SUPPLEMENTARY INFORMATION

Inference of Seed Bank Parameters In Two Wild Tomato Species Using Ecological and Genetic Data

Aurélien TELLIER, Stefan J. Y. LAURENT, Hilde LAINER, Pavlos PAVLIDIS, Wolfgang STEPHAN

Section of Evolutionary Biology, Department of Biology II, University of Munich (LMU), Grosshaderner Str. 2, 82152 Planegg-Martinsried, Germany

*Corresponding author: tellier@bio.lmu.de

Section 1: Ecological data

Estimates of the deme census sizes

We found that the indicated census sizes of demes varied depending on the investigator (few being reported), years of sampling and on the size of the populations. Sometimes, only qualitative estimates of census size were given. For example, several demes (around 20 in both species) were referred to as being "large", "huge", or "very large", and for populations containing more than 100 plants, the counting is often not precise. For our calculations, accessions with undefined census sizes were first removed from the dataset, and then added sequentially with assigned values between 100 and 500 (based on the fact that the largest populations observed contain 400 to 500 plants) to create various observed datasets. We fitted the exponential regression to the distributions of census sizes using the software R for each of those datasets (*lm* function on the log transform of the census size distribution; Figure S1). For each dataset we calculated the mean census size N_{cs} as the inverse of the exponential coefficient and obtained ranges of values for the mean N_{cs} of 44 to 185 for *S. peruvianum* and 33 to 154 for *S. chilense* (Table S1).

A second possibility was to calculate the arithmetic mean of the sample of census sizes for each dataset. This represents the maximum-likelihood of the mean N_{cs} for an exponential distribution. The 95% confidence intervals are obtained by multiplying this mean by (1-1.96$\sqrt{n}$ or 1+1.96$\sqrt{n}$) where n is the sample size. We obtained values ranging from 29 (lower confidence limit value) to 101 (highest confidence limit value) for *S. chilense*, and from 51 to 142 for *S. peruvianum*. These values are imbedded in the range of our above conservative estimates (Table S1).

We also tested if a Poisson and a power-law function may explain our distribution of census sizes.

First, we analyzed our census data using a linear mixed-effects model (function *lmer*() from the *lme4* package in R). To deal with count data, we assumed a Poisson distribution (for family) and year of sampling, altitude and geographical province as random effects (investigators could not be included due to the paucity of assignments). The model analysis was run with one, two or three random effects, to obtain the estimate of the mean λ of the Poisson distribution. We show in Figure S1a and S1b the cumulative density functions for Poisson distributions with the mean estimated based on the models with three random effects which had the highest log-likelihood.

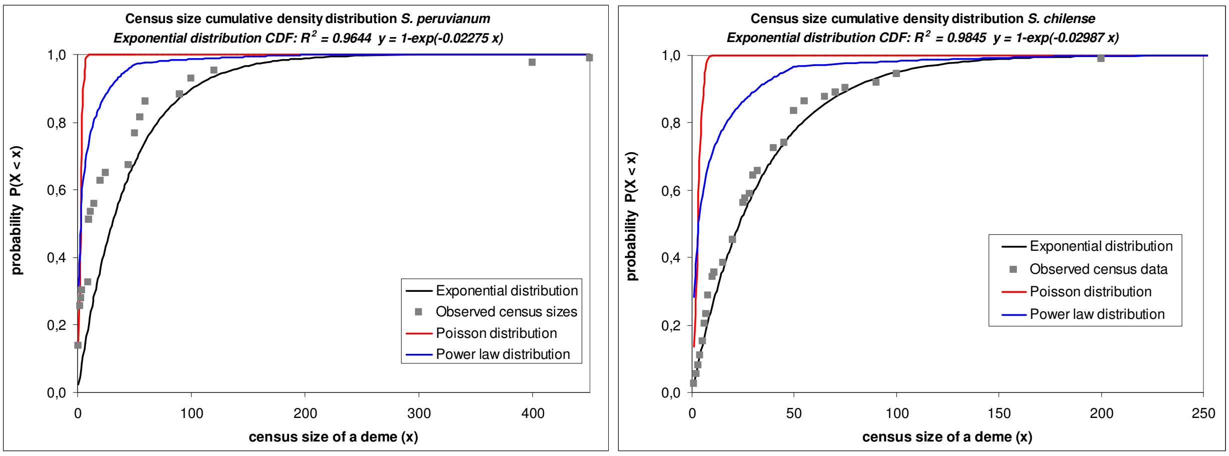
Second, we tested if our census sizes are distributed following a power law distribution as used in plant ecology studies (1, 2). The power law function is characterized by larger frequencies of very high population sizes compared to the exponential distributions. Two methods were used to estimate the coefficient α of the power law function: the function *power.law.fit* in the R package *igraph*, and the likelihood estimate described in ref. 3. These methods yielded respectively different estimates of the power law coefficients: 1.245 or 2.564 for *S. peruvianum* and 1.18 or 2.124 for *S. chilense* assuming the x_{min} parameter to be one. In Figure S1A and S1B we show the best power law CDF for each species.

Note that neither the power law function nor the Poisson distribution is a better fit than the exponential distribution. Compared to other plant ecological studies, these two wild tomato species show an excess of populations with small to medium sizes (up to 100 plants), a lack of populations with sizes between 100 and 500, and no populations with larger size than 500. This made the power law function unsuitable for regression on our observed census sizes.

Figure S1: Exponential regression for the CDF of deme census sizes for (A) *S. peruvianum* and (B) *S. chilense*. The estimates of mean census size per deme are respectively of (A) 44 and (B) 33. The coefficient of regression and the equation of the best fitting regression for the exponential distribution are indicated. The coefficients of the Poisson distributions (red lines) are $\lambda_{peruvianum}$ = 3.376 and $\lambda_{chilense}$ = 3.49, and for the power law distributions (blue lines) $\alpha_{peruvianum}$ = 1.245 and $\alpha_{chilense}$ = 1.18.

Figure S1A

Figure S1B



Census size cumulative density distribution *S. peruvianum*

Exponential distribution CDF: R^2 = 0.9644 y = 1-exp(-0.02275 x)

Census size cumulative density distribution *S. chilense*

Exponential distribution CDF: R^2 = 0.9845 y = 1-exp(-0.02987 x)

For simplicity, we fix thereafter for each coalescent simulation, the census size per deme to be equal for all demes. This value has a prior defined in Table S1.

Ecological data and geographical range of each species

The range area of *S. peruvianum* was estimated to 80,961 km^2, and the percentage of niche filling as 22.4% (4). *S. chilense* shows a smaller range of 62,401 km^2, and the percentage of niche filling was 31.5% (4). Assuming that the total number of observed accessions for *S. peruvianum* and *S. chilense*, 118 and 135, respectively, fill only 22.4 and 31.5% of their potential niche, we estimated the number of physical demes to be around 526 for *S. peruvianum* and 428 for *S. chilense* (Table S1).

Table S1: Summary of the key ecological data for the two wild tomato species from the TGRC (Tomato Genetics Ressource Center, UC Davis, USA) collection (in bold the values used in prior definitions below).

Species	Total number of accessions in the TGRC database	Number of populations with census size available	% of niche filling[a]	Estimated number of demes in the species range	Estimated range of mean census size per deme N_{cs}
S. peruvianum	118	75	22.4	**526**	**44 – 185**
S. chilense	135	107	31.5	**428**	**33 – 154**

[a] values from Nakazato *et al.* 2010 (4).

Methods of reconstructing the ecological range of a species are biased. For example, Nakazato *et al.* assume that only one population is present in a radius of 20km around a sampled accession from TGRC, but also predict that the ecological range of *S. peruvianum* and *S. chilense* should extend east of the Andes, where these species are never found (4). It is thus hard to predict the direction of the bias (over- or under-estimation). Our choice of the number of demes per species means that the sampling of the TGRC reflects the distribution of between 1/4 and 1/5 of all *S. peruvianum* populations, and between 1/3 and 1/4 of the total number of *S. chilense* populations.

Section 2: Sampling and DNA sequencing

Plant material and sequences for the population sample

The population sample is composed of the sequences previously obtained (5, 6; Table S2). Note that we do not use the Arequipa (*S. peruvianum*) and the Antofogasta (*S. chilense*) populations studied in refs 5, 6, because it was shown, on the basis of the frequency spectrum of alleles, that these populations experienced some demographic events, most likely bottlenecks or admixture (5, 6).

Table S2: List of the population samples of the two studied *Solanum* species.

Species	Population	Location	Coordinates (latitude, longitude)
S. peruvianum	Tarapaca (LA2744)	Northern Chile	18°33'S, 70°09'W
	Nazca	Southern Peru	14°51'S, 74°44'W
	Canta	Central Peru	11°31'S, 76°41'W
S. chilense	Tacna	Southern Peru	17°53'S, 70°07'W
	Moquegua	Southern Peru	17°04'S, 70°52'W
	Quicacha	Southern Peru	15°37'S, 73°48'W

Where applicable, the TGRC accession numbers are indicated. *S. chilense* and *S. peruvianum* populations have been described in ref. 5.

Plant material and sequences for the species-wide sample

We selected one plant per 14 accessions of *S. peruvianum* and 10 accessions of *S. chilense* from the TGRC, chosen to be distributed uniformly over the range of both species (Table S3). One allele for each of the seven loci was sequenced per plant of the species-wide sample.

Genomic DNA was extracted from tomato leaves using the DNeasy Plant Mini Kit (Qiagen GmbH, Hilden, Germany). PCR amplification was performed with High Fidelity Phusion Polymerase (Finnzymes, Espoo, Finland), and all PCR products were examined with 1% agarose gel electrophoresis. Generally, direct sequencing was performed on PCR products to identify homozygotes and obtain their corresponding sequences. For heterozygotes, a dual approach of both cloning before sequencing and direct sequencing was used to obtain the sequences of both alleles. The first allele present in at least three clones was chosen. Sequencing reactions were run on an ABI 3730 DNA analyser (Applied Biosystems and HITACHI, Foster City, USA).

One allele was sequenced for each individual, and a total of 14 (*S. peruvianum*) and 10 (*S. chilense*) sequences were obtained for each locus. Contigs of each locus were first built and edited using the Sequencher program (Gene Codes, Ann Arbor, USA) and adjusted manually in MacClade 4 (version 4.0 for OS X). These new sequences are deposited in GenBank (accession numbers JF736670-JF736839).

Table S3: List of the species-wide samples with the TGRC accession numbers from the two *Solanum* species.

Species	Accessions	Location	Coordinates (latitude, longitude)
S. peruvianum	LA0153	Central Peru	09°57'S, 78°13'W
	LA0111	Central Peru	10°48'S, 77°44'W
	LA1616	Central Peru	12°05'S, 76°55'W
	LA1913	Central Peru	14°23'S, 75°12'W
	LA2834	Central Peru	14°46'S, 74°49'W
	LA0446	Southern Peru	15°47'S, 74°23'W
	LA1336	Southern Peru	16°12'S, 73°37'W
	LA1951	Southern Peru	16°25'S, 73°08'W
	LA1333	Southern Peru	16°34'S, 72°38'W
	LA3218	Southern Peru	16°57'S, 72°05'W
	LA1954	Southern Peru	17°01'S, 72°05'W
	LA2964	Southern Peru	17°59'S, 70°50'W
	LA4125	Northern Chile	19°18'S, 69°25'W
	LA2732	Northern Chile	19°24'S, 69°36'W
S. chilense	LA1930	Southern Peru	15°17'S, 74°36'W
	LA1960	Southern Peru	17°05'S, 70°52'W
	LA1958	Southern Peru	17°15'S, 71°15'W
	LA1969	Southern Peru	17°32'S, 70°02'W
	LA3355	Southern Peru	18°03'S, 70°18'W
	LA2778	Northern Chile	18°23'S, 69°33'W
	LA2932	Northern Chile	22°29'S, 70°10'W
	LA2748	Northern Chile	21°12'S, 69°30'W
	LA2750	Northern Chile	22°05'S, 70°12'W
	LA2930	North-Central Chile	25°24'S, 70°24'W

Genes sequenced

Table S4: Chromosome location, putative function, and sizes of coding and non-coding regions of the seven studied loci in *S. peruvianum* and *S.chilense*.

Locus	Chromosome	Putative protein function	Non-coding region	Coding region	
				synonymous	non-synony-mous
CT066	10	Arginine decarboxylase	0	335	1008
CT093	5	S-adenosylmethionine Decarboxylase proenzyme	359	263	765
CT166	2	Ferredoxin-NADP reductase	823	118	322
CT179	3	Tonoplast intrinsic protein D-type	234	174	404
CT198	9	Submergence induced protein 2-like	359	90	242
CT251	2	At5g37260 gene	348	348	974
CT268	1	Receptor-like protein kinase	0	404	1476

The number of sites in each category was estimated with the method of Yang and Nielsen (7) and is based on the alignment of sequences for the pooled sequences in *S. peruvianum*.

Note that we have used 7 out of the 8 loci studied in Arunyawat *et al*. (2007) removing the locus CT208 because it shows “in *S. chilense* an intriguing geographic pattern of nucleotide diversity, in that levels of nucleotide variation gradually diminish from north to south, with essentially no variation in the southernmost sample” (5). Further analysis, confirmed that the 7 loci used here do not show any deviation from the expected evolution under purifying selection (8).

Population genetics analysis of the sequence data

We present here a summary of statistics from the new sequence data of the species-wide sample obtained by analysis with DnaSP v5.1 (9) and libsequence C++ library (10).

Table S5a: Summary statistics at 7 loci for the species-wide sample of *S. peruvianum* (for 14 sequences).

Locus	Number of segregating sites S_{sw}	Population mutation rate[b] θ_{w_sw}	Tajima's *D* at all sites D_{sw}	Tajima's *D* at silent sites D_{silent_sw}	Tajima's *D* at synonymous sites D_{syn_sw}
CT066	58	18.24 (0.0136)	−1.04	−1.12	−1.12
CT093	39	12.26 (0.0088)	−1.62	−1.42	−1.25
CT166	59	18.55 (0.0147)	−1.6	−1.55	−1.72
CT179	54	16.98 (0.019)	−0.58	−0.63	−0.82
CT198	59	18.55 (0.0268)	−0.61	−0.69	−0.75
CT251	94	29.56 (0.0177)	−0.87	−0.79	−0.83
CT268	83	26.1 (0.0139)	−0.79	−0.18	−0.18
average across loci[a]	63.71	**20.04** (0.0153)	−1.02	−0.91	**−0.95**

[a] arithmetic average across loci. [b] θ_w is given per locus and per site (in brackets).

Table S5b: Summary statistics at 7 loci for the species-wide sample of *S. chilense* (for 10 sequences).

Locus	Number of segregating sites S_{sw}	Population mutation rate[b] θ_{w_sw}	Tajima's *D* at all sites D_{sw}	Tajima's *D* at silent sites D_{silent_sw}	Tajima's *D* at synonymous sites D_{syn_sw}
CT066	43	15.2 (0.0113)	0.064	0.44	0.44
CT093	21	7.42 (0.0053)	−1.26	−1.05	−0.64
CT166	48	16.97 (0.0134)	−0.39	−0.32	−0.67
CT179	39	13.79 (0.0153)	−0.72	−0.72	−0.24
CT198	25	8.84 (0.0128)	−1.02	−1.02	0.02
CT251	24	8.48 (0.0051)	−0.34	−0.53	−0.39
CT268	50	17.67 (0.0094)	0.005	0.29	0.29
average across loci[a]	35.71	**12.62** (0.0097)	−0.52	−0.41	**−0.17**

[a] arithmetic average across loci. [b] θ_w is given per locus and per site (in brackets).

A summary of the single populations is given in Table S6a and S6b (for 10 to 12 sequences per population), see refs 5, 6.

Table S6a: Summary statistics at 7 loci for the population samples of *S. peruvianum*.

Locus	Population mutation rate for population Tarapaca[b] θ_{W_TAR}	Population mutation rate for population Nazca[b] θ_{W_NAZ}	Population mutation rate for population Canta[b] θ_{W_CAN}	Tajima's *D* at all sites for population Tarapaca D_{TAR}	Tajima's *D* at all sites for population Nazca D_{NAZ}	Tajima's *D* at all sites for population Canta D_{CAN}	Fixation index among populations F_{ST}
average across loci[a]	**14.51** (0.0111)	**13.42** (0.0103)	**17.2** (0.0132)	**−0.26**	**−0.25**	**−0.71**	**0.13**

[a] arithmetic average across loci. [b] θ_w is given per locus and per site (in brackets).

Table S6b: Summary statistics at 7 loci for the population samples of *S. chilense*.

Locus	Population mutation rate for population Moquegua[b] θ_{W_MOQ}	Population mutation rate for population Tacna[b] θ_{W_TAC}	Population mutation rate for population Quicacha[b] θ_{W_QUI}	Tajima's *D* at all sites for population Moquegua D_{MOQ}	Tajima's *D* at all sites for population Tacna D_{TAC}	Tajima's *D* at all sites for population Quicacha D_{QUI}	Fixation index among populations F_{ST}
average across loci[a]	**13.23** (0.0101)	**11.88** (0.009)	**12.35** (0.0095)	**−0.04**	**0.06**	**0.13**	**0.17**

[a] arithmetic average across loci. [b] θ_w is given per locus and per site (in brackets).

A summary of the pooled populations is given in Table S7a and S7b (for 30 to 36 sequences per species) see refs 5, 6.

Table S7a: Summary statistics at 7 loci for the pooled sample of *S. peruvianum*.

Locus	Number of segregating sites S_{pooled}	Population mutation rate[b] $\theta_{\mathrm{w\_pooled}}$	Tajima's D at all sites D_{pooled}	Tajima's D at silent sites $D_{\mathrm{silent\_pooled}}$	Tajima's D at synonymous sites $D_{\mathrm{syn\_pooled}}$
average across loci[a]	**90.57**	**22.37** (0.0171)	−1.09	−1.01	**−0.91**

[a] arithmetic average across loci. [b] θ_w is given per locus and per site (in brackets).

Table S7b: Summary statistics at 7 loci for the pooled sample of *S. chilense*.

Locus	Number of segregating sites S_{pooled}	Population mutation rate[b] $\theta_{\mathrm{w\_pooled}}$	Tajima's D at all sites D_{pooled}	Tajima's D at silent sites $D_{\mathrm{silent\_pooled}}$	Tajima's D at synonymous sites $D_{\mathrm{syn\_pooled}}$
average across loci[a]	**70**	**17.13** (0.0131)	−0.55	−0.34	**−0.04**

[a] arithmetic average across loci. [b] θ_w is given per locus and per site (in brackets).

Figure S2: Mean Tajima's *D* values across seven loci for all sites, silent and synonymous sites for both species: *S. peruvianum* in black, and *S. chilense* in grey.
The rectangles indicate the value of Tajima's *D* for the species-wide sample (SW), and the diamond for the pooled sample (pooling of the three populations per species).

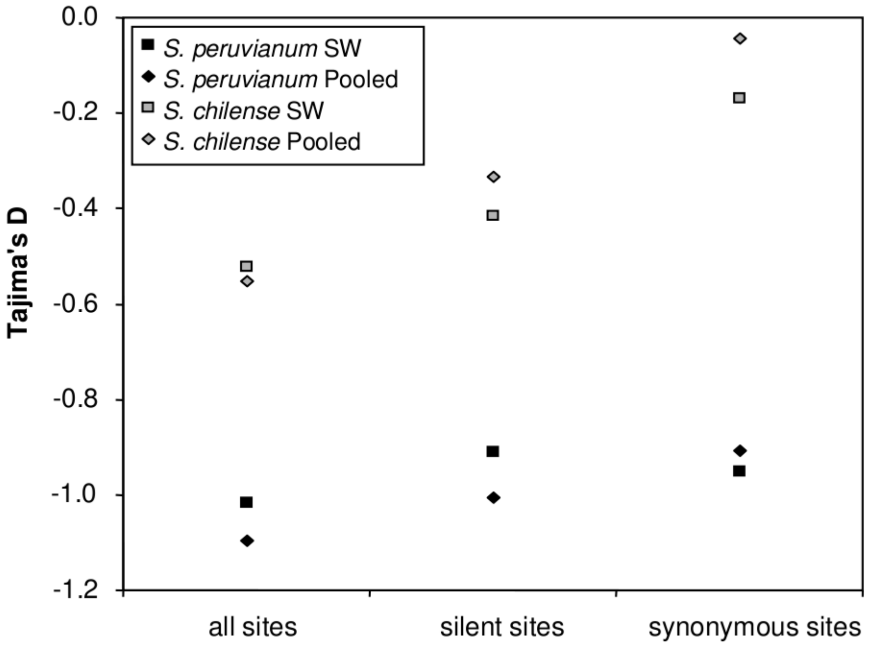
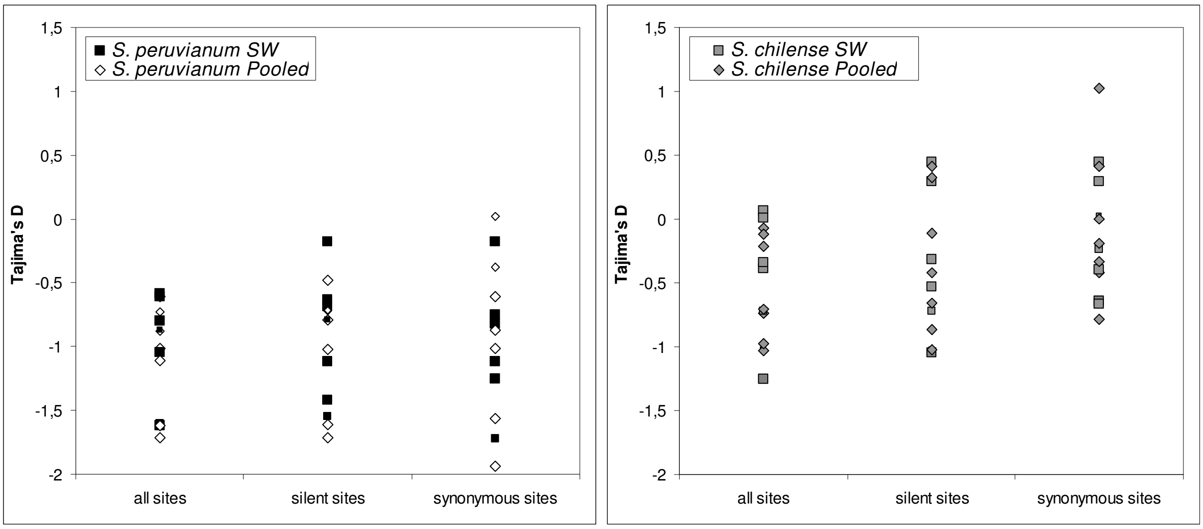


Figure S3: Tajima's *D* values for each of the seven loci for all sites, silent and synonymous sites for both species: A) for *S. peruvianum*, and B) *S. chilense*.

The rectangles indicate the value of Tajima's *D* for the species-wide sample (SW), and the diamond for the pooled sample (pooling of the three populations per species).

Figure S3A Figure S3B



Tajima's *D* for each locus indicates that the scattered and pooled samples do not differ significantly in their frequency spectra.

Section 3: Model of coalescence for population with seed bank

Population genetics modeling

We summarize here the theory on coalescence with seed bank. We model a neutral seed bank with haploid Wright-Fisher type dynamics for a single population with constant size (11). The population of plants is composed at each generation of *N* individuals, with a proportion b_i, $i = 1, \ldots, m$, coming from seeds produced *i* generations ago. In other words, seeds are allowed to remain in the seed bank for up to *m* generations. At a given generation, each individual is drawn from a pool of seeds build up during the previous *m* generations. Each individual is obtained with the probability b_1 from the seeds produced at the previous generation, b_2 from the seeds produced two generations ago, …, and b_{m} from the seeds produced *m* generations ago. The rate of coalescence in a population with a seed bank is (11):

$$\beta_1^2 \binom{r}{2} \qquad [1]$$

with *r* being the number of ancestral lineages at any point in time, and β_1 the seed bank rescaling coalescent rate.

$$\beta_1 = 1 / \sum_{i=1}^{m} i b_i \qquad [2]$$

where $\sum_{i=1}^{m} i b_i$ is the expected value of the seed bank age distribution. Similarly, the mutation rate γ along an ancestral line in the coalescent is (11):

$$\gamma = \frac{\beta_1}{2}(b_1\theta_1 + b_2\theta_2 + \ldots + b_m\theta_m). \qquad [3]$$

Where θ_j is the population mutation rate for individuals produced by age *j* seeds (*j*=1, …, *m*) $b_1 = b(11)$.

We make further biological assumptions to implement in our simulation program the modified rates of coalescence, mutation, migration and recombination under seed bank (the parameters of our coalescent model are in Table S8).

1) Seed germination is a memoryless process modeled as a geometric process in time. We suppose that the germination rate of a given seed is *b*. Each individual is obtained thus with the probability $b_i = b(1-b)^{i-1}$ from the seeds produced at generation *i*.

For clarity, this means that each individual is obtained with the probability:

$b_1 = b$ from the seeds produced at the previous generation,

$b_2 = b(1-b)$ from the seeds produced two generations ago, …, and $b_m = b(1-b)^{m-1}$ from the seeds produced m generations ago.

Assuming a geometric germination rate, β_1 from eq. 2 can thus be explicitly written as:

$$\beta_1 = 1 / \sum_{i=1}^{m} ib(1-b)^{i-1} \ . \qquad [4]$$

Eq. 4 can be approximated as $\beta_1 ; \dfrac{b(1-(1-b)^{m+1})}{1-(1+bm)(1-b)^m}$ if m is sufficiently large.

The rate of coalescence implemented in our program is thus $\beta_1^2 \binom{r}{2}$ using β_1 from eq. 4.

2) We enforce the condition that the sum of the germination probabilities over m generations should be equal to 1 (11). This condition is incorporated in our program. Furthermore, the mutation rate under a seed bank model assumes that the mutation rate does not depend on the age of seeds. The population mutation rate with seed bank as implemented in our program is thus (explicitly from eq. 3):

$$\gamma = \frac{\beta_1}{2}\theta(b + b(1-b) + ... + b(1-b)^{m-1}) = \frac{\beta_1}{2}\theta \qquad [5]$$

where θ is the population mutation rate without seed bank (based on the census size N).

It has been suggested that aging of seeds can lead to an increase of the mutation rate, with most of the new mutations being deleterious (12, 13). However, a recent meta-analysis did not reveal high levels of genetic diversity accumulating in the soil seed bank (14). Reviewing different plant species, no evidence was found for genetic differences between the standing crop and the seed bank (14). In species where such differences were found, they were likely to be the result of local selection acting as a filter on the alleles present in the seed bank (14). As we are interested only in analyzing neutral evolutionary processes, we chose to keep the mutation rate constant among all seed ages assuming that no selection is acting in the seed bank, but see (12, 13). The assumptions behind equations 4 and 5 are similar to those of Nunney (15) and Vitalis *et al.* (16) describing the expected heterozygosity in a population with seed bank.

3) We multiply the recombination rate per nucleotide r also by β_1 (eq. 4). This is because recombination only occurs in a lineage when a plant is above ground and produces seeds.

4) We also rescale the migration rate (κ) between demes in a metapopulation due to the seed bank. We assume here that only pollen migrates between demes, and that this occurs only when plants are above ground. The migration rate (κ) is thus also multiplied by β_1.

5) A key assumption in this model is to assume that every generation the number of individuals (N from ref. (11)) is equal to N_{cs} in each deme. In other words, each generation above ground and each generation in the seed bank has the same census size equal to N_{cs}. The census size used in our model is calculated above from ecological data. This approximation holds as long as the variation between years in census sizes is moderate (15).

Moreover, when there was no seed bank, *i.e.* when all seeds germinated the year after being produced (b = 1), we verified that equation 1 and simulations are equivalent to the classic Wright-Fisher model with non-overlapping generations (as implemented in *ms* (17)).

Table S8: List of parameters and compound parameters in the model with metapopulation, demography and seed bank.

	Parameter name	Parameter definition
Estimated parameters	N_{cs}	Census size of each deme in the metapopulation
	b	Germination rate
	m	Maximum time seeds can spend in the seed bank (in generations)
	κ	Migration rate between demes (without seed bank rescaling)
	n_{d}	Number of demes in the metapopulation (effective number)
	μ	Mutation rate per nucleotide per generation
	t_{event}	Time of the population split in generations
	$S_{\mathrm{current}} / S_{\mathrm{anc}}$	Ratio of current to ancestral metapopulation sizes
Compound parameters (not estimated)	β_1	Rescaling parameters of the seed bank (eq. 4)
	θ	Population mutation rate without seed bank per deme: $\theta = 4\times N_{\mathrm{cs}}\times\mu$
	γ	Population mutation rate with seed banks (eq. 5)
	r	Local crossing-over rates per nucleotide per generation, obtained for each locus from (18), without seed bank

Section 4: Description of the ABC procedure

Approximate Bayesian Computation

Simulations were conducted on a 64-bit Linux cluster with 510 nodes. Source code is available upon request.

Parameter estimation

The RMSE indicates the percentage of variation unexplained by the PLS components and is constructed by comparing the simulated parameter values with the ones predicted using a given number of PLS components (19). We chose the number of components for the parameter estimation procedure such that additional components do not decrease the RMSE of any parameter of the model. The retained PLS components are used to transform the observed and the simulated datasets. The rejection step consists in computing δ between simulated and observed sets of summary statistics and to retain the 2,500 simulations closest to the observed data. Finally, we estimate posterior distributions of the parameters by applying the locally weighted multivariate regression method (20) implemented in the ABCest program (21). We estimate the marginal posterior probability distribution of each demographic parameter using the kernel density estimation method implemented in the R core package and report the mode and the 95% credibility intervals of these distributions. To avoid the posterior distributions to exceed the upper and lower bound of our prior distributions we transform the data as $z = \log[\tan(1/x)]$, where x is the original dataset and z is the transformed data (22).

Stepping-stone model

Our stepping-stone model features 526 and 428 demes for *S. peruvianum* and *S. chilense*, respectively. It is a linear one-dimensional array with absorbing edges. Migration occurs symmetrically at rate κ between two adjacent demes. The sampled demes (populations and species wide) are equally distributed over the whole range, *i.e.* every 30 to 40 demes. Each metapopulation edge consists of 5 demes which are not sampled to avoid boundaries effect (*ms* command available upon request).

Section 5: Demography and models without seed bank (ABC analysis Part 1)

Table S9a: Summary of prior boundaries of the ABC chosen for each tested model in *S. peruvianum*

Model	Parameters	Min	Max
All models	μ	5×10^{-9}	10^{-8}
	N_{cs}	44	185
	$\log(\kappa)$	-4	-2
Seed bank + constant population size	b	0.01	0.5
	t_{fusion}	0	200
Seed bank + expansion	b	0.01	0.5
	t_{exp}	0	200
	$S_{\mathrm{current}} / S_{\mathrm{anc}}$	1	100
Seed bank + crash	b	0.01	0.5
	t_{crash}	0	200
	$S_{\mathrm{current}} / S_{\mathrm{anc}}$	0.1	1
No seed bank + expansion	t_{exp}	0	200
	$S_{\mathrm{current}} / S_{\mathrm{anc}}$	100	1
No seed bank + crash	t_{crash}	0	200
	$S_{\mathrm{current}} / S_{\mathrm{anc}}$	0.04	1
No seed bank + expansion + stepping-stone model	t_{exp}	0	200
	$S_{\mathrm{current}} / S_{\mathrm{anc}}$	1	100

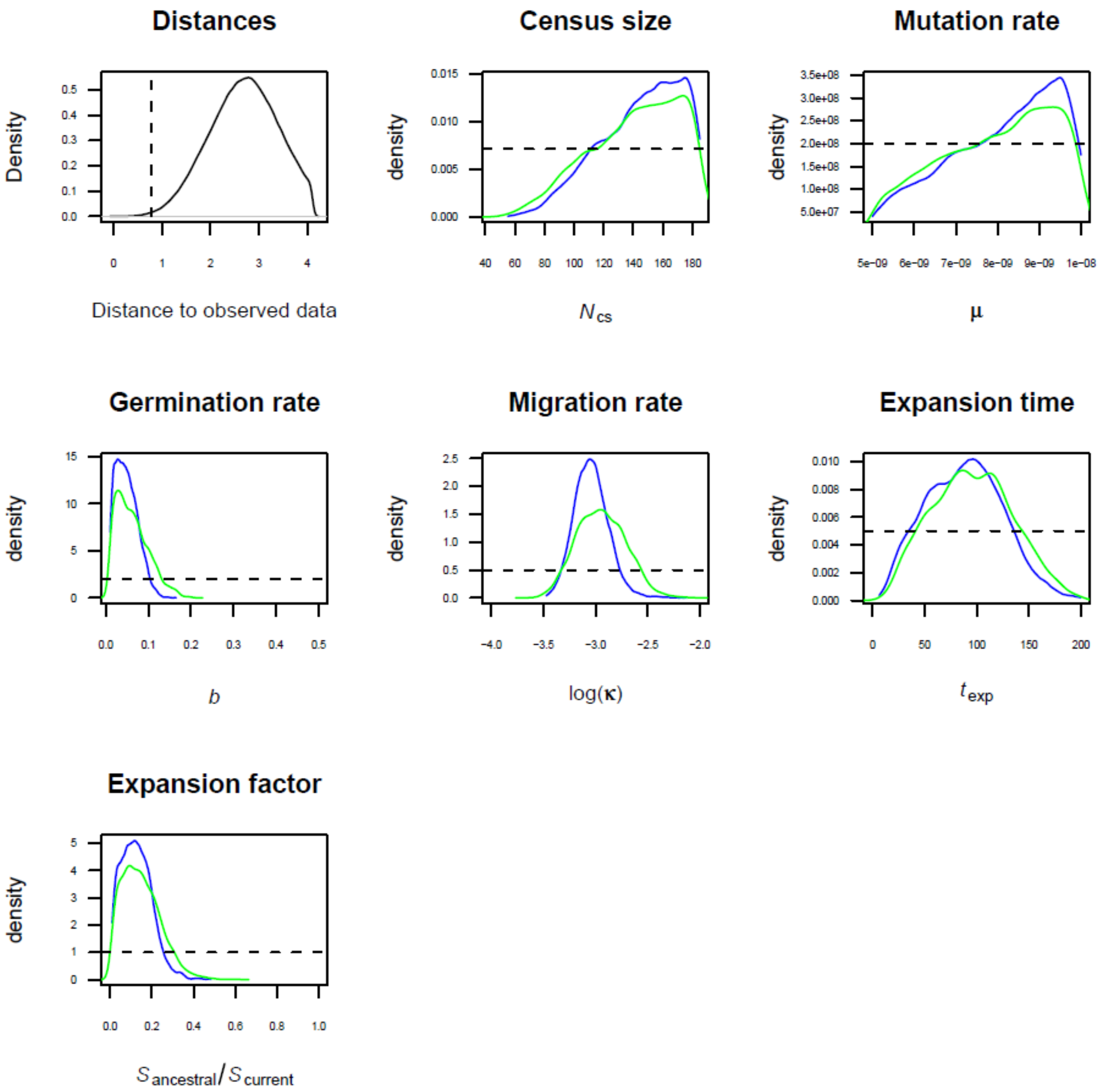
Table S9b: Summary of prior boundaries of the ABC chosen for each tested model in *S. chilense*

Model	Parameters	Min	Max
All models	μ	5×10^{-9}	10^{-8}
	N_{cs}	33	154
	$\log(\kappa)$	-4	-2
Seed bank + constant population size	b	0.01	1
	t_{fusion}	0	200
Seed bank + expansion	b	0.01	1
	t_{exp}	0	200
	$S_{\mathrm{current}} / S_{\mathrm{anc}}$	100	1
Seed bank + crash	b	0.01	1
	t_{crash}	0	200
	$S_{\mathrm{current}} / S_{\mathrm{anc}}$	0.1	1
No seed bank + expansion	t_{exp}	0	200
	$S_{\mathrm{current}} / S_{\mathrm{anc}}$	100	1
No seed bank + crash	t_{crash}	0	200
	$S_{\mathrm{current}} / S_{\mathrm{anc}}$	0.04	1
No seed bank + expansion + stepping-stone model	t_{exp}	0	200
	$S_{\mathrm{current}} / S_{\mathrm{anc}}$	100	1

Section 6: Parameter estimates (ABC analysis Part 2)

Figure S4: Posterior distributions of the parameters of an island model with 526 demes under demographic expansion for *S. peruvianum*.

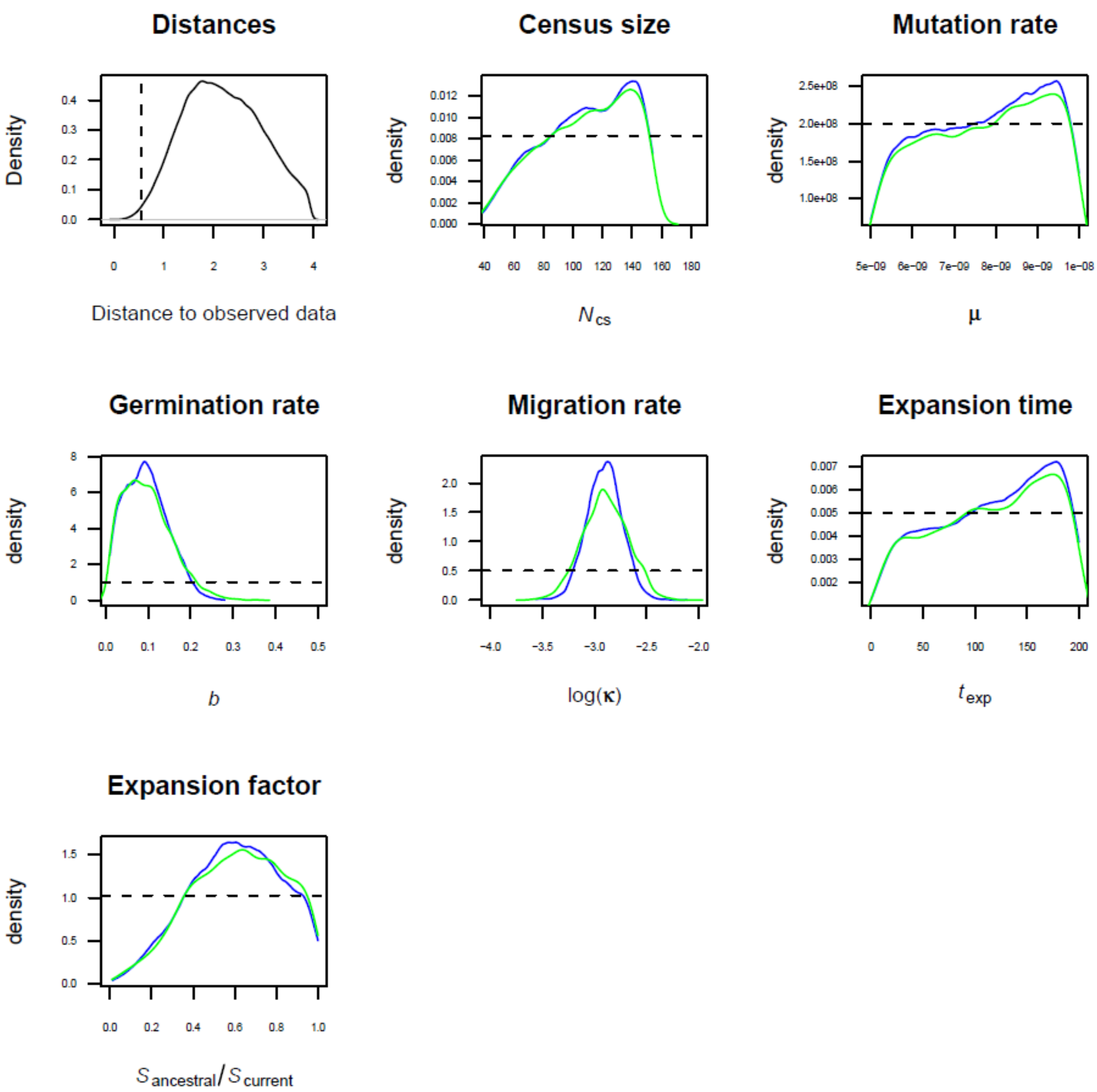
The top left panel represents the distribution of Euclidean distances (δ) with the dotted line indicating the proportion of retained simulations (2,500 of 2,000,000). The other panels represent respectively the posterior distributions for census size per deme (N_{cs}), mutation rate (μ), germination rate (b), migration rate (log(κ)), time of expansion (t_{exp}) in units of $4N_e$ of a given deme, and the expansion factor (S_{anc} / $S_{current}$). The prior uniform distribution is indicated as a dashed line, the green line is the posterior distribution based on the rejection algorithm, and the blue line is the posterior distribution after the regression adjustment.



For clarity of graphical representation, S_{anc} / $S_{current}$ is represented in Figures S4. The expansion ($S_{current}$ / S_{anc}) is estimated to be 8.99 fold in *S. peruvianum*.

Figure S5: Posterior distributions of the parameters of an island model with 428 demes under population expansion for *S. chilense*.

The top left panel represents the distribution of Euclidean distances (δ) with the dotted line indicating the proportion of retained simulations (2,500 of 2,000,000). The other panels represent the posterior distributions for census size per deme (N_{cs}), mutation rate (μ), germination rate (b), migration rate (log(κ)), time of expansion (t_{exp}) in units of $4N_e$ of a given deme, and the expansion factor (S_{anc} / $S_{current}$). The prior uniform distribution is indicated as a dashed line, the green line is the posterior distribution based on the rejection algorithm, and the blue line is the posterior distribution after the regression adjustment.



For clarity of graphical representation, S_{anc} / $S_{current}$ is represented in Figures S5. The expansion ($S_{current}$ / S_{anc}) is estimated to be 1.7 fold in *S. chilense*.

Table S10: Summary of the prior and posterior distributions of each parameter. The prior distributions are uniform between the lower and upper bound. The posterior distributions are summarized as the mode and the boundaries of the 95% credibility interval (CI 0.025 – CI 0.975).

Species	Parameter	Prior		Posterior		
		Lower bound	Upper bound	Mode	CI 0.025	CI 0.975
S. peruvianum	Germination rate (b)	0.01	0.5	0.027	0.011	0.103
	Migration rate (κ)	10^{-4}	10^{-2}	9.39×10^{-4}	4.61×10^{-4}	1.91×10^{-3}
	Time of expansion (t_{event})	0	200	106.24	25.87	157.25
	Expansion ratio ($S_{\text{current}} / S_{\text{anc}}$)	1	100	8.99	3.55	68.1
	Census size per deme (N_{cs})	44	185	173.7	88.29	183.25
	Mutation rate (μ)	5×10^{-9}	10^{-8}	9.29×10^{-9}	5.26×10^{-9}	9.94×10^{-9}
S. chilense	Germination rate (b)	0.01	1	0.093	0.016	0.2
	Migration rate (κ)	10^{-4}	10^{-2}	1.34×10^{-3}	6.12×10^{-4}	2.71×10^{-3}
	Time of expansion (t_{event})	0	200	166.26	10.01	196.5
	Expansion ratio ($S_{\text{current}} / S_{\text{anc}}$)	1	100	1.68	1.02	6.8
	Census size per deme (N_{cs})	33	154	137.62	50.63	152.03
	Mutation rate (μ)	5×10^{-9}	10^{-8}	8.52×10^{-9}	5.19×10^{-9}	9.91×10^{-9}

The time is given in generations, in units of 4N_e of a given deme (including the seed bank, as implemented in our version of Hudson's *ms*). For simplicity, assuming b = 0.2, we calculate for example that t_{exp} = 200 equals to a demographic event occurring 3.4×10^6 generations ago. The time of the expansion we infer here are much older than the rough previous estimate of divergence time between these species of 550,000 generations ago (23). These species are short lived perennials, and the generation time has been estimated between one and seven years, most likely around 3 to 5 years.

Statistical analysis with varying number of demes. We show in Figure 3 the joint posterior distributions for germination rates and number of demes. Figure 3 and the estimates of the modes of b were performed using the R package *loc2plot* (20) available from M.A. Beaumont webpage (http://www.rubic.rdg.ac.uk/~mab/). The Hotelling T-square test implemented in the R package *rrcov* (24) was used to compare the two bivariate distributions.

Section 7: Principal Component Analysis on S. peruvianum simulated datasets

We conducted a Principal Component Analysis with the software R (function *prcomp*) based on 1,000 random simulated datasets of *S. peruvianum* best demographic model (Figure S4). The PCA analysis is based on simulated values before transformation with the PLS (19).

Figure S6: PCA on 2,000 simulations for *S. peruvianum*. PCA axes 1 and 2 are shown in A) and PCA axes 3 and 4 in B). The percentage of the total variance explained is indicated for each of the four main PCA axes. For clarity, the simulated datasets are not indicated, and we show in blue the representation of the summary statistics and in red the parameters of the model.

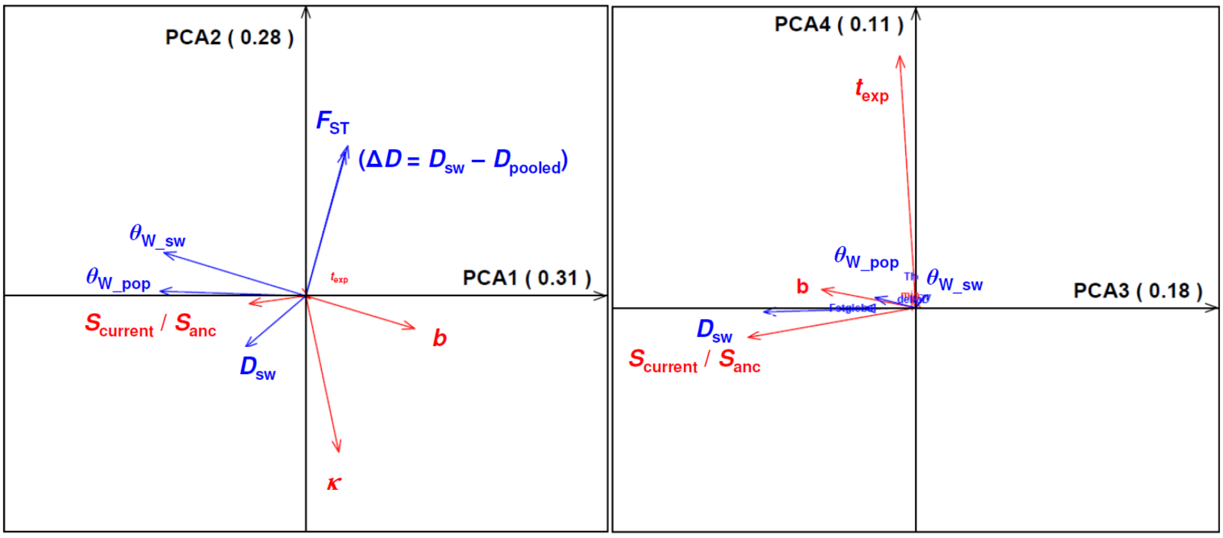


Figure S6A

Figure S6B

The PCA components are based on the parameters of the model because these have uniform distributions, and thus explain the major part of the variance. The main parameters of the model correlated with most summary statistics are thus found in Figure S6A, *i.e.* b and migration rate (κ). The two parameters explaining a smaller amount of the model variance such as the ratio of expansion ($S_{current}$ / S_{anc}) and the time of expansion (t_{exp}), barely visible on Figure S6A, explain mainly axes 3 and 4 of the PCA, respectively.

The major parameter of the model, b, is inversely correlated to the amount of genetic diversity per population (θ_{W_pop}) and at the species wide level (θ_{W_sw}) as evident from equations [4,5] above, but does not influence the level of fixation (F_{ST}; contrary to expectations from ref. (16)). The migration rate κ is correlated to F_{ST}, but also to ΔD the difference between Tajima's D_{sw} and D_{pooled}. This new summary statistic developed here is thus correlated in a metapopulation to the level of spatial structuring (25). The ratio of the species past expansion is indicated by D_{sw}. However, none of our statistics correlate with the time of expansion, explaining the low power of inference on this parameter (see large CI in Figure S4 and Table S10).

Section 8: Influence of deme size distribution on genetic diversity in a metapopulation

We tested here the influence of the distribution of deme sizes in a metapopulation on the overall genetic diversity observed in a species-wide sample. We simulated, using Hudson's *ms*, an island-model with 100 demes linked by equal effective number of migrants K ($K = 4N_1\kappa$) where N_1 is the effective size of the first deme which is fixed (see below). The species-wide sample is composed of 20 demes sampled over the whole range of the metapopulation (*i.e.* every five demes), and sequences are obtained for 100 independent loci assuming no intra-locus recombination.

Two models are compared. Model 1 has all demes with similar effective population size equal to N_1 (as assumed in our other models above). In model 2, each deme size varies and its value is randomly picked from an exponential distribution with mean N_1.

Model 2 is simulated by setting deme 1 with θ_1 for effective population mutation rate ($\theta_1 = 4N_1\mu$) and drawing a set of deme sizes from an exponential distribution (with mean value N_1) using the R software. The genetic diversity is computed as the average π, *i.e.* mean pairwise differences between sequences, across the 100 loci in the species-wide sample using the *sample_stats* program provided with *ms*.

We fixed $\theta_1 = 0.5$, *i.e.* $N_1 = 25{,}000$ with a mutation rate of $\mu = 5\times10^{-9}$ per site per generation for 1,000bp loci, which would translate for example into demes having a census size of 250 with a germination rate $b = 0.1$. We performed 500 simulations for each type of model drawing independent exponential distributions for each simulation. We vary the number of migrants between demes between $K = 0.1$ ($\kappa = 10^{-6}$) and $K = 100$ ($\kappa = 10^{-3}$). We checked that the total census size of the metapopulations are not significantly different between model 1 and 2 over the 500 simulations (Student t-test, $P > 0.5$).

In Figure S7, the genetic diversity represented by π is always higher in the island model 1 with all demes being of equal size, compared to the model 2 with demes sizes being exponentially distributed. This difference is higher at higher migration rates, for which the diversity is overall smaller in the metapopulation (comparing panels A to E). This demonstrates that when migration is weak, a large genetic diversity is generated by the metapopulation structure, independently of the deme sizes. When the deme size is exponentially distributed, the variance of genetic drift is higher due to the presence of very small demes, which then reduces the species-wide genetic diversity in the metapopulation compared to the situation with demes of equal size.

Figure S7: Density distributions of π values (mean pairwise differences between sequences) over 500 simulations for model 1 with all demes with equal sizes (black solid line) and for model 2 with deme size exponentially distributed (black dotted line). The mode of each distribution is indicated. The effective number of migrants between demes per generation varies: A) $K = 0.1$, B) $K = 0.5$, C) $K = 1$, D) $K = 10$, E) $K = 100$.

Figure S7A

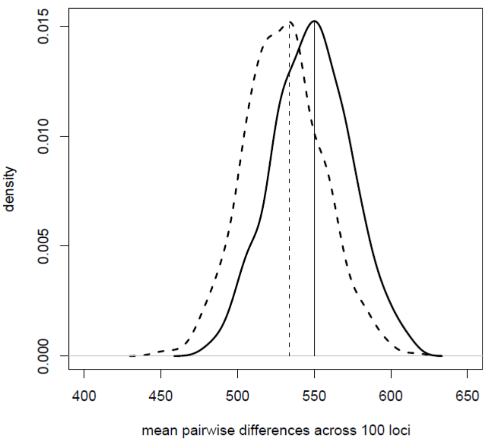


Figure S7B

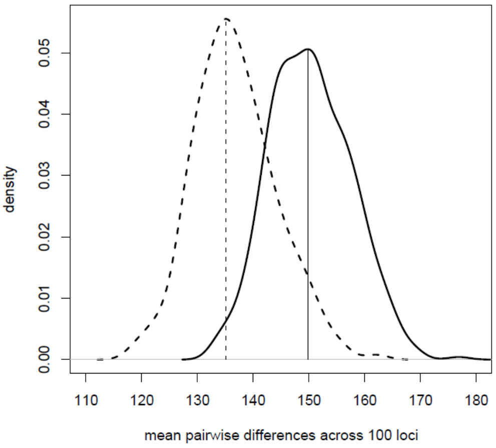


Figure S7C

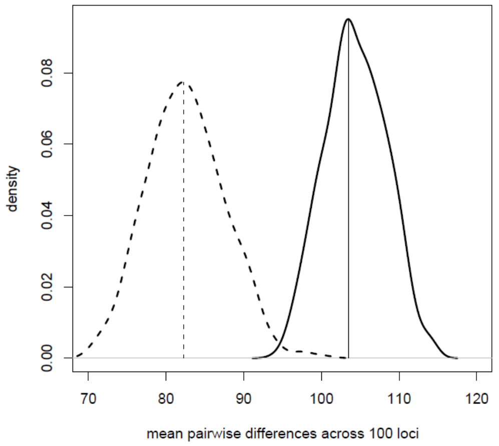


Figure S7D

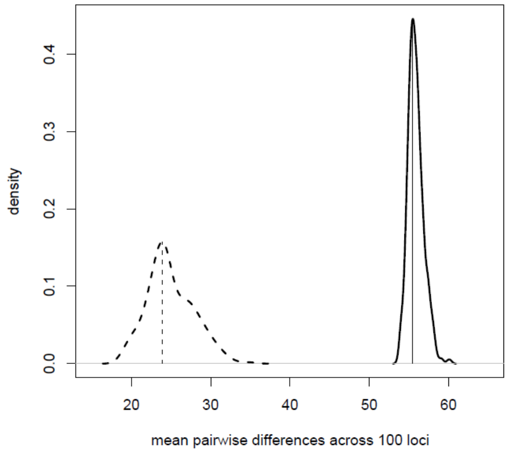
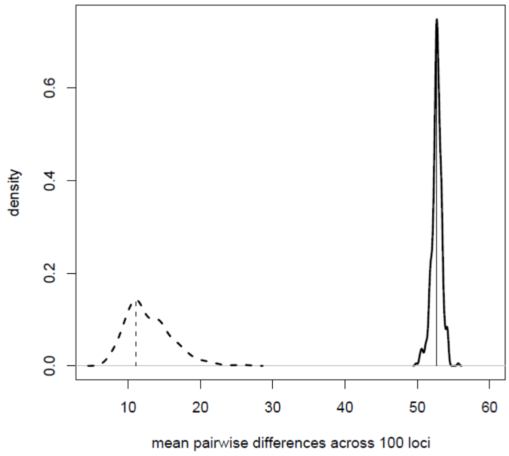


Figure S7E



References cited in the SI

1. Kefi S, et al. (2007) Spatial vegetation patterns and imminent desertification in Mediterranean arid ecosystems. *Nature* 449:213-217.
2. Scanlon TM, Caylor KK, Levin SA, Rodriguez-Iturbe I (2007) Positive feedbacks promote power-law clustering of Kalahari vegetation. *Nature* 449:209-212.
3. Clauset A, Shalizi CR, Newman MEJ (2009) Power-Law Distributions in Empirical Data. *SIAM Rev.* 51:661-703.
4. Nakazato T, Warren DL, Moyle LC (2010) Ecological and geographic modes of species divergence in wild tomatoes. *Am. J. Bot.* 97:680-693.
5. Arunyawat U, Stephan W, Städler T (2007) Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Mol. Biol. Evol.* 24:2310-2322.
6. Städler T, Arunyawat U, Stephan W (2008) Population genetics of speciation in two closely related wild tomatoes (*Solanum* section *lycopersicon*). *Genetics* 178:339-350.
7. Yang ZH, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17:32-43.
8. Tellier A, et al. (2011) Fitness effects of derived deleterious mutations in four closely related wild tomato species with spatial structure. *Heredity* 107: 189-199.
9. Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451-1452.
10. Thornton K (2003) libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* 19:2325-2327.
11. Kaj I, Krone SM, Lascoux M (2001) Coalescent theory for seed bank models. *J. Appl. Proba.* 38:285-300.
12. Levin DA (1990) The seed bank as a source of genetic novelty in plants. *Am. Nat.* 135:563-572.
13. Whittle CA (2006) The influence of environmental factors, the pollen : ovule ratio and seed bank persistence on molecular evolutionary rates in plants. *J. Evol. Biol.* 19:302-308.
14. Honnay O, Bossuyt B, Jacquemyn H, Shimono A, Uchiyama K (2008) Can a seed bank maintain the genetic variation in the above ground plant population? *Oikos* 117:1-5.
15. Nunney L (2002) The effective size of annual plant populations: The interaction of a seed bank with fluctuating population size in maintaining genetic variation. *Am. Nat.* 160:195-204.

16. Vitalis R, Glemin S, Olivieri I (2004) When genes go to sleep: The population genetic consequences of seed dormancy and monocarpic perenniality. *Am. Nat.* 163:295-311.

17. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337-338.

18. Stephan W, Langley CH (1998) DNA polymorphism in *Lycopersicon* and crossing-over per physical length. *Genetics* 150:1585-1593.

19. Wegmann D, Excoffier L (2010) Bayesian Inference of the Demographic History of Chimpanzees. *Mol. Biol. Evol.* 27:1425-1435.

20. Beaumont MA, Zhang WY, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* 162:2025-2035.

21. Excoffier L, Estoup A, Cornuet JM (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* 169:1727-1738.

22. Hamilton G, Stoneking M, Excoffier L (2005) Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proc. Natl. Acad. Sci. U.S.A.* 102:7476-7480.

23. Städler T, Roselius K, Stephan W (2005) Genealogical footprints of speciation processes in wild tomatoes: Demography and evidence for historical gene flow. *Evolution* 59:1268-1279.

24. Willems G, Pison G, Rousseeuw PJ, Van Aelst S (2002) A robust hotelling test. *Metrika* 55:125-138.

25. Städler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P (2009) The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* 182:205-216.