Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication

Carneiro et al.

Supplementary Materials

www.sciencemag.org

Methods

Tables S1 to S7

Figs. S1 to S7

Databases S1 to S5

Full reference list

List of Supplementary material:

Table S1. Summary statistics of the rabbit OryCun2.0 assembly.

Table S2. Comparison between rabbit genome assembly versions OryCun1.0 and OryCun2.0.

Table S3. Summary of RNA-sequencing data used for annotation of the rabbit genome.

Table S4. Sample details of wild and domestic rabbits used in this study both for whole-genome resequencing and DNA sequence capture using hybridization on microarrays. Average sequence coverage per sample is provided.

Table S5. Summary of SNPs and Insertions/Deletions detected in the rabbit genome.

Table S6. Distributions of SNP counts in the different delta allele frequency bins for conserved non-coding elements, UTRs, coding sequences and introns. Deviations from expected values were tested with a standard X^2–analysis (d.f.=1). M-values were calculated using the average frequency of the corresponding annotation category as reference.

Table S7. Summary of electrophoretic mobility shift assays using nuclear extracts from ES-cell derived neural stem cells (ES) or from mouse P19 embryonic carcinoma cells before (un-diff) or after neuronal differentiation (diff). "-" = no shift, "+" = shifted, "++" = lower band intensity, "+++" = whole band/complex disappeared and "?" = unclear.

Fig. S1. Genetic relatedness between populations sampled in this study **(A)** Heat map (color code to the right) of identity scores based on comparing resequencing data with the assembly. The x-axis represents genome coordinates with chromosome 1 to the left and chromosome X to the right. The first row (R) represents the reference rabbit against itself, rows 8-21 correspond to locations marked with red dots in Fig. 1A, ordered according to a northeast to southwest transection. **(B)** Strong correlation in allele frequencies between domestic and French wild rabbits; RAF French and RAF Domestic represent the frequency of the reference allele in wild French and domestic rabbits, respectively.

Fig. S2. Distributions and scatterplot of heterozygosity and F_{ST} in 50 kb windows across the rabbit genome. **(A)** Distribution of pooled heterozygosity in pools of domestic rabbits. The arrow indicates the upper heterozygosity threshold for opening a sweep. **(B)** Distribution of F_{ST} values observed in the contrast domestics vs. wild French. The arrow indicates the lower F_{ST} threshold for opening a sweep. **(C)** Scatter plot of F_{ST} and pool heterozygosity. Red dots represent windows fulfilling the sweep opening requirement ($F_{ST} \geq 0.35$ and heterozygosity ≤ 0.05).

Fig. S3. Demographic model and magnitude of the domestication bottleneck estimated using the targeted capture dataset. **(A)** Schematic representation of the coalescent model used to represent the main events of the demographic history of wild and domestic rabbits. See Supplementary Methods for a detailed description of all parameters and assumptions. **(B)** Diagram representing the multilocus maximum-likelihood (ML) estimate for the magnitude of the domestication bottleneck (k_{dom}). Multilocus ML (y axis) of the parameter k_{dom} (x axis) was estimated as the product of the likelihood for each locus.

Fig. S4. Comparison between observed and simulated data under the estimated demographic model. Observed and simulated data are represented in red and black, respectively. The first two panels (**A** and **B**) illustrate the relationship between values of both π and θ_w in wild rabbits from France (x axis) and domestic rabbits (y axis). The two histograms on the bottom row (**C** and **D**) illustrate the distribution of F_{ST} and percentage of shared mutations between the two populations.

Fig. S5. Gene overrepresentation analysis using the GREAT software based on genes associated with high delta allele frequency SNPs ($\Delta AF \geq 0.8$) at non-coding conserved sites. **(A)** Gene

Ontology categories within Biological Processes, **(B)** MGI mouse expression data, and **(C)** MGI mouse phenotypes. Each row in the heat maps represents one specific category and colors on that row indicate the proportion of shared genes in relation to the other categories (ordered in the same way on the x-axis as on y-axis). To the left of each cluster, the type of terms in that cluster is summarized. Bars immediately left of heat-maps visualize the significance-, enrichment- and number of genes of each significant term (P= -log10 Bonferroni-corrected P-value; E=Enrichment: N=number of genes). For P, E, and N, the ranges are indicated below the plot. The full results are presented in Database S3.

Fig. S6. Selective sweep at IMMP2L on chromosome 7. Heterozygosity plots for wild (red) and domestic (black) rabbits together with plots of F_{ST} values and high ΔAF (HΔAF) SNPs with ΔAF>0.75. Putative sweep regions detected with the F_{ST} -Heterozygosity outlier approach and SweepFinder are marked with horizontal bars. Gene annotations in sweep regions are indicated. IMMP2L was not predicted by Ensembl but by our in-house predictions based on RNA-sequencing data. The human consensus model features all human IMMP2L transcripts in a collapsed fashion. Asterisks (*) represent ENSOCUT000000. Red dots indicate the location of the high ΔAF insertion/deletion in a conserved element for IMMP2L.

Fig. S7. Results for all 17 SNPs functionally examined by electrophoretic mobility shift assays (EMSA). Results of EMSA using nuclear extracts from ES-cell derived neural stem cells or from mouse P19 embryonic carcinoma cells before (un-diff) or after neuronal differentiation (diff) are presented. WT=wild-type allele; D=domestic, the most common allele in domestic rabbits. Cold

probes at 100-fold excess were used to verify specific DNA-protein interactions. The results are summarized in Table S7.

Database S1. Regions inferred to have been targeted by directional selection using: 1) a F_{ST}-*H* outlier approach contrasting genetic diversity between wild and domestic rabbits, 2) allele frequency spectra (*SweepFinder*), and 3) an explicit demographic model contrasting genetic diversity between wild and domestic rabbits (capture arrays data).

Database S2. *P*-values obtained using coalescent simulations of the demographic scenario inferred in this study for the *SweepFinder* and targeted capture analyses. The obtained value for the magnitude of the domestication bottleneck in this study (k_{dom} = 1.3) was lower than a previous estimate (k_{dom} = 2.8). Under the model used to infer directional selection in the domesticated lineage using the targeted capture dataset, the estimation from this study describes a stronger bottleneck and thus renders our selection tests conservative. However, this is not the case for the *SweepFinder* analysis and *P*-values using both bottlenecks estimates are provided.

Database S3. Full results of overrepresentation analysis using SNPs at conserved non-coding sited with ΔAF>0.80 performed using the GREAT tool.

Database S4. Missense SNPs showing a delta allele frequency difference of 0.90 or higher between wild and domestic rabbits.

Database S5. Duplications and deletions showing allele frequency differences between wild and domestic rabbits according to an ANOVA analysis.

Supplementary Methods

Genome assembly and annotation

Genome sample collection. Eight adult female rabbits of the Thorbecke New Zealand white partially inbred line were obtained from Covance Research Products. This entire line was destroyed in a fire in 2005, and thus is no longer available. Sequence heterozygosity of these samples was assessed at up to 267 random nuclear loci through PCR and Sanger sequencing at the Broad Institute. Heterozygosity was calculated based on high quality heterozygous single nucleotide polymorphisms (SNPs) inferred from the chromatograms. The sample with the lowest heterozygosity was selected as reference individual to whole genome sequencing.

Genome sequencing and assembly. An initial 2X assembly of the rabbit genome (OryCun1) was constructed using paired-end reads from 4 kb plasmids and 40 kb fosmids from a single female rabbit (Table S2). This assembly was described in Lindblad-Toh *et al.* (*8*). The assembly described in this paper (OryCun2) used the sequencing reads from OryCun1 as well as additional paired-end reads from 4 kb plasmids, 10 kb plasmids, 40 kb fosmids, and bacterial artificial chromosomes (BACs). All sequenced reads derive from the same individual rabbit, except for those from BACs (see below). Genome assembly was performed using the software package Arachne 2.0 (*20*). The 6.55X OryCun2 assembly is comprised of 5.04X 4 kb reads, 1.14X 10 kb reads, 0.33X 40 kb reads, and 0.04X BAC-derived reads. This assembly has an N50 contig size of 64.65 kb (i.e. half of all bases reside in a contiguous sequence of 64.65 kb or more), an N50 scaffold size of 35.92 Mb and a total assembly size of 2.66 Gb (Table S1). It shows a heterozygosity of 1/3,506 bp, a GC content of 43.8% and is 16.7% repetitive as defined by 48-

mer counts. OryCun2 used 92.3% of all reads produced, comprising 94.3% of all bases produced. It is of similar quality to other Sanger-based draft assemblies.

OryCun2 was anchored to chromosomes using a cytogenetically anchored microsatellite map (*21*). 364 BACs were end-sequenced, resulting in the anchoring of 99 scaffolds, comprising 82% of the genome assembly, or 2.178 Gb. Of the 2.178 Gb anchored, 238 Mb were only ordered, while 1.940 Gb were ordered and oriented.

BAC library. The white blood cells of a New Zealand white rabbit of unknown sex were embedded in agarose plugs at the concentration of 10^7 per ml, treated with ESP buffer (50mg/ml proteinase K, 1% Sarkosyl, and 0.5M EDTA, pH 8.0), and rinsed with TE until suitable for enzymatic digests. The DNA was partially digested with EcoRI and EcoRI methylase, size selected, and cloned into the pBACe3.6 vector as described (*22*). The ligated DNA was then transformed into DH10B electro-competent cells (Invitrogen). The library was arrayed into 322 384-well plates. The average insert size of 269 randomly selected clones is 175 kb, and about 92% of these clones are greater than 140 kb. This library represents about 7-fold coverage of the rabbit genome.

RNA sequencing, transcriptome assembly, and annotation. A panel of 19 RNA samples derived from New Zealand white rabbits was used. Ten RNA samples (nine different tissues from a single female and testis from a male) were purchased from the company Zyagen while the other nine samples were isolated from INRA 1077 New Zealand white rabbits (Table S3). Nineteen strand-specific dUTP libraries (*23*) were produced from Oligo dT polyA-isolated RNA. The libraries were sequenced using Illumina Hi-Seq instruments, producing 76 bp single end reads (3-4 Gb of sequence/tissue). All nineteen RNA-seq datasets were assembled via the genome-independent RNA-seq assembler Trinity (*24*).

The genome assembly was annotated both by the Ensembl gene annotation pipeline (Ensembl release 73, Sept. 2013) and by a novel methodology using both RNA-seq and orthologous annotation in human. The Ensembl gene annotation pipeline created gene models using UniProt protein alignments and RNA-seq data. This pipeline produced 24,964 transcripts arising from 19,293 protein coding genes and 3,375 short non-coding transcripts. The custom pipeline created gene models using the same RNA-seq panel, and older Ensembl rabbit (Genebuild 71.3) and human annotations (Genebuild 71.37). It produced 19,118 high-confidence protein-coding genes, 881 low-confidence protein-coding genes, 1,318 spliced antisense transcripts, 2,243 unspliced antisense loci, 2,746 high confidence lncRNA genes and 48,794 low-confidence non-coding transcripts. Our analysis of rabbit domestication used Ensembl annotations as well as the custom pipeline for annotation of UTRs, non-coding RNA and non-coding conserved elements.

Genome resequencing and data analyses

Sampling. To identify regions of the genome likely to have been targeted by selective breeding at domestication and shared across breeds, we obtained whole genome sequence data for both domestic and wild animals using a pool-sequencing approach (Table S4). Our sampling of domestic rabbits followed that of Carneiro *et al.* (*3*), and includes individuals from breeds representing a wide range of phenotypes and for which historical records indicate a mostly unrelated origin. This sampling design focusing on divergent breeds should assure an approximation to the ancestral genetic diversity prior to breed formation and captured in the early years of rabbit domestication. In total, we sampled six domestic breeds (Table S4). From the wild, we sampled individuals from three localities in southern France (i.e. the ancestral population that gave origin to domestic rabbits) and from 11 localities within the ancestral range

of rabbits in the Iberian Peninsula, including localities within the range of both rabbit subspecies (Table S4). The number of individuals sampled for each breed and locality varied from 10-20. A single snowshoe hare (*Lepus americanus*) individual was sampled as outgroup.

Resequencing, alignment and SNP calling. Paired-end sequencing libraries were generated from domestic and wild rabbit DNA pools using standard procedures. The resulting libraries were sequenced as 2X 76bp paired-end reads using Genome Analyzer II (Illumina), to a coverage of ~10X per pool (Table S4). Sequence reads were then mapped to the reference genome assembly (OryCun2) with the Burrows-Wheeler alignment tool (*25*) using default parameters, except for the q-parameter, indicating the base quality cut-off for soft-clipping reads, which was set to 5. We then marked duplicate reads using Picard (*26*). The mapping distance distribution of paired-end reads revealed that a large proportion of paired-end reads had been generated from fragments smaller than the length of the two reads (76bp x 2). In order not to bias allele frequency estimations by counting overlapping parts of the same molecule twice, we merged overlapping paired-end reads into a single read by using a custom python script. For mismatching bases in overlapping parts, the highest quality base and its quality value were retained and for overlapping bases that agreed the base qualities were rescaled ($Phred_{OL}=Phred_{R1}+Phred_{R2}$) to reflect the increased confidence of base calling.

SNP calling was performed using Samtools (*27*) (version 0.1.19-44428cd) and the output was further filtered using the mpileup2snp option of VarScan v2.3.3 (*28*). This approach enabled the detection of more than 61 x 10^6 raw SNPs, a set further processed by using a custom script requiring that: i) the sum of read depths (RD) across all populations $100 \geq RD \leq 350$ for autosomes and $100 \geq RD \leq 270$ for the X chromosome; ii) the least abundant SNP allele was observed in at least three independent reads at a frequency ≥ 0.01; iii) the most commonly

observed variant allele constituted ≥ 85% of the total count of all three possible variant alleles; iv) the reference allele was represented by at least one read in the dataset; v) that the variant allele was not observed in the reference individual; and vi) the average Phred scaled base quality of the variant allele was ≥15. Application of these filters retained 50,165,386 SNPs for downstream analyses and the allele counts for each of the sequenced rabbit pools were extracted using mpileup from the Samtools package.

Read alignments from the Snowshoe hare (*Lepus americanus*) individual were interrogated at rabbit SNP positions to infer the ancestral and derived states of alleles. For a position to be assessed it was required that either the rabbit reference or variant allele was the only observed allele in the hare and that this allele was supported by ≥3 reads with an average mapping quality of ≥20.

Genome-wide identity scores from pool-sequencing data. To visualize genetic relatedness between populations we calculated identity scores (IS) of individual SNPs in relation to the reference assembly using the reference allele frequencies (AF_{REF}) observed for each sequenced pool as well as for the resequenced reference individual (IS= AF_{REF}). Identity scores for 50 kb windows presented along the genome in Fig. S1A were calculated as the average IS of all SNPs in a window.

Genome-wide estimates of nucleotide diversity from pool-sequencing data. To estimate levels of nucleotide diversity (π) across all sampled populations we used the PoPoolation package (*29*), which implements corrections for biases introduced by pooling and sequencing errors. We obtained genome-wide values of π by averaging estimates computed along each chromosome in 50 kb non-overlapping windows. We required per position a minimum coverage of 4, a maximum coverage of 30, at least two reads supporting the minor allele in polymorphic

positions, and a Phred quality score of 20 or higher. Windows with less than 60% of positions passing these quality filters were discarded.

Identification of regions consistent with directional selection in the domesticated lineage. The merits of model-based methods using simulations versus outlier methods based on genome-wide empirical distributions of summary statistics for detecting genomic regions targeted by directional selection have frequently been discussed. Here, we use both these approaches. We started by using an outlier-based approach that is free from any assumptions regarding domestic rabbits' demographic history and we searched for regions showing unusual levels and patterns of genetic diversity using statistics summarizing heterozygosity and differentiation. We then estimated a neutral demographic scenario for rabbit domestication taking advantage of individual genotypes resulting from an additional dataset where we resequenced to high coverage 5,000 fragments enriched by DNA hybridization on microarrays. This null demographic model was then used to search for signatures of selection in the domesticated lineage using methods relying on different aspects of the data, including the allele frequency spectra, heterozygosity, and genetic differentiation. We described these approaches in detail below.

Inference of selection using heterozygosity and F_{ST}. In order to define candidate regions having undergone directional selection during rabbit domestication we started by using an approach combining heterozygosity estimates for the six sequenced domestic breeds with estimates of F_{ST} between domestic and French wild rabbits. We calculated pooled heterozygosity (H_P) for 50 kb windows iterated every 25 kb along each chromosome. H_P was calculated independently for the three main populations considered in this study: 1) six sequenced pools of domestic rabbits (H_{PDOM}); 2) three pools of wild rabbits from France (WF); and 3) eleven pools of wild rabbits sampled across the Iberian Peninsula (WI). For each SNP in each pool combination, we

determined the major allele from the sum of observed reference and variant alleles and then proceeded to calculate the average frequency of the major allele (MajFreq) and minor allele (MinFreq) in the individual pools included in that particular pool combination: H_P was then calculated for individual SNPs in each sequenced pool by $H_P = 2(\text{MajFreq} * \text{MinFreq})$. H_{PDOM} of individual SNPs were calculated as the mean H_P of the individual pools in which read depth ≥ 3 and H_{PDOM} of 50 kb windows was calculated as the average H_P of all SNPs in that window. F_{ST} was calculated for individual SNPs between domestic rabbits and wild rabbits from France using the formula by Weir & Cockerham (*30*) and F_{ST} of 50 kb windows were then calculated as the average values of all SNPs in a window.

For selective sweep calling based on extreme H_{PDOM} and F_{ST} values of 50 kb windows we consulted the distributions of H_{PDOM} and F_{ST} (Fig. S2) and selected the joint criterion of $H_{PDOM} \leq 0.05 + F_{ST} \geq 0.35$ to open a selective sweep and then extended the sweep to each side for as long as windows fulfilled either $H_{PDOM} \leq 0.05$ or $F_{ST} \geq 0.35$. In order not to excessively fragment predicted selective sweeps, regions separated by two or fewer windows not meeting the above extension criteria were collapsed into a single putative sweep region.

Inference of selection using the site frequency spectrum and a demographic model. To detect evidence of directional selection we used an additional approach (*SweepFinder*) that uses a likelihood framework to search for local deviations in the site frequency spectrum when compared to the remainder of the genome, evaluating for each location a selective sweep with a neutral model (*7*). The genomic background frequency spectrum was obtained for each chromosome independently and the likelihood ratio between neutral and selective models was calculated for a different grid size for each chromosome in order to obtain estimates approximately every 10 kb. Following Williamson *et al.* (*31*), we included SNPs that were

monomorphic within the targeted subpopulation (i.e. domestic rabbits), but variable in the combined sample (i.e. domestic and wild rabbits from France). This should increase the power to detect recent population-specific sweeps in the domesticated lineage that have eliminated most genetic variation, while accounting for mutation rate heterogeneity across the genome. The number of chromosomes sampled for each position was equal to the number of reference and alternative allele counts. The ancestral state of mutations was polarized using the French wild rabbits (i.e. the most common allele was picked as the ancestral allele).

A null distribution is required to infer the statistical significance of the selective sweep hypothesis. To create a null distribution of the likelihood ratio statistic we used neutral coalescent simulations according to a non-equilibrium demographic model based on historical records and incorporating the estimated magnitude of the domestication bottleneck obtained from genetic data (Fig. S3A) (see below for details). We performed 1,000 simulations of 4Mb segments for a total of 200 chromosomes (the total number of chromosomes of domestic rabbits for the six breeds) and a SNP density equal to the observed data. Sequencing pooled DNA results in an additional source of error associated with sampling with replacement of the sequenced alleles. The average coverage (~60X) for domestic rabbits in our dataset was much lower than the total number of chromosomes in the pools (200 chromosomes of domestic origin), and thus allele counts for each position are likely to mostly represent different chromosomes. Nevertheless, in order to more closely mimic the pool sequencing data structure we resampled alleles in the simulated data for each position by randomly drawing values from the empirical distribution of coverage in the observed data. The background frequency spectrum was obtained for each simulation independently and, similarly to the observed data, the likelihood ratio was calculated every 10 kb. Sweep regions were considered significant at a $P<0.001$ (i.e. likelihood ratio values higher than

the top value observed in our simulated data) and the borders of these regions were extended by aggregating genomic positions while P<0.01. Sweep regions separated by less than 10 genomic positions with *P*-values not meeting the above criteria were collapsed into a single region.

The magnitude of the domestication bottleneck (k_{dom}) estimated in this study (Fig. S3B) describes a more stringent bottleneck (k_{dom} = 1.3) when compared to a previous estimate (k_{dom} = 2.8) (*3*). The power to uncover regions consistent with directional selection using the allele frequency spectra has been shown to vary with the magnitude of the bottleneck, and not always in the more intuitive direction (*31*). For example, CLR values and their variance are lower in stronger bottlenecks than that in bottlenecks of intermediate strength or even in equilibrium models. Although the previous estimate in rabbits was based on a much smaller dataset (16 fragments and biased towards the X-chromosome) when compared to the 5,000 fragments used here, we created an additional null distribution using the same non-equilibrium demographic model but incorporating this less stringent estimate. The distribution is slightly inflated but 63% of regions still displayed CLR values that were not observed in the null distribution and the remaining regions were highly significant ($P \leq 0.003$). Given that our current estimate is based on a much larger dataset, and thus likely to be more robust and more representative of genome-wide patterns of genetic diversity, we list in the main manuscript all regions identified using this estimate but both *P*-values are available in Database S2. Genes within regions robust to these different demographic models are the most promising candidates to follow up with functional studies. For instance, regions containing GRIK2 and SOX2 (Fig. 2) are amongst these regions.

A potential confounding factor in both our sweep analysis is artificial selection associated with breed formation. These later events in the domestication process could potentially result in selection being inferred for genomic regions not associated with the initial steps of domestication.

We examined this by testing whether two genes (*ASIP* and *TYR*) that control breed-characteristic coat colour phenotypes overlap regions detected in our genomic scan. Three of the six breeds carried mutations at these loci, New Zealand (*TYR*), Champagne d'Argent (*ASIP*) and Dutch (*ASIP*) (*32, 33*). Although these phenotypes follow Mendelian inheritance patterns and certainly represent some of the strongest signatures of selection imposed by the process of breed formation, none of the genes were found within the regions inferred to be under selection. This finding, although anecdotal, suggests that our catalog of regions targeted by directional selection should be highly enriched for genes selected before breed formation, validating the utility of our approach.

Confirmation of selective sweeps using a targeted capture dataset. To confirm our selective sweep regions obtained using the pool sequencing approach we generated an additional dataset by sequencing genomic regions enriched by DNA hybridization on microarrays (*34*) and focused on a slightly different set of breeds (six in total and three in common with the pool sequencing data) and wild individuals from France (five localities in total and none in common with the pool sequencing data; Table S4). Published sequence data for the same genomic regions from six wild rabbits from the Iberian Peninsula (*35*) were used in data analysis.

The full methodological details are given elsewhere (*35*) so here we provide a brief description. We designed a custom Agilent array for the selective enrichment of 6 Mb of intronic sequence throughout the rabbit genome (5000 fragments of 1.2 kb). This array was designed initially for the study described above and thus the sequenced fragments are located randomly with regard to gene function or to the location of the sweeps inferred from the pool sequencing approach. All sequencing runs of the barcoded Illumina sequencing libraries were performed on an Illumina GAII platform using a combination of single-end and paired-end 76 bp reads. As

before, read mapping to the rabbit reference genome was performed using the Burrows-Wheeler alignment tool using default parameters and PCR duplicates were removed from further analysis. SNP and genotype/consensus calling were also carried out using Samtools. SNPs with a minimum quality of 20, minimum mapping quality of 20 (root mean square), and distancing 10 bp from indel polymorphisms were kept for individual genotype calling. Homozygote and heterozygote genotypes were accepted according to the algorithm implemented in Samtools if the total effective sequence coverage was equal or higher than 8X and genotype quality equal or higher than 20, otherwise that specific genotype was coded as missing data. Consensus calling for positions where no SNPs were identified was performed using these same criteria, otherwise that specific position was coded as missing data. The average effective coverage per individual (i.e. after quality filtering and duplicates removal) was 30X±4.3, and >91% of targeted positions were covered by ≥8 reads in all individuals (Table S4).

This dataset consisting of DNA sequence variation in individual (rather than pooled) wild and domesticated rabbits allows a formal investigation of the main demographic events in the recent history of domestic rabbits, and its impact upon levels and patterns of genetic diversity across the genome. We carried out computer simulations of the coalescent process, according to the model depicted in Fig. S3A, using a modified version of the computer program ms (*36*). Our main goal was to estimate the magnitude of the domestication bottleneck which is described as $k_{dom} = Nb_{dom}/d_{dom}$, where Nb_{dom} is the size of the bottlenecked population and d_{dom} the duration of the bottleneck in generations. Similar models have been applied previously to other domestic species (*37-42*) and also to rabbits (*3*). Our model consisted of three populations and describes two bottleneck events occurring from an ancestral source population. First, we incorporated the recent colonization of France from the Iberian Peninsula after the last glacial maximum and

consequent bottleneck (*3, 43*), in which an ancestral population (rabbits from the Iberian Peninsula) of constant size (N_a) gives rise at time $t2_{fr}$ to a small founder population (Nb_{fr}). More recently, the bottlenecked population at time $t1_{fr}$ ($d_{fr} = t2_{fr} - t1_{fr,}$ where d_{fr} is the duration of the bottleneck in generations) expands into its current size (Np_{fr}). Second, we incorporated the domestication bottleneck from French wild populations. The overall configuration of this second event is similar to the colonization of France and consists of similar parameters (Nb_{dom}, Np_{dom}, d_{dom}, and k_{dom}). Several summary statistics were calculated both for observed and simulated data. We summarized levels and patterns of genetic diversity using two standard estimators of the neutral mutation parameter: Watterson's θ_w (*44*), the proportion of segregating sites in a sample, and π (*45*), the average number of pairwise differences per sequence in a sample. Genetic differentiation between populations was described by means of the fixation index (F_{ST}) (*46*). All these statistics were calculated for each 1.2 kb fragment independently.

We simulated data varying the magnitude of the domestication bottleneck (k_{dom}). To find the best fitting model we compared summary statistics computed on the observed data to those computed on simulated data for each fragment independently. The observed missing data for each fragment was incorporated in the simulations and calculations of the summary statistics (*36*). Owing to the complex and mostly unknown demography of the natural populations of rabbits, our model is certainly a simplified scenario. Therefore, we conditioned the simulations using rejection sampling (*47*). Briefly, we simulated the population of rabbits from the Iberian Peninsula as the ancestral population, and forced the simulations to fit the data observed in French wild rabbits. Simulations were kept if the values obtained were within 20% of the observed values of π, θ_w and F_{ST}, and run until 1,000 genealogies were recorded. This should, in principle, help remove uncertainty associated with the demographic scenario prior to

domestication that is not historically documented, which might otherwise have negatively impacted our estimation of k_{dom} (*39*). The best-fitting value of k_{dom} for each fragment was then estimated as the proportion of 1,000 simulations whose summary statistics (θ_w and F_{ST}) were contained within 20% of the observed values in domestic rabbits. The overall maximum-likelihood across loci was estimated as the product of likelihoods for each locus. Loci incompatible with the multilocus estimate of k_{dom} (see below) were removed and the k_{dom} estimate updated until no deviations were found.

We used several key assumptions. First, previously published effective population size estimates for wild rabbits from Iberian Peninsula and France were used as current population sizes (N_a = 1,000,000 and Np_{fr} = 500,000 (*3, 48*)), and the current population size of domestic rabbits was assumed to be 50,000. Second, variation in mutation rate among fragments was incorporated in the simulations from the variation in the population mutation parameter ($\theta_w = 4N_e\mu$ for autosomal and $\theta_w = 3N_e\mu$ for X-linked loci) estimated from the wild rabbits from the Iberian Peninsula (i.e. the ancestral population). Third, using available historical evidence (*43, 49*) and assuming a generation time of 1 year (*50*), we assumed a split time between wild rabbits from Iberian Peninsula and wild rabbits from France of 10,000 generations (i.e. after the last glacial maximum), and the initial domestication event 1,400 generations ago. We have previously shown that because the short time scale since the initial rabbit domestication provides little opportunity for new mutations to have a substantial impact, several fold changes in N_e or split times result in little or no qualitative change in the estimation of k_{dom} 2*5*). Fourth, the parameter k is the ratio of Nb and d, which are positively correlated (*37, 38*). Therefore, we fixed d_{dom} at 500 generations and used 35 different values of Nb_{dom} so that k_{dom} ranged between 0.1 and 15. Because our previous estimate of k_{dom} based on a smaller dataset was 2.8 (*3*), we used a finer grid

of values between 0.5 and 3.6. Fifth, we included the population recombination parameter ($\theta = 4N_e r$) in our simulations assuming the genome-wide average recombination rate in rabbits (~1cM/Mb) (*21*) and estimates of N_e described above. Finally, for computational efficiency we used a previous estimate of $k_{fr} = 2$ for the magnitude of the bottleneck associated with the colonization of France (*3*). Our rejection sampling approach should, in principle, attenuate the potential uncertainty associated with this parameter.

Although several features of the demography of both wild and domestic rabbits may not have been accurately captured by our model, the estimated model generated shifts in genetic diversity between wild rabbits from France and domestic rabbits (measured using both π and θ_w) as well as levels of differentiation between populations (F_{ST} and percentage of shared mutations) that were qualitatively similar to those estimated from the observed data (Fig. S4). Average values for all statistics were also similar between observed [$\pi_{dom}/(\pi_{dom} + \pi_{fr}) = 0.355$; $\theta_{dom}/(\theta_{dom} + \theta_{fr}) = 0.337$; $F_{ST} = 0.141$; % shared = 0.470] and simulated data [$\pi_{dom}/(\pi_{dom} + \pi_{fr}) = 0.376$; $\theta_{dom}/(\theta_{dom} + \theta_{fr}) = 0.354$; $F_{ST} = 0.090$; % shared = 0.535], and even more similar when loci incompatible with genome-wide background levels and patterns of genetic diversity were removed as described below from the observed data [$\pi_{dom}/(\pi_{dom} + \pi_{fr}) = 0.393$; $\theta_{dom}/(\theta_{dom} + \theta_{fr}) = 0.367$; $F_{ST} = 0.091$; % shared = 0.519]. Overall, our model is generally consistent with observed patterns of sequence diversity.

If directional selection resulted in a sweep of a favorable allele associated with a domestication trait, we expect to find one, or both of the following signatures: 1) a loss of heterozygosity in the genomic region under selection significantly greater than the loss in genetic diversity produced by the domestication bottleneck; and 2) higher differentiation between wild and domestic populations due to the potential for directional selection to generate elevated levels

of population differentiation in structured populations. Both these signatures under the described model make no assumptions as to whether selection happened on new or on previously existing genetic variation. In order to identify regions under selection, we used the multilocus value of k_{dom} and interrogated for each locus individually whether levels of genetic diversity (θ_w) and differentiation (F_{ST}) were significantly lower and higher ($P \leq 0.05$), respectively, than expected from the null demographic model alone. Given that these two signatures, although partially interrelated because both depend on a within-population component of variation, may respond differently to distinct forms of selection, we performed a significance test for each summary statistic independently (Database S2). Unlike the previous implementation based on *SweepFinder*, this method using a more stringent k_{dom} and focused on reductions of heterozygosity and increased differentiation renders our test conservative. Due to the small number of individuals used and the relatively short size of the sequenced fragments, we note that *P*-values associated with individual fragments are not robust to multiple testing. However, our main intent with this additional screen for selection was not to attach great significance to individual loci but to corroborate the findings of the pool-sequencing approach (see main manuscript). In fact, under the specified *P*-value threshold 250 fragments (5000 * 0.05) are expected to be significant just by chance alone, but we identified 936 such fragments. This suggests that our list contains false positives but should be highly enriched for regions displaying levels and patterns of genetic diversity that are incompatible with the estimated demographic model.

Analyses of allele frequency differences between domestic and wild rabbits at individual SNPs. For estimations of allele frequencies of single SNPs we required minimum read depth sums of 20, 10, and 40 reads across the six domestic population pools, the three French wild

pools and the 11 wild Iberian pools, respectively. This filter retained 34,293,238 SNPs where reliable allele frequencies could be estimated. The per-SNP absolute allele frequency difference (ΔAF) between domestic and wild rabbits was then calculated using the formula: *ΔAF= abs(RefAFdom-mean(RefAFfrench+ RefAFiberian)).* We next binned SNPs by ΔAF in steps of 0.05 (i.e. ΔAF=0–0.05, 0.05–0.10, etc. until 0.95-1.00) and intersected these binned SNPs with coding exons, UTRs, introns and non-coding elements identified as under evolutionary constraint in mammals using a rate-based score (omega) (*8*). Coding exons and introns were based on rabbit Ensembl version 73. UTRs of Ensembl gene models were frequently either missing or were truncated in relation to those annotated in human. For the intersections of ΔAF SNPs with UTRs we therefore utilized a custom gene prediction pipeline less reliant on cross-species homology and more reliant on our rabbit RNA-sequencing data. This pipeline defined UTRs as elements of predicted protein coding exons where no open reading frame was identified. In order not to include retained introns due to incomplete splicing in our UTR set we removed sequences extending an Ensembl intron by more than 90% of intron length. For intersections with ΔAF binned SNPs we also used elements identified as under evolutionary constraint in analysis of 29 mammalian genomes (*8*) (Omega set) that were lifted over from hg18 to OryCun2 using the liftOver tool (*51*) from the University of Santa Cruz (UCSC) and the resulting rabbit elements were then stripped of those overlapping Ensembl version 73 exons. For each ΔAF bin, the proportions of SNPs falling into coding exons, UTRs, introns and the 29 mammals conserved elements were then determined by intersections using the software bedops (*52*).

Coding variants and assessment of functional significance. To investigate the location of SNPs within genes and their potential coding properties we started by using ANNOVAR (*53*) and relied on Ensembl 73 annotations. We specifically focused on missense SNPs showing a

ΔAF>0.90 between wild and domestic rabbits. To infer whether these missense mutations are likely to have functional significance we used several approaches. First, we analyzed sequence conservation for each SNP among 100 vertebrates and placental mammals using the UCSC (University of California Santa Cruz) genome browser with rabbit coordinates obtained with the liftover tool. To identify potential artefacts due to errors in gene predictions, we used the Basic Local Alignment Search Tool (BLAST) (*54*) for alignment of proteins to the non-redundant database in UniProtKB (UniProt Knowledgebase), aligned the orthologs, and visually inspected the alignments. For example, this information allowed us to identify Ensembl gene models that we believe represent nonfunctional retrogenes due to errors in gene prediction, which was further supported by lack of expression data for those genes using the RNA-seq data generated to annotate the reference genome (see above). Second, the nature of each mutation was assessed using standalone versions of SIFT v4.0.5 (*55*) and PolyPhen2 v2.2.2 (*56*) using default parameters. Finally, we inferred the ancestral or derived state of each allele using the outgroup *Lepus americanus* as mentioned above.

Detection of insertions/deletions. We called small insertions/deletions (indels) with GATK (*57*) using the INDEL model of UnifiedGenotyper with default parameters. In total, 9,331,686 indels in relation to the reference genome were discovered. Because indels are often a product of read misalignment, sequence reads overlapping called indel positions were realigned with the GATK pipeline. We could validate that every initially called indel was recalled again and although restricted by previously discovered indel positions, no novel indels were reported for the realigned dataset. However, the number of alleles counted had been readjusted giving a better estimation of the allelic frequencies at indel positions. In order to calculate the frequency spectra

of the distribution, we extracted counts of reads supporting the observed alleles from the GATK vcf files and calculated the reference allele frequency for each sequenced pool.

In order to limit the effect of misalignment on called indels we discarded those where the resequenced reference individual had any reads supporting a non-reference allele (n=1,223,493). Then, we established restrictive depth filters to distinguish real indels from putative artefacts. Only indel positions that were supported by a read depth of reads within a 50% range from the average for each population, inclusive the reference individual, were analysed and counted and 1,112,286 positions were removed by this filter. In this analysis, different median depth thresholds were used for autosomes, X-chromosome and the mitochondrion. All the unplaced (Un) scaffolds were treated as autosomal. We ran GATK with default parameters to report a maximum of six allelic variants. From those resulting indels (n= 6,995,907), we found 6,380,621 bi-allelic, 502,852 tri-allelic, 103,723 tetra-allelic, 8,550 penta-allelic, and 161 hexa-allelic indels.

There were 1,413,933 indels that supported different major alleles in the two wild sub-species included in this study (*O. c. cuniculus* and *O. c. algirus*). These positions were discarded from further analyses. We then grouped the remaining indels (n= 5,581,974) into 5% bins of absolute allele frequency difference (ΔAF) between wild and domestic rabbits to create the background genome-wide distribution. We intersected indels in these ΔAF bins with Ensembl v73 UTRs and coding sequences as well as elements under evolutionary constraint in mammals and identified 2,331, 251 and 9,213 indels falling in these three categories, respectively. Sweeps detected by *SweepFinder* and the F_{ST}-Het combination were merged in one unique dataset and we enlarged these loci by taking 20 kb surrounding them and intersected with indels. Thus, we discovered 15,923 indels overlapping sweeps.

Detection of structural changes. To identify duplications/deletions that could differ between wild and domestic rabbits we performed an analysis of variance (ANOVA) on depth of coverage using the package implemented in R *(58)*. We scanned the genome in 1 kb non-overlapping windows and all depths were normalized against the Flemish giant breed, which had higher average depth compared to the other breeds. The contrast was made between all domestic breeds and wild populations. For each window we also calculated a *M-value* to measure fold-coverage difference between domestic and wild populations as follows:

$$M - value = log_2(\frac{\overline{Domestic\ depth}}{\overline{Wild\ depth}})$$

In total, 2,165,484 windows were analysed from which 710 windows were filtered out due to low coverage. 29,792 windows were significant with a $P \leq 0.001$. Using the *P*-value of the ANOVA analysis together with *M-values*, we merged significant windows to identify regions indicating the same signature of duplication/deletion using the following criteria. First, we opened a locus when a given window fulfilled two criteria: 1) $P \leq 0.001$; and 2) absolute (*M-value*) ≥ 0.6. Second, we continued scanning by setting a dynamic cut-off for |*M-value*|. It was redefined when a given window had higher |*M-value*| than the starting threshold (|*M-value*| ≥ 0.6). To merge adjacent windows, we employed a more lenient cut-off for the *M-value* (≥ 0.5 * redefined |*M-value*|) and *P*-value (≤ 0.05). Third, we allowed two tandem outliers and closed the locus when three or more consecutive outliers were detected. Finally, we rescanned windows upstream of the opening site using the criteria mentioned in previous steps.

Gene overrepresentation analyses. In order to assess enrichment of gene ontology terms and other ontology terms such as those associated with murine phenotypes and gene expression

among ΔAF≥0.8 SNP loci we utilized the software *Genomic Regions Enrichment of Annotations Tool* (GREAT) (*11*). In order not to inflate significances due to inclusion of SNPs in strong linkage disequilibrium we selected only one SNP per 50 kb from the set of 1635 SNPs in conserved elements with ΔAF ≥0.80, leaving 1,071 SNPs. We next lifted this selected set of rabbit SNPs over to the human assembly hg18 and used the resulting human coordinates as input for GREAT, all genes with a Transcription Start Site (TSS) within 1 Mb from a high ΔAF SNP were included in the analysis. Overrepresentation was tested using hypergeometric tests, and Bonferroni corrected *P*-values were used to evaluate statistically significant overrepresentation of terms. For extraction of data in order to condense the GREAT output for visualization in Fig. S5, we first required ≥6 genes associated with a term and then further filtered the GREAT output requiring a Bonferroni $P < 0.05$ and an enrichment >1.7 for GO Biological Process as well as for mouse phenotype gene association terms (MGI Phenotypes) in Fig. S5 and a Bonferroni $P < 1x10^{-8}$ and an enrichment >2.2 for MGI expression. Next we determined the fractions of genes shared in each pairwise contrast of associated terms (in relation to the term containing fewest genes) and this matrix of gene sharing fractions was then subjected to hierarchical clustering (Pearson correlation) to construct heat maps. Annotations summarizing clustered terms to the left of heat-maps were manually inferred from included terms to represent a description of the individual terms in each cluster.

Electrophoretic mobility shift assays (EMSA)

Cells and preparation of nuclear extracts. Mouse ES-cell derived neural stem cells were obtained by *in vitro* differentiation of R1 ES cells using the protocol by Conti *et al.* (*59*). The resulting neural stem cells were propagated in N2B27 medium supplemented with EGF and FGF2 (*60*). After dissociation into single cells by trypsinization, cell pellets were stored at -70°C

for further analysis. The P19 embryonic carcinoma cells were maintained in Alpha Modified Eagle's Medium (αMEM, Invitrogen) supplemented with 10% heat-inactivated fetal bovine serum and penicillin (0.2 U/mL)/streptomycin (0.2 μg/ml)/L-glutamine (0.2 μg/ml) (Gibco) at 37ºC in a 5% CO2 humidified atmosphere. In order to induce the neuronal differentiation of P19, the cells were cultured in αMEM growth medium containing 1 μM all-trans-Retinoic Acid (RA, Sigma Aldrich R-2625) in bacterial garde Petri dishes to promote the aggregation of embryonic bodies (EB). After 48h the EBs were plated in adherent culture plates in RA-free growth medium and after 48h the differentiated P19 cells were harvested for preparation of nuclear extracts.

The nuclear extracts were prepared from embryonic stem cell-derived neural stem cells (ES-NSC) and P19 cells using NucBuster Protein Extraction Kit (Novagen).

EMSA. Seventeen selected SNPs were functionally assessed using EMSA to reveal their potential to affect DNA-protein interaction. 5'Biotin-labelled probes were synthesized (Integrated DNA Technologies) and can be obtained upon request. Double-stranded probes were generated by annealing single-stranded complementary oligonucleotides in 1X NEB2 buffer (New England Biolabs). Two μg of the nuclear extracts were added to the binding reaction (10X binding buffer, 30.1 mM KCl, 2 mM MgCl2, 7.5% Glycerol, 0.1 mM EDTA, 0.063% NP-40 and 1μg/ml Poly(dI•dC)) and then preincubated for 20 min on ice. For competition reactions, 20 pmol of unlabeled double-stranded oligos were added to the reaction. Thereafter, 20 fmol of biotinylated oligos were added to the reactions and incubated for 20 min at RT. The DNA-protein complexes were separated by electrophoresis on 5% non-denaturing polyacrylamide gel at 100 V for 1.5 h at RT in 0.5×TBE running buffer. The separated complexes were transferred to nylon membrane (Perkin Elmer) at 45 V for 1 h in cold 0.5×TBE. The DNA-protein complexes were crosslinked

using a transilluminator with 312 nm UV bulbs. The biotinylated probes were detected using Lightshift Electrophoretic Mobility-Shift Assay kit (Thermo Scientific).

References

1. C. Darwin, *On the origins of species by means of natural selection or the preservation of favoured races in the struggle for life*. (John Murray, London, 1859).

2. J. A. Clutton-Brock, *Natural History of Domesticated Mammals*. (Cambridge University Press, Cambridge, 1999).

3. M. Carneiro, S. Afonso, A. Geraldes, H. Garreau, G. Bolet *et al.*, The genetic structure of domestic rabbits. *Mol. Biol. Evol.* **28**, 1801-1816 (2011).

4. Materials and methods are available as supplementary material on Science Online.

5. M. Carneiro, F. W. Albert, J. Melo-Ferreira, N. Galtier, P. Gayral *et al.*, Evidence for Widespread Positive and Purifying Selection Across the European Rabbit (Oryctolagus cuniculus) Genome. *Mol. Biol. Evol.* **29**, 1837-1849 (2012).

6. J. Maynard-Smith, J. Haigh, The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23-35 (1974).

7. R. Nielsen, S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**, 1566-1575 (2005).

8. K. Lindblad-Toh, M. Garber, O. Zuk, M. F. Lin, B. J. Parker *et al*., A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476-482 (2011).

9. M. M. Motazacker, B. R. Rost, T. Hucho, M. Garshasbi, K. Kahrizi *et al.*, A defect in the ionotropic glutamate receptor 6 gene (GRIK2) is associated with autosomal recessive mental retardation. *Am. J. Hum. Genet.* **81**, 792-798 (2007).

10. K. Takahashi, S. Yamanaka, Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676 (2006).

11. C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar *et al.*, GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotech.* **28**, 495-501 (2010).

12. C.-J. Rubin, M. C. Zody, J. Eriksson, J. R. S. Meadows, E. Sherwood *et al.*, Whole genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587-591 (2010).

13. C.-J. Rubin, H. Megens, A. Martinez Barrio, K. Maqbool, S. Sayyab *et al.*, Strong signatures of selection in the domestic pig genome. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 19529-19536 (2012).

14. M.V. Olson, When less is more: Gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.* **64**, 18–23 (1999).

15. P. V. Tran, C. J. Haycraft, T. Y. Besschetnova, A. Turbe-Doan, R. W. Stottmann *et al.*, THM1 negatively modulates mouse sonic hedgehog signal transduction and affects retrograde intraflagellar transport in cilia. *Nat. Genet.* **40**, 403-410 (2008).

16. K. Agger, P. A. C. Cloos, J. Christensen, D. Pasini, S. Rose *et al.*, UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. *Nature* **449**, 731-734 (2007).

17. H. Deng, K. Gao, J. Jankovic, The genetics of Tourette syndrome. *Nat. Rev. Neurol.* **8**, 203-213 (2012).

18. J. K. Pritchard, J. K. Pickrell, G. Coop, The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* **20**, 208-215 (2010).

19. L. Andersson, Molecular consequences of animal breeding. *Curr. Opin. Genet. Dev.* **23**, 295-301 (2013).

20. D. B. Jaffe, J. Butler, S. Gnerre, E. Mauceli, K. Lindblad-Toh *et al.*, Whole-Genome Sequence Assembly for Mammalian Genomes: Arachne 2. *Genome Res.* **13**, 91-96 (2003).

21. C. Chantry-Darmon, C. Urien, H. De Rochambeau, D. Allain, B. Pena *et al.*, A first-generation microsatellite-based integrated genetic and cytogenetic map for the European rabbit (*Oryctolagus cuniculus*) and localization of angora and albino. *Anim. Genet.* **37**, 335-341 (2006).

22. K. Osoegawa, P. Y. Woon, B. Zhao, E. Frengen, M. Tateno *et al.*, An Improved Approach for Construction of Bacterial Artificial Chromosome Libraries. *Genomics* **52**, 1-8 (1998).

23. J. Z. Levin, M. Yassour, X. Adiconis, C. Nusbaum, D. A. Thompson *et al.*, Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Meth.* **7**, 709-715 (2010).

24. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644-652 (2011).

25. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).

26. http://picard.sourceforge.net

27. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

28. D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan *et al.*, VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568-576 (2012).

29. R. Kofler, P. Orozco-terWengel, N. De Maio, R. V. Pandey, V. Nolte *et al.*, PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One* **6**, e15925 (2011). doi:10.1371/journal.pone.0015925.

30. B. S. Weir, C. C. Cockerham, Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358-1370 (1984).

31. S. H. Williamson, M. J. Hubisz, A. G. Clark, B. A. Payseur, C. D. Bustamante *et al.*, Localizing recent adaptive evolution in the human genome. *PLoS Genet.* **3**, e90 (2007). doi:10.1371/journal.pgen.0030090.

32. B. Aigner, U. Besenfelder, M. Muller, G. Brem, Tyrosinase gene variants in different rabbit strains. *Mamm. Genome* **11**, 700-702 (2000).

33. L. Fontanesi, L. Forestier, D. Allain, E. Scotti, F. Beretti *et al.*, Characterization of the rabbit agouti signaling protein (ASIP) gene: Transcripts and phylogenetic analyses and identification of the causative mutation of the nonagouti black coat colour. *Genomics* **95**, 166-175 (2010).

34. E. Hodges, M. Rooks, Z. Xuan, A. Bhattacharjee, D. Benjamin Gordon *et al.*, Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat. Protoc.* **4**, 960-974 (2009).

35. M. Carneiro, F. W. Albert, S. Afonso, R. Pereira, H. Burbano *et al.*, The genomic architecture of population divergence between subspecies of the European rabbit. *PLoS Genet.* (in press).

36. P. Pavlidis, S. Laurent, W. Stephan, msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. *Mol. Ecol. Resour.* **10**, 723-727 (2010).

37. A. Eyre-Walker, R. L. Gaut, H. Hilton, D. L. Feldman, B. S. Gaut, Investigation of the bottleneck leading to the domestication of maize. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4441-4446 (1998).

38. M. I. Tenaillon, J. U'Ren, O. Tenaillon, B. S. Gaut, Selection versus demography: a multilocus investigation of the domestication process in maize. *Mol. Biol. Evol.* **21**, 1214-1225 (2004).

39. S. I. Wright, I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley *et al.*, The effects of artificial selection on the maize genome. *Science* **308**, 1310-1314 (2005).

40. Q. Zhu, X. Zheng, J. Luo, B. S. Gaut, S. Ge, Multilocus analysis of nucleotide variation of Oryza sativa and its wild relatives: severe bottleneck during domestication of rice. *Mol. Biol. Evol.* **24**, 875-888 (2007).

41. A. Haudry, A. Cenci, C. Ravel, T. Bataillon, D. Brunel *et al.*, Grinding up wheat: a massive loss of nucleotide diversity since domestication. *Mol. Biol. Evol.* **24**, 1506-1517 (2007).

42. H. S. Yu, Y. H. Shen, G. X. Yuan, Y. G. Hu, H. E. Xu *et al.*, Evidence of selection at melanin synthesis pathway loci during silkworm domestication. *Mol. Biol. Evol.* **28**, 1785-1799 (2011).

43. G. Queney, N. Ferrand, S. Weiss, F. Mougel, M. Monnerot, Stationary distributions of microsatellite loci between divergent population groups of the European rabbit (Oryctolagus cuniculus). *Mol. Biol. Evol.* **18**, 2169-2178 (2001).

44. G. A. Watterson, On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256-276 (1975).

45. M. Nei, *Molecular Evolutionary Genetics*. (Columbia University Press, New York, 1987).

46. R. R. Hudson, M. Slatkin, W. P. Maddison, Estimation of Levels of Gene Flow from DNA-Sequence Data. *Genetics* **132**, 583-589 (1992).

47. G. Weiss, A. von Haeseler, Inference of population history using a likelihood approach. *Genetics* **149**, 1539-1546 (1998).

48. M. Carneiro, N. Ferrand, M. W. Nachman, Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (Oryctolagus cuniculus). *Genetics* **181**, 593-606 (2009).

49. C. Callou, *De la garenne au clapier: etude archeozoologique du lapin en Europe Occidentale*, (Memoir Mus Natl Hist, 2003), pp. 352.

50. R. C. Soriguer, *El conejo: papel ecológico y estrategia de vida en los ecosistemas mediterráneos* (In Proc.: XV Congreso Internacional de Fauna Cinegética y Silvestre, Trujillo, Spain, 517-542, 1983).

51. D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs *et al.*, The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51-54 (2003).

52. S. Neph, M. S. Kuehn, A. P. Reynolds, E. Haugen, R. E. Thurman *et al.*, BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919-1920 (2012).

53. K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010). doi:10.1093/nar/gkq603.

54. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).

55. N. L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider *et al.*, SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452-457 (2012). doi:10.1093/nar/gks539.

56. I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova *et al.*, A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248-249 (2010).

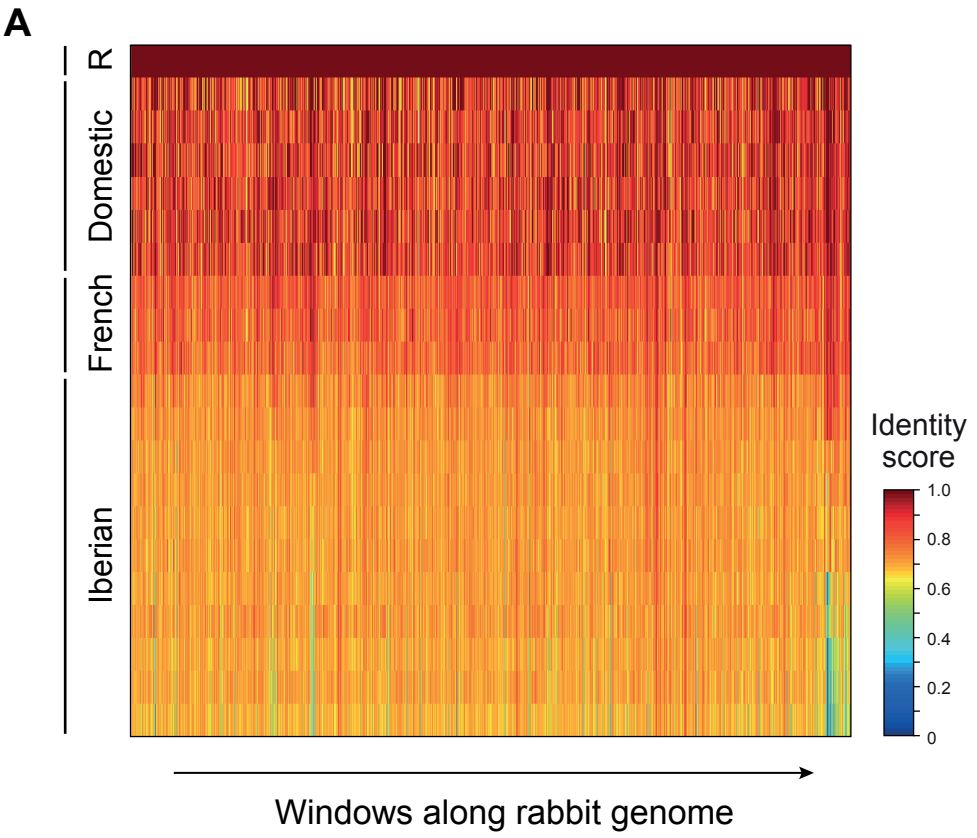
57. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491-498 (2011).

58. R Core Team, A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2013

59. L. Conti, S. M. Pollard, T. Gorba, E. Reitano, M. Toselli *et al.*, Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *Plos Biol.* **3**, e283 (2005). doi:10.1371/journal.pbio.0030283

60. Q. L. Ying, M. Stavridis, D. Griffiths, M. Li, A. Smith, Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nat. Biotechnol.* **21**, 183-186 (2003).

61. E. M. Gertz, A. A. Schaffer, R. Agarwala, A. Bonnet-Garnier, C. Rogel-Gaillard et al., Accuracy and coverage assessment of Oryctolagus cuniculus (rabbit) genes encoding immunoglobulins in the whole genome sequence assembly (OryCun2.0) and localization of the IGH locus to chromosome 20. *Immunogenetics* **65**, 749-762 (2013).



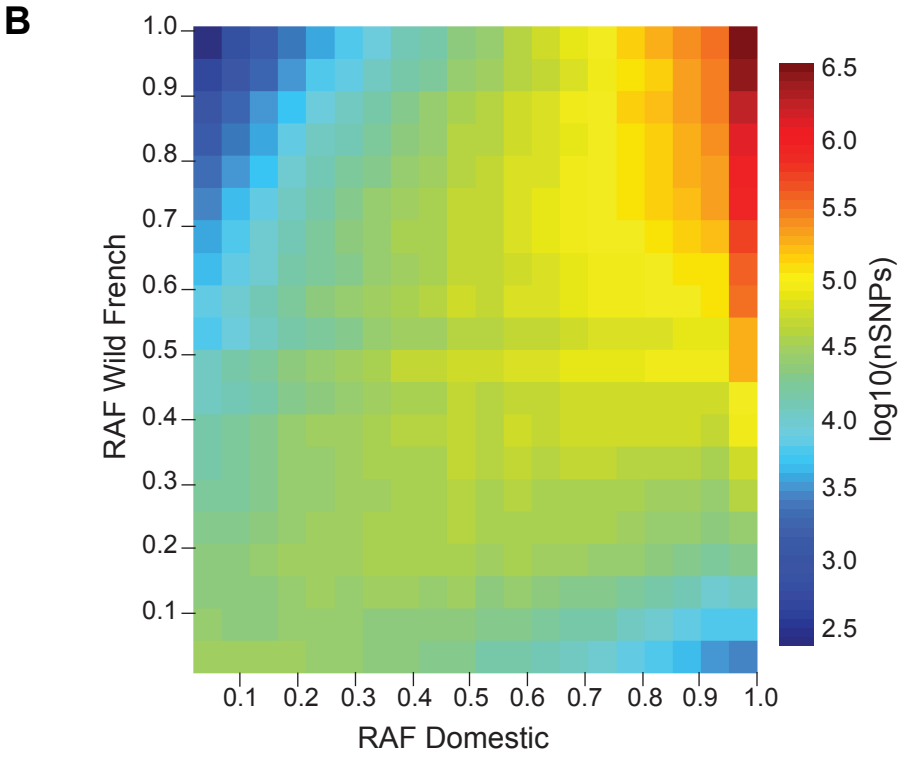


Fig. S1: Genetic relatedness between populations sampled in this study. **(A)** Heat map (color code to the right) of identity scores based on comparing resequencing data with the assembly. The x-axis represents genome coordinates with chromosome 1 to the left and chromosome X to the right. The first row (R) represents the reference rabbit against itself, rows 8-21 correspond to locations marked with red dots in Fig. 1A, ordered according to a northeast to southwest transection. **(B)** Strong correlation in allele frequencies between domestic and French wild rabbits; RAF French and RAF Domestic represent the frequency of the reference allele in wild French and domestic rabbits, respectively.

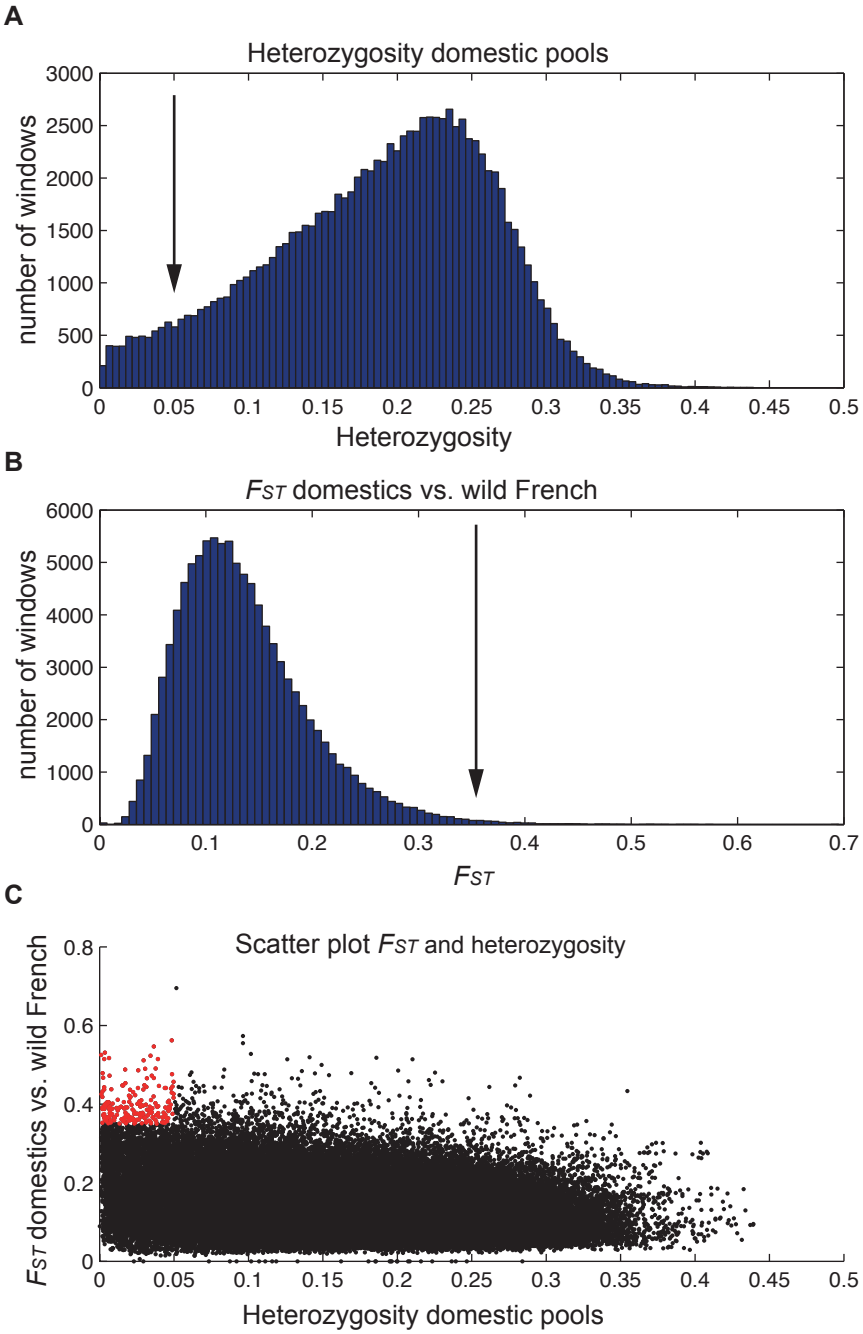


Fig. S2: Distributions and scatterplot of heterozygosity and F_{ST} of 50 kb windows. **(A)** Distribution of pool heterozygosity in the domestic pools sequenced. The arrow indicates the upper heterozygosity treshold for opening a sweep. **(B)** Distribution of F_{ST} values observed in the contrast domestics vs. wild French. The arrow indicates thelower F_{ST} treshold for opening a sweep. **(C)** Scatter plot of F_{ST} and pool heterozygosity. Red dots represent windows fulfilling the sweep opening requirement (F_{ST} >= 0.35 and heterozygosity <= 0.05)

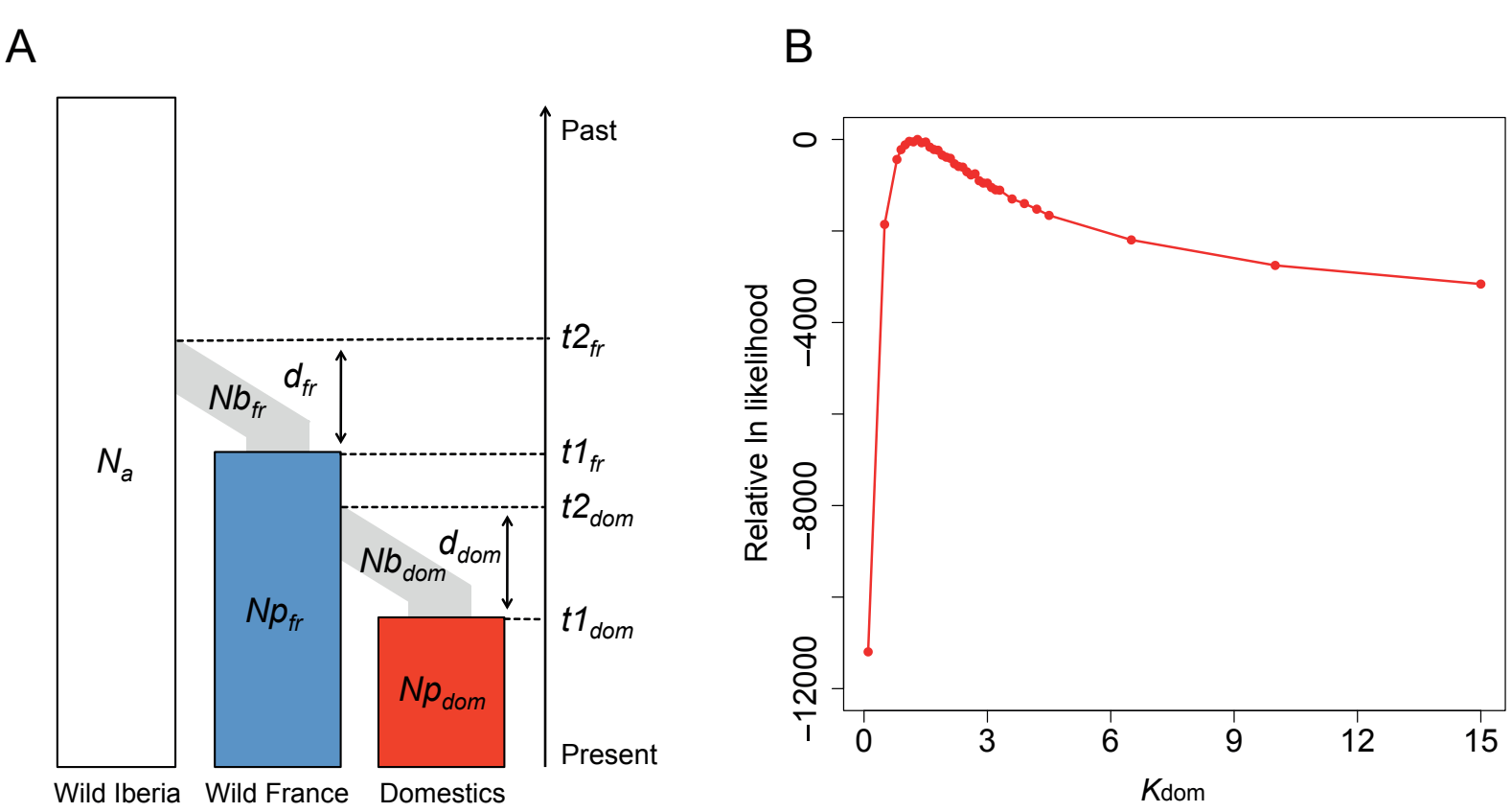


Fig. S3: Demographic model and magnitude of the domestication bottleneck estimated using the targeted capture dataset.**(A)** Schematic representation of the coalescent model used to represent the main events of the demographic history of wild and domestic rabbits. See Supplementary Methods for a detailed description of all parameters and assumptions. **(B)** Diagram representing the multilocus maximum-likelihood (ML) estimate for the magnitude of the domestication bottleneck (k_{dom}). Multilocus ML (y axis) of the parameter k_{dom} (x axis) was estimated as the product of the likelihood for each locus.

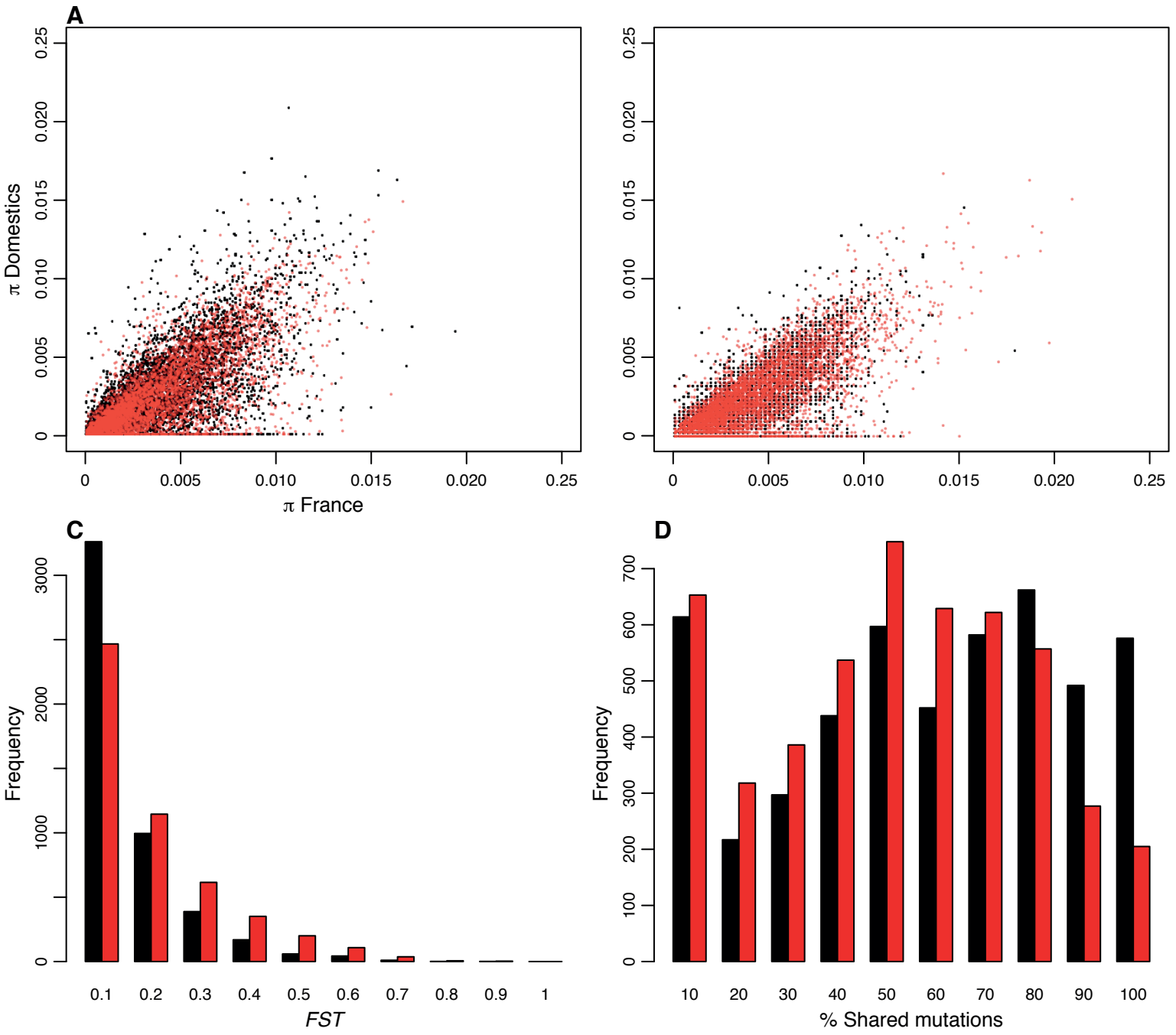


Fig. S4: Comparison between observed and simulated data under the estimated demographic model. Observed and simulated data are represented in red and black, respectively. The first two panels (**A and B**) illustrate the relationship between values of both π and θw in wild rabbits from France (x axis) and domestic rabbits (y axis). The two histograms on the bottom row (**C and D**) illustrate the distribution of *FST* and percentage of shared mutations between the two populations.

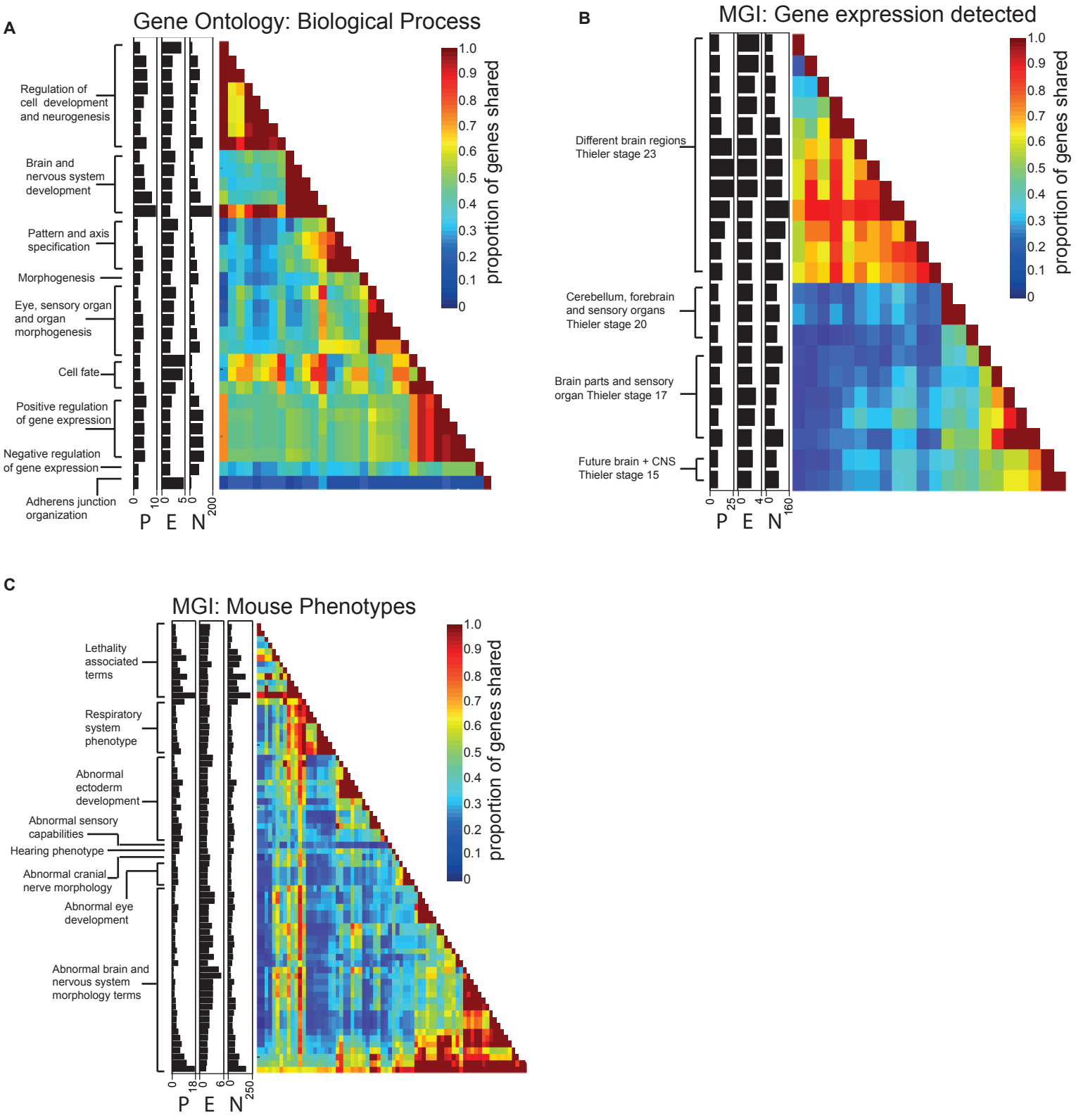


Fig. S5: Gene overrepresentation analysis using the GREAT software based on genes associated with high delta allele frequency SNPs (ΔAF≥0.8) at non-coding conserved sites. **(A)** Gene Ontology categories within Biological Processes, **(B)** MGI mouse expression data and **(C)** MGI mouse phenotypes. Each row in the heat-maps represents one specific category and colors on that row indicate the proportion of shared genes in relation to the other categories (ordered in the same way on the x-axis as on y-axis). To the left of each cluster of enriched terms the types of terms in that group is indicated in text. Bars immediately left of heat-maps visualize the significance-, enrichment- and number of genes of each significant term (P= -log10 Bonferroni-corrected P-value; E= fold enrichment: N=number of genes). The ranges for P, E, and N are indicated below the plot. The full results are presented in Database S3.

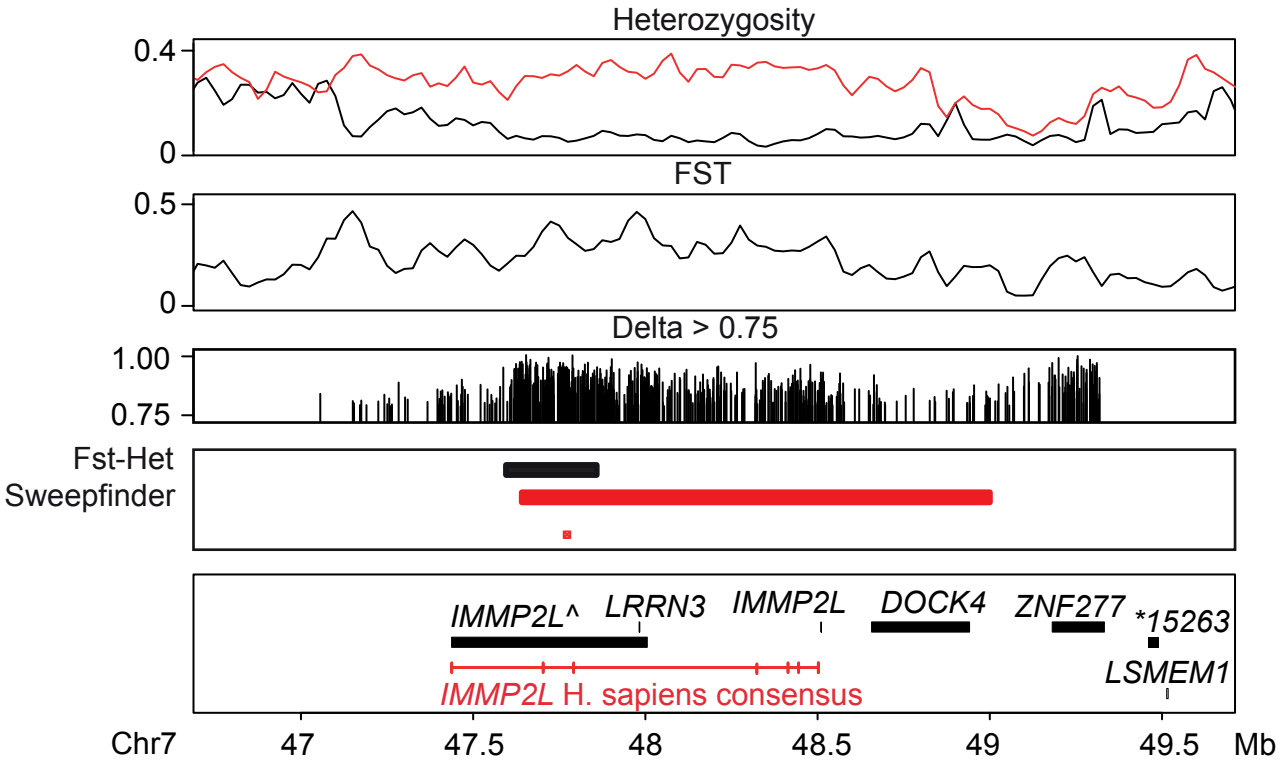


Fig. S6: Selective sweep at *IMMP2L* on chromosome 7. Heterozygosity plots for wild (red) and domestic (black) rabbits together with plots of F_{ST} values and high ΔAF (HΔAF) SNPs with ΔAF>0.75. Putative sweep regions detected with the F_{ST}-Heterozygosity outlier approach and SweepFinder are marked with horizontal bars. Gene annotations in sweep regions are indicated. *IMMP2L* was not predicted by Ensembl but by our in-house predictions based on RNA-sequencing data. The human consensus model shows all human *IMMP2L* transcripts in a collapsed fashion. Asterisks (*) represent ENSOCUT000000. Red dots indicate the location of the high ΔAF insertion/deletion in a conserved element for *IMMP2L*.

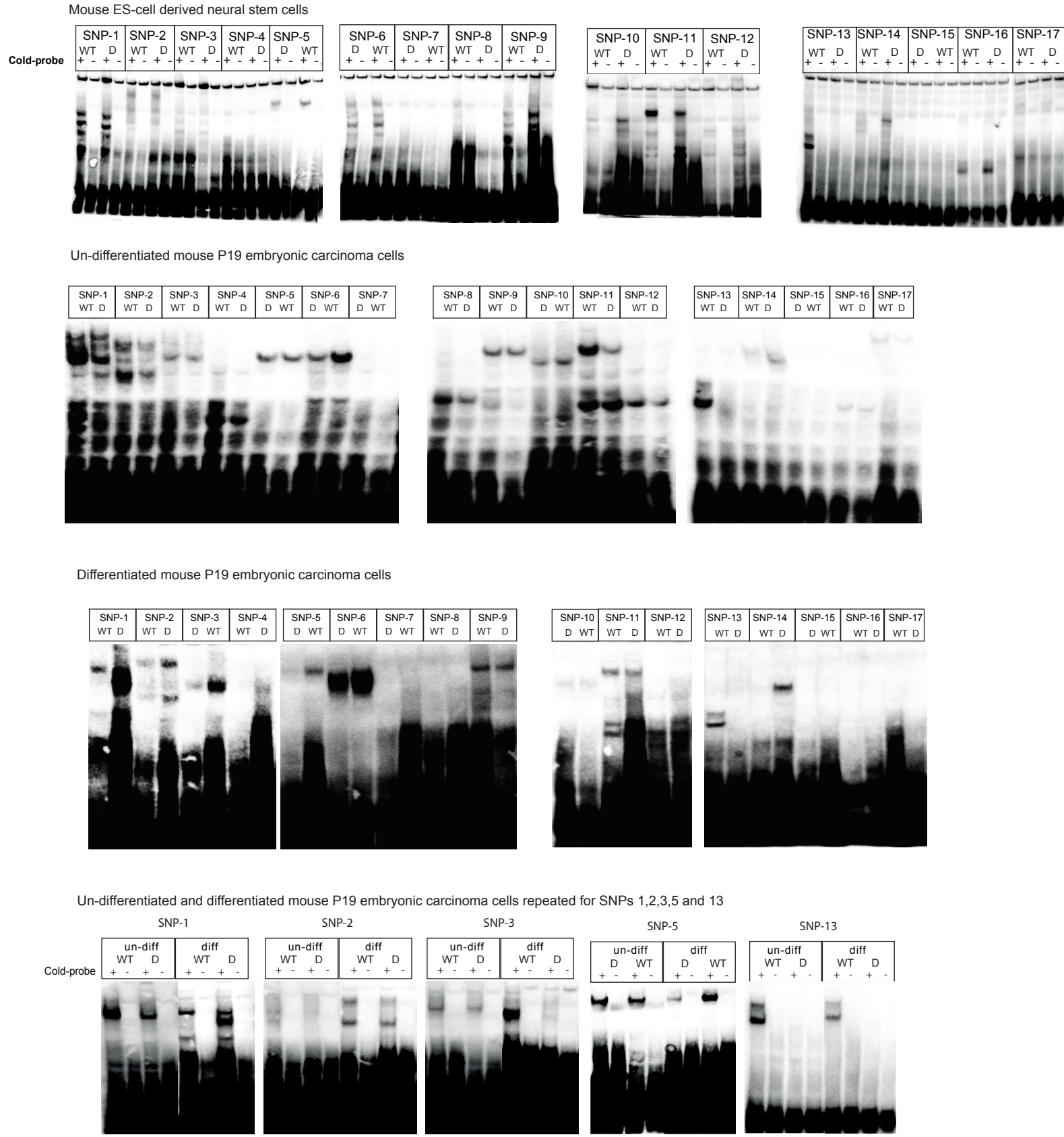


Fig. S7: Results for all 17 SNPs functionally examined by electrophoretic mobility shift assays (EMSA). Results of EMSA using nuclear extracts from ES-cell derived neural stem cells or from mouse P19 embryonic carcinoma cells before (un-diff) or after neuronal differentiation (diff) are presented. WT=wild-type allele; D=domestic, the most common allele in domestic rabbits. Cold probes at 100-fold excess were used to verify specific DNA-protein interactions. The results are summarized in Table S7.

Table S1. Summary statistics of the rabbit OryCun2.0 assembly

	OryCun2.0
Coverage (Q20)	6.55X
Contig N50 (kb)	64.7
Scaffold N50 (Mb)	35.9
Assembly size with gaps (Gb)	2.66
Assembly size without gaps (Gb)	2.60
Number of anchored scaffolds	99
% of assembly in anchored scaffolds	82.0
Sequence in anchored scaffolds (Gb)	2.18

Table S2. Comparison between rabbit genome assembly versions OryCun1.0 and OryCun2.0.

Assembly	Coverage (Q20)	Contig N50 (kb)	Scaffold N50 (Mb)	Assembly size (with gaps)	Assembly size (without gaps)	Number anchored scaffolds	% assembly anchored scaffolds	Sequence in anchored bases (Gb)
OryCun1.0	1.95X	2.84	0.05	3.69	2.08	N/A	N/A	N/A
OryCun2.0	6.55X	64.65	35.92	2.66	2.60	99	82	2.18

Table S3. Summary of RNA-sequencing data for annotation of the rabbit genome

Tissue Site	Pass Filter Reads	Assembled Transcripts	Median Transcript Length (bp)	Source
Adult adrenal gland	85 775 334	29 672	1380	INRA
Blood	103 242 794	25 451	1226	Zyagen
Brain	118 014 338	59 029	1362	Zyagen
Fetal adrenal gland – gestation day 28	92 146 626	36 404	1465	INRA
Heart	100 862 716	35 557	1433	Zyagen
Kidney	108 298 292	39 556	1403	Zyagen
Liver	102 233 272	30 711	1435	Zyagen
Lung	93 531 078	41 731	1479	Zyagen
Mammary gland – gestation day 3	89 869 204	57 763	1336	INRA
Mammary gland - gestation day 14	84 998 914	58 136	1377	INRA
Mammary gland – lactation day 16	89 377 970	27 691	1403	INRA
Ovary	110 264 790	42 350	1381	Zyagen
Peripheral blood mononuclear cells	83 915 910	28 031	1493	INRA
Placenta – female fetus – gestation day 28	86 728 588	30 992	1388	INRA
Placenta – male fetus – gestation day 28	76 666 886	28 420	1389	INRA
Skeletal muscle	96 519 628	28 805	1489	Zyagen
Skin	87 995 594	43 294	1462	Zyagen
Spleen	87 432 290	48 398	1426	INRA
Testis	117 649 896	63 676	1191	Zyagen

Table S4. Sample details of wild and domestic rabbits used in this study both for whole genome resequencing and DNA sequence capture using hybridization on microarrays. Average sequence coverage per sample is provided.

Population	Breed/Location	ID Fig.1c	Sample size	Sequencing method	Coverage
Domestic	Belgian hare		17 (pool)	WGS	10.4
Domestic	Champagne d'argent		16 (pool)	WGS	11.2
Domestic	Dutch		13 (pool)	WGS	10.1
Domestic	Flemish giant		18 (pool)	WGS	12.1
Domestic	French lop		20 (pool)	WGS	11.9
Domestic	New Zealand white		16 (pool)	WGS	11.1
Domestic	Thorbecke (reference)		1 (individual)	WGS	10.9
Hare (outgroup)	Missoula, Montana		1 (individual)	WGS	10.2
Wild France	Caumont	FRW1	10 (pool)	WGS	10.7
Wild France	La Roque	FRW2	10 (pool)	WGS	11.5
Wild France	Villemolaque	FRW3	10 (pool)	WGS	12.0
Wild Iberian	Huelva	IW11	16 (pool)	WGS	11.0
Wild Iberian	Milmarcos	IW10	11 (pool)	WGS	10.7
Wild Iberian	S. Agustin de Guadalix	IW9	20 (pool)	WGS	11.0
Wild Iberian	Mora	IW8	13 (pool)	WGS	11.9
Wild Iberian	Toledo	IW7	14 (pool)	WGS	10.7
Wild Iberian	Mazarambroz	IW6	16 (pool)	WGS	11.3
Wild Iberian	Urda	IW5	16 (pool)	WGS	11.4
Wild Iberian	Carrión de Calatrava	IW4	17 (pool)	WGS	11.7
Wild Iberian	Calzada de Calatrava	IW3	16 (pool)	WGS	12.1
Wild Iberian	Pedroche	IW2	11 (pool)	WGS	11.2
Wild Iberian	Zaragoza	IW1	12 (pool)	WGS	6.1
Wild France	Aveyron		1 (individual)	Seq. capture	27.3
Wild France	Fos-su-Mer		2 (individual)	Seq. capture	31.7; 31.1
Wild France	Herauld		1 (individual)	Seq. capture	30.5
Wild France	Lancon		2 (individual)	Seq. capture	33.3; 34.3
Wild France	Vaucluse		1 (individual)	Seq. capture	34.0
Domestic	Belgian hare		1 (individual)	Seq. capture	35.5
Domestic	Champagne d'argent		1 (individual)	Seq. capture	31.3
Domestic	Angora		2 (individual)	Seq. capture	27.9; 30.7
Domestic	French lop		1 (individual)	Seq. capture	36.5
Domestic	Flemish giant		2 (individual)	Seq. capture	24.3; 20.8
Domestic	Rex		1 (individual)	Seq. capture	26.7

* WGS= whole genome resequencing, Seq. Capture = Targeted resequencing after sequence capture

Table S5. Summary of SNPs and Insertions/Deletions detected in the rabbit genome

Type	Total count	In coding sequence[1]	Missense[2]	Conserved non-coding[3]
SNPs	50,165,386	154,489	48,453	719,911
Indels	5,581,974	2,812	-	82,289

[1]In coding exons based on Ensembl v73 gene models
[2]Amino acid alteration predicted by the software Annovar (*28*)
[3]Located within elements under evolutionary constraint but outside of Ensembl v73 coding exons

Table S6. Distributions of SNP counts in the different delta allele frequency bins for conserved non-coding elements, UTRs, coding sequences and introns.

	All	**Conserved non-coding sites**				**UTRs**				**Coding**				**Introns**			
ΔAF bin*	Observed	Observed	Expected	*P***	M***	Observed	Expected	*P***	M***	Observed	Expected	*P***	M***	Observed	Expected	*P***	M***
0.95-1.0	1,030	27	15	**1.80E-03**	**0.8**	22	14	**3.13E-02**	**0.7**	14	6	**1.05E-03**	**1.2**	325	304	1.51E-01	**0.1**
0.90-0.95	6,987	140	102	**1.50E-04**	**0.5**	113	98	1.27E-01	**0.2**	56	42	**3.03E-02**	**0.4**	2,089	2,061	4.63E-01	0.0
0.85-0.90	22,968	448	335	**4.99E-10**	**0.4**	378	321	**1.36E-03**	**0.2**	147	140	5.53E-01	**0.1**	6,951	6,774	**1.04E-02**	0.0
0.80-0.85	55,438	1,021	809	**5.98E-14**	**0.3**	875	775	**2.97E-04**	**0.2**	391	337	**3.17E-03**	**0.2**	16,671	16,351	**2.88E-03**	0.0
0.75-0.80	108,850	1,918	1,588	**7.29E-17**	**0.3**	1,720	1,522	**3.20E-07**	**0.2**	646	661	5.58E-01	0.0	32,177	32,105	6.32E-01	0.0
0.70-0.75	184,708	3,009	2,694	**9.74E-10**	**0.2**	2,782	2,583	**8.04E-05**	**0.1**	1,161	1,122	2.43E-01	0.0	54,993	54,479	**8.73E-03**	0.0
0.65-0.70	287,540	4,703	4,194	**2.42E-15**	**0.2**	4,397	4,022	**2.60E-09**	**0.1**	1,812	1,747	1.19E-01	**0.1**	85,412	84,808	**1.35E-02**	0.0
0.60-0.65	417,269	6,560	6,086	**9.32E-10**	**0.1**	6,304	5,836	**6.85E-10**	**0.1**	2,592	2,535	2.56E-01	0.0	124,361	123,071	**1.19E-05**	0.0
0.55-0.60	578,178	8,953	8,433	**1.17E-08**	**0.1**	8,375	8,087	**1.26E-03**	**0.1**	3,521	3,513	8.92E-01	0.0	172,726	170,530	**2.40E-10**	0.0
0.50-0.55	771,964	11,789	11,259	**4.86E-07**	**0.1**	10,936	10,797	1.78E-01	0.0	4,762	4,691	2.98E-01	0.0	229,011	227,687	**9.51E-04**	0.0
0.45-0.50	1,003,765	15,031	14,640	**1.13E-03**	0.0	14,266	14,039	5.37E-02	0.0	6,046	6,099	4.96E-01	0.0	299,028	296,055	**7.65E-11**	0.0
0.40-0.45	1,294,445	18,986	18,880	4.37E-01	0.0	18,459	18,105	**8.06E-03**	0.0	7,810	7,865	5.34E-01	0.0	387,925	381,789	**2.84E-32**	0.0
0.35-0.40	1,650,687	23,783	24,076	5.71E-02	0.0	23,277	23,087	2.08E-01	0.0	9,629	10,030	**5.91E-05**	-0.1	495,609	486,861	**2.07E-50**	0.0
0.30-0.35	2,095,342	30,227	30,561	5.43E-02	0.0	29,643	29,306	4.74E-02	0.0	12,263	12,732	**3.06E-05**	-0.1	632,561	618,010	**1.10E-107**	0.0
0.25-0.30	2,640,617	37,522	38,514	**3.54E-07**	0.0	36,784	36,933	4.35E-01	0.0	15,408	16,045	**4.55E-07**	-0.1	793,553	778,835	**8.75E-88**	0.0
0.20-0.25	3,317,366	47,090	48,385	**3.02E-09**	0.0	45,569	46,398	**1.06E-04**	0.0	19,111	20,157	**1.47E-13**	-0.1	984,798	978,439	**1.92E-14**	0.0
0.15-0.20	4,268,059	60,926	62,251	**8.81E-08**	0.0	59,111	59,695	1.61E-02	0.0	24,976	25,933	**2.51E-09**	-0.1	1,255,095	1,258,840	**7.03E-05**	0.0
0.10-0.15	5,451,393	79,151	79,510	2.00E-01	0.0	75,301	76,245	**5.76E-04**	0.0	32,760	33,123	**4.54E-02**	0.0	1,597,012	1,607,858	**2.27E-24**	0.0
0.05-0.10	6,017,447	89,894	87,766	**4.62E-13**	0.0	83,900	84,162	3.63E-01	0.0	37,931	36,563	**7.17E-13**	0.1	1,757,751	1,774,813	**1.58E-52**	0.0
0-0.05	4,119,252	58,999	60,080	**8.88E-06**	0.0	57,428	57,614	4.35E-01	0.0	27,334	25,029	**2.28E-48**	0.1	1,186,573	1,214,951	**1.87E-206**	0.0

* Binned delta allele frequencies for the contrast domestic vs. wild rabbits. Domestic and Wild reference allele frequencies (dRAF and wRAF) and ΔAF were calculated as: dRAF= mean(RAF Domestics), wRAF= (mean(RAF Wild French)+mean(RAF Wild Iberian))/2, ΔAF=abs(dRAF-wRAF)

** Statistical significance of deviations from expected values were tested with a standard X2–analysis (d.f.=1)

*** M-value (log2 fold change observed vs. expected SNP count). M-values were calculated using the average frequency of the corresponding annotation category as reference.

P-values in bold font indicate those <= 0.05. M-values in bold font indicate those with M >= 0.1 and with P <0.05

Table S7. Summary of electrophoretic mobility shift assays using nuclear extracts from ES-cell derived neural stem cells (ES or from mouse P19 embryonic carcinoma cells before (un-diff) or after neuronal differentiation (diff).

SNP	Chr	Pos	Gene*	**Shifted bands**			**Difference domestic vs. wild**		
				ES	P19 un-diff.	P19 diff.	ES	P19 un-diff.	P19 diff.
1	chr14	77728853	*SOX2*	+	+	+	-	-	++
2	chr14	77734115	*SOX2*	+	+	+	-	-	++
3	chr14	78187181	*SOX2*	+	+	+	-	-	++
4	chr14	78228121	*SOX2*	-	-	-	-	-	-
5	chr18	48839466	*PAX2*	+	+	+	-	-	+++
6	chr1	5538210	*KLF4*	+	+	+	-	-	-
7	chr1	5609176	*KLF4*	-	-	-	-	-	-
8	chr14	77386300	*SOX2*	+	+	?	-	?	-
9	chr14	77458138	*SOX2*	+	+	+	++	++	++
10	chr14	77700191	*SOX2*	+	+	+	-	-	-
11	chr14	77885074	*SOX2*	+	+	+	++	++	++
12	chr14	78046337	*SOX2*	+	+	+	-	-	-
13	chr14	78227621	*SOX2*	+	+	+	+++	+++	+++
14	chr14	78223293	*SOX2*	+	+	+	+++	+++	+++
15	chr1	5609251	*KLF4*	-	-	-	-	-	-
16	chr14	78256018	*SOX2*	+	-	-	-	-	-
17	chr14	78290160	*SOX2*	+	-	+	-	-	?

"-" = no shift, "+" = shifted, "++" = lower band intensity, "+++" = whole band/complex disappeared and "?" = unclear