Supporting Information

Hoffman et al. 10.1073/pnas.1318945111

SI Methods

Restricted Site Associated DNA Library Construction. Total genomic DNA was extracted from mouse skin samples and seal kidney samples stored in 95% (vol/vol) ethanol at −20 °C, using a modified phenol-chloroform protocol (1). Eight hundred nanograms of DNA from each sample was individually digested with 20 units SbfI, followed by the ligation of P1 adapters with unique 5-base barcodes for each individual in a restriction site associated DNA (RAD) library. To minimize errors during sequence demultiplexing, at least 2 bases differed between each of the P1 adapter barcodes. Uniquely barcoded samples were pooled and then sheared to ~400 bp on a Covaris S2 sonicator. For each library, fragments in the size range 300–700 bp were excised from an agarose gel. Following end repair and A-tailing, a P2 paired-end adapter (P2 top oligo 5′-/5Phos/CTCAGGCATCACTCGATTCCTCCGAGAACAA-3′ and P2 bottom oligo 5′-CAAGCAGAAGACGGCATACGACGGAGGAATCGAGTGATGCCTGAG*T-3′, where * denotes a phosphorothioate bond; both oligos of the P1 adapters were also modified with a phosphorothioate bond at the same position) was ligated to the size-selected DNA. This template was subjected to 16–17 cycles of PCR enrichment, followed by agarose gel size selection of the 300 to 700-bp fraction. Three oldfield mouse (*Peromyscus polionotus*) and four harbor seal (*Phoca vitulina*) RAD libraries were prepared, each comprising a pool of 14 and 20 individuals, respectively. Each library was paired-end sequenced on an Illumina HiSeq2000 flow cell. DNA quantification was carried out using a Qubit fluorometer (Invitrogen). Agencourt AMPure XP magnetic beads (Beckman Coulter) were used for all reaction clean-up stages.

Bioinformatic Analyses. FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used for initial sequence quality assessment. Fluidics problems were encountered during the sequencing of the harbor seal RAD libraries. Consequently, the last 36 bases of the paired-end reads required trimming due to very low sequence quality. These libraries were sequenced a second time and the two datasets combined for subsequent analyses. Stacks process_radtags.pl (2) was used to filter the raw fastq sequences and to demultiplex the samples according to the P1 barcode. At this stage, 6 oldfield mouse and 20 harbor seal samples with very low numbers of sequences were removed from the datasets. The failure of these libraries was due to the inadvertent use of faulty P1 adapters for these samples.

Our pipeline for obtaining single nucleotide polymorphism (SNP) genotypes from the Illumina sequence data involved: (*i*) clustering of sequences into RAD contigs using Stacks version 0.9999 (2); (*ii*) using the resulting contigs as a reference genome for mapping the sequences within BWA version 0.6.2 (3); and (*iii*) SNP calling using the GATK UnifiedGenotyper version 2.1.13 (4). The GATK UnifiedGenotyper (4) uses a Bayesian genotype likelihood model outputting accurate posterior probabilities of there being a segregating variant allele at each locus as well as for the genotype of each sample. Thus, our pipeline allowed us to take advantage of the more sophisticated and statistically more rigorous genotyping and SNP calling framework implemented within the GATK UnifiedGenotyper compared with Stacks.

i) Clustering of sequences into RAD contigs using Stacks. Sequences from all individuals were combined to create a "superparent". Stacks denovo_map.pl (2) was then used with the superparent acting as a pseudoparent to de novo assemble read 1 sequences into RAD tags. To remove potentially spurious and uninformative tags, the tags generated above were filtered to include only those present in at least two individuals. The paired-end reads corresponding to each of the remaining tags were assembled into contigs using the Stacks sort_read_pairs.pl and exec_velvet.pl scripts (2). Stacks constructs individual tags based on read 1 sequence similarity but does not take into account the paired-end sequence. Therefore, as an additional quality-control step aimed at eliminating any tags potentially comprising more than one locus, only tags for which a single contig was assembled from the paired-end reads were retained. A reference genome was then constructed from these tag sequences together with their corresponding paired-end contigs, padded out with Ns corresponding to the average size of the RAD library sequenced.

ii) Mapping sequences using the reference genome. The original demultiplexed paired-end fastq files were mapped back to the reference genome using BWA (3) with default parameters. SAMtools (5) was used for SAM and BAM file manipulation. Picard MarkDuplicates version 1.89 (http://picard.sorceforge.net) was used to remove PCR duplicates. Individual BAM files for each sample were merged into a single file.

iii) Genotype calling. Genotypes were called using the GATK UnifiedGenotyper (3) with default parameters (6) except with –hets 0.01 to reflect the higher levels of polymorphism found in these mammals compared with humans. As linked SNPs are non-independent, only tags containing a single polymorphic SNP were retained for subsequent analyses. This measure also guards against the inclusion of false-positive SNPs assembled from paralogous loci (7).

Calculation of RAD-Based Heterozygosity and Relatedness. An individual's heterozygosity was calculated as the total number of heterozygous tags divided by the number of tags for which the individual was called. Because not all individuals were called for the same loci, we then standardized individual heterozygosity values by the mean average observed heterozygosity in the population of the subset of loci successfully typed in the focal individual [standardized multilocus heterozygosity (sMLH)] (8). Pairwise relatedness (RAD allele sharing) was calculated as the total number of identical alleles between individuals (zero, one, or two per tag) divided by twice the number of tags considered.

Filtering of Genotypes. To maximize the signal-to-noise ratio, we explored a range of filtering thresholds based on genotype quality (GQ), low coverage (LC), and mapping quality (MQ). The oldfield mouse pedigree was used to assess the impact of such filtering by measuring the strength of correlation between (*i*) pedigree-based inbreeding coefficient *f* and RAD heterozygosity and (*ii*) pedigree-based relatedness and RAD allele sharing (Figs. S3 and S4, respectively). Studies often use a GQ threshold >30. However, applied to our data we observed a strong systematic bias whereby individuals with lower sequence depth of coverage, and therefore fewer called SNPs, tended to be called as highly heterozygous. This probably reflects the fact that a heterozygote genotype can be called with higher confidence at low depth of coverage, but to be confident of a homozygote call requires a larger number of reads. To further explore this bias at low sequence coverage, we randomly subsampled the oldfield mouse RAD sequences across all individuals to mimic 50% and 25% of the actual coverage (Figs. S3 and S4, respectively).

Filtering based on MQ thresholds had a negligible impact on the data. However, the r^2 between pedigree-based *f* and RAD heterozygosity declined with a combination of increasing GQ

and decreasing LC threshold, and this pattern was exacerbated when average sequencing depth was reduced by subsampling the data (Fig. S3). The correlation between pedigree-based relatedness and RAD allele sharing was also optimal for filtering thresholds of GQ ≥ 1 and a LC ≥ 2, as was the number of RAD tags retained (Fig. S4). Consequently, to provide the best balance between SNP quality and the number of tags retained for analysis, we filtered genotypes obtained from GATK using a GQ ≥ 1, corresponding to the maximum-likelihood genotype, and a LC ≥ 2 to generate the final oldfield mouse SNP dataset. The same filtering criteria were subsequently used for the harbor seals.

Computation of g_2 with Large Numbers of Loci. ***Notations.*** In the following we denote H_{ik} an indicator variable that takes a value of 1 if individual k is heterozygous at locus i and 0 otherwise. When there are missing data, some of these values are unknown; in that case we denote $\tilde{H}_{ik}$ a variable that takes a value of 1 if the individual k is known to be heterozygous at locus i and 0 if its genotype at locus i is either unknown or known to be homozygous. x_{ik} denotes a constant that takes a value of 1 if the datum is missing for individual k at locus i and 0 otherwise. The number of individuals in the sample is N, whereas the number of loci is L. We denote $\tilde{h}_k = \sum_{i=1}^{L} \tilde{H}_{ik}$ the number of known heterozygous loci in an individual k, $\mu_i = E(H_i)$ the expectation of true heterozygosity at locus i in the population, and m_i and m_{ij} the proportions of individuals with missing data at locus i and at both locus i and locus j, respectively. We also define $\tilde{\mu}_i = (1 - m_i)\mu_i$ the expected proportion of individuals that can be successfully scored and found heterozygous in the sample. The ms are considered as constants characteristic of a sample, whereas the μs are population characteristics. Hats (^) denote estimates based on data from a sample, rather than true values of population parameters.

Estimation of g_2. The estimates of g_2 presented by David et al. (9) and implemented in the RMES software are (correcting for typographical errors)

$$\hat{g}_2 = \frac{\sum_{i=1}^{L}\sum_{j\neq i}\left(\sum_{k=1}^{N}\tilde{H}_{ik}\tilde{H}_{jk}\right)}{\sum_{i=1}^{L}\sum_{j\neq i}\frac{1}{(N-1)}\left(\sum_{k_1=1}^{N}\sum_{k_2\neq k_1}\tilde{H}_{ik_1}\tilde{H}_{jk_2}\right)} - 1 \quad \textbf{[S1]}$$

in the absence of missing data, which becomes

$$\hat{g}_2 = \frac{\sum_{i=1}^{L}\sum_{j\neq i}\frac{1}{N(1-m_i-m_j+m_{ij})}\left(\sum_{k=1}^{N}\tilde{H}_{ik}\tilde{H}_{jk}\right)}{\sum_{i=1}^{L}\sum_{j\neq i}\frac{1}{N(N-1)(1-m_i-m_j+m_im_j)-N(m_{ij}-m_im_j)}\left(\sum_{k_1=1}^{N}\sum_{k_2\neq k_1}\tilde{H}_{ik_1}\tilde{H}_{jk_2}\right)} - 1$$

in the presence of missing data.

These estimates are impractical when the number of loci is high because of the double summations over all pairs of loci. With 15,000 loci, the double summations take of the order of 0.2×10^9 computation steps (which then have to be multiplied by N^2 as there are also double summations over individuals). To reduce computation time, we can look for an estimate of g_2 that takes a more computationally tractable form. The basic assumption behind this computation (which also underlies the previous estimates) is that the distribution of true heterozygosity is the same in missing data as in nonmissing data. In such conditions it can be shown that

$$g_2 = \frac{1 + (B - C)/A}{1 + \bar{\alpha}} - 1,$$

where

$$A = \sum_{i=1}^{L}\sum_{j\neq i}\tilde{\mu}_i\tilde{\mu}_j = \left(\sum_{i=1}^{L}\tilde{\mu}_i\right)^2 - \sum_{i=1}^{L}\tilde{\mu}_i^2,$$

$$B = VAR\left(\tilde{h}_k\right)$$

$$C = \sum_{i=1}^{L} VAR\left(\tilde{H}_i\right)$$

$$\bar{\alpha} = \left[\sum_{i=1}^{L}\sum_{j\neq i}\tilde{\mu}_i\tilde{\mu}_j\alpha_{ij}\right] \Big/ A,$$

with

$$\alpha_{ij} = \frac{m_{ij} - m_im_j}{(1-m_i)(1-m_j)}.$$

The αs represent the extent to which missing loci are clustered within individuals; in the absence of clustering (i.e., missing data occur independently at all loci), they would be zero; however, it is possible that some individuals may have more missing data than others on average, for example because they would have a lower coverage in the RAD sequences. $\bar{\alpha}$ is the weighted average of the α_{ij}s.

Unbiased estimators of the $\tilde{\mu}_i$s are simply found by averaging over individuals as $\hat{\tilde{\mu}}_i = \frac{1}{N}\sum_{i=1}^{N}\tilde{H}_{ik}$.

Unbiased estimators of A, B, and C can then be obtained as

$$\hat{A}=\frac{N}{N-1}\left[\left(\sum_{i=1}^{L}\hat{\bar{\mu}}_i\right)^2-\sum_{i=1}^{L}\hat{\bar{\mu}}_i^2\right]-\frac{\hat{J}}{N-1},$$

with $\hat{J}=\frac{1}{N}\sum_{k=1}^{N}\tilde{h}_k^2-\sum_{i=1}^{L}\hat{\bar{\mu}}_i$,

$$\hat{B}=\frac{1}{N-1}\left[\sum_{k=1}^{N}\tilde{h}_k^2-\frac{1}{N}\left(\sum_{k=1}^{N}\tilde{h}_k\right)^2\right]$$

$$\hat{C}=\frac{N}{N-1}\left(\sum_{i=1}^{L}\hat{\bar{\mu}}_i-\sum_{i=1}^{L}\hat{\bar{\mu}}_i^2\right).$$

All these equations do not require double summation over loci and can therefore be computed in a reasonable time. However, computing $\bar{\alpha}$ in principle requires this double summation. To avoid this, one can assume that the α_{ij}s do not vary a lot between pairs of loci. With this approximation we obtain the following estimate of $\bar{\alpha}$,

$$\hat{\bar{\alpha}}=\frac{\sum_{k=1}^{N}\hat{M}_k-N\left(\sum_{i=1}^{L}\hat{\bar{\mu}}_i\frac{m_i}{1-m_i}\right)^2+N\sum_{i=1}^{L}\left(\hat{\bar{\mu}}_i\frac{m_i}{1-m_i}\right)^2}{(N-1)\hat{A}+\hat{J}},$$

in which

$$\hat{M}_k=\left(\sum_{i=1}^{L}\frac{\hat{\bar{\mu}}_i x_{ik}}{1-m_i}\right)^2-\sum_{i=1}^{L}\left(\frac{\hat{\bar{\mu}}_i x_{ik}}{1-m_i}\right)^2.$$

This does not require a double summation over loci and the final estimate of g_2 reads

$$\hat{g}_2=\frac{1+\left[\hat{B}-\hat{C}\right]/\hat{A}}{1+\hat{\bar{\alpha}}}-1.$$

This estimate is slightly biased because the ratio of expectations differs from the expectation of a ratio; however, with reasonable conditions (say, when individuals have on average 10 or more successfully scored, heterozygous loci), the bias is very small (9). The SD can be obtained by bootstrapping over individuals. All of the computations were done using a Mathematica program, available upon request (a Windows executable will be made available in the near future).

Bayesian Analysis of Population Structure. Population structure can potentially generate spurious associations between heterozygosity and fitness (10, 11). Consequently, we used the program Structure 2.3.4 (12) to conduct a Bayesian cluster analysis of the harbor seal RAD dataset, comprising 60 individuals genotyped at 14,585 SNPs. Structure uses a maximum-likelihood approach to determine both the most likely number of distinct genetic groups (K) in a sample and the probability of membership of every individual to each of these K groups. We ran three independent runs for K = 1–5 using 100,000 Markov chain Monte Carlo iterations after a burn-in of 100,000 steps and specifying an admixture model with correlated allele frequencies. To be able to detect even a very weak signature of population structure, we implemented the LOCPRIOR model, using age class as prior information to assist the clustering (13). LOCPRIOR tends to outperform the standard model when populations are weakly differentiated, providing more accurate estimates of K together with improved group membership coefficients.

SI Results

Oldfield Mouse. RAD sequencing of 36 oldfield mice generated 265 million paired-end reads, of which 231 million contained appropriate barcodes and the RAD restriction site and passed initial quality filtering, resulting in an average of 6.4 million paired-end reads per individual, varying between 2.5 million and 11 million (Fig. S1). These reads were assembled into 79,360 contigs, which, after eliminating those with multiple paired-end contigs, were reduced to 63,129. Average contig length including the paired end was 479 bp (±117-bp SD). A total of 16,060 contigs (25.4%) mapped to the mouse genome, using an e-value threshold of $1e^{-10}$ to reveal a broad genomic distribution (Fig. S2*A*). Following SNP calling and filtering as described in *SI Methods*, 13,198 RAD tags were retained for downstream analysis, each containing a single biallelic SNP. Most of these SNPs were called in the majority of individuals (Fig. S5).

Harbor Seal. RAD sequencing of 60 harbor seals generated 374 million paired-end reads, of which 280 million contained appropriate barcodes and the RAD restriction site and passed initial quality filtering, resulting in an average of 3.6 million paired-end reads per individual, varying between 0.6 and 9.6 million (Fig. S1). These data were assembled into 126,121 contigs, of which 83,148 were retained after filtering out those with multiple paired-end contigs. Average contig length including the paired end was 538 bp (±135-bp SD). A broad genomic distribution was inferred by mapping 44,961 contigs (35.6%) to the dog genome (Fig. S2*B*). Following SNP calling and filtering, 14,585 RAD tags were retained for downstream analyses, most of which were present in the majority of sequenced individuals (Fig. S5).

Differences in heterozygosity between old and young seals could potentially arise if the two age classes originated from separate populations rather than from the same panmictic population. Consequently, we subjected our harbor seal RAD dataset to Bayesian cluster analysis, using the program Structure (12). The average log-likelihood value increased from K = 1 to K = 2 and thereafter leveled off (Fig. S8*A*). Although this appears to indicate support for the presence of more than one genetic cluster, inspection of the membership probabilities to each of the inferred clusters shows that no signal of population structure is present, with all individuals being predominantly assigned to a single cluster (each bar is almost entirely blue in color) and the remaining clusters making negligible contributions (Fig. S8 *B*–*E*).

1. Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab Press, Cold Spring Harbor, NY) 2nd Ed.
2. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) *Stacks*: Building and genotyping loci *de novo* from short-read sequences. *G3* 1(3):171–182.
3. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics* 26:589–595.
4. McKenna A, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303.
5. Li H, et al. (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25:2078–2079.
6. DePristo M, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–498.
7. Sanchez CC, et al. (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics* 10:599.
8. Coltman DW, Pilkington JG, Smith JA, Pemberton JM (1999) Parasite-mediated selection against inbred Soay sheep in a free-living, island population. *Evolution* 53(4):1259–1267.
9. David P, Pujol B, Viard F, Castella V, Goudet J (2007) Reliable selfing rate estimates from imperfect population genetic data. *Mol Ecol* 16:2474–2487.
10. Slate J, Pemberton JM (2006) Does reduced heterozygosity depress sperm quality in wild rabbits (*Oryctolagus cuniculus*)? *Curr Biol* 16:R790–R792.
11. Luquet E, et al. (2011) Heterozygosity-fitness correlations among wild populations of European tree frog (Hyla arborea) detect fixation load. *Mol Ecol* 20:1877–1887.

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

13. Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Mol Ecol Res* 9:1322–1332.

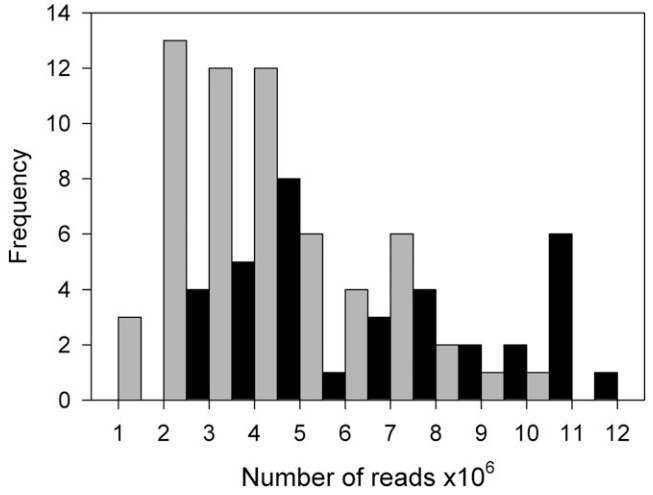

Fig. S1. Distribution of the number of RAD sequence reads obtained across samples. Oldfield mice and harbor seals are denoted by solid and shaded bars, respectively.

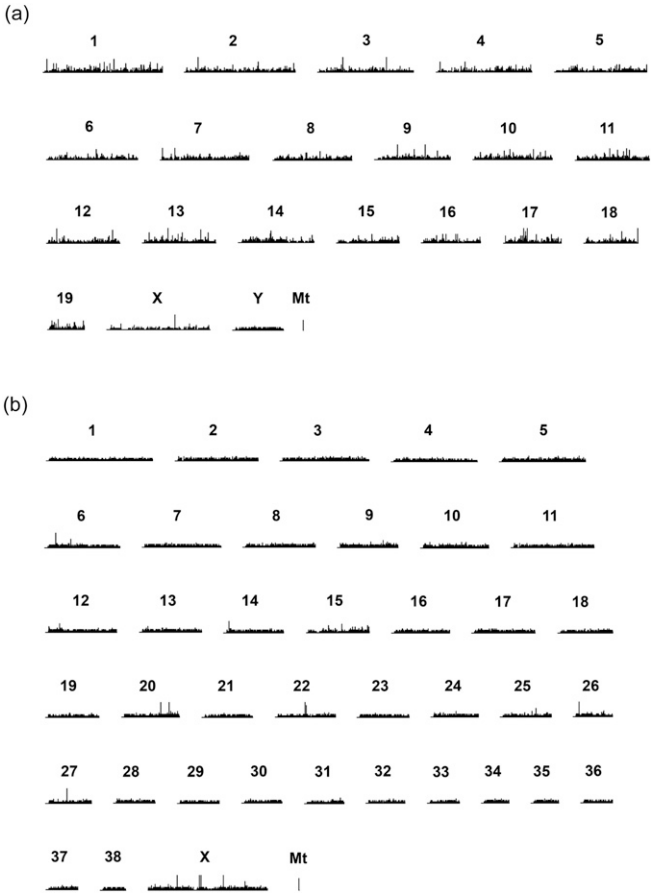

Fig. S2. Inferred chromosomal distributions of (*A*) oldfield mouse and (*B*) harbor seal contigs (details in *SI Methods*). The oldfield mouse contigs were BLASTed against the mouse (*Mus musculus*) genome and the harbor seal contigs were BLASTed against the dog (*Canis familiaris*) genome, using an *e*-value cutoff of $1e^{-10}$. Chromosomal distributions are based on a bin size of 1,000 bp with the *x* axis being scaled relative to the largest chromosome and the maximal *y* axis being 10 contigs per bin.

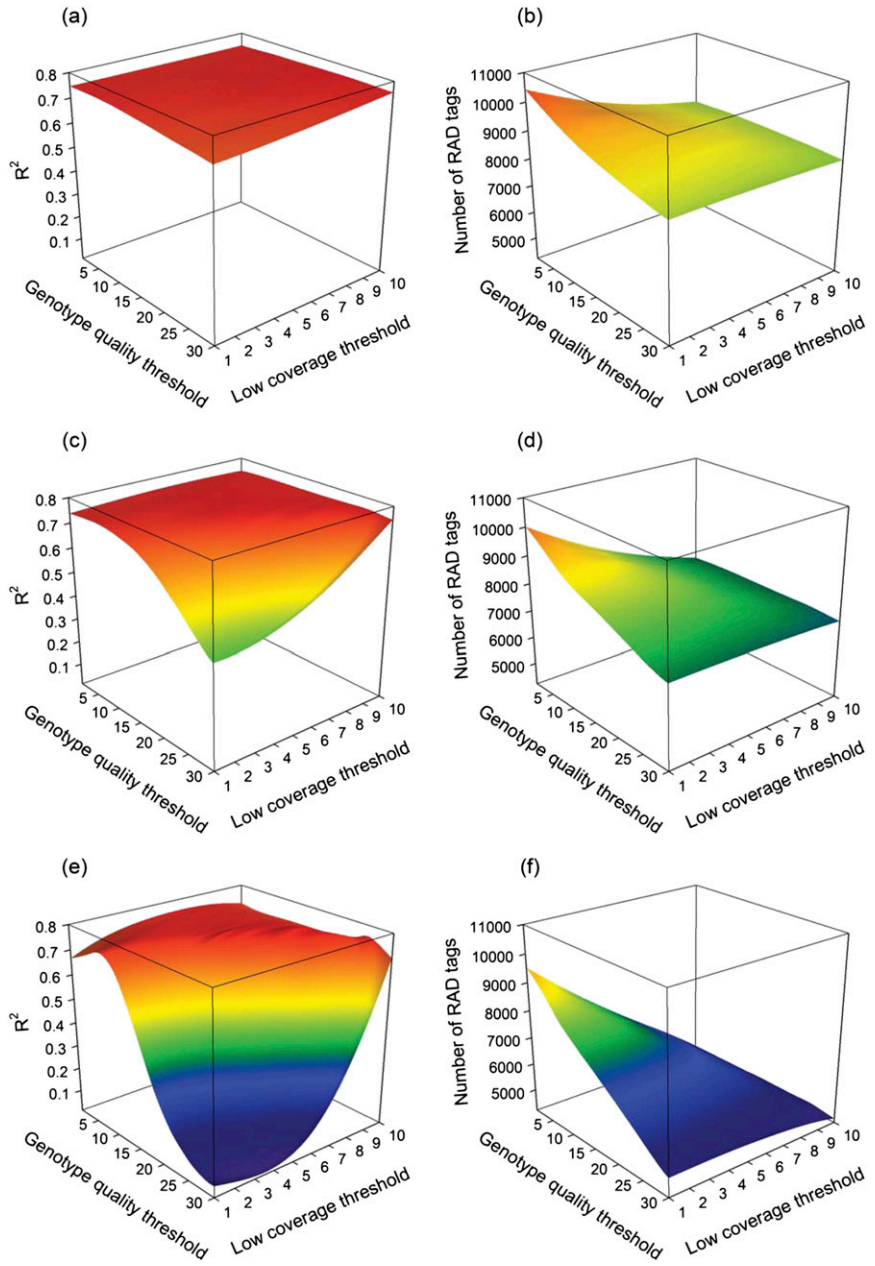

Fig. S3. The effect of varying two genotype filtering thresholds, genotype quality and low-coverage cutoff, on the strength of the relationship between pedigree-based inbreeding coefficient *f* and RAD heterozygosity in the oldfield mouse. *A*, *C*, and *E* show the r^2 between pedigree *f* and RAD heterozygosity based on 100%, 50%, and 25% of the sequence data, respectively, across all 36 individuals. *B*, *D*, and *F* show the corresponding median numbers of RAD tags retained for analysis. Local regression, implemented using the "locfit" package in R, was used to fit smoothed splines to the raw datasets.

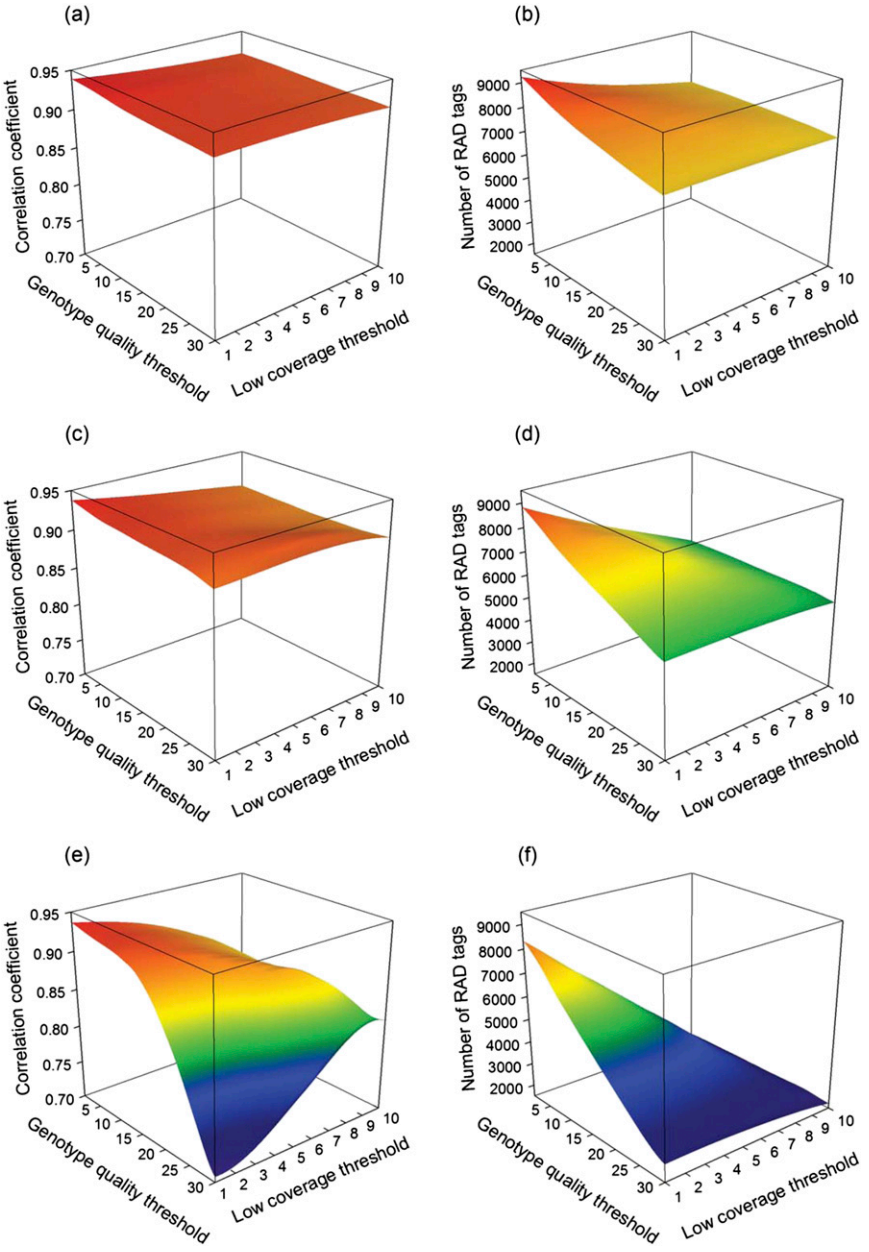

Fig. S4. The effect of varying two genotype filtering thresholds, genotype quality and low-coverage cutoff, on the strength of the relationship between pedigree-based relatedness and RAD allele sharing in the oldfield mouse. *A*, *C*, and *E* show the correlation coefficient between the two relatedness measures based on 100%, 50%, and 25% of the sequence data, respectively. *B*, *D*, and *F* show the corresponding median numbers of RAD tags retained for analysis. Note that fewer RAD tags in general are retained for allele sharing because each tag needs to be called in more than one individual to be counted. Local regression, implemented using the locfit package in R, was used to fit smoothed splines to the raw datasets.

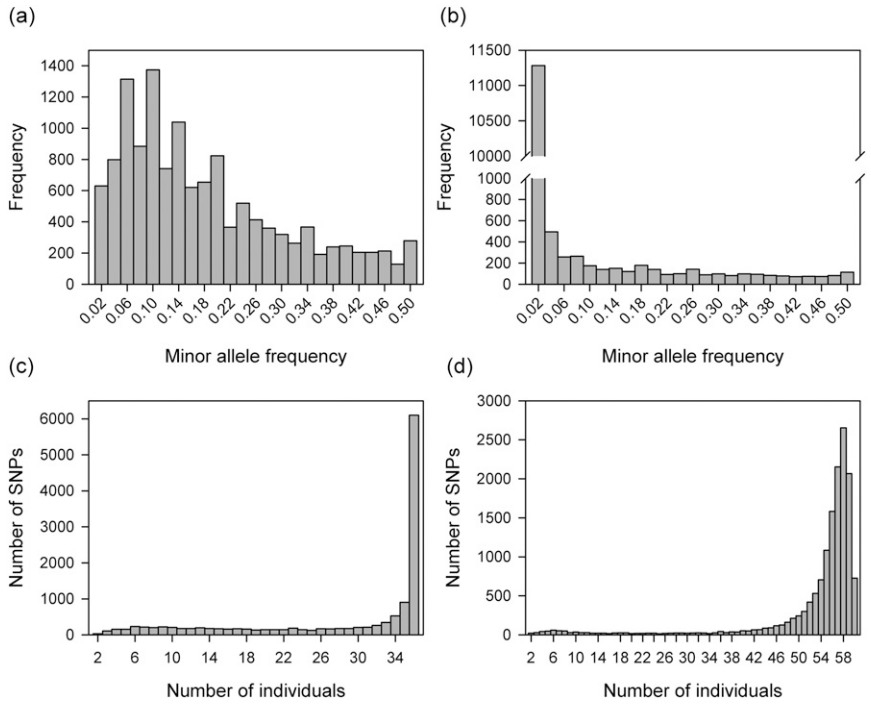

Fig. S5. Summary of 13,198 SNPs called in the oldfield mouse (*A* and *C*) and 14,585 SNPs called in the harbor seal (*B* and *D*). *A* and *B* show the distribution of SNP coverage across individuals. The majority of SNPs were called in most of the individuals. *C* and *D* show corresponding minor allele frequency (MAF) distributions.

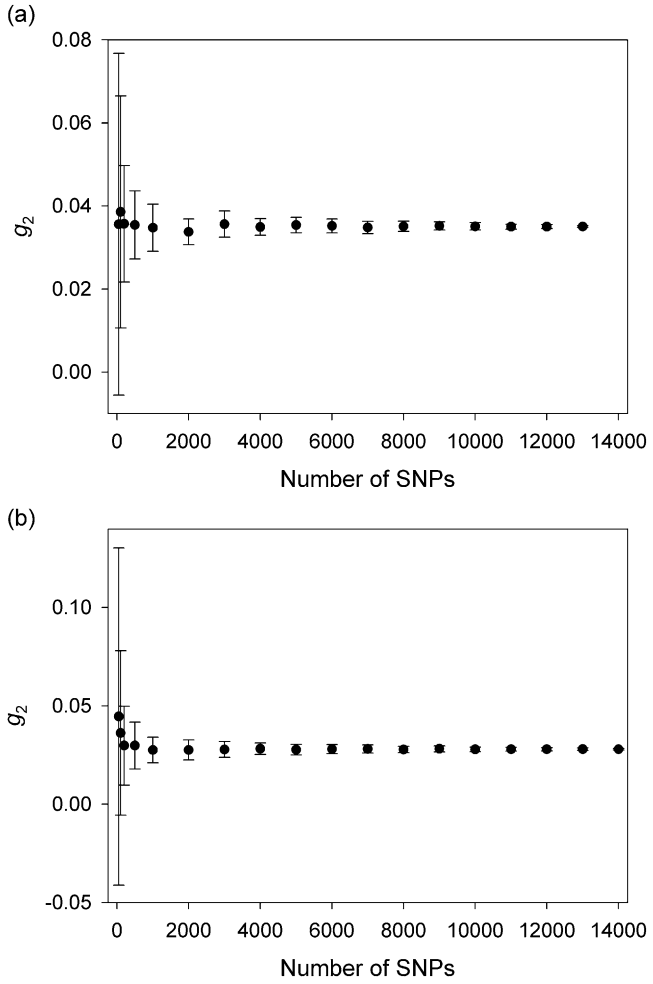

Fig. S6. Relationship between the number of randomly subsampled SNPs and g_2 (±SD) in (*A*) oldfield mice and (*B*) harbor seals.

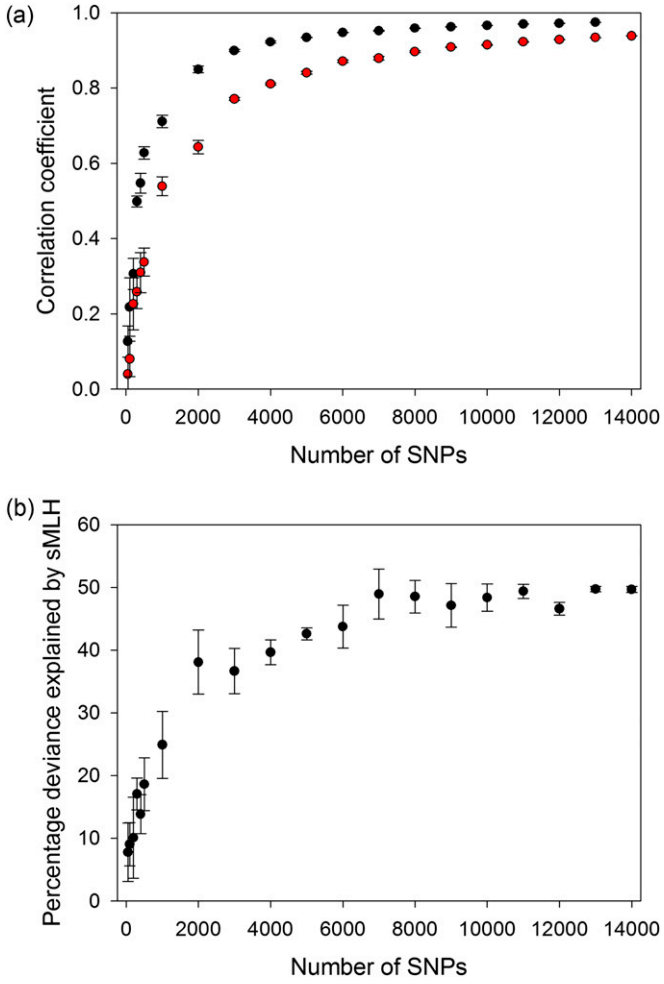

Fig. S7. Relationship between the number of randomly subsampled SNPs and (*A*) the mean (± SEM) correlation in sMLH among markers, with black circles denoting oldfield mice and red circles harbor seals, and (*B*) the mean (± SEM) percentage deviance explained by sMLH in the generalized linear model (GLM) of parasite infection in harbor seals.

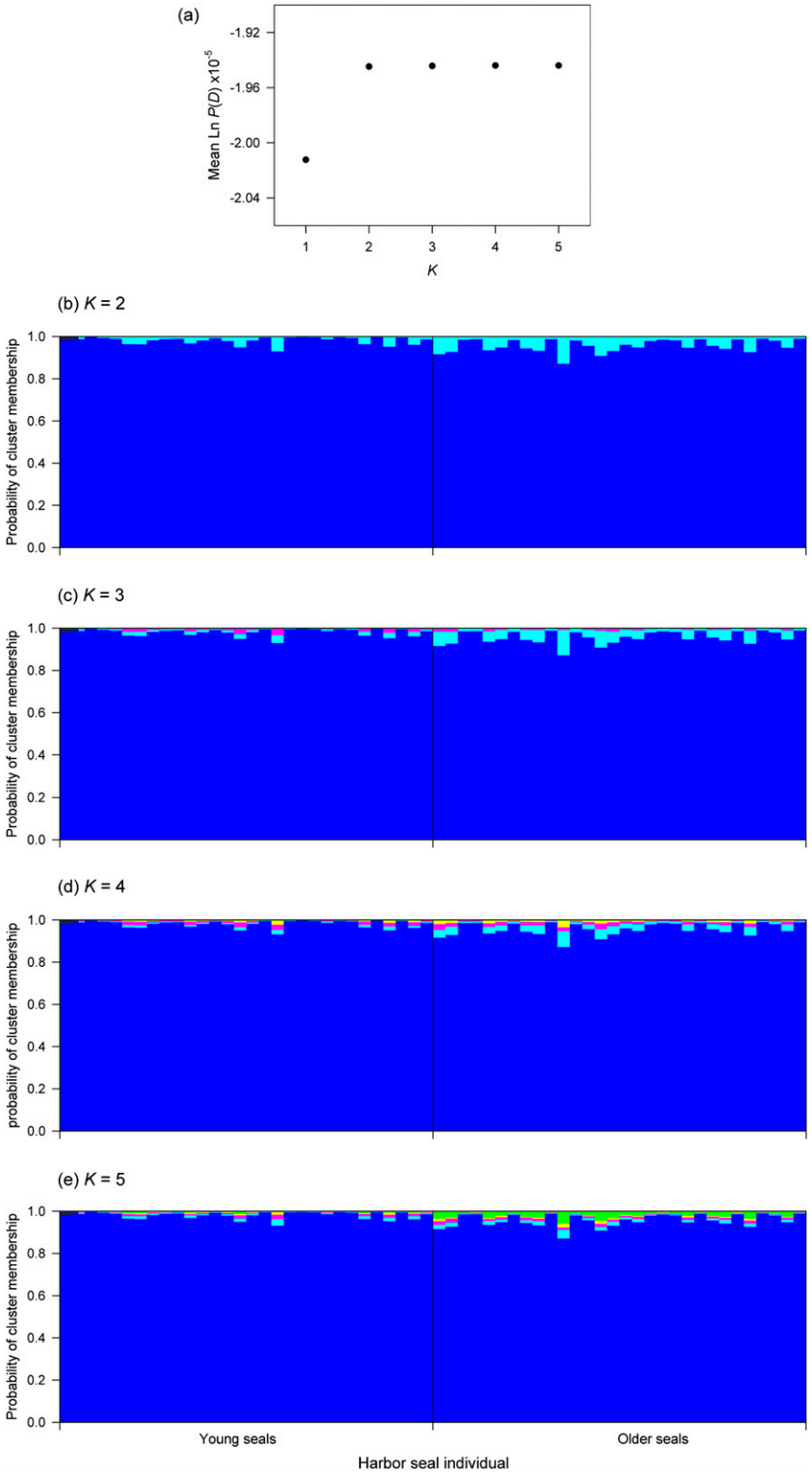

Fig. S8. Results of the Structure analysis of the harbor seal RAD dataset. (*A*) Average log-likelihood value values based on three replicates for each value of *K*, the hypothesized number of clusters in the data. (*B*–*E*) Clustering results shown separately for $K = 2$–5. Each individual is represented by a vertical bar partitioned into different segments, the lengths of which indicate the probability of membership in the different clusters.

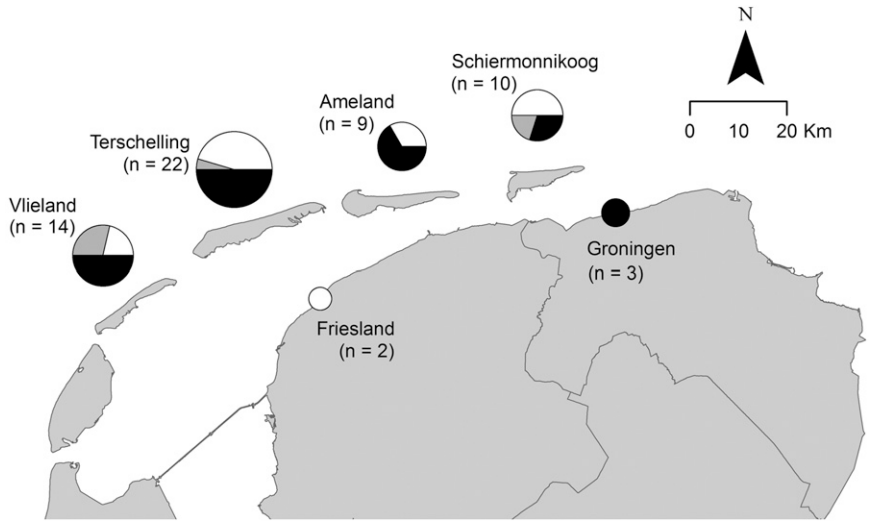

Fig. S9. Spatial distribution of the harbor seals in our sample set. Young seals with lungworm, young seals without lungworm, and old seals without lungworm are denoted by open, shaded, and solid sections, respectively. The diameter of each pie chart corresponds to the number of seals tested per location.