Supplementary Text S1

The old split time estimate between northern and southern Sweden in relation to previous results

The split time between northern and southern Sweden (estimated from the secondaryContact6 model) is $\sim$ 150 kya (125-180 kya). This is much older than what was estimated by François et al. (2008) (approx. 14 kya), but in the range of the split time between Spanish and Italian *A. thaliana* as estimated by Mathew et al. (2013) (83 kya), considering that F_{ST} for those two populations is smaller than between northern and southern Sweden. Beck et al. (2008) use data from 475 individuals, analyzing a chloroplast sequence region (no recombination). They fit the mismatch distribution to one modelled under a hypothesis of historical demographic expansion and thereby inferred the timing of the expansion. They arrive at an old age of the expansion (122 kya, lower CI: 74 kya, upper CI: 156 kya), which lies within the last interglacial (Eemian interglacial from about 114 - 130 kya). Similar to Beck et al. (2008), we also estimate a population expansion for the southern Swedish population at that time (1.6-fold expansion).

Our lower bound for the split time between northern and southern Sweden assuming an upper estimate of mutation rate is 100 kya and thus clearly predates the last glacial maximum about 20 kya, suggesting that northern Sweden and southern Sweden have been colonized from different refugia. Selection and population structure can lead to a bias of estimated split times. However, when we account for selection, isolation by distance and sequencing errors, the estimated split time does not change much, and excluding the 22 sweep regions of Long et al. (2013) makes it even older (see Table S4).

Comparison to estimates from the literature

François et al. (2008) acknowledge the fact that northern Sweden (+Finland) is highly diverged and therefore does not fit a simple model of range expansion. Here is a list of aspects of the estimation procedure of François et al. (2008) that are different from ours and that might explain the discrepancy:

- They use intergenic and intronic data, which are more effected by direct selection than synonymous data (Kim et al., 2007).
- They use the "mean number of distinct haplotypes" and "mean number of private haplotypes" to estimate split time and migration rate. We use the joint Site Frequency Spectrum (jSFS).
- Their model choice is quite restrictive. They fix N_e for Central Europe to 135,000 and to $135,000 \times 1/4$ for northern Europe (we estimate a similar ancestral N_e, but 1/10 in northern Sweden, not 1/4). They do not state how they derive the value of 1/4. However, this might have a strong influence on estimation results.
- They assume symmetric backward migration (we reject symmetrical migration).
- They assume a population expansion some time ago (model C) for both the Central European and the Northern European population ("The growth scenario was assumed to be the same in the two populations, with only the population sizes differing"). We only find an expansion in southern Sweden, but a size reduction at split time in northern Sweden.
- Their approach is highly sequential, first settling on "model C", which assumes a population size expansion with certain parameters. Then this "model C" is used to calculate a split time (model C is assumed to be true in both the Central European and the Northern European population), and then this split time is used to estimate migration rate. We fit parameters simultaneously, which is more appropriate when parameters are not independent from each other.

- As an example of such independence of parameters, the effect of older split times on their statistics can be reverted by higher migration rates (see François et al. (2008), Figures 6 and 7). They do not provide a confidence interval for their split time estimate, although we expect it to be large. E.g. the confidence interval for the time of expansion in their "model C" is quite broad (from 5 kya to 117 kya).
- To compute split time in generations, they use an estimate of population recombination rate of 0.3 (per kb) to scale the coalescent time estimates. This leads to a relatively large mutation rate (2.2×10^{-8} per bp per generation). The estimate from mutation accumulation experiments (7×10^{-8}; Ossowski et al. (2010)) or estimates from a phylogenetic approach (0.38-0.86 x 10-8; Huang et al. (2012)) are about 3 times smaller.

Mathew et al. (2013) use the joint Site Frequency Spectrum to estimate the split time between Spanish and Italian *A. thaliana*, using part of the 80 genomes from Cao et al. (2011) and also obtain a fairly old time of 83 kya. They interpret their result as indicating that the southern European populations have split long before the last glacial maximum, and that it is unlikely that the ancestors of both populations survived the last glaciation in a common refugium. Similar to us, they assume a mutation rate of 7×10^{-8}. Furthermore, they use three different "selection classes": first and second codon (FS), third codon (TR) and noncoding (NC), but don't find different results for those classes. Also, using a finite site model instead of an infinite site model does not influence results. They assume a very simple model of no size change (neither at split time nor subsequently) and symmetric migration rates. The only parameters they estimate are migration rate, theta, split time and some rate heterogeneity parameter for mutation rate across the genome. They arrive at a migration rate of 3.4 and a split time of 0.16 ($2N_e$). We find a split time between northern Sweden and southern Sweden of 0.3 ($2N_e$).

A very old split time between northern and southern Sweden also makes sense in the light of the large genetic distance between those populations, given their small geographic distance. In Long et al. (2013) it is shown that F_{ST} between northern and southern Sweden (distance: 800km) is about the same as F_{ST} between samples from Spain and Central Asia (distance: 6400km).

Results for fitting the model of François et al. (2008) to our jSFS data with $\delta a \delta i$

The François et al. (2008) model starts with a population size of 73,750, then splits into two populations with sizes 59,000 (South) and 14,750 (North) for a duration of 1,500 generations. Note that $N_e(North) = 1/4 \times N_e(South)$. Next both populations undergo exponential growth at the same rate for 7,000 generations, resulting in population sizes of 135,000 (South) and 33,750 (North), followed by a period of constant size until present. Finally, there is constant and continuous migration with rate 3 (scaled by 135,000), which in units used in $\delta a \delta i$ is a migration rate of 3.28 (normalized by $2N_0$, where N_0 is 73,750).

The ($\delta a \delta i$-scaled) parameters for the split time and the migration rate are 0.092 and 3.28, respectively. The log likelihood of the model given our jSFS data (without any optimization) is -35,491. Compare this to the log likelihood of our secondaryContact6 model of -2,597.

Fitting the François et al. model to our jSFS data by optimizing the split time and migration rate instead of assuming that they are fixed, results in a split time of 0.1984 and a migration rate of 1.81 (again in $\delta a \delta i$-units, which are $2N_0$ for times and $2N_0$ for migration rates). Assuming a mutation rate of 7×10^{-8}, we would estimate ancestral N_e to be 138,515 instead of 73,750. This results in a split time of about 55,000 years, clearly older than the 13,500 years of François et al., but also much younger than our estimates. This indicates that the restrictive model specification of François et al. has a strong influence on the split time estimation. The fit is better than before (LL = -4,560), but still worse than secondaryContact6 (LL = -2,597), which has unequal migration rates, a population size ratio of 10 between North and South, and secondary contact.

References

Beck, J. B., Schmuths, H., and Schaal, B. A. Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. *Molecular Ecology*, 17(3): 902–915, 2008.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., and Weigel, D. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, 43(10):956–963, 2011.

François, O., Blum, M. G. B., Jakobsson, M., and Rosenberg, N. A. Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genet*, 4(5):e1000075, 2008.

Huang, C.-C., Hung, K.-H., Wang, W.-K., Ho, C.-W., Huang, C.-L., Hsu, T.-W., Osada, N., Hwang, C.-C., and Chiang, T.-Y. Evolutionary rates of commonly used nuclear and organelle markers of *Arabidopsis* relatives (Brassicaceae). *Gene*, 499(1):194–201, 2012.

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., Ecker, J. R., Weigel, D., and Nordborg, M. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics*, 39 (9):1151–1155, 2007.

Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B. J., Korte, A., Nizhynska, V., Voronin, V., Korte, P., Sedman, L., Mandáková, T., Lysak, M. A., Seren, Ü., Hellmann, I., and Nordborg, M. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nature Genetics*, 45(8):884–890, 2013.

Mathew, L. A., Staab, P. R., Rose, L. E., and Metzler, D. Why to account for finite sites in population genetic studies and how to do this with Jaatha 2.0. *Ecology and Evolution*, 3(11):3647–3662, 2013.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D., and Lynch, M. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, 327(5961):92–94, 2010.

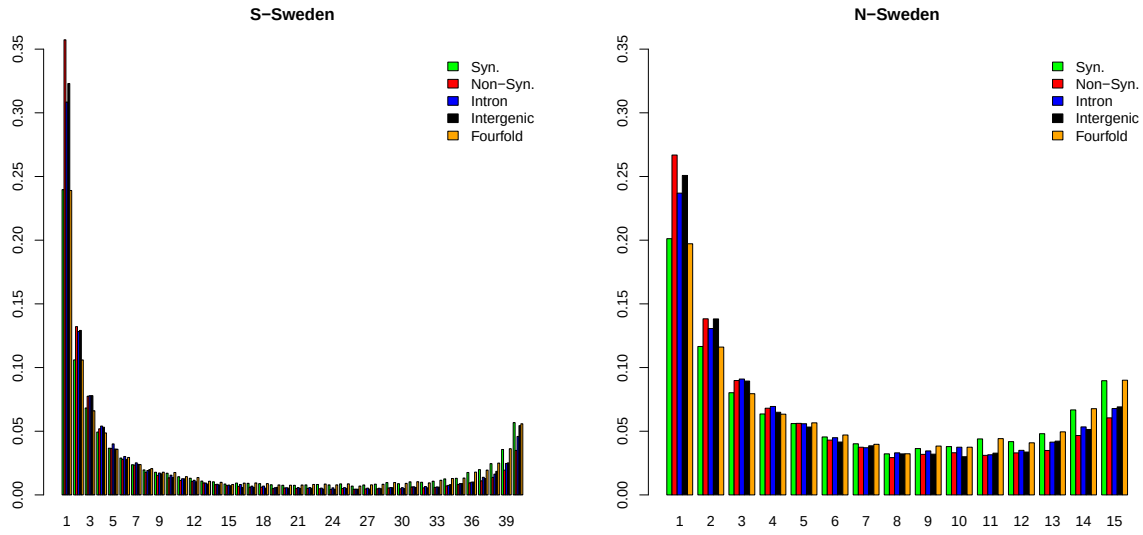

Figure S1: **Site frequency spectrum for 5 classes of sites.** SFS for synonymous, non-synonymous, intronic, intergenic and fourfold degenerate SNPs, for the southern Swedish sample and the northern Swedish sample.

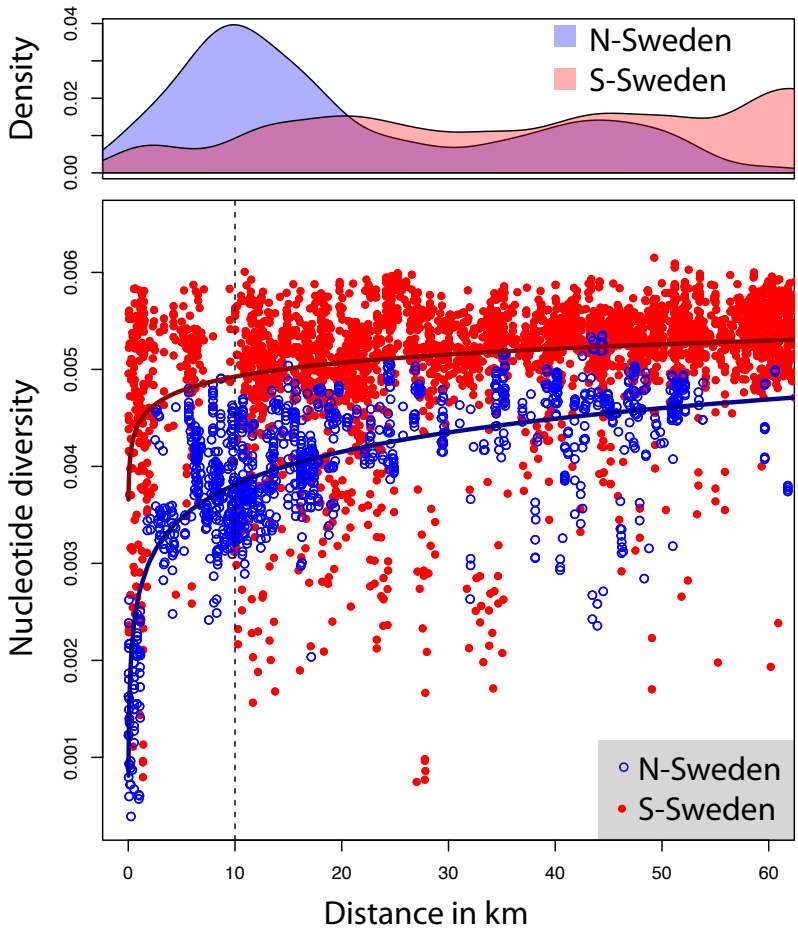

Figure S2: **Pairwise nucleotide diversity (the average number of pairwise differences between chromosomes per base pair) as a function of geographical distance.** A function of the form nucleotide diversity $= b_0 + b_1 \times$ log(geographical distance) is fitted to the data from northern Sweden and southern Sweden. A density plot of pairwise distances is plotted on top. There is significantly lower nucleotide diversity in northern Sweden compared to southern Sweden, even after accounting for geographic distance.

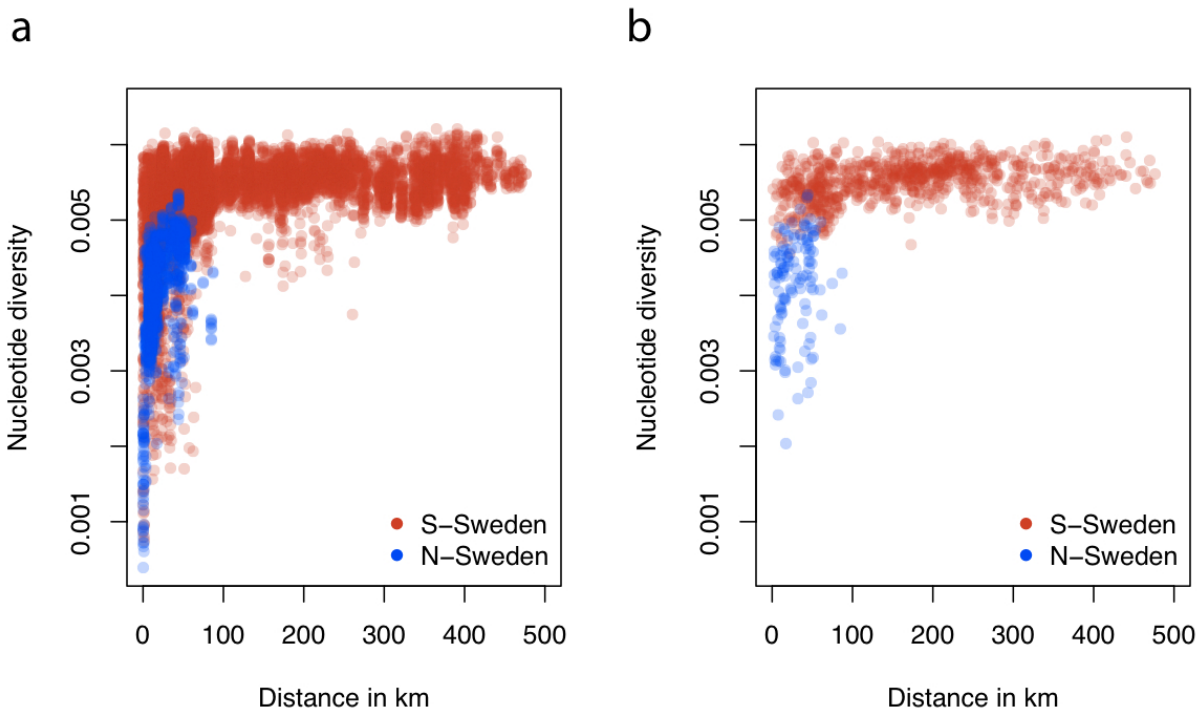

Figure S3: **Pairwise nucleotide diversity vs. physical distance in northern Sweden and southern Sweden.** a) Before subsampling, there is a clear pattern of isolation by distance. b) After subsampling, the pattern of isolation by distance is strongly reduced.

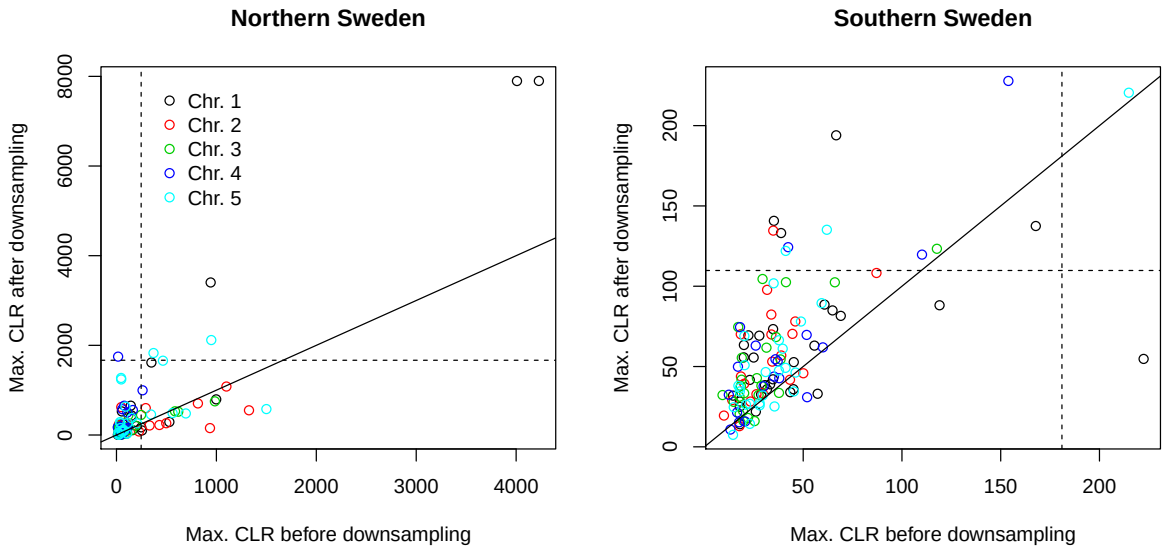

Figure S4: **The effect of subsampling on CLR peaks.** Correlation of largest CLR value in 1Mb windows before sub-sampling and after sub-sampling. Dashed lines indicate 99% significance thresholds.

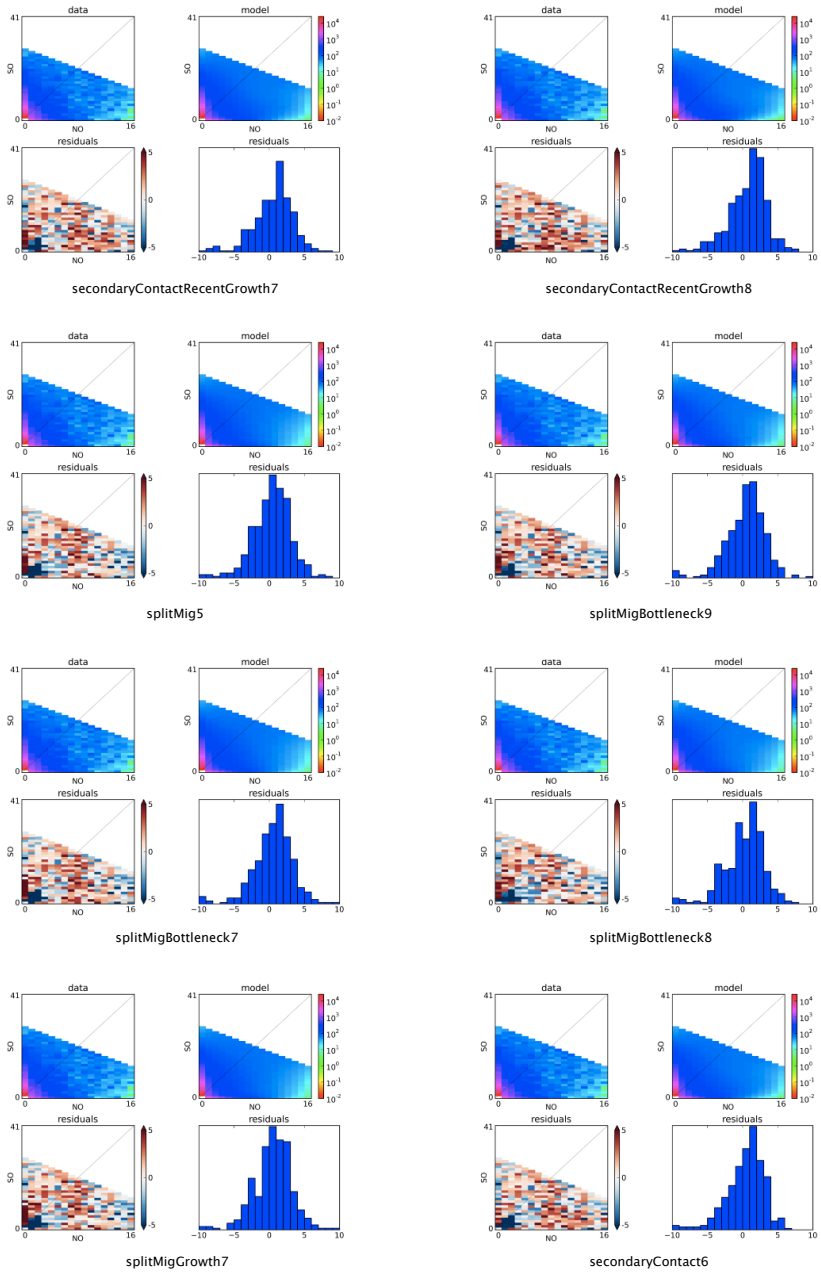

Figure S5: **Expected joint site frequency spectrum and residual plot of "data minus model" of the 8 models that have AIC<4000.** In all 8 cases, overall fit is good, however there are too many polymorphisms that are at low frequency in both populations (dark blue color in residual plots).

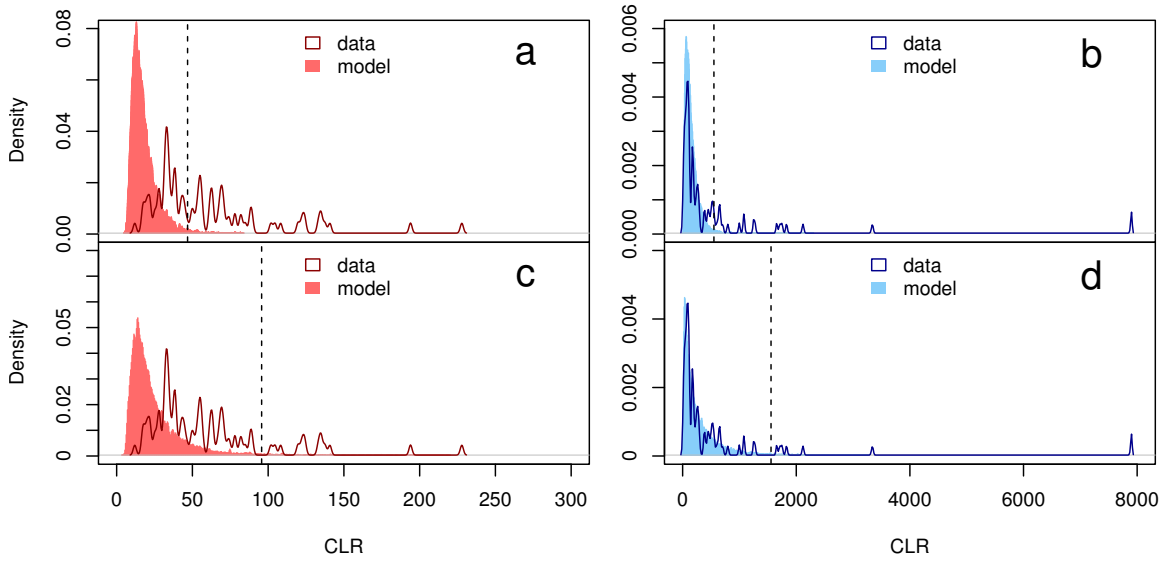

Figure S6: **Distribution of maximum CLR values in 1Mb windows, from simulations and data.** Simulations based on the splitMig5 model for southern Sweden (a) and northern Sweden (b). Simulations based on the splitMigBottleneck8 model for southern Sweden (c) and northern Sweden (d). The dashed line indicates the 99% statistical cutoff (47, 551, 96 and 1556 in a, b, c and d respectively).

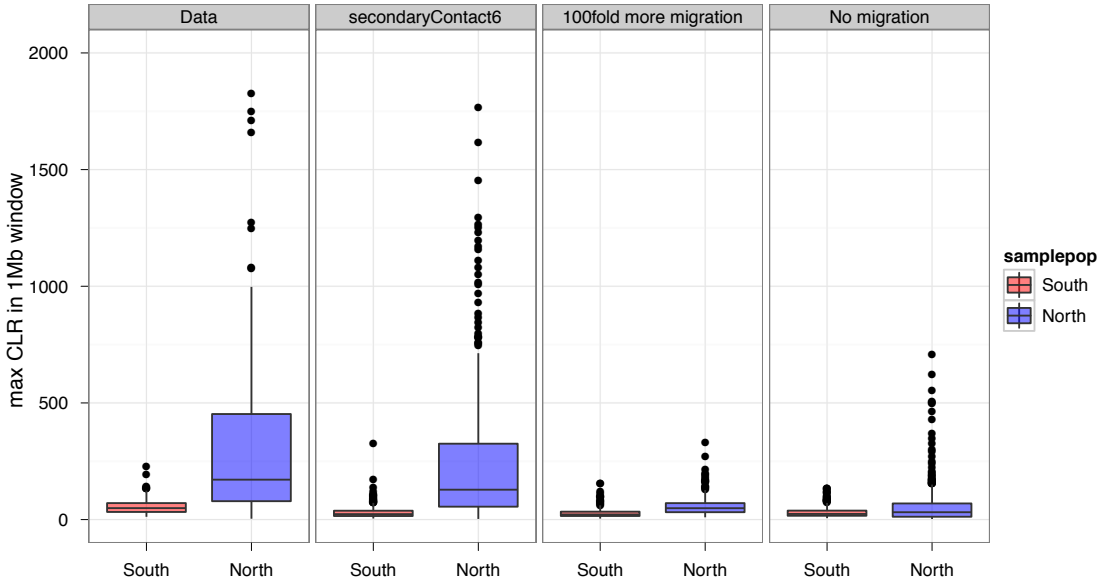

Figure S7: **Distribution of maximum CLR values from 1Mb windows for data and simulations.** The difference in the CLR distribution between North and South is a function of migration rate. The right amount of migration is crucial to generate the pattern that is observed in the data, indicating that sample size differences and effective population size difference can not explain the difference in the distribution of the CLR values between North and South alone.

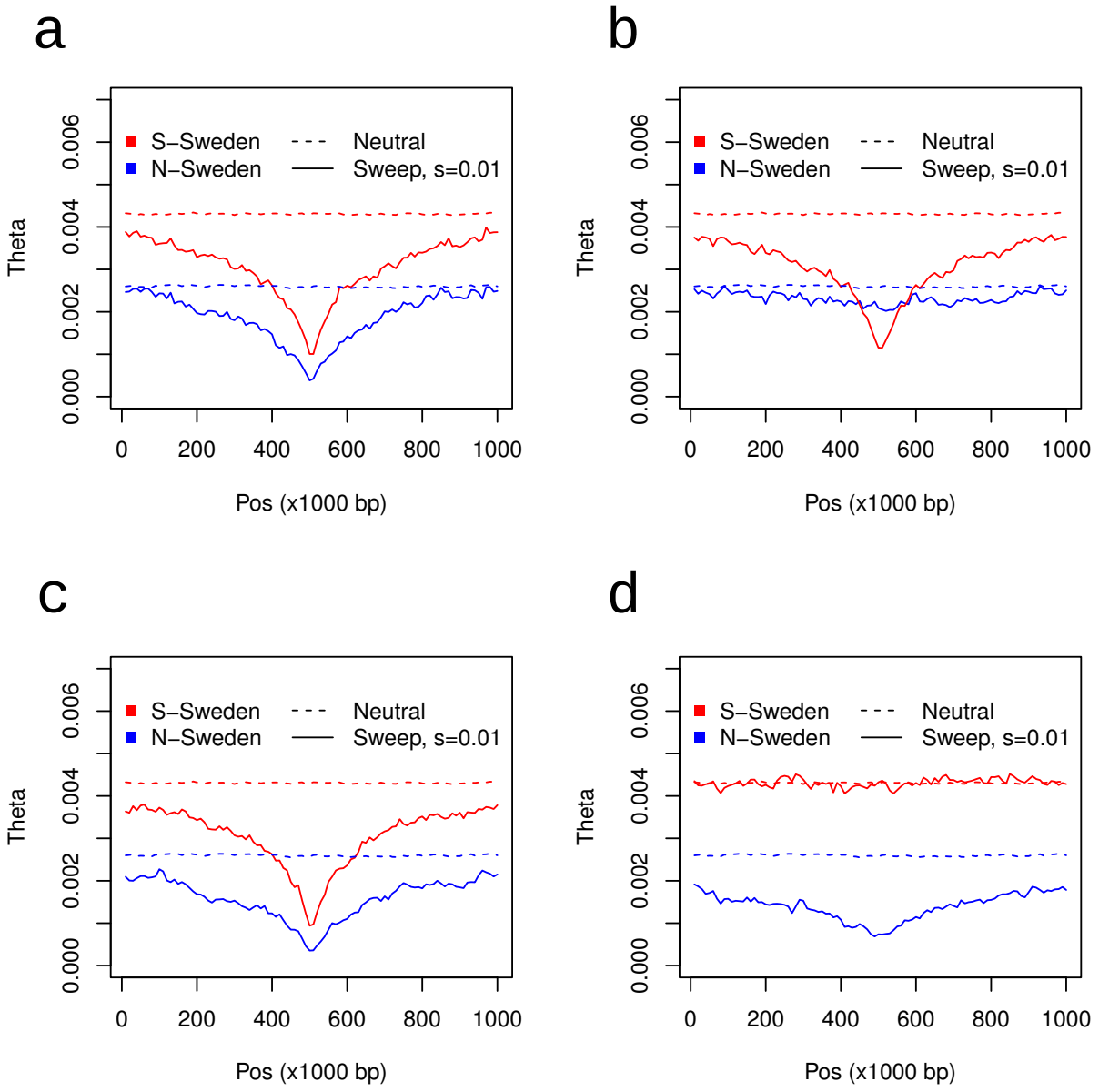

Figure S8: **Reduction in diversity (Watterson's theta) across a 1Mb long sequence for selective sweeps ($2N_e s = 200$) in the middle of the sequence.** Results for northern Sweden are in blue, for southern Sweden in red. The simulated sweeps start at 0.05 coalescent times in the past. a) Global selection, selected mutation starts in southern Sweden. b) Local selection in southern Sweden, selected mutation starts in southern Sweden. c) Global selection, selected mutation starts in northern Sweden. d) Local selection in northern Sweden, selected mutation starts in northern Sweden.

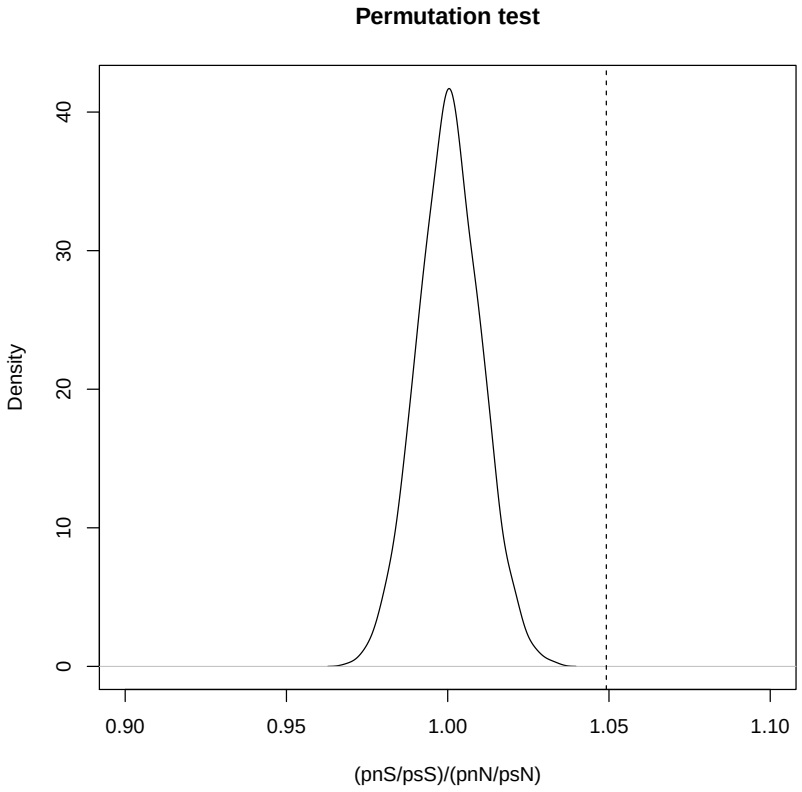

Figure S9: **Permutation test for the (pnS/psS)/(pnN/psN) ratio.** pnS and psS are the non-synonymous and synonymous polymorphisms in southern Sweden, pnN and psN are the non-synonymous and synonymous polymorphisms in northern Sweden. The null distribution is calculated by permutation of the labels "North" and "South" and recalculating the ratio for a total of 1,000 replications. The ratio given the true labeling is 1.049 and therefore clearly an outlier of this null distribution. To account for differences in sample size, the southern Swedish sample was down-sampled to the same size as in northern Sweden (16).

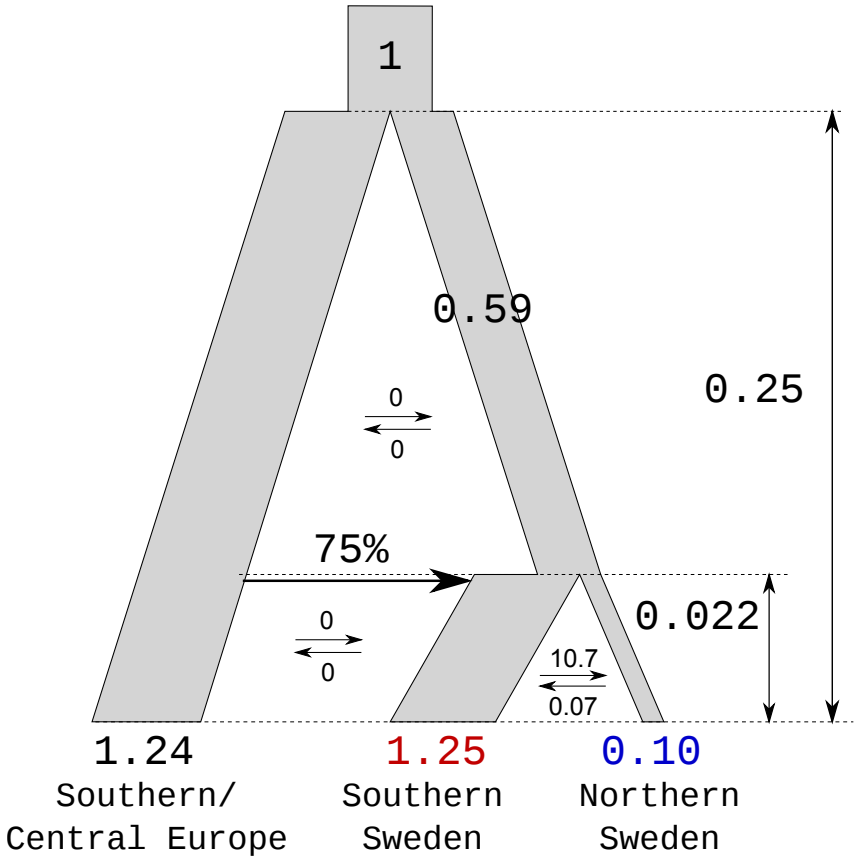

Figure S10: **Model scheme and parameter estimates for the splitAdmixture9 model.** This model assumes that the strong differentiation between southern Sweden and northern Sweden is caused by a recent admixture event from Central Europe. Migration rates are given as $2N_0m$, times in units of $4N_0$ and population sizes in units of N_0. The admixture event happens almost immediately after the split into northern and southern Sweden, and the admixture proportion of 75% means that 75% of the southern Swedish population at that time is replaced by immigrants. Note that the estimated split time between Southern/Central Europe and the ancestral Swedish population is still old (132,000 years).

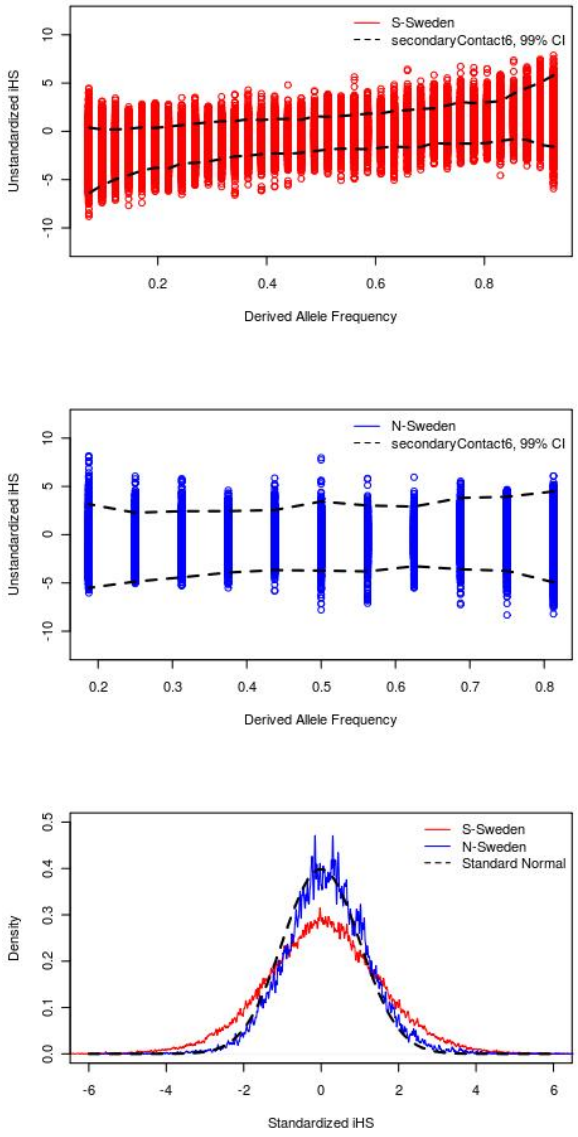

Figure S11: **The iHS distribution for southern Sweden and northern Sweden, compared to simulations from the secondaryContact6 model.** The thicker tail in the distribution of the unstandardized iHS compared to the simulations in southern Sweden suggests the occurrence of partial selective sweeps. For significance testing, SNPs were LD pruned ($r^2 < 0.1$) to avoid violation of test assumptions due to auto-correlation. We found a significant deviation of the standardized iHS from normality in southern Sweden (Anderson-Darling normality test, $p = 4.0 \times 10^{-7}$), but not in northern Sweden ($p = 0.48$). Note that the standardized iHS is using the mean and standard deviation of simulated values of iHS from the secondaryContact6 model. That way, iHS outliers are outliers relative to the simulation-based distribution of iHS.

Table S1: **Parameter range for inference.** Lower and upper limit of parameter values for finding the maximum likelihood parameter combination with $\delta a \delta i$. Starting parameters were sampled uniformly from log transformed parameters in the defined ranges.

Parameter	Unit	Lower limit	Upper limit
Time	$2N_0$	0	10
Migration rate	$2N_0 m_{i,j}$	0	100
Population size	N_0	0.001	200
Bottleneck strength	Relative reduction	0.001	0.999

Table S2: **Analysis of the fit of a two vs. a ten deme models.** Four different types of Models with two or ten demes were simulated and summary statistics (F_{ST} and the diversity ratio between North and South) were calculated. A + indicates the average summary statistic produced by the model fits the observed in the data (for some parameter combination). Model A: range expansion, Model B: lower population size in northern Sweden, Model C: higher migration rate from North to South, Model D: range expansion with South sampled from a deme in the middle of the expansion. Note that two deme models fit well and better than most ten deme models. Therefore we decided to explore the two deme models further.

Model	# demes	TajD S	TajD N	F_{ST}	Diversity N/S	Best fitting parameters
A	10		+	+	+	Bottleneck size: 1000; Migration rate: 10
A	2	+	+	+	+	Bottleneck size: 10; Migration rate: 1
B	10	+		+	+	Ratio demesize N vs. S: 0.3; Migration rate: 70
B	2	+	+	+	+	Ratio demesize N vs. S: 0.3; Migration rate: 6
C	10	+	+	+	+	Migration asymmetry N-S: 0.3; Migration rate: 30
C	2		+	+	+	Migration asymmetry N-S: 0.05; Migration rate: 3
D	10			+	+	Bottleneck size: 10; Migration rate: 10
D	2	+		+	+	Bottleneck size: 8; Migration rate: 5

Table S3: **Robustness of model selection.** The second best fitting model (splitMigBottleneck8) and a simpler model (splitMig5) where compared to the best fitting model (secondaryContact6). Data is generated under a certain model (rows) and it is counted how often the secondaryContact6 model (first column), the splitMig5 model (second column) or the splitMigBottleneck8 model (third column) is the best fitting model (lowest AIC). All steps of model fitting and AIC calculation were done in the same way as it was done for the actual data.

Selected model	secondaryContact6	splitMig5	splitMigBottleneck8
secondaryContact6	15	0	5
splitMig5	6	1	13
splitMigBottleneck8	5	0	15

Table S4: **Robustness of time estimates regarding filtering of data.** Times were estimated by fitting the expected jSFS of the secondaryContact6 model to the jSFS of the data after filtering steps. The various steps of filtering did not reduce the time estimates of the split between northern Sweden and southern Sweden or the time of the secondary contact strongly.

Type of data	Split time (kya)	Lower CI	Upper CI	Secondary contact (kya)	Lower CI	Upper CI
Unfiltered, all plants	154	141	167	59	50	69
Plants subsampled (distance >2km to avoid IBD)	133	122	143	35	27	43
Plants subsampled + only Synonymous sites + no singletons	153	124	182	39	19	59
Plants subsampled + only Synonymous sites + no singletons + excluding 22 sweep regions	189	150	228	48	21	74

Table S5: **Commands that were used with the coalescent simulation software msms for the simulations in figures 6, 7, 8 and 10.**

Figure 6	secondaryContact6: msms -ms 57 1 -t 3000 -r 200 -I 2 41 16 -n 1 1.655290 -n 2 0.168389 -ma x 2.451840 11.804420 x -ema 0.078107 2 x 0 0 x -ej 0.308381 2 1 -en 0.308381 1 1
Figure 7	Standard Neutral Model South: msms -ms 41 1 -t 4500 -r 200 Standard Neutral Model North: msms -ms 16 1 -t 2500 -r 200 secondaryContact6: msms -ms 57 1 -t 3000 -r 200 -I 2 41 16 -n 1 1.655290 -n 2 0.168389 -ma x 2.451840 11.804420 x -ema 0.078107 2 x 0 0 x -ej 0.308381 2 1 -en 0.308381 1 1
Figure 8	secondaryContact6: msms -ms 57 1 -t 3000 -r 200 -I 2 41 16 -n 1 1.655290 -n 2 0.168389 -ma x 2.451840 11.804420 x -ema 0.078107 2 x 0 0 x -ej 0.308381 2 1 -en 0.308381 1 1 100fold more migration: msms -ms 57 1 -t 3000 -r 200 -I 2 41 16 -n 1 1.655290 -n 2 0.168389 -ma x 245.1840 1180.4420 x -ema 0.078107 2 x 0 0 x -ej 0.308381 2 1 -en 0.308381 1 1 No migration: msms -ms 57 1 -t 3000 -r 200 -I 2 41 16 0 -n 1 1.655290 -n 2 0.168389 -ej 0.308381 2 1 -en 0.308381 1 1
Figure 10	a) + c) Global selection: msms -ms 57 1 -t 3000 -r 200 -I 2 41 16 -n 1 1.655290 -n 2 0.168389 -ma x 2.451840 11.804420 x -ema 0.078107 2 x 0 0 x -ej 0.308381 2 1 -en 0.308381 1 1 -N 10000.0 -SFC -SI 0.05 2 0.0 0.000296931509778 -Sc 0 -1 200.0 -Sp 0.5 a) + c) Local selection North: msms -ms 57 1 -t 3000 -r 200 -I 2 41 16 -n 1 1.655290 -n 2 0.168389 -ma x 2.451840 11.804420 x -ema 0.078107 2 x 0 0 x -ej 0.308381 2 1 -en 0.308381 1 1 -N 10000.0 -SFC -SI 0.05 2 0.0 0.000296931509778 -Sc 0 2 200.0 -Sp 0.5 b) + d) Global selection: msms -ms 57 1 -t 3000 -r 200 -I 2 41 16 -n 1 1.655290 -n 2 0.168389 -ma x 2.451840 11.804420 x -ema 0.078107 2 x 0 0 x -ej 0.308381 2 1 -en 0.308381 1 1 -N 10000.0 -SFC -SI 0.12 2 3.02061874354e-05 0.0 -Sc 0 -1 50.0 -Sp 0.5 b) + d) Local selection South: msms -ms 57 1 -t 3000 -r 200 -I 2 41 16 -n 1 1.655290 -n 2 0.168389 -ma x 2.451840 11.804420 x -ema 0.078107 2 x 0 0 x -ej 0.308381 2 1 -en 0.308381 1 1 -N 10000.0 -SFC -SI 0.12 2 3.02061874354e-05 0.0 -Sc 0 1 50.0 -Sp 0.5

Table S6: **Ranks for the significant sweep regions of table 4 in the main text.** CLR for N-Sweden and S-Sweden, F_{ST} and respective genome-wide ranks. Ranks are calculated from non-overlapping 1Mb windows, *i.e.* the 1Mb window that contains the smallest value has rank 120.

Chr	**Pos** x10^3[a]	**CLR** North	**CLR** South	F_{ST}[b]	Rank **CLR** North	Rank **CLR** South	Rank F_{ST}
1	11,417	97	133	0.223	77	6	71
1	12,855	249	194	0.313	44	2	49
1	19,020	1845	56	0.453	2	46	19
1	20,009[c]	6217	127	0.679	1	49	3
1	24,521	79	141	0.160	94	3	97
2	13,549	34	135	0.139	106	5	106
3	14,961	175	123	0.148	54	7	102
4	5,552	160	228	0.111	61	1	113
4	6,637	1748	55	0.246	5	50	64
4	9,374	245	120	0.156	65	9	99
5	2,228	1658	89	0.545	6	92	7
5	5,780	110	122	0.320	90	8	47
5	6,748	2118	69	0.539	3	36	9
5	19,815	633	135	0.679	29	4	2
5	26,166	1829	25	0.373	4	104	34

[a] positions of the putative sweep regions are rounded to kb. [b] largest value of 100kb windows within 1Mb around the CLR peak. [c] sweep mutation is a transposition that is collapsed to a single bp for calculation of F_{ST} and CLR.

Table S7: Power and mean signal strength for a soft sweep starting from standing variation with 5% and 1% frequency, assuming the secondaryContact6 demographic model. The start time of the sweep is in units of $4N_0$ generations.

Start freq.	s	Start time of sweep	Population	Origin of sel. Mutation	Mean CLR North	Power North	Mean CLR South	Power South
0.05	0.0025	0.12	Global	South	200	0	63	0.12
0.05	0.0025	0.12	Global	North	402	0.01	252	0.67
0.05	0.0025	0.12	South	South	278	0	42	0.05
0.05	0.0025	0.12	South	North	292	0.01	155	0.36
0.05	0.0025	0.12	North	South	310	0	32	0.02
0.05	0.0025	0.12	North	North	709	0.07	32	0.01
0.05	0.01	0.05	Global	South	336	0.04	36	0.03
0.05	0.01	0.05	Global	North	535	0.02	529	0.75
0.05	0.01	0.05	South	South	301	0.01	35	0
0.05	0.01	0.05	South	North	314	0	818	0.9
0.05	0.01	0.05	North	South	584	0.07	31	0.02
0.05	0.01	0.05	North	North	601	0.08	30	0
0.01	0.0025	0.12	Global	South	226	0	168	0.43
0.01	0.0025	0.12	Global	North	522	0.04	275	0.75
0.01	0.0025	0.12	South	South	213	0.01	146	0.42
0.01	0.0025	0.12	South	North	257	0.01	199	0.54
0.01	0.0025	0.12	North	South	315	0.01	30	0.01
0.01	0.0025	0.12	North	North	574	0.06	24	0.01
0.01	0.01	0.05	Global	South	363	0	74	0.08
0.01	0.01	0.05	Global	North	589	0.06	901	0.91
0.01	0.01	0.05	South	South	280	0.02	146	0.2
0.01	0.01	0.05	South	North	350	0.03	879	0.96
0.01	0.01	0.05	North	South	568	0.08	31	0.01
0.01	0.01	0.05	North	North	1008	0.23	28	0