Supplementary Information 3: *Antirrhinum* Hybrid Zone Data

In the main text we apply our method to a data set from a hybrid zone between two sub-species of the model plant *Antirrhinum majus*. In the following, we describe sampling, genotyping and filtering criteria we used for selecting data for our analysis.

Sampling and Data Collection

As part of a long-term project examining wild pedigrees and geographic clines, each year from 2009 to 2014 we sampled plants from a hybrid zone between *Antirrhinum majus pseudomajus* and *Antirrhinum majus striatum* in the Spanish Pyrenees located in val di Ribes near the village of Planoles. Individual plants are primarily found within 100 m either side of two roughly parallel roads that run up the valley (Fig. 10). The sampling concentrated on a ≈ 4 km transect encompassing the center of the phenotypic and genetic clines involved in magenta and yellow flower pigmentation (ROSEA; Whibley *et al.* (2006), and SULFUREA; Desmond et al 2017, respectively[1]) and some of the flanking regions in which the pure subspecies mostly reside.

The following data were obtained for each plant:

- A global positioning system (GPS) coordinate
- Leaf material (for DNA extraction)

The collection occurred between May and July, for individuals with open flowers only. Individual's geographic coordinate was collected using a GPS device (Trimble GeoXT datalogger) with a mean accuracy of ≈ 2 m. Four to six fresh leaves from each plant were stored in individual glassine envelope bags, which were placed within a plastic bag containing silica gel (Fisher Scientific) for drying the leaf tissue. Components of the magenta and yellow color of the flowers were scored in the field according to Whibley *et al.* (2006).

SNP Genotyping

The KASP genotyping platform (LGC genomics) was used to genotype single nucleotide polymorphisms (SNPs) across the *Antirrhinum* genome. In total, we designed ≈ 240 SNP at a subset of polymorphic and divergent loci, but here report just on a subset of 60 polymorphic loci. The remaining markers that clearly violate our model assumptions have been filtered out (see below).

Candidate loci were identified using a draft *A. majus* reference genome (≈ 630 Mb across eight linkage groups; courtesy of Yongbiao Xue, BGI) and allele frequencies obtained from whole-genome Illumina PoolSeq of six pools of n = 50 individuals located along a transect through the hybrid zone (unpublished data). All potential SNP loci were identified across the genome with a custom Python script SNPextract.py (`https://github.com/dfield007/genomics_general`) which

[1] Add citation in final version, this paper is accepted but not published yet

identified SNPs positions suitable for KASP genotyping platform (LGC Genomics). The script was run with the following parameters: (i) 30 < depth < 300 in all pools at the focal SNP (to reduce the probability of false positives and paralogs), (ii) 30 < depth < 300 for sequences 50bp upstream and downstream of focal SNP, (iii) <3 other SNPs within 50bp (to ensure primer efficiency), and (iv) biallelism (a KASP requirement). We also selected loci on the basis of being polymorphic in the hybrid zone ($0.3 < \bar{p} < 0.7$) and selected one locus randomly every couple of mapping units (cM) to maximize marker independence (Figure S1). For each candidate, the script extracted the 100bp sequence surrounding each candidate polymorphic site required to design the SNP primers. DNA extractions and SNP genotyping were carried out by LGC Genomics. Replicate DNA extractions and genotyping confirmed relatively low error rates of the KASP platform (mean error rate < 0.1% per locus).

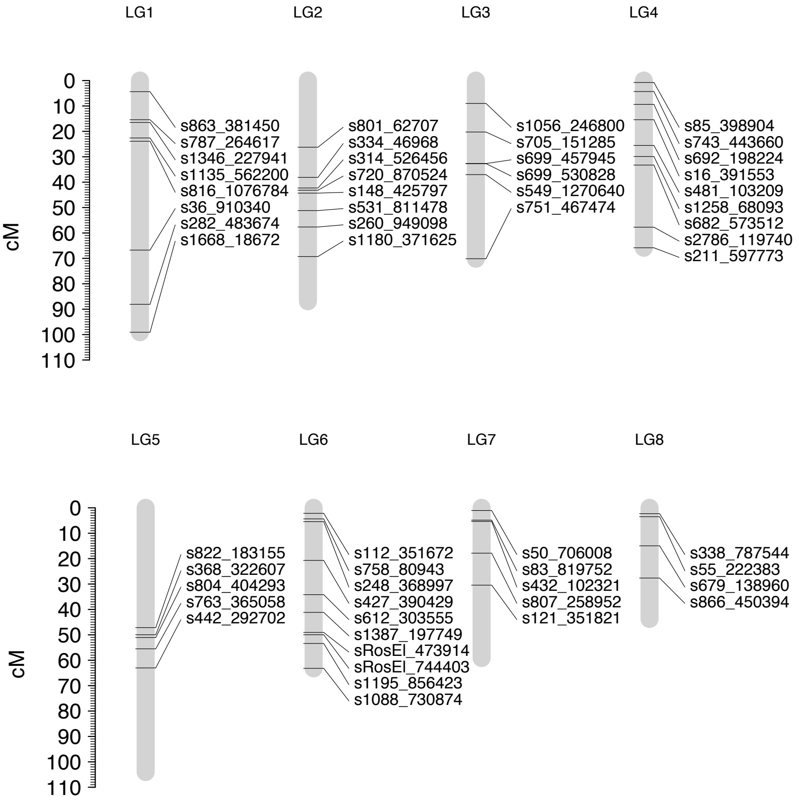

Figure S1: Position of KASP SNPs used in our analysis on draft *Antirrhinum majus* linkage map. Positions of 56 loci (black lines) indicated, the remaining 4 SNP loci could not be placed on the linkage map.

Data Filtering

Our method requires all individuals to have no missing genotype data. Starting with $n = 13722$ individuals in the core hybrid zone ($\approx \pm 2$km around flower color transition), we first removed individuals with more than 8 missing genotypes ($n = 246$). Next, we identified individuals with at most one genotype mismatch, and deleted duplicate individuals ($n = 1087$) to remove intentionally or non-intentionally regenotyped plants. For the remaining data ($n = 12389$), we imputed missing genotypes. For this, we first calculated the mean allele frequency per marker averaged over all individuals; and then binomially draw two alleles for missing genotypes at random with the corresponding calculated mean allele frequency. As only a fraction 0.84% of all genotypes had to be inputed, this step does not significantly affect the results of the inference method.

Before applying our method, we filtered markers based on the following 4 criteria:

1. **Geographical Variation**: We removed markers that were correlated to the x or the y coordinate, as such large scale variations could originate from deeper time scales or be the traces of divergent selection or could also be the remnants of secondary contact. We chose a cut-off value of $R^2 = 0.015$.
2. **Linkage Disequilibrium**: We further filtered markers with strong linkage disequilibrium, because our method assumes that different markers contain independent information. We iteratively pruned markers that were correlated more than $R^2 = 0.03$ with any other marker.
3. **Minor Allele Frequency**: We removed all markers with a minor allele frequency below $\bar{p} = 0.15$, as rare markers can have a dominating influence on pairwise measures of relatedness, but here we aim to base inference on the independent information of many markers. Most markers in our dataset have intermediate allele frequency near $p = 0.5$, and the overall allele frequency distribution is relatively narrow (Fig. S2) .
4. **Deviations from Hardy-Weinberg Equilibrium**: We tested for a significant deficit or surplus of heterozygous from random mating expectations. These deviations can have multiple reasons, for instance failed genotyping, strong geographic structure or non-random mating. For filtering, we first calculated local allele frequency estimates by weighting all other individuals with a two-dimensional symmetric Gaussian. After testing the fit of several standard deviations σ, we found that $\sigma = 500$ meters gave the best predictions for local allele frequencies. We first calculated this expected mean allele-frequency for every marker and every individual with this Gaussian. Based on these local frequencies, we then obtained the expect number of heterozygous and homozygous sites for each marker. Using a χ^2-test, we calculated p-values for deviations from the expected numbers. We then filtered markers that had a $\bar{p}$-value below a cutoff of 10^{-5}.

After filtering, we were left with a dataset of $n = 12389$ individuals and 60 SNP markers. To ensure that there is no bias towards low or high frequencies, we flipped the 0 and 1 state for every marker with probability 0.5.

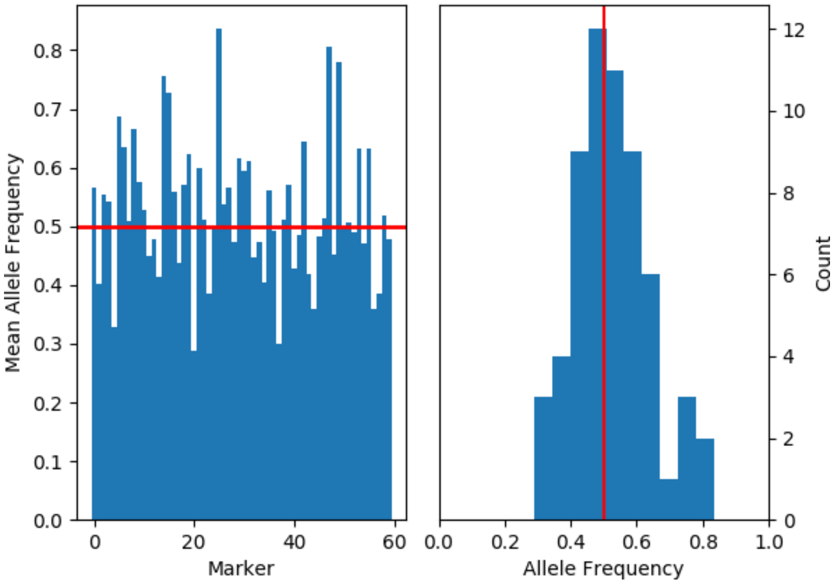

Figure S2: Allele frequency distribution. This figure depicts the distribution of the mean allele frequencies of the 60 markers (standard deviation 0.117). Left: Mean allele frequency ordered by marker. Right: Histogram of mean allele frequencies.

Data Availability

The detailed oligo-sequences for SNP genotyping, filtered genotype and geographic data are available at `https://github.com/hringbauer/BarrierInfer/tree/master/DataHZ`. The Python scripts used for data filtering are freely accessible at `https://github.com/hringbauer/BarrierInfer/tree/master/SNPCleaningScripts`.

Heterogeneity of Isolation by Distance

Our method assumes a uniform isolation by distance pattern in two dimensions. To confirm that this is not grossly violated for the *Antirrhinum* data, we calculated IBD patterns and investigated them for heterogeneity, both with respect to absolute position and angle (Fig. S3). Our analysis indicates that there are some spatial fluctuations of isolation by distance. However, they are mostly within the uncertainty estimates obtained by bootstrapping over genetic markers; so there is no indication of gross violations of the model assumptions.

Power Simulation

To test whether our method has sufficient power to detect a strong barrier to gene flow, we simulated a dataset similar to the *Antirrhinum* dataset. We used the same simulation engine described in the main text. We used 60 markers with a standard deviation $\sigma(\bar{p}) = 0.117355$, as in the filtered hybrid zone data set.

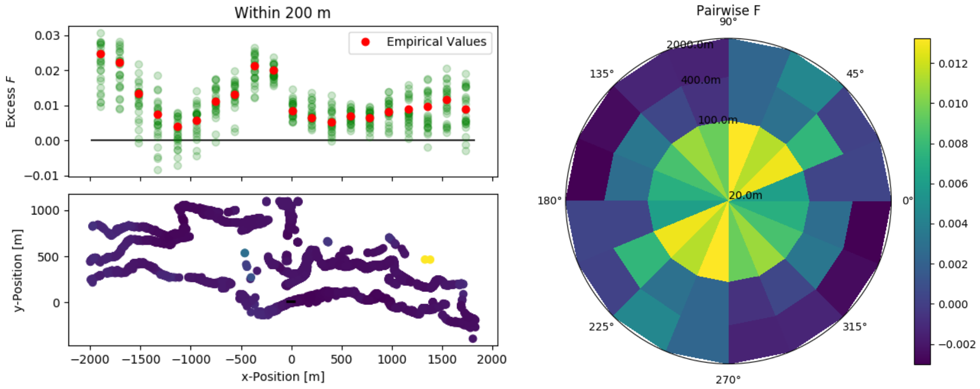

Figure S3: Heterogeneity of isolation by distance estimates. As a test statistic for the chance of recent co-ancestry F we used $\frac{(p_1-\bar{p})(p_2-\bar{p})}{4}$. This statistic should be a good estimator for F for intermediate mean allele frequencies ($\bar{p} \approx 0.5$). In the figures we depict the excess deviation compared to the average over all pairs. Left: Spatial heterogeneity and 25 bootstraps over loci: We depict mean excess F for all pairs with distance less than 200 meters. Right: Excess F when pairs are binned into 16 angular bins and three distance bins ($20-100$, $100-400$, $400-2000$ meter).

We simulated a population of a 60 × 40 demes one dispersal unit apart, with 16 diploid individuals per deme (thus Nbh=201.06). We sampled one individual per deme, and simulated a strong barrier to gene flow ($\gamma = 0.02$). This synthetic dataset has a similar isolation by distance pattern as the hybrid zone data set (see Fig. S4). Running our inference scheme on this dataset of 2400 samples indicates that there is sufficient power to infer the presence of a strong barrier (Fig. S5). At the true position of the barrier, the fit as well as 20 bootstraps over markers estimate a strong barrier to gene flow. For most other putative locations, no strong barrier is estimated. There is variation of bootstrap estimates which indicates that power is limited, but in total only a small number of bootstrap fits estimates a strong barrier. We stress that these power simulations are done for an idealized scenario in which our model assumptions hold.

Literature Cited

Whibley, A. C., N. B. Langlade, C. Andalo, A. I. Hanna, A. Bangham, C. Thébaud, and E. Coen, 2006 Evolutionary paths underlying flower color variation in Antirrhinum. Science **313**: 963–966.

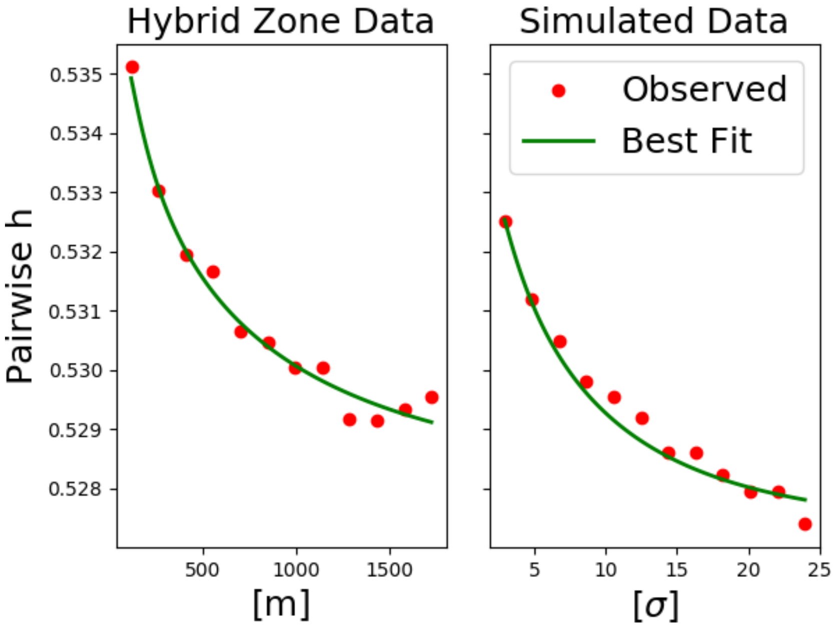

Figure S4: IBD of data simulated with hybrid zone parameters. The plot shows pairwise homozygosity for pairwise distance bins. The small difference in absolute values is due to the randomness with which mean allele frequencies for the synthetic data set were drawn. Left: Hybrid zone data (fit: Nbh = $192.20, m = 0.00839, s = 0.528088$). Right: Synthetic data set (fit: Nbh = $150.6, m = 0.0056, s = 0.52735$), pairwise distance is measured in standard deviations σ of the dispersal kernel.

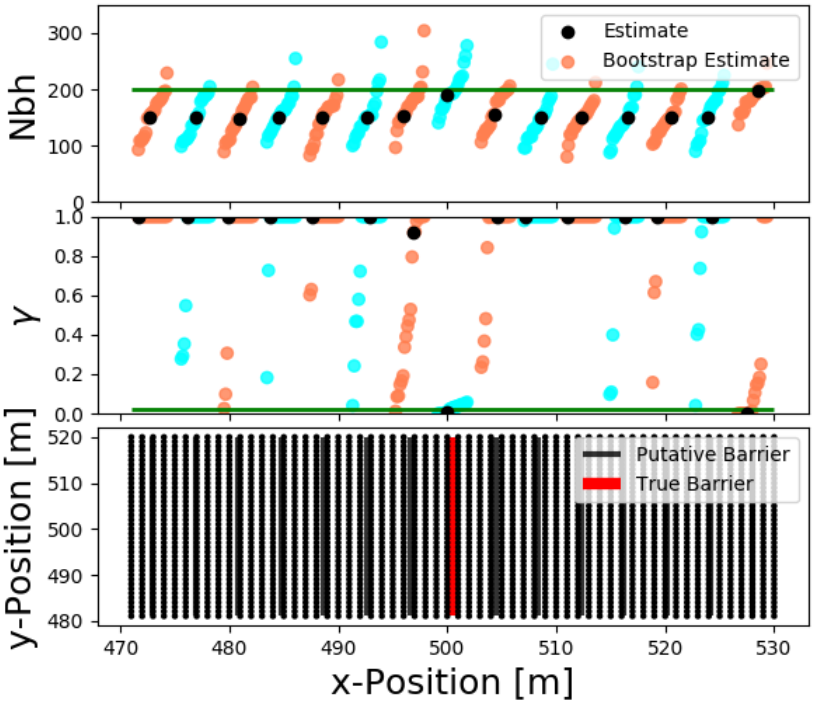

Figure S5: Power of inference scheme on simulated data set. Black dots indicate best fit estimates, colored dots are estimates after bootstrapping over genetic markers. Marker number, mean allele frequency distribution and demographic parameters were chosen to approximate the parameters from the *Antirrhinum* hybrid zone. We simulated a strong barrier ($\gamma = 0.02$), and run the inference scheme for multiple putative barrier locations (indicated by black lines).