Sjölund MJ, González-Díaz P, Moreno-Villena JJ & Jump AS (2019) Gene flow at the leading range edge - the long-term consequences of isolation in European Beech (Fagus sylvatica L. Kuhn). Journal of Biogeography. https://doi.org/10.1111/jbi.13701

Journal of Biogeography

Supporting Information

Gene flow at the leading range edge - the long-term consequences of isolation in European Beech (*Fagus sylvatica* L. Kuhn)

M. Jennifer Sjölund, Patricia González-Díaz, Jose J. Moreno-Villena, and Alistair S. Jump

Appendix S1 Approaches for identifying regional population structure at the leading edge

Individual-based assignment methods were performed on the adult cohort using GENELAND 4.0.4 (Guillot et al., 2005) a spatially explicit Bayesian clustering model. Seedlings were excluded from cluster analysis as significant deviations from Hardy-Weinberg equilibrium were found in six out of the 14 sites, as well as significant isolation-by-distance in seedlings (slope 0.026, $p < 0.01$) deviating from model assumptions for both clustering programs (STRUCTURE methods described below). Although isolation-by-distance was present, it was weaker in adults (slope 0.012, $p < 0.05$). Site GAR was removed from analysis as it was exclusively assigned to a cluster which displayed significant deviations from Hardy-Weinberg equilibrium and linkage disequilibrium.

Runs were performed in GENELAND for 500,000 Markov Chain Monte-Carlo (MCMC) iterations with a thinning of 500, and a burn-in of 200. To determine the initial number of *K*, the uncorrelated allele frequency model with a spatial prior was used, with *K* varying from 1 to 13. Using a spatial prior allows the identification of genetic discontinuities associated with barriers to gene flow and potentially isolation (Francois & Durand 2010). Since primers for *F. sylvatica* are known to be subject to null alleles (Chybicki & Burczyk, 2009), the null allele model was implemented as recommended by (Guillot, Santos, & Estoup, 2008). Runs were performed 10 times for each model to compare average posterior probabilities for each value of *K*. To check compliance of inferred clusters with modelling assumptions (Guillot, Lebloit, Coulon, & A.C., 2009), we performed tests for gametic disequilibrium within the three inferred clusters and genetic differentiation between pairs of clusters in FSTAT 2.9.3.2 (Goudet, 1995).

To refine cluster membership, we used the correlated allele frequency model with *K* fixed at the value obtained from the uncorrelated allele frequency model. Setting *K* as a variable in the correlated model can lead to its overestimation (Guillot et al. 2014) as larger values are not sufficiently penalised, resulting in the inference of spurious sub-populations, which occurred in preliminary tests. The correlated model is better at detecting low differentiation from recent ecological events, although it is more sensitive to departures from model assumptions (Guillot, 2012). Post-processing analysis was performed on the correlated allele model output to assess the level of admixture using 500,000 iterations with a burn-in of 200. Admixture and substructure within subsets of the westerly and easterly clusters were analysed further using the same protocol as above.

To validate our results, we analysed the data using a second Bayesian clustering program, STRUCTURE 2.3.4 (Pritchard, Stephens, & Donnelly, 2000), as recommended by (Guillot et al., 2009). Repeats of 10 runs were performed for each *K* value, set from 1 to 10, with each run consisting of 500,000 MCMC iterations, with a burn-in period of 100,000, using the correlated allele frequency model (Falush, Stephens & Pritchard (2003) and the admixture ancestry model. Unlike GENELAND, geographic coordinates cannot be implemented as a spatial prior in this program. We assessed the mean log-likelihood values for each *K* to identify their convergence and the true number of clusters in the data. The value of *K* was validated using the method of Evanno, Regnaut, and Goudet (2005)using STRUCTURE HARVESTER 0.6.94 (Earl & VonHoldt, 2012) by calculating ΔK, a statistic related to the

second-order rate of change in the log probability of the data. Post-processing of Q-matrices was performed in CLUMPP 1.1.2 (Jakobsson & Rosenberg, 2007) with graphics created in DISTRUCT 1.1 (Rosenberg, 2004)

References

Chybicki, I. J., & Burczyk, J. (2009). Simultaneous estimation of null alleles and inbreeding coefficients. *Journal of Heredity*, *100*, 106–113.

Earl, D. A., & VonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*, 359–361.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, *14*, 2611–2620.

Falush, D., Stephens, M., & Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics 164,* 1567-1587.

Francois, O., & Durand, E. (2010). Spatially explicit Bayesian clustering models in population genetics. *Molecular Ecology Resources*, 10: 773-784.

Goudet, J. (1995). FSTAT (version 1.2): A computer program to calculate F-statistics. *Journal of Heredity*, *86*, 485–486.

Guillot, G., Estoup, A., & Cosson, J. F. (2005). A spatial statistical model for landscape genetics. *Genetics Society of America*, *170*, 1261–1280.

Guillot, G., Santos, F., & Estoup, A. (2008). Analysing georeferenced population genetics data with Geneland: a new algorithm to deal with null alleles and a friendly graphical user interface. *Bioinformatics Applications Note*, *24*, 1406–1407.

Guillot, G., Lebloit, R., Coulon, A., & A.C., F. (2009). Statistical methods in spatial genetics. *Molecular Ecology*, *18*, 4734–4756.

Guillot, G. (2012). Population genetic and morphometric data analysis using R and the Geneland program The Geneland development group. The Geneland development group.

Guillot, G., Santos, F., & Estoup, A. (2014). Package 'Geneland': Detection of structure from multilocus genetic data.

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, *23*, 1801–1806.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959.

Rosenberg, N. A. (2004). DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes*, *4*, 137–138.

Appendix S2

Figure S2.1 Distribution of beech in Europe. Distribution map of Beech (Fagus sylvatica) EUFORGEN 2009, www.euforgen.org.

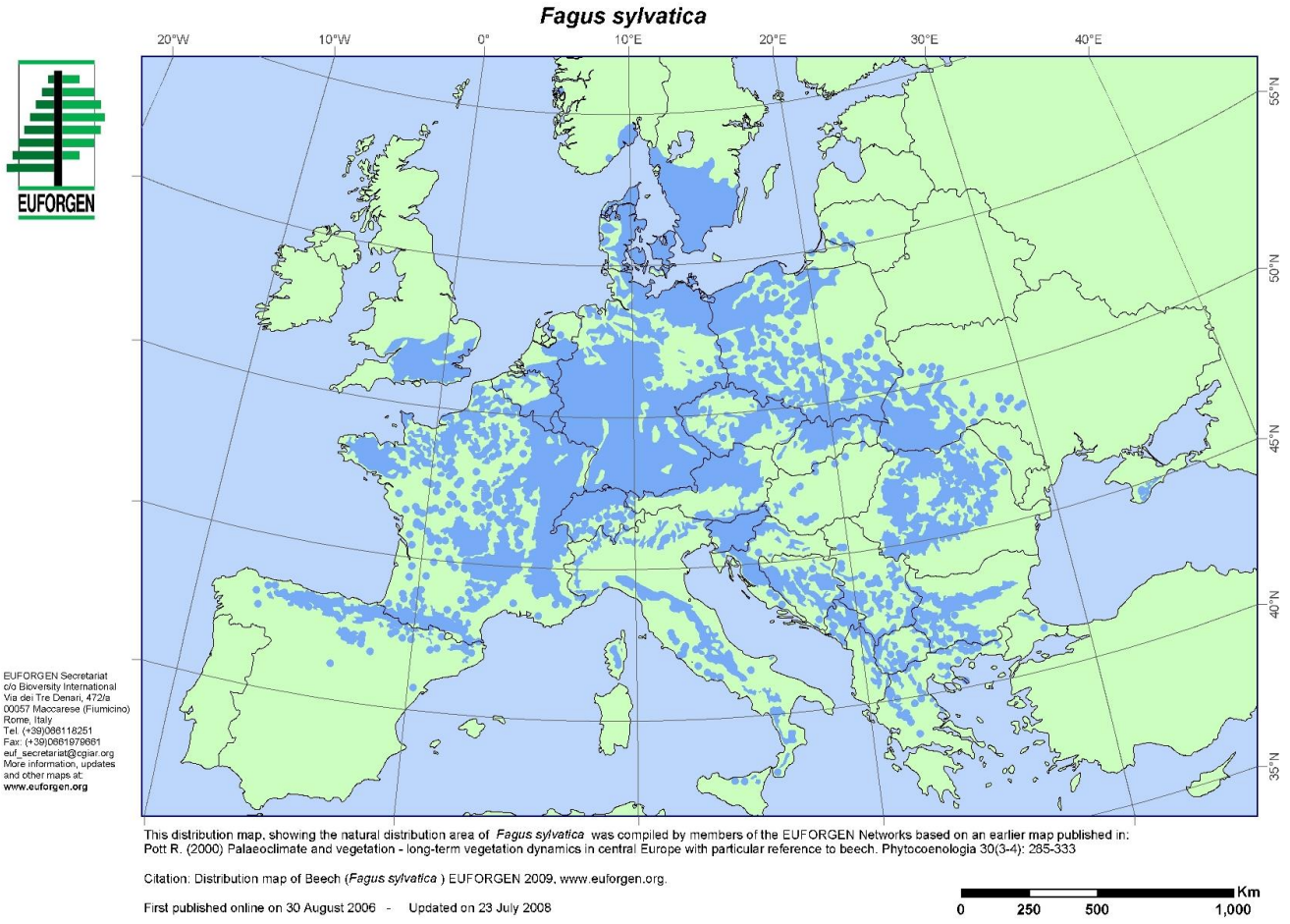


Figure S2.2 Examples of buffer zones. Sites, from left to right with decreasing beech area (ha), are SOD, GUL, STO, and OMB.

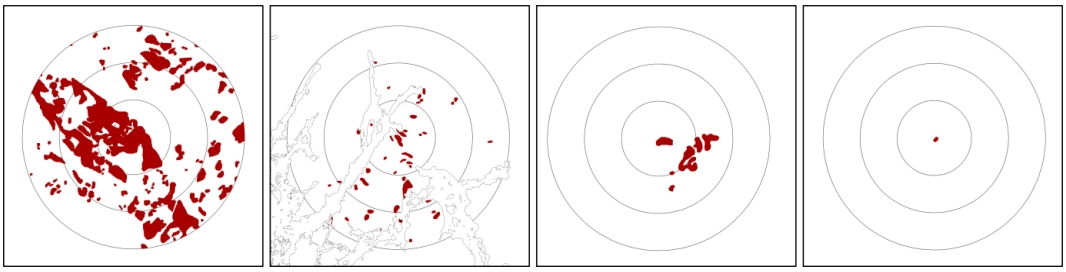


Figure S2.3 Maps of posterior probability of cluster membership for each population in GENELAND describing posterior probabilities of belonging to cluster 1, 2, and 3, in order from left to right. Colour lightness increases with the probability of population membership and contour lines represent the spatial position of genetic discontinuities.

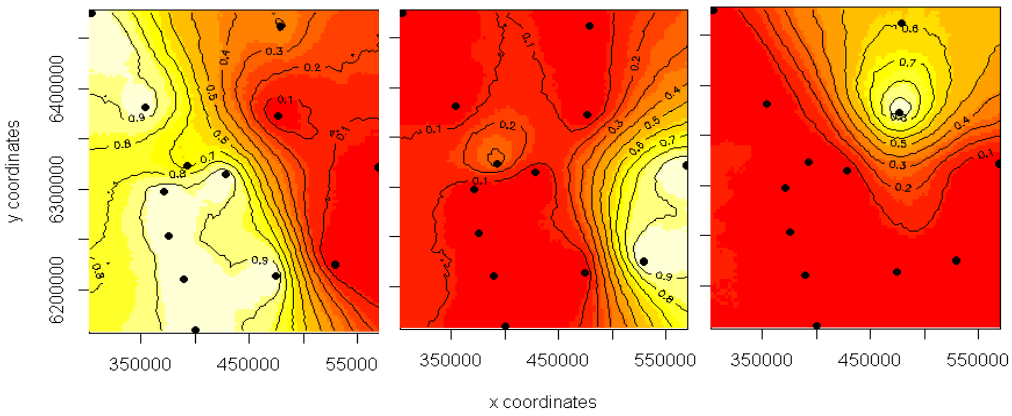


Figure S2.4 Maps of posterior probability of cluster membership for the subset of clusters 2 and 3 in GENELAND. The subset includes sites TRO and HOR from the south-eastern cluster (2), and MAT and OMB from the north-eastern cluster (3). Colour lightness increases with the probability of population membership and contour lines represent the spatial position of genetic discontinuities.

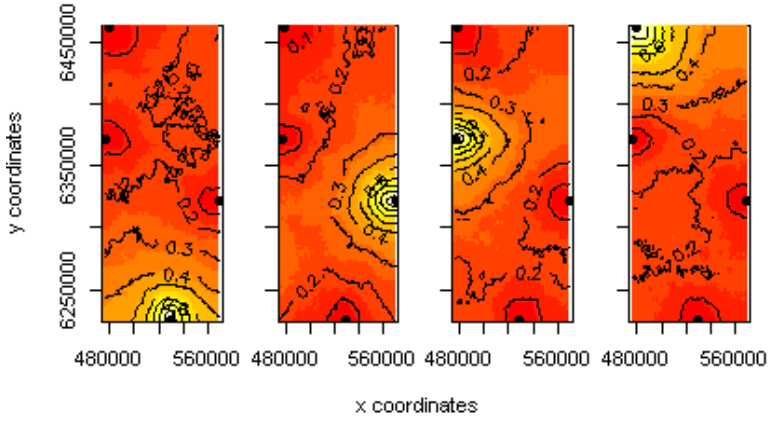


Figure S2.5 Identifying the number of *K* in the data from analysis in STRUCTURE. We assessed the convergence of mean values for the log probability of the data (Ln P(D)) and used the Evanno et al. (2005) method to determine the number of clusters. Data consists of the adult cohort at 13 sites.

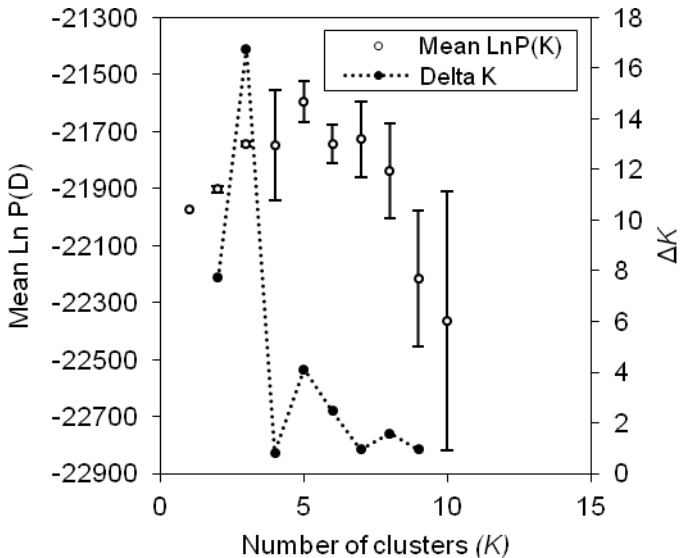
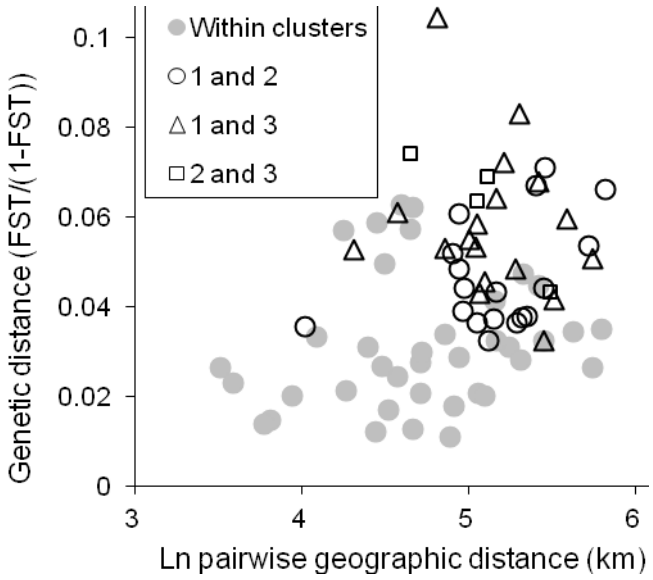


Figure S2.6 The relationship between geographic and genetic distance partitioned into comparisons within and between GENELAND clusters. Grey circles are comparisons of sites within the same cluster. Open symbols represent comparison of sites in different clusters, i.e. between cluster 1 and 2 (open circles), 1 and 3 (open triangles), and 2 and 3 (open squares).



Appendix S3

Table S3.1 Details of study sites and isolation indices

				Area (ha)				Distance (km)	
Site name	**Code**	**Lat. Long.**	**Elev.**	**5km**	**10km**	**15km**	**Site Bound-ary**	**CB**	**BB**
Söderåsen	SOD	N56.0264 E13.2235	321	3649.95	5650.5	8152.86	6181.65	3.758	0.067
Ryssberget	RYS	N56.0674 E14.5903	73	2861.01	3810.39	5593.13	8396.37	4.451	0.338
Häckeberga Sjön	HAC	N55.5733 E13.4131	16	2342.50	2858.21	2628.66	1651.73	1.657	0.083
Osbecks Bokskoggar	OSB	N56.4119 E12.9794	47	906.21	2952.2	2859.55	319.90	1.747	0.065
Tromtö	TRO	N56.1684 E15.4698	31	575.27	1656.88	3423.32	190.37	1.180	0.053
Biskorpstorp	BIS	N56.8023 E12.8940	151	727.13	1069.22	2792.19	182.91	2.285	1.561
Flahult	FLA	N56.9738 E13.8226	109	244.19	306.35	918.71	55.92	1.401	1.070
Gullmarsberg	GUL	N58.3745 E11.6514	149	432.07	674.18	360.48	96.71	0.654	0.217
Stoms Ås	STO	N57.5509 E12.5560	195	511.56	744.54	0.01	213.50	3.076	1.873
Måråsa	MAR	N57.0382 E13.2366	156	239.51	223.73	540.27	65.25	1.186	0.827
Hornsö Ekopark	HOR	N57.0338 E16.1402	148	623.69	84.54	8.44	361.28	1.465	0.377
Garpäror	GAR	N58.4955 E13.8352	103	119.92	65.86	0.00	117.60	1.596	1.139
Mattarp	MAT	N57.4927 E14.6146	96	75.38	0.00	3.98	60.31	4.864	4.290
Omberg Ekopark	OMB	N58.2976 E14.6473	169	32.94	0.00	0.00	32.94	50.804	50.429

The first three letters of the site names were abbreviated to give a site code (*Code*). Geographic coordinates (*Lat. Long.*) are presented in decimal degrees with elevation (*Elev*.) in metres. The sites are ordered in terms of the sum area of beech within all the buffer zones. Area-based measurements of beech in 5km, 10km, 15km exclusive buffer zones, and site boundary are grouped under *Area (ha)*, with the centre to boundary (*CB*) and boundary to boundary (*BB*) distance-based measures grouped under *Distance (km)*.

Table S3.2 Results of the bottleneck analysis. P-values indicate the significance of $H_S > H_{eq}$ (where, H_s is gene diversity, and H_{eq} is the heterozygosity expected under mutation-drift equilibrium) using the Wilcoxon test, which is indicative of a recent reduction in population effective size.

		Wilcoxon P-value	
Site name	**Code**	**Adults**	**Seedlings**
Söderåsen	SOD	0.51709	0.382324
Ryssberget	RYS	0.991943	0.86084
Häckeberga Sjön	HAC	0.949219	0.839844
Osbecks Bokskoggar	OSB	0.996582	0.92627
Tromtö	TRO	0.991943	0.989502
Biskorpstorp	BIS	0.997559	0.998779
Flahult	FLA	0.995361	0.938477
Gullmarsberg	GUL	0.997559	0.998779
Stoms Ås	STO	0.681152	0.711426
Måråsa	MAR	0.989502	0.86084
Hornsö Ekopark	HOR	0.999268	0.999756
Garpäror	GAR	0.973145	0.793457
Mattarp	MAT	0.449219	0.415527
Omberg Ekopark	OMB	0.998779	0.839844