Online supporting information

Evolution of strong reproductive isolation in plants: broad-scale patterns and lessons from a perennial model group

HUIYING SHANG[1,2*], JAQUELINE HESS[1,3], MELINDA PICKUP[4], DAVID L. FIELD[1,5*], PÄR K. INGVARSSON[6], JIANQUAN LIU[7], CHRISTIAN LEXER[1]

Original research paper

Theme issue on "The evolution of strong reproductive isolation during speciation"

Author affiliations:

[1]Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria

[2]Vienna Graduate School of Population Genetics, Vienna, Austria

[3]Helmholtz Centre for Environmental Research, Halle (Saale), Germany

[4]Institute of Science and Technology (IST), Klosterneuburg, Austria

[5]Edith Cowan University, Perth, Australia

[6]Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden

[7]Key Laboratory for Bio-resources and Eco-environment, College of Life Science, Sichuan University, Chengdu, China

[*]Corresponding author: Huiying Shang, Department of Botany and Biodiversity Research, Faculty of Life Sciences, University of Vienna, Rennweg 14, A-1030 Vienna, Austria; Email: huiying.shang@univie.ac.at

David L. Field, Edith Cowan University, Perth, Australia; Email: d.field@ecu.edu.au

Online supporting methods

Plant literature survey: categorisation of gene flow. The type of genetic marker used (Allozymes, RFLPs, AFLPs, SSRs, SNPs) and number of loci involved (4-1000's) varied markedly between the 127 studies for which we were able to assign a gene flow category. Following Abbott (2017) [1] we categorized gene flow into four categories: very low, low, high and variable. There are, of course, caveats to any single approach/measure for classifying gene flow. Moreover, the diversity of marker and analysis types across the different studies precluded the use of a single quantitative number to classify gene flow. To categorize

each species pair, we used information on the frequency of hybrids and backcrosses (numbers of each hybrid class (F1s, F2s and backcrosses to each parental type) based on STRUCTURE, NewHybrids, Hybrid index) and models of gene flow (IM models such as Migrate, IMa2, Lamarc). Where available F_{ST} (between populations adjacent to the hybrid zone) was also used as an indicator of gene flow between taxa (and $N_em = 1 – F_{ST}/4 * F_{ST}$ [2, 3]), although F_{ST} was interpreted with caution and does not equal gene flow. Allocation to gene flow categories was first based on hybrid frequency and the presence of backcrosses. Then, interpretations/conclusions of gene flow from individual studies, Abbott (2017) [1] (supporting information Table S7) and F_{ST} values (where available) were considered in allocating species pairs to one of four categories. If there was insufficient information, a gene flow category was not assigned:

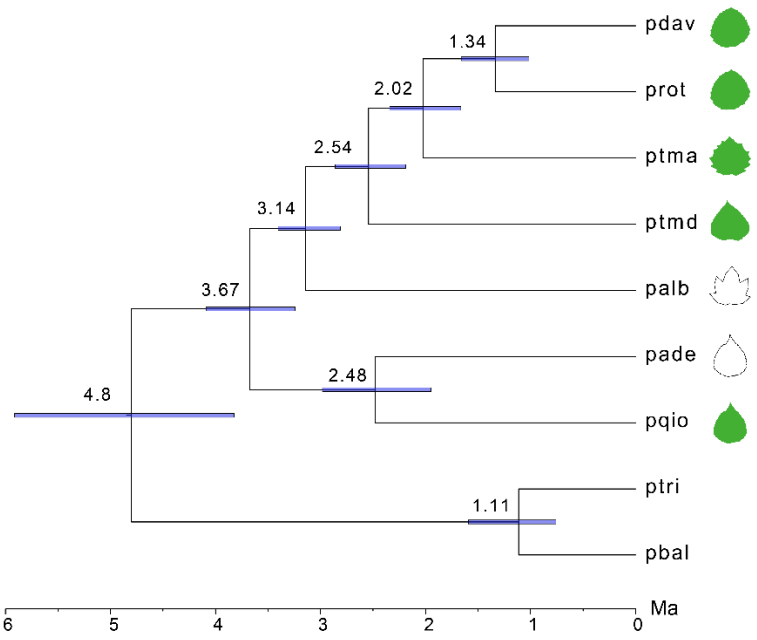
Very low = very few hybrids observed and backcrosses and advanced generation hybrids absent (or very low frequency). Generally high F_{ST} ($F_{ST} > 0.3$)
Low = low frequency of hybrids, backcrosses and advanced generation hybrids. Generally high F_{ST} ($F_{ST} > 0.2$)
High = high frequency of hybrids and the presence of backcrosses and advanced generation hybrids. Generally low F_{ST} ($F_{ST} < 0.2$)
Variable = patterns of gene flow varied among hybrid zones (applicable when multiple hybrid zones were studied, often in relation to ecological gradients).

Whole genome sequencing and data processing. Paired-end resequence reads for all individuals were first analysed using FastQC (http:// www.bioinformatics.babraham.ac.uk/projects/fastqc/) to detect poor quality data. Then Trimmomatric [4] was used to remove adapters and low-quality reads with the command "TruSeq3-SE.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:36". The reads of each individual were then mapped to the *P. trichocarpa* reference genome [5] using BWA (Version: 0.7.15-r1140) with the end-to-end alignment option [6]. Picard package v2.5 AddOrReplaceReadGroups was used to add group names and MarkDuplicates was used to remove duplicate reads (http://broadinstitute.github.io/picard/). Subsequently, reads in insertion/deletion (indel) regions were identified and realigned using the Genome Analysis Toolkit (GATK v3.6) RealignerTargetCreator and IndelRealigner [7]. Single nucleotide polymorphisms (SNPs) and genotypes were called using the GATK Unified Genotyper and base recalibration using BaseRecalibrator [8]. We used the high-quality SNPs with genotype quality above 20 as reference SNPs. All sites were then filtered with Python script available at Github (https://github.com/Huiying123). SNPs were discarded with genotype quality lower than 20, depth lower than 5X, depth higher three times the mean depth and sites with missing data above 50%. As a result, 7,026,036 SNPs were retained and used for subsequent analysis.

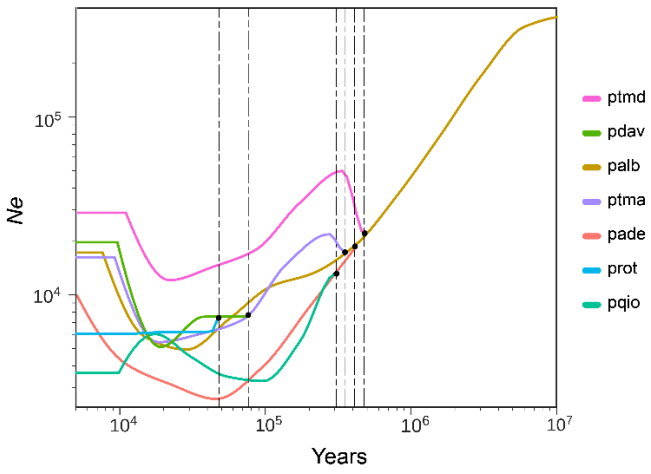
Estimation of a species tree and divergence times from whole-genome phylogenomic data. To establish a phylogenomic framework for this study, we constructed a species tree using MP-EST version1.5 [9], an approach based on the multi-species coalescent. We first separated the whole genome resequencing data according to the locations of orthologue genes. For each gene, a gene tree was generated in RAxML version 8.0 [10] using the GTRGAMMAI model with 100 bootstrap replicates. The coalescent species tree was then estimated from the 26,041 rooted orthologue gene trees by maximizing a pseudo-likelihood function in MP-EST. Each analysis started with a random number seed and 10 independent tree searches within each run. To evaluate the bootstrap support of the species tree, we extracted half of the gene trees randomly as input trees to generate the species tree and repeated this step 100 times.

Divergence times among *Populus* taxa were estimated using MCMCTree software in the PAML package [11]. The program is based on a Bayesian algorithm for species divergence time estimation using fossil constraints. We based our estimation procedure on four-fold degenerate sites and the species tree generated from MP-EST. A molecular clock was assumed according to divergence time estimates for *P. alba* and *P. tremula* (2.8-3.2Ma) available from a previous study [12]. We used 100 million years per unit time for the analysis and the root age was set to < 0.1. The mutation rate of *Populus* was estimated to be 2.5×10^{-9} per site per year and we assumed a generation time of 15 years [13]. Therefore, we set the rgene_gamma = (1, 4) and the rate drift parameter sigma2_gamma = (1, 0.4) under the GTR model. The analysis was run 200,000 times with a burn-in of 10,000 iterations and a sample frequency of 100. We did this twice to check for convergence using TRACER for the posterior distribution. Effective sample size values exceeded 200 for each run.

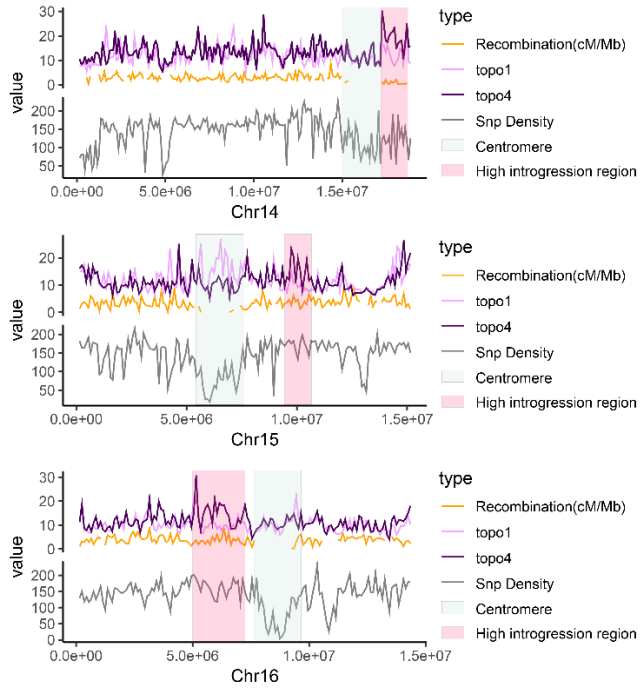
Online supporting figures and tables



Supporting Information Fig. S1. Species tree of all nine *Populus* taxa sampled and sequenced for this study (including outgroups) inferred using the coalescent-based method implemented in MP-EST. Divergence time was estimated with MCMCtree using four-fold degenerate sites. The blue bars along nodes indicate the 95% confidence intervals of divergence time. Species abbreviations follow **Fig. 2** of the main paper.



Supporting Information Fig. S2. The dynamic effective population size (Ne) and divergence times of each species inferred by SMC++. Ne is shown along the vertical axis, and divergence time in years is shown along the horizontal axis on a log scale. Bifurcation sites indicated by dashed vertical lines (separation points of coloured lines) indicate the divergence times of species pairs.



Supporting Information Fig. S3. Putatively introgressed regions along chromosomes of early-stage speciation taxa. Weightings for species topology (topo1, pink) and introgression topology (topo4, purple) (**Fig. 4d**) are shown along exemplary chromosomes. The population-scaled recombination rate of *P. tremula* (orange) and SNP density (grey) are shown in 100kb windows. Grey boxes indicate approximate centromeric regions. Pink boxes exemplify regions with consistently increased weightings for the introgression topology (topo4), potentially pointing to locally introgressed chromosome segments.

Supporting Table S1. Sampling locations and sample IDs for all sequenced individuals.

ID	Species Name	Location (N, E)(Coordinate system:WGS84)		Elevation(m)	Number
MaoKS-CX-2014-083A	*Populus rotundifolia Griff. var. duclouxiana (Dode)*	29.8194	102.2435	2148.66	1
MaoKS-CX-2014-177	*Populus rotundifolia Griff. var. duclouxiana (Dode)*	29.7356	96.0565	3068.15	1
MaoKS-CX-2014-261A	*Populus rotundifolia Griff. var. duclouxiana (Dode)*	27.1423	99.3916	2670.89	1
LiuJQ-QTP-2013-123	*Populus rotundifolia Griff. var. duclouxiana (Dode)*	29.5170	94.8716	2963.13	1
MaoKS-CX-2014-056	*Populus davidiana Dode*	31.5549	102.4176	3362.57	1
LiuJQ-MZL-2013-221	*Populus davidiana Dode*	41.0005	123.1636	321.20	1
LiuJQ-MZL-2013-302	*Populus davidiana Dode*	45.4693	130.9263	407.00	1
LiuJQ-MZL-2013-425	*Populus davidiana Dode*	38.7521	105.9355	1899.92	1
LiuJQ-MZL-2013-167	*Populus davidiana Dode*	39.2277	114.7393	1445.70	1
MaoKS-CX-2014-311	*Populus adenopoda Maxim.*	25.8204	107.3589	918.42	1
MaoKS-CX-2014-320	*Populus adenopoda Maxim.*	27.6683	107.2082	846.08	1
LiuJQ-MZL-2013-055	*Populus adenopoda Maxim.*	29.3404	109.5691	830.00	1
LiuJQ-MZL-2013-063	*Populus adenopoda Maxim.*	32.3771	113.3022	905.92	1
LiuJQ-F-2015-01	*Populus adenopoda Maxim.*	32.7559	105.2528	916.02	1
LiuJQ-Tian-2015-001	*Populus qiongdaoensis* T.Hong et P.Luo	19.1167	109.0925	212.66	1
LiuJQ-Tian-2015-002	*Populus qiongdaoensis* T.Hong et P.Luo	19.1162	109.0923	218.45	2
pop2014-3	*Populus alba*	47.4614	87.8041	500.00	1
pop2014-5	*Populus alba*	47.3817	87.8045	500.00	1
pop2014-15	*Populus alba*	47.3484	87.8669	500.00	1
pop2014-24	*Populus alba*	47.7175	86.8830	482.00	1
pop2014-26	*Populus alba*	47.0117	86.2693	485.00	1
pop2014-45	*Populus tremula*	47.9115	88.1265	993.00	1
pop2014-48	*Populus tremula*	47.9623	88.1792	1236.00	1

pop2014-50	*Populus tremula*	47.9669	88.1836	1272.00	1
pop2014-52	*Populus tremula*	47.9679	88.1852	1305.00	1
pop2014-53	*Populus tremula*	47.9768	88.2001	1333.00	1

Supporting Table S2. Sequencing statistics for all sequenced individuals.

Species name	ID	Raw reads	Cleaned reads	Mapping rate	Average depth	SNP
P. adenopoda	pade0121	51641058	47807522	92.27%	22.18	5082714
	pade31109	69700010	64542562	91.64%	27.11	5873211
	pade32018	76076568	70859083	87.30%	34.34	5984982
	pade5508	62788985	58356536	91.70%	29.76	5538752
	pade6307	85708024	79607881	92.35%	32.61	6310055
P. alba	palb01	43330912	38009829	92.35%	15.55	7298529
	palb02	56211103	48897168	91.03%	19.64	7497659
	palb03	41657930	35537631	90.63%	14.67	6921131
	palb04	49828854	43107169	91.09%	17.69	8094748
	palb05	40179026	32073180	89.82%	12.63	6885215
P. balsamifera	pbal01	61933611	47410704	95.58%	20.39	3316707
	pbal02	66532895	62076405	96.92%	26.54	3402179
P. davidiana	pdav16709	38789412	33915862	92.29%	32.90	7545093
	pdav22110	79929667	74313895	92.71%	38.46	7898376
	pdav30211	92659451	86296246	92.88%	42.87	7734377
	pdav42521	102786028	95584067	92.12%	27.81	8202549
	pdav5607	67161988	62385748	91.11%	15.43	7341735
P. qiongdaoensis	pqioT0103	100321484	92067364	91.47%	42.60	7324558
	pqioT0202	76744734	70556396	91.24%	32.44	6612612
	pqioT0205	82550486	76492429	91.48%	35.56	6704599

P. rotundifolia	prot083A04	47365375	40546868	91.81%	18.44	7046880
	prot12319	78975717	73288740	92.22%	33.01	6973289
	prot17718	60670864	56256163	91.48%	25.19	6891117
	prot261A13	53047205	46089672	86.66%	19.34	5908876
P. tremula	ptma01	37910238	33716847	91.15%	14.29	7627370
	ptma02	63839360	56861868	90.79%	23.24	8815487
	ptma03	25993725	23350515	92.26%	10.43	6980171
	ptma04	49584091	42905174	90.28%	17.68	8327853
	ptma05	57261669	51099511	91.97%	21.92	8863568
P. tremuloides	ptmd01	99361734	92255121	91.70%	28.51	9785394
	ptmd02	96230649	88156718	91.09%	28.35	9890484
	ptmd03	96387182	88430839	91.98%	29.22	9944656
	ptmd04	108934535	100107534	91.23%	31.81	9951773
	ptmd05	130684419	118019244	91.26%	37.69	10145944
P. trichocarpa	ptri01	99715650	95227609	94.03%	38.16	2088232
	ptri02	98048861	91057729	94.72%	34.96	2126432

Supporting Table S3. Divergence time estimates in millions of years (Mya) obtained with MCMC tree within the PAML software package.

Splits	Posterior mean(Mya)	95% HPD CI(Mya)	HPD-CI-width(Mya)
(ptri,pbal)-((pade,pqio),(palb, (ptmd, (ptma, (pdav, prot)))))	4.8	(3.62, 6.06)	2.44
(pade,pqio)-(palb, (ptmd, (ptma, (pdav, prot))))	3.67	(3.17, 4.15)	0.98
palb- (ptmd, (ptma, (pdav, prot)))	3.14	(2.81, 3.41)	0.6
ptmd- (ptma, (pdav, prot))	2.54	(2.16, 2.88)	0.71
ptma-(pdav, prot)	2.02	(1.66, 2.36)	0.7
pdav-prot	1.34	(0.98, 1.66)	0.68
pade-pqio	2.48	(1.85, 3.09)	1.24
ptri-pbal	1.11	(0.70, 1.62)	0.93

Supporting Table S4. Spearman rank correlation and linear regression statistics for relationships between average TWISST tree topology weightings for each chromosome and average recombination rate or gene density in windows of 100kb along each chromosome. Results for the topologies discussed in the main text are indicated by bold red type. For footnote see overleaf.

Spearman's correlation analysis between average weighting of each topology and recombination or gene density

	Ancient introgression		Recent introgression	
Topology	Average Weighting and recombination	Average Weighting and gene density	Average Weighting and recombination	Average Weighting and gene density
topo1	r=0.5; p=0.0310	r=-0.297; p=0.217	**r=-0.688; p=0.001546**	**r=0.735; p=0.0003408**
topo2	r=0.516; p=0.0255	r=-0.250; p=0.301	r=0; p=1	r=-0.22; p=0.3662
topo3	r=0.660; p=0.00273	r=-0.507; p=0.0267	r=0.289; p=0.2286	r=-0.3418; p=0.152
topo4	**r=0.258 p=0.2852**	**r=0.225; p=0.3545**	**r=-0.854; p=2.2e-16**	**r=0.793; p=5.029e-05**
topo5	**r=0.163; p=0.503**	**r=-0.228; p=0.3487**	r=-0.342; p=0.1518	r=0.283; p=0.2405
topo6	**r=-0.653; p=0.00312**	**r=0.728; p=0.00041**	r=-0.291; p=0.2257	r=0.0413; p=0.8667
topo7	r=0.711; p=0.000927	r=-0.643; p=0.00297	r=0.812; p=2.485e-05	r=-0.842; p=6.213e-06
topo8	r=0.656; p=0.00292	r=-0.774; p=0.00010	r=0.849; p=2.2e-16	r=-0.718; p=0.000525
topo9	r=0.681; p=0.00179	r=-0.887; p=4.29e-07	r=0.884; p=2.2e-16	r=-0.731; p=0.000376
topo10	r=0.311; p=0.1953	r=-0.146; p=0.5513	r=0.796; p=6.04e-05	r=-0.793; p=5.029e-05

topo11	r=-0.844; p=2.2-16	r=0.605; p=0.00611	r=0.658; p=0.00282	r=-0.687; p=0.00115
topo12	r=0.153; p=0.5313	r=-0.266; p=0.2705	r=0.867; p=2.2e-16	r=-0.698; p=0.000896
topo13	r=0.598; p=0.00794	r=-0.682; p=0.0013	r=-0.705; p=0.00105	r=0.783; p=7.16e-05
topo14	r=-0.788; p=9.61e-05	r=0.695; p=0.00096	r=0.739; p=0.000455	r=-0.784; p=7.16e-05
topo15	r=0.646; p=0.00355	r=-0.787; p=6.31e-05	r=0.830; p=3.22e-06	r=-0.713; p=0.000617

Linear regression analysis between average weighting of each topology and recombination or gene density

	Ancient introgression										**Recent introgression**									
	Average Weighting and recombination					Average Weighting and gene density					Average Weighting and recombination					Average Weighting and gene density				
Topology	slope	intercept	R2	R2.Ad	P	slope	intercept	R2	R2.Ad	P	slope	intercept	R2	R2.Ad	P	slope	intercept	R2	R2.Ad	P
topo1	0.004	0.026	0.322	0.282	0.011	-0.001	0.05	0.253	0.209	0.028	**-0.015**	**0.161**	**0.517**	**0.488**	**0.001**	**0.005**	**0.056**	**0.495**	**0.465**	**0.001**
topo2	0.002	0.035	0.174	0.125	0.076	-0.001	0.046	0.083	0.029	0.231	0	0.071	0.003	-0.056	0.825	0	0.071	0.005	-0.054	0.776
topo3	0.004	0.016	0.451	0.419	0.002	-0.001	0.039	0.325	0.285	0.011	0.001	0.063	0.039	-0.017	0.417	0	0.071	0.042	-0.015	0.401
topo4	**0.001**	**0.129**	**0.015**	**-0.043**	**0.613**	**0.001**	**0.124**	**0.072**	**0.017**	**0.267**	**-0.023**	**0.195**	**0.761**	**0.747**	**0**	**0.007**	**0.049**	**0.56**	**0.535**	**0**
topo5	**0.002**	**0.124**	**0.021**	**-0.037**	**0.554**	**-0.001**	**0.136**	**0.028**	**-0.029**	**0.491**	-0.002	0.071	0.132	0.081	0.126	0	0.064	0.008	-0.051	0.722
topo6	**-0.02**	**0.278**	**0.53**	**0.502**	**0**	**0.007**	**0.144**	**0.502**	**0.472**	**0.001**	-0.003	0.093	0.072	0.017	0.268	0	0.082	0.001	-0.057	0.878
topo7	0.004	0.029	0.498	0.469	0.001	-0.001	0.054	0.388	0.352	0.004	0.007	0.024	0.711	0.694	0	-0.002	0.073	0.628	0.607	0
topo8	0.006	0.018	0.501	0.472	0.001	-0.002	0.057	0.517	0.489	0.001	0.007	0.02	0.724	0.708	0	-0.002	0.061	0.407	0.372	0.003
topo9	0.006	0.011	0.548	0.522	0	-0.002	0.051	0.563	0.538	0	0.008	0.032	0.831	0.821	0	-0.002	0.077	0.453	0.421	0.002
topo10	0.001	0.032	0.216	0.169	0.045	0	0.039	0.053	-0.002	0.342	0.007	0.019	0.695	0.677	0	-0.002	0.064	0.533	0.505	0
topo11	-0.009	0.105	0.697	0.679	0	0.002	0.05	0.377	0.341	0.005	0.007	0.024	0.574	0.549	0	-0.002	0.069	0.42	0.386	0.003
topo12	0.001	0.034	0.082	0.028	0.233	-0.001	0.045	0.17	0.122	0.079	0.005	0.035	0.758	0.744	0	-0.001	0.066	0.439	0.406	0.002
topo13	0.005	0.023	0.457	0.425	0.001	-0.002	0.058	0.411	0.377	0.003	-0.011	0.132	0.525	0.497	0	0.004	0.057	0.579	0.554	0
topo14	-0.01	0.117	0.599	0.576	0	0.003	0.055	0.366	0.328	0.006	0.006	0.034	0.557	0.53	0	-0.002	0.077	0.574	0.549	0
topo15	0.004	0.023	0.467	0.436	0.001	-0.002	0.05	0.594	0.57	0	0.006	0.024	0.714	0.697	0	-0.002	0.063	0.522	0.494	0

Footnote to Supporting Table S4

The R-squared (R2) statistic: a measure of how well the model fits the data, expressed as percentage of the response variable variation explained by a linear model. The value is a biased estimate based on the sample size. R2.Ad: unbiased R2. P-value: the ability to reject the null hypothesis. If the p-value is less than 0.05 or 0.01, corresponding respectively to a 5% or 1% chance of rejecting the null hypothesis when it is true. Calculated by F-statistics.

Online supporting references

[1] Abbott, R. J. 2017 Plant speciation across environmental gradients and the occurrence and nature of hybrid zones. *J Syst Evol* 55, 238-258. (DOI:10.1111/jse.12267).

[2] Rieseberg, L. H., Church, S. A. & Morjan, C. L. 2004 Integration of populations and differentiation of species. *New Phytol* 161, 59-69. (DOI:10.1046/j.1469-8137.2003.00933.x).

[3] Wright, S. 1931 Evolution in Mendelian Populations. *Genetics* 16, 97-159.

[4] Bolger, A. M., Lohse, M. & Usadel, B. 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120. (DOI:10.1093/bioinformatics/btu170).

[5] Tuskan, G. A. & Difazio, S. & Jansson, S. & Bohlmann, J. & Grigoriev, I. & Hellsten, U. & Putnam, N. & Ralph, S. & Rombauts, S. & Salamov, A., et al. 2006 The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science (New York, N.Y.)* **313**, 1596-1604. (DOI:10.1126/science.1128691).

[6] Li, H. 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In *arXiv*:1303.3997v2 (q-bio.GN)

[7] DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., et al. 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491. (DOI:10.1038/ng.806).

[8] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303. (DOI:10.1101/gr.107524.110).

[9] Liu, L., Yu, L. & Edwards, S. V. 2010 A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol Biol* **10**, 302. (DOI:10.1186/1471-2148-10-302).

[10] Stamatakis, A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313. (DOI:10.1093/bioinformatics/btu033).

[11] Yang, Z. 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences : CABIOS* **13**, 555-556.

[12] Christe, C., Stolting, K. N., Paris, M., Frasmall yi, U. C., Bierne, N. & Lexer, C. 2017 Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. *Mol Ecol* **26**, 59-76. (DOI:10.1111/mec.13765).

[13] Wang, J., Street, N. R., Scofield, D. G. & Ingvarsson, P. K. 2016 Variation in linked selection and recombination drive genomic divergence during allopatric speciation of European and American aspens. *Mol Biol Evol* **33**, 1754-1767. (DOI:10.1093/molbev/msw051).