}, which measures genetic divergence between pairs of three groups of populations, South, Mid and North, was calculated using a sliding-window approach with window size of 10 Kbp and moving step of 5 Kbp by VCFtools; (iii) a combination of H12 and H2/H1 [36], which measures haplotype homozygosity and can distinguish hard from soft selective sweeps, were calculated in windows of 200 SNPs (~15 Kbp) for common SNPs with MAF higher than 5% separately for South, Mid and North populations. As the mean LD (r^2) in *P. tremula* decays to less than 0.1 within 10 Kbp (Supplementary Fig. S18A and [26]), the use of ~15 Kbp windows should be large enough to differentiate the footprint of selective sweeps from those caused by neutral processes. The H12 and H2/H1 values were then averaged using a sliding window method with window size of 10 Kbp and moving step of 5 Kbp; (iv) a composite likelihood ratio statistic (CLR) [37], which contrasts the likelihood of the null hypothesis based on the genome-wide site frequency spectrum with the likelihood of a model where the site frequency has been altered by a recent selective sweep, was

computed using `SweepFinder2` [38] separately for South, Mid and North populations. `SweepFinder2` is most efficient when information on the ancestral and derived states is available for SNPs and we therefore polarized SNPs as described in Section 5.2. The small fraction of SNPs (~3.7%) that could not be polarized were excluded from further analysis using `SweepFinder2`. CLRs were calculated using non-overlapping windows with a spacing of 2 Kbp, and the empirical site frequency spectrum across the whole *P. tremula* genome was estimated using the `-f` option in `SweepFinder2` after including all polymorphic sites in the genome (a total of 8,007,303 SNPs). As recommended by [39] we only used sites that were polymorphic or that represented fixed substitutions in each group of populations to scan for sweeps. To determine whether there are significant differences of the above statistics between the 700 Kbp region around *PtFT2* gene on chromosome 10 and genome-wide estimates, we use the same strategy to divide the genome into the windows with the same size for each test and calculated the above statistics across the genome (Results are shown in Supplementary Fig. S7 and Table S5). Significance for the above statistical measurements was evaluated using Mann-Whitney tests.

To assess the scale of a genomic region that is affected by a selective sweep, we ran coalescent simulations modeling a selective sweep in the Northern populations. Simulations were run assuming that the selected site was located at the centre of the simulated region. Parameters for the simulations were taken from ABC calculations dating the selective sweep inferred in the North populations (See Section 5.7). Briefly, we used a scaled population mutation rate ($4N_e\mu$) of 0.0081/bp, which corresponds to the average observed diversity in the North populations. Similarly we set the scaled population recombination rate ($4N_er$) to 0.0019 to match the genome-wide ratio of $r/\mu=0.229$ in *P. tremula* [26]. Analyses of the simulated data using `SweepFinder2` showed that a single selective sweep often yields multiple significant peaks across a region spanning up to, and even exceeding, 100 Kbp (95% quartile: 148221 bp; Supplementary Fig. 19).

5.7 Dating the selective sweep in the North populations

To date the inferred selective sweep in the North populations we used the Approximate Bayesian Computation (ABC) method described in [40] to jointly estimate

s (the strength of selection on the beneficial mutation causing the sweep) and *T* (the time since the beneficial allele fixed) assuming a model of selection from a *de novo* mutation (hard selective sweep). We simulated 5×10^5 independent selective sweep events using the coalescent simulation program `msms` [41]. For the coalescent simulations, the ancestries of samples were traced backwards in time using standard coalescent methods and allowing for recombination. Selection was modeled at a single site by applying forward simulations, assuming additive selection so that the fitness of heterozygous and homozygous genotypes carrying the selected (derived) allele were 1 + s/2 and 1 + s, respectively. We simulated a chromosome region consisting of L=25000 sites and assumed a diploid effective population size of N_e=92000, a mutation rate of $\mu=3.75\times10^{-8}$ per base pair per generation[42], and a recombination rate of $r=0.729\times10^{-8}$ per base pair per generation. Together these parameters yielded a scaled population mutation rate equal to $\Theta=4N_e\mu L=86.27$ and a scaled population recombination rate $\rho=4N_erL=19.76$. For each simulation, values for both *s* and *T* were drawn from uniform prior distributions, $\log_{10}(T)\sim U(-4,-0.5)$ and $\log_{10}(s)\sim U(-4,-0.5)$.

6. Greenhouse and field experiments

6.1 Gene expression of *PtFT2* under active growth and during growth cessation

Samples used for the expression analysis of *PtFT2* were collected from both climate chamber and the field (Sävar, 63.4°N, Umeå) conditions. For treatment in the climate chamber, two southern clones (SwAsp018, 56.2°N, Ronneby; SwAsp023, 56.2°N, Ronneby) and two northern clones (SwAsp100, 63.9°N, Umeå; SwAsp112, 65.6°N, Luleå) were selected. Plants were grown under 23h day lengths for one month and then transferred to 19h day length condition for 2 weeks before the start of sampling. Field samples were collected in the Sävar common garden in early July, 2014 and samples were taken from two southern clones (SwAsp005, 56.7°N, Simlång; SwAsp023, 56.2°N, Ronneby) and two northern clones (SwAsp100, 63.9°N, Umeå; SwAsp116, 65.6°N, Luleå). Leaves were harvested from three different clonal replicates to serve as biological repeats, flash-frozen in in liquid nitrogen and stored at -80°C until sample preparation. Samples were collected at 2h intervals for a total

period of 24 h. RNA extraction for all samples was performed using a CTAB-LiCl method [43]. CDNA synthesis was performed using the iScript cDNA Synthesis Kit (BIO-RAD) according to the manufacturer's instructions. Quantitative real-time PCR analyses were performed using a Roche LightCycler 480 II instrument, and the measurements were obtained using the relative quantification method [44]. We used primers qFT2F (5'-AGCCCAAGGCCTACAGCAGGAA-3') and qFT2R (5'-GGGAATCTTTCTCTCATGAT-3') for amplifying the transcript of *FT2* and qUBQF (5'-GTTGATTTTTGCTGGGAAGC-3') and qUBQR (5'-GATCTTGGCCTTCACGTTGT-3') for *UBQ* as the internal control.

6.2 Field experiment with transgenic *PtFT2* lines

Construction of the *PtFT2* RNAi lines are described in detail in [45]. Transformed plants were planted together with wild type (WT) controls in a common garden at Våxtorp, Halland (latitude 56.4N, longitude 13.1E) in 2014. 18 replicates of each line were planted in a complete randomized block design together with six WT controls per block. Starting in 2015, data were collected on growth cessation, bud formation and bud set for all trees in the common garden. From early August plants were visually inspected roughly every five days and top shoots were scored according to a pre-determined scoring sheet (Supplementary Fig. 20) and classified as active growth (score 3), growth cessation (score 2), bud formation (score 1) and bud set (score 0). Scoring was continued until all plants had completely senesced in late October. Bud scoring data was converted to Julian date of bud set and analysed using the following linear model:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where μ is an overall mean, α_i is the effect of treatment i (where i is either *PtFT2* RNAi or WT) and β_j is the effect of block j and ε_{ij} are individual residual errors.

7. Data availability

7.1 Sequence and SNP data

The raw sequencing reads have been deposited in NCBI's short-read archive, SRA, under accession number PRJNA297202. In total we identified 8,007,303 SNPs and 4,425,109 SNPs with MAF high than 5% and VCF files with SNPs (original and

phased) and SNP annotations are available to download from ftp://plantgenie.org/Data/PopGenIE/Populus_tremula/v1.1/VCF/.

7.2 Other data

The following data is available from zendo.org (https://doi.org/10.5281/zenodo.844372)

- Bud set genetic values (BLUPs) for all clones used in the GWAS

7.3 BASH, Perl, Python and R scripts

All scripts used for the analysis described are available in an online repository at https://github.com/parkingvarsson/PhotoperiodLocalAdaptation

8. References

1. Luquez, V. *et al.* Natural phenological variation in aspen (Populus tremula): the SwAsp collection. *Tree Genetics & Genomes* **4,** 279–292 (2008).
2. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
3. Sundell, D. *et al.* The Plant Genome Integrative Explorer Resource: PlantGenIE.org. *The New phytologist* **208,** 1149–1156 (2015).
4. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
5. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
6. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43,** 491–498 (2011).
7. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4,** Unit 4.10–4.10.14 (2009).
8. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6,** 80–92 (2012).
9. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
10. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genetics* **2,** e190 (2006).
11. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* (1984). doi:10.2307/2408641

12. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

13. Oksanen, J. *et al. vegan: Community Ecology Package*. https://cran.r-project.org/web/packages/vegan/

14. Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, É. & Blum, M. G. B. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. *Mol Biol Evol* **33,** 1082–1093 (2016).

15. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. **100,** 9440–9445 (2003).

16. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25,** 1965–1978 (2005).

17. Frichot, E. & François, O. LEA: an R package for landscape and ecological association studies. *Methods in Ecology and Evolution* **6,** 925–929 (2015).

18. Frichot, É., Schoville, S. D., Bouchard, G. & François, O. Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol Biol Evol* **30,** 1687–1699 (2013).

19. Hall, D. *et al.* 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* **61,** 2849–2860 (2007).

20. Ingvarsson, P. K., Garcia, M. V., Luquez, V., Hall, D. & Jansson, S. Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 Locus in European aspen (Populus tremula, Salicaceae). *Genetics* **178,** 2217–2226 (2008).

21. Gilmour, A. R., Gogel, B. J., Cullis, B. R. & Thompson, R. *ASReml User Guide Release 3.0* (vsni.co.uk, 2009).

22. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44,** 821–824 (2012).

23. Vilhjálmsson, B. J. & Nordborg, M. The nature of confounding in genome-wide association studies. *Nat Rev Genet* **14,** 1–2 (2013).

24. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84,** 210–223 (2009).

25. Grabherr, M. G. *et al.* Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* **26,** 1145–1151 (2010).

26. Wang, J., Street, N. R., Scofield, D. G. & Ingvarsson, P. K. Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. *Genetics* **202,** 1185–1200 (2016).

27. Wang, Z. *et al.* Phylogeny reconstruction and hybrid analysis of *Populus* (Salicaceae) based on nucleotide sequences of multiple single-copy nuclear genes and plastid fragments. *PLoS ONE* **9,** e103645 (2014).

28. Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **173,** 419–434 (2006).

29. Remington, D. L. *et al.* Structure of linkage disequilibrium and phenotypic

associations in the maize genome. *PNAS* **98,** 11479–11484 (2001).

30. Supple, M. A. *et al.* Genomic architecture of adaptive color pattern divergence and convergence in Heliconius butterflies. *Genome Res.* **23,** gr.150615.112–1257 (2013).
31. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4,** e72 (2006).
32. Ferrer-Admetlla, A., Liang, M., Korneliussen, T. & Nielsen, R. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol Biol Evol* **31,** 1275–1291 (2014).
33. Szpiech, Z. A. & Hernandez, R. D. Selscan: an efficient multi-threaded program to perform EHH-based scans for positive selection. *Mol Biol Evol* **31,** 2824–2827 (2014).
34. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).
35. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15,** 356 (2014).
36. Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. *PLoS Genetics* **11,** e1005004 (2015).
37. Kim, Y. & Stephan, W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160,** 765–777 (2002).
38. DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I. & Nielsen, R. SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics* **32,** 1895–1897 (2016).
39. Huber, C. D., DeGiorgio, M., Hellmann, I. & Nielsen, R. Detecting recent selective sweeps while controlling for mutation rate and background selection. *Mol Ecol* **25,** 142–156 (2016).
40. Ormond, L., Foll, M., Ewing, G. B., Pfeifer, S. P. & Jensen, J. D. Inferring the age of a fixed beneficial allele. *Mol Ecol* **25,** 157–169 (2016).
41. Ewing, G. & Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26,** 2064–2065 (2010).
42. Ingvarsson, P. K. Multilocus patterns of nucleotide polymorphism and the demographic history of Populus tremula. *Genetics* **180,** 329–340 (2008).
43. Xu, M., Zang, B., Yao, H. S. & Huang, M. R. Isolation of high quality RNA and molecular manipulations with various tissues of Populus. *Russian journal of plant physiology* **56,** 716–719 (2009).
44. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-ΔΔC(T)) Method. *Methods* **25,** 402–408 (2001).
45. Böhlenius, H. *et al.* CO/FT regulatory module controls timing of flowering and seasonal growth cessation in trees. *Science* **312,** 1040–1043 (2006).