Genome-Wide Footprints of Pig Domestication and Selection Revealed through Massive Parallel Sequencing of Pooled DNA

Andreia J. Amaral[1]*, Luca Ferretti[2,3], Hendrik-Jan Megens[1], Richard P. M. A. Crooijmans[1], Haisheng Nie[1], Sebastian E. Ramos-Onsins[2,3], Miguel Perez-Enciso[2,3,4], Lawrence B. Schook[5], Martien A. M. Groenen[1]

1 Animal Breeding and Genomics Centre, Wageningen University, Wageningen, The Netherlands, 2 Department of Animal Science and Food Technology, Universitat Autonoma de Barcelona, Bellaterra, Spain, 3 Animal Science Department, Centre for Research in Agricultural Genomics, Bellaterra, Spain, 4 Life and Medical Sciences, Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain, 5 Institute for Genomic Biology, University of Illinois, Urbana, Illinois, United States of America

Abstract

Background: Artificial selection has caused rapid evolution in domesticated species. The identification of selection footprints across domesticated genomes can contribute to uncover the genetic basis of phenotypic diversity.

Methodology/Main Findings: Genome wide footprints of pig domestication and selection were identified using massive parallel sequencing of pooled reduced representation libraries (RRL) representing ~2% of the genome from wild boar and four domestic pig breeds (Large White, Landrace, Duroc and Pietrain) which have been under strong selection for muscle development, growth, behavior and coat color. Using specifically developed statistical methods that account for DNA pooling, low mean sequencing depth, and sequencing errors, we provide genome-wide estimates of nucleotide diversity and genetic differentiation in pig. Widespread signals suggestive of positive and balancing selection were found and the strongest signals were observed in Pietrain, one of the breeds most intensively selected for muscle development. Most signals were population-specific but affected genomic regions which harbored genes for common biological categories including coat color, brain development, muscle development, growth, metabolism, olfaction and immunity. Genetic differentiation in regions harboring genes related to muscle development and growth was higher between breeds than between a given breed and the wild boar.

Conclusions/Significance: These results, suggest that although domesticated breeds have experienced similar selective pressures, selection has acted upon different genes. This might reflect the multiple domestication events of European breeds or could be the result of subsequent introgression of Asian alleles. Overall, it was estimated that approximately 7% of the porcine genome has been affected by selection events. This study illustrates that the massive parallel sequencing of genomic pools is a cost-effective approach to identify footprints of selection.

Introduction

Animal domestication is the process whereby animals adapt to humanized environments through a process of selection resulting in a different phenotype of the domestic animal compared to its wild counterpart [1].

Artificial selection is perhaps the best understood aspect of the domestication process [2]. Understanding how domestication has shaped the patterns of genetic variation is important, since domestication can reflect rapid evolution triggered by human-generated pressures. Previous studies have shown that domestication was associated with selective pressures on specific genes related to eg. growth [3] and coat color [4,5] and that artificial selection might have contributed to reduce polymorphism levels and increase linkage disequilibrium in some domesticated species [6–9]. However, the degree to which adaptive evolution has affected DNA polymorphism genome wide has not been extensively studied and few studies have determined which types of selection are most prevalent in domesticated animals.

The domestication of the European wild boar started around 13,000 B.P. and most likely was triggered by the introduction of domestic pigs by Middle East Neolithic farmers [10]. The process of selection for different environments within the European continent might have resulted in the generation of a wide variety

of domestic pig breeds with divergent phenotypes [11]. Behavioral phenotypes would also have been modified due to changes in social structure, reproduction and adaptation towards humans [1]. Today most breeds are created in captivity but, during the beginning of domestication, selected animals would freely breed among human settlements and would have had a flexible diet [12,13] resembling extensive pig production systems which still exist today (eg. Iberian pig). Thus, pig domestication can be defined as a mixture of weak selection (applied postzygotically with no conscious wish to alter a breed) and strong selection (applied prezygotically guided by a predetermined goal) for favorable traits [14,15]. Nowadays, a small number of breeds dominate pig production worldwide, i.e. Large White, Landrace, Duroc and Pietrain which have distinct phenotypes generated by selection applied with different intensities for desired production attributes.

Genomic regions under selection are expected to display extreme levels of DNA polymorphisms in comparison with unselected regions. Such regions displaying footprints of selection can be studied by determining the allelic variation of SNPs (single nucleotide polymorphisms). In human populations, previous studies have revealed footprints of recent positive selection involving genes linked to resistance to malaria [16], dairy farming [17], and brain development [18]. Likewise, studies have detected signatures of recent balancing selection in humans in response to disease [19] and to the need for kin recognition and survival [20]. However, these studies were based on SNP genotypes that had initially been identified by an ascertainment (or SNP discovery) process. This ascertainment creates a bias due to the fact that the sample size used for the SNP discovery panel is often small, causing a frequency-specific distortion in SNP discovery. Consequently, the frequency spectrum obtained from the two-tier sampling differs from the frequency spectrum obtained by sequencing the entire study sample [21]. This will affect any statistical measures that rely on the site frequency spectrum, including nucleotide diversity, Tajima's D and F_{st}. Alternatively, massive parallel sequencing (MPS), also called next generation sequencing, is considerably less expensive and is able to produce sequence data orders of magnitude higher than can be obtained by traditional Sanger capillary sequencing [22,23]. Consequently, MPS provides the opportunity to obtain genome-wide estimates of the genetic diversity of a species without the effect of ascertainment bias.

Previously, within a large scale SNP discovery project [24], a large data set of ∼380 million sequences were generated from pooled reduced representation libraries (RRLs) of Large White, Landrace, Duroc and Pietrain and the wild boar using the Genome Analyzer (GA, Illumina). Here we investigated whether signatures of selection due to domestication or recent breeding practices can be detected using this data set. We identified genomic regions with extreme values of nucleotide diversity and with extreme values of genetic differentiation among pig populations. Statistics were developed to characterize nucleotide diversity and genetic differentiation in pooled GA sequence data generated using MPS.

Our results suggest a prevalence of positive selection in genomic regions known to harbor genes affecting coat color, behavior, muscle development, and metabolism following domestication. Furthermore, these results indicate that the olfactory receptors and the immune system (MHC genes) have most likely undergone a process of balancing selection both in the domestic European pig breeds and in the wild boar.

Results

Sequence analysis

Approximately 380 million GA sequences (hereafter referred to as 'reads') were generated from RRLs obtained from pooled DNA of five pig populations (Duroc, Landrace, Large White, Pietrain and the wild boar; hereafter referred to as 'sampled populations') with divergent phenotypes (Figure 1; Table S1). The sampled animals were non-related (no shared grandparents) and represent current global breeds. Raw reads were preprocessed to remove reads with errors (see Methods) and the remaining 200 million reads were aligned with the pig genome reference assembly (Build 8). Since reads were obtained from RRLs of pooled DNAs, reads were aligned to the reference genome, forming clusters, for which only the information of the population of origin was available (Figure 1). A total of approximately 2% of the porcine genome met alignment quality parameters (see Methods), and the average sequencing depth ranged from 7.5X for wild boar to 10X for Duroc (Table 1). SNPs identified included 70% transitions and 30% transversions (Figure S1) and rare variants (SNPs observed in only one read) were nearly absent (Figure S1). The correlation between the GC content and the total number of aligned bases per
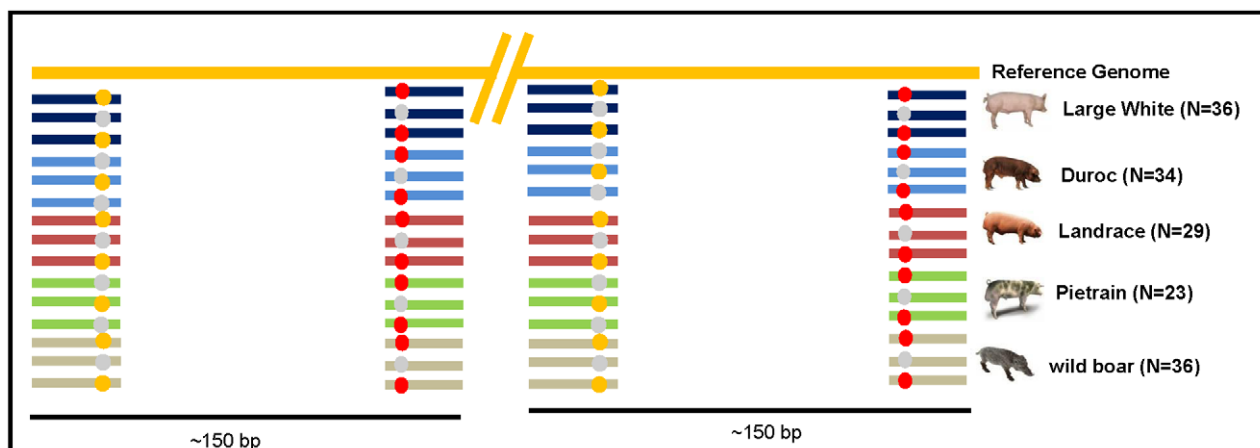
Figure 1. Schematic drawing showing the expected alignment between reads and the reference genome. The colored bars represent the reads; each color corresponds to a different pig population. The reads originated from ∼150 bp fragments of RRL libraries of pooled DNA for each pig population. Therefore, for each read the identification is only available at the population level. Colored dots on the reads represent SNP positions.
doi:10.1371/journal.pone.0014782.g001

Table 1. Summary statistics of sequence filtering and alignment of overall chromosomes in pig breeds and in wild boar.

Population	Total raw sequences	Total filtered sequences	Total aligned length	Average Sequence depth
Large White	81,501,174	45,233,951	40,580,400	7.8
Landrace	88,002,147	43,341,083	42,832,864	8.3
Pietrain	71,561,104	42,242,606	51,566,246	9.7
Duroc	75,925,390	36,424,674	37,176,400	10.0
wild boar	65,053,290	28,723,892	29,229,683	7.5
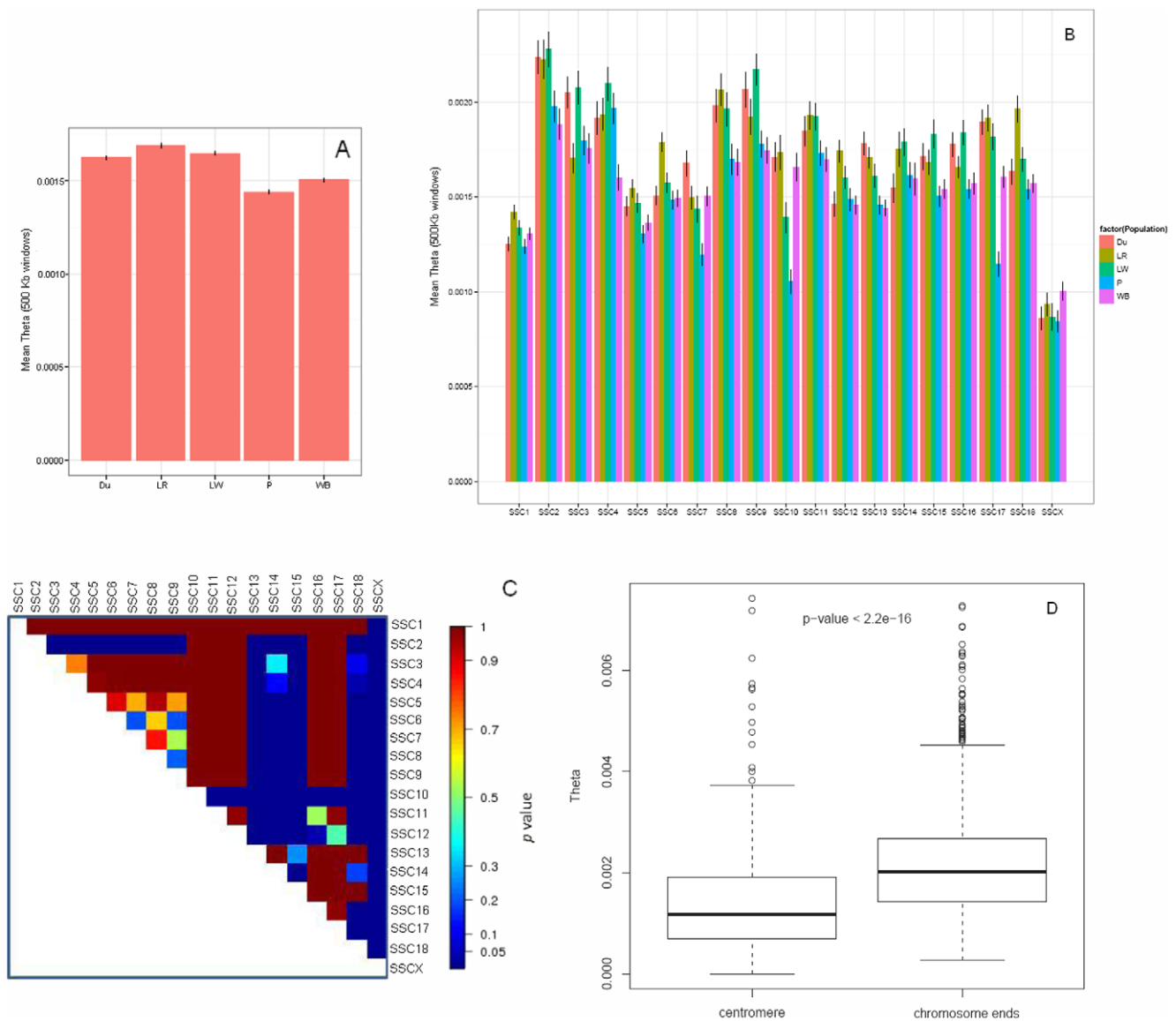
Figure 2. Average nucleotide diversity per chromosome and sampled population. Nucleotide diversity was estimated in non-overlapping 500 Kb windows ($\hat{\theta}_W$). A- Average $\hat{\theta}_W$ overall chromosomes per population. B- Average $\hat{\theta}_W$ per chromosome for each population. Vertical lines represent standard errors of the mean. C- Heatmap representing the p-values obtained from comparing the average $\hat{\theta}_W$ per chromosome. D- Boxplot showing that centromeres and chromosome ends behave different in terms of nucleotide diversity.
doi:10.1371/journal.pone.0014782.g002

Figure 3. Genome-wide variation of nucleotide diversity in the studied populations. Each dot represents the observed nucleotide variation over a window of 500 Kb.
doi:10.1371/journal.pone.0014782.g003

Table 2. Number of regions per breed with significant values of nucleotide diversity ($\hat{\theta}_W$).

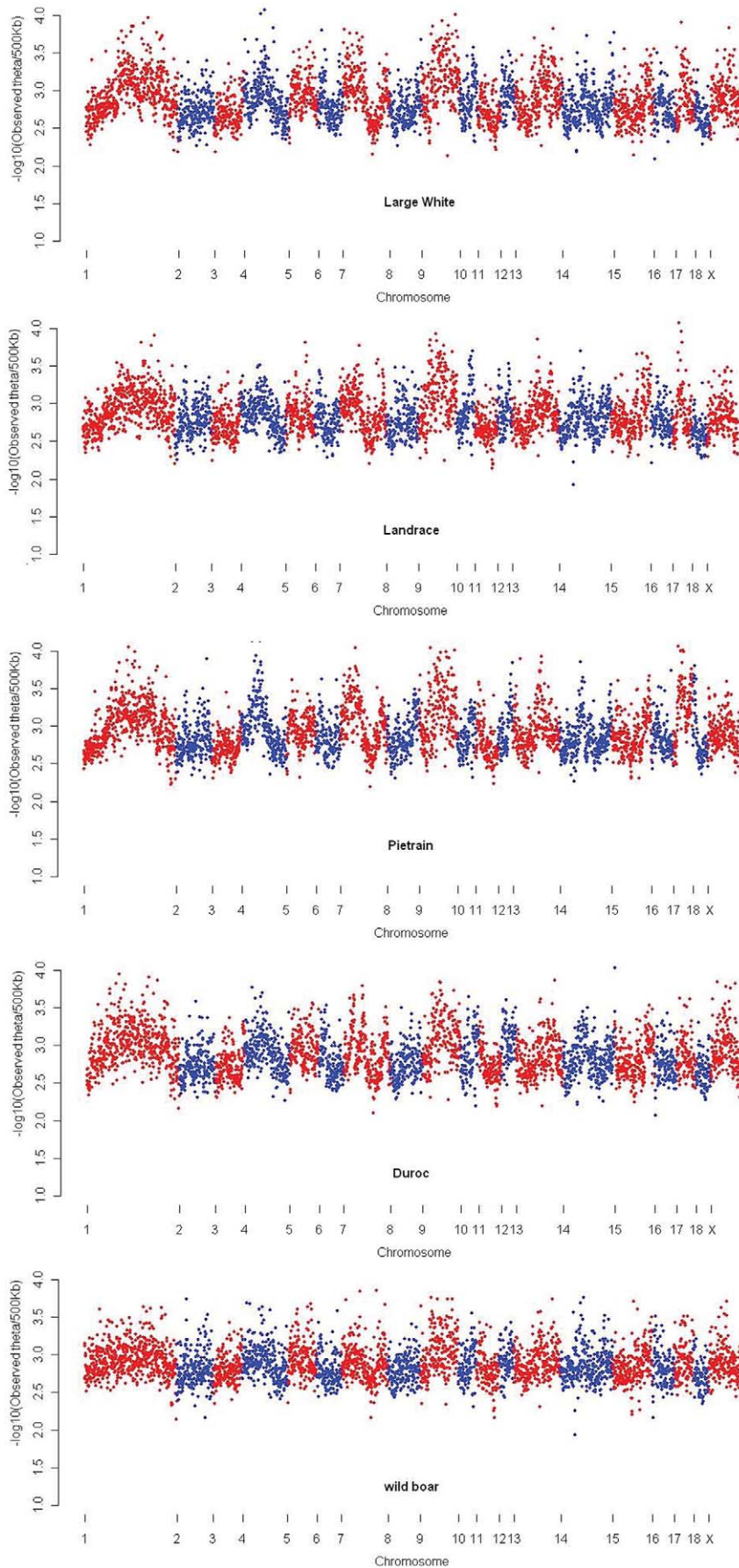| Breed	TWHT	TWLT	LW	LR	P	Du	WB
Large White (LW)	391	317		124 (39)	119 (38)	88 (28)	32 (10)
Landrace (LR)	446	354	155 (35)		122 (34)	87 (25)	34 (10)
Pietrain (P)	478	408	164 (34)	200 (42)		86 (21)	32 (8)
Duroc (Du)	421	331	137 (33)	150 (36)	147 (35)		27 (8)
wild boar (WB)	226	111	91 (40)	100 (44)	107 (47)	90 (40)	

TWHT- total number of windows with a significant high value of $\hat{\theta}_W$. TWLT- total number of windows with a significant low value of $\hat{\theta}_W$. Values shown above the diagonal are the number of regions with low $\hat{\theta}_W$ values that were shared between sampled populations (parentheses enclose the percentage, calculated over the TWLT between the pairs compared); values shown below the diagonal are the number of regions with high $\hat{\theta}_W$ values that were shared between sampled populations (the percentage in parentheses is calculated over the TWHT between the pairs compared).
doi:10.1371/journal.pone.0014782.t002

cluster ranged between 0.67 and 0.72 (p-value <0.0001) for all sampled populations.

Genome-wide estimates of nucleotide diversity in pig

The Watterson's estimator of nucleotide diversity ($\hat{\theta}_W$) was modified in order to take into account important characteristics of the data (pooled DNA, sequencing errors and the lack of rare variants). Nucleotide diversity was estimated in non-overlapping windows of 500 Kb for each chromosome and for each population. The Landrace breed displayed the highest average of $\hat{\theta}_W$ over all the chromosomes, whereas the Pietrain breed displayed the lowest average of $\hat{\theta}_W$ (Figure 2A). The average of $\hat{\theta}_W$ per chromosome is shown for each population in Figure 2B. In particular, the average of $\hat{\theta}_W$ for the Pietrain breed on SSC8, SSC15, and SSC18 was significantly lower (p-value <0.0001) than the average of $\hat{\theta}_W$ observed in the Landrace and Large White breeds (Figure 2B). Considering the comparison of the average of $\hat{\theta}_W$ between chromosomes, significant differences (p-value<0.05) were also observed between autosomes (Figure 2C). A significant decrease of $\hat{\theta}_W$ (p-value<0.0001) was observed on SSC X (mean $\hat{\theta}_W = 0.0008$) compared to the autosomes (0.0010<mean $\hat{\theta}_W$ <0.022) (Figure 2C). For metacentric and submetacentric

chromosomes the values of $\hat{\theta}_W$ were significantly higher at the chromosome ends than at the centromeres (Figure 2D). The variation of $\hat{\theta}_W$ along the chromosomes was similar across all populations (Figure 3).

Identification of candidate regions of recent selection

In order to identify outlier estimates of $\hat{\theta}_W$ potentially representing candidate regions that might have been under selection, 95% confidence intervals (CI) of $\hat{\theta}_W$ were obtained for each window by performing neutral coalescent simulations with recombination (see Methods). Genomic regions with $\hat{\theta}_W$ values outside the boundaries established by the CI were classified as candidate regions for recent selection. The number of candidate regions putatively under selection identified in the different populations is summarized in Table 2. The Pietrain breed showed the highest number of regions with a significant low value of $\hat{\theta}_W$ whereas the wild boar had the lowest. Approximately 70% of putatively selected regions were observed in only one of the sampled populations (Table 2).

Approximately 36% of the candidate regions with a value of $\hat{\theta}_W$ below the lowest C.I. boundary (LT) were shared among the white coated breeds, Pietrain, Large White, and Landrace. The Duroc (red coated) breed shared fewest regions with the other breeds (Table 2). Only 10% were shared between any of the domesticated breeds and the wild boar population (black coated) (Table 2). The candidate regions under selection that were shared between populations were also investigated through correspondence analysis. The white coat color breeds Landrace, Large White and Pietrain share more candidate regions under selection and therefore cluster together (Figure 4A). The Duroc breed apparently has undergone a divergent effect of positive selection when compared with the other breeds since this breed contributed most to the space arrangement of the plot shown in Figure 4A.

Approximately 40% of the candidate regions with values of $\hat{\theta}_W$ above the highest C.I. boundary (HT) were shared among the domestic breeds but only 20% were shared with the wild boar population. The correspondence analysis of the HT regions in all the breeds and the wild boar showed that the wild boar shared many regions with the white breeds (Figure 4B). The Duroc, Landrace, and Pietrain breeds represented the domestic breeds that contributed most to the spatial arrangement of the plot (Figure 4B) and the ones that shared the lowest number of HT regions.
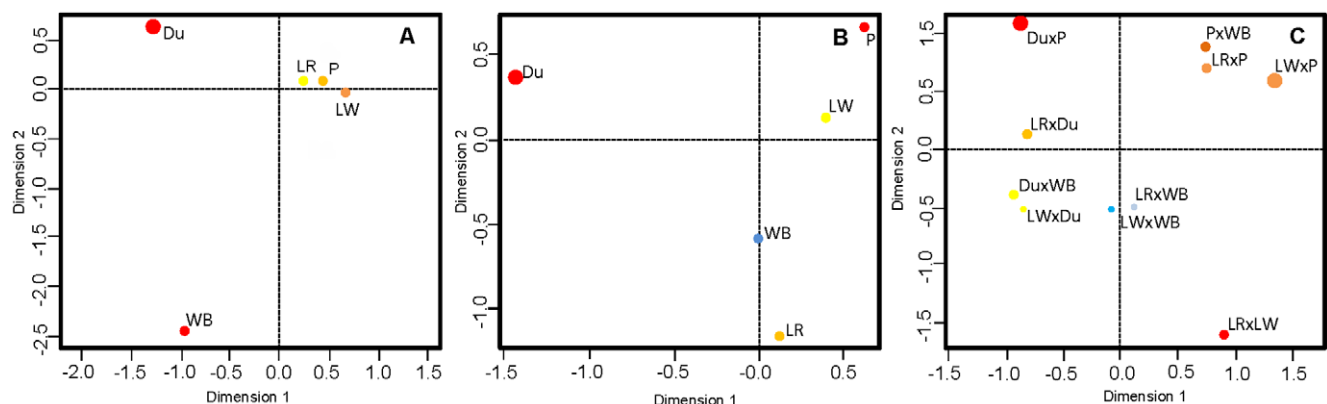
Figure 4. Correspondence analysis of population *vs.* genomic regions under selection and observed $\hat{\theta}_W$. A- LT regions (regions with $\hat{\theta}_W$ smaller than the lowest 95% C.I. boundary); B- HT regions (regions with $\hat{\theta}_W$ larger than the highest 95% C.I. boundary); C- Genomic regions with significant high F_{st} values. Different color and point size indicate the relative contribution of each population to the space arrangement in the plot. Brown color and larger size indicates highest contribution, light blue and smaller size indicates smallest contribution.
doi:10.1371/journal.pone.0014782.g004

Table 3. Examples of genes putatively under selection.

	Population	$\hat{\theta}_W$	95% C.I.
KIT	Large White	0.00034	0.0005–0.002
	Landrace	0.00021	0.0006–0.002
	Pietrain	0.00025	0.0004–0-001
	Duroc	0.00067	0.0005–0.002
	wild boar	0.00078	0.0004–0.002
TRIM26	Large White	0.00487	0.001–0.004
	Landrace	0.00446	0.001–0.003
	Pietrain	0.00341	0.001–0.003
	Duroc	0.00370	0.001–0.003
	wild boar	0.00289	0.0006–0.003
OR4K13	Large White	0.00813	0.0008–0.003
	Landrace	0.00615	0.0006–0.002
	Pietrain	0.00340	0.0005–0.002
	Duroc	0.00863	0.0007–0.003
	wild boar	0.00681	0.0003–0.002
MAPK1	Large White	0.00029	0.0005–0.002
	Landrace	0.00043	0.0008–0.002
	Pietrain	0.00063	0.0006–0.002
	Duroc	0.00055	0.0004–0.002
	wild boar	0.00644	0.0003–0.002
PPP1R1B	Large White	0.00080	0.0011–0.003
	Landrace	0.00065	0.0011–0.0027
	Pietrain	0.00119	0.0012–0.0026
	Duroc	0.00057	0.0010–0.0027
	wild boar	0.00073	0.0007–0.0026
LRRTM2	Large White	0.00137	0.0014–0.0040
	Landrace	0.00098	0.0014–0.0038
	Pietrain	0.00105	0.0015–0.0035
	Duroc	0.00167	0.0012–0.0037
	wild boar	0.00153	0.0009–0.0036

Values represent the $\hat{\theta}_W$ on the window overlapping the gene and the respective 95% confidence interval.
doi:10.1371/journal.pone.0014782.t003

Genes overlapping candidate regions of selection

After comparing the genomic locations of the identified candidate regions of recent selection with the available annotation of the pig genome, genes were identified which putatively have been under selection due to pig domestication (Table S2).

We have identified LT regions overlapping with the *KIT* gene, responsible for the white color, on SSC8 in the Large White, Landrace, and Pietrain breeds, but not in the red coated Duroc breed or the black coated wild boar (Table 3; Figure 5A). A systematic overlap of LT regions with genes related to growth and muscle development (eg. *MAPK1* gene - SSC 14:51.5–52 Mb*)* was observed in domestic pig breeds (Table 3). Moreover, a consistent pattern was observed in which genes related with brain development and neuronal functions (eg. *PPP1R1B* gene - SSC12:15–15.5 Mb) overlapped LT regions in the domestic pig breeds but not in the wild boar population (Table 3; Table S2). In HT regions, we have detected several genes from the olfactory receptor complex and from the *SLA* locus located in SSC7, where the strongest signals correspond to the *TRIM26* gene (zinc finger

protein 173) (Table 3; Figure 5A) and to the *OR4K13* gene (olfactory receptor 4K13) (Figure 5B). The *TRIM26* gene from the *SLA* locus is located between positions 24–24.5Mb on SSC7 and significantly high values of $\hat{\theta}_W$ were observed in the Duroc, Pietrain, Landrace and Large White breeds. The *OR4K13* gene, from the olfactory receptor gene family overlaps with the genomic region between 86–86.5 Mb and displayed significant high values of $\hat{\theta}_W$ in the studied populations.

Widespread signals of breed differentiation

Genomic regions affected by selection can also potentially be identified as outliers in the extreme tails of the empirical distribution of a measure of genetic differentiation, e.g., F_{ST}. Such an approach has the advantage of being free of theoretical predictions regarding population structures or demography [25]. To this end, the measure F_{ST} as defined by Nei [26] was adapted to accommodate the characteristics of the data, namely reads of pools, ascertainment bias against rare variants and sequencing errors. The estimates of F_{ST} varied along chromosomes (Figure S2). Most F_{ST} values were significantly different from 0 (p-value<0.05) (Figure S2) displaying an overall mean of 0.122 with a standard deviation of 0.187. The F_{ST} values for all pair-wise population comparisons are shown in Table 4. The Duroc breed displayed the highest F_{ST} value compared to the other breeds, having the largest divergence with the Pietrain. The Landrace and Large White breeds were the populations most similar to the wild boar. Genomic regions that displayed high genetic differentiation between populations - F_{ST} values in the 95% quartile with significant p-values (<0.05) – represented candidate regions of positive selection and were selected for further analysis. A correspondence analysis was performed on these identified genomic regions in order to investigate whether the studied pairs of populations shared these regions (Figure 4C). The Landrace *vs.* wild boar and Large White *vs.* wild boar clustered together. In contrast, the Pietrain *vs.* Duroc clustered far apart from the other populations *vs.* Pietrain.

Types of biological processes under selection

To investigate whether certain GO categories or KEGG pathways were enriched in the identified candidate regions of recent selection, a gene set enrichment analysis was performed. Genes overlapping HT and LT regions were classified using gene ontology (GO) categories and functional pathway (KEGG) categories. If a genomic region contained multiple genes for the same GO/KEGG category, this category was only counted once thus avoiding GO/KEGG categories being significant due to a cluster of genes from the same gene family.

The results suggest an overrepresentation of genes related to muscle development, growth and melanogenesis in LT regions of domestic pig breeds and overrepresentation of genes related to immune defense in the wild boar (raw p-value<0.05) (Table S3). Furthermore an overrepresentation of genes related to metabolism was observed in the LT regions of all the studied populations (raw p-value<0.5). These results however were no longer significant after correcting for multiple testing (0.3<adjusted p-value<0.9). Within the HT regions, there was a significant overrepresentation (adjusted p-value<0.0001) of genes related to olfaction and other sensorial capabilities (Table 5). Within genomic regions extremely differentiated between the Duroc and Pietrain breeds, a significant enrichment of genes related to olfaction (adjusted p-value = 2×10^{-3}) was observed. In fact, from a total of 1,577 genes located within the HT regions for all the studied populations, 144 genes encode for olfactory receptors.
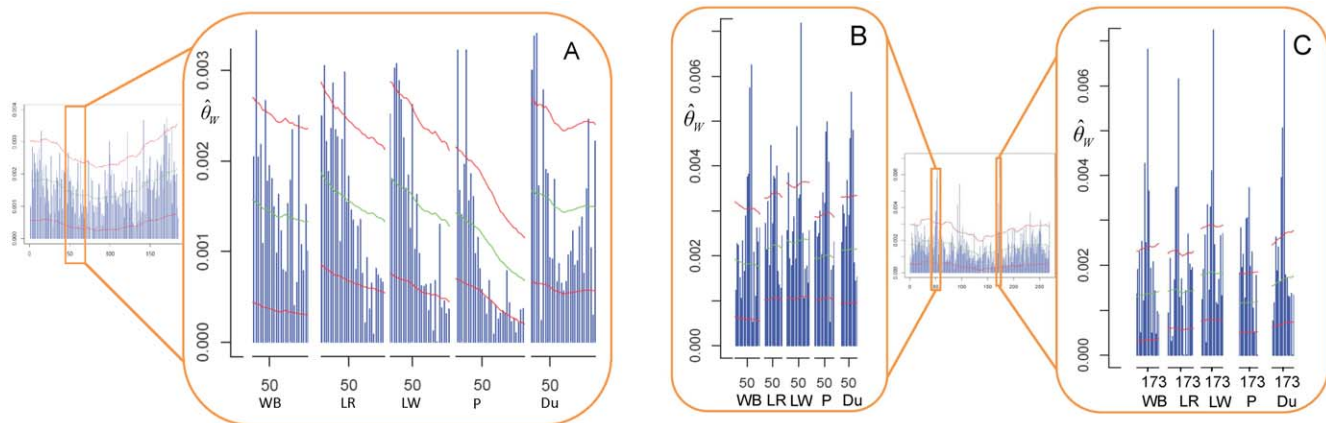
Figure 5. Variation of $\hat{\theta}_W$ along the wild boar SSC8 and SSC7. Blue bars represent point estimates for each 500 Kb window. Red lines represent confidence interval limits with a significance level of 95%. Green lines represent the average $\hat{\theta}_W$ per window. The insets are enlargements of the orange boxes that show details of the variation of $\hat{\theta}_W$ in genomic regions that deviate from the standard neutral model in all sampled populations. A- Detail of genomic regions with a significantly low $\hat{\theta}_W$ and that potentially contains the *KIT* gene. B- Detail of genomic regions with a significantly high $\hat{\theta}_W$ and that potentially contains the *TRIM26* gene member of the *SLA* locus. C- Detail of genomic regions with a significantly low $\hat{\theta}_W$ and that potentially contains the *OR4K13* gene.
doi:10.1371/journal.pone.0014782.g005

Discussion

By performing MPS of RRLs obtained from DNA pools we identified candidate regions within the porcine genome that putatively have been under selection. MPS on RRLs was shown to be a simple, cost-effective approach applicable to any species, even in the absence of a complete reference genome. Using a coalescent-based estimation of the uncertainty of the observed nucleotide diversity we have identified regions with a value of $\hat{\theta}_W$ deviating significantly from the expectations and which are therefore candidate regions of selection. Genomic regions that had a $\hat{\theta}_W$ value significantly smaller than the expectations represented regions where the frequency of the favored allele in a haplotype with low diversity could have increased very rapidly and thus were considered as potential candidate regions of recent positive selection. In contrast, in genomic regions with a $\hat{\theta}_W$ value significantly larger than the expectations an increase of variability was observed along with an increase of intermediate allele distributions and these regions were therefore considered as potential candidate regions for the occurrence of balancing selection [27]. Because demographic processes can mimic selection, the interpretation of outlier values of nucleotide diversity is not trivial. In fact, the occurrence of population growth can also result in a high occurrence of low-frequency alleles and population

subdivision also results in the high occurrence of intermediate-frequency alleles. Furthermore, the occurrence of copy number polymorphisms can also result in the increase of nucleotide diversity [27,28]. Therefore, throughout the discussion of our results the identified outlier regions will always be referred as candidate regions of selection. Nevertheless, this study provides the first genome-wide map of putative signatures of selection in the porcine genome. These results, together with knowledge regarding the functional aspects of the regions, provide new insights into the molecular nature of animal domestication and selection.

Variation of nucleotide diversity between pig breeds and wild boar

Overall, and in agreement with previous studies [29], the Pietrain breed demonstrated the lowest $\hat{\theta}_W$ which is indicative of a small effective population size compared to other breeds and a low contribution to the genetic diversity of the *Sus scrofa* species [30]. It is worth noting that the wild boar population had low values of $\hat{\theta}_W$ as compared to the Large White, Landrace, and Duroc breeds. This low level of $\hat{\theta}_W$ in the wild boar may be attributed to the severe decline of the European wild boar populations that over time might have lead to high levels of linkage disequilibrium and also to a smaller effective population size [6]. As reported previously for other species, we observed lower $\hat{\theta}_W$ on SSCX compared to the autosomes, probably reflecting the smaller effective population size for this chromosome [31]. The observed significant increase of $\hat{\theta}_W$ in the chromosome ends in comparison to the centromeric areas is consistent with previous studies in humans [32,33] and other mammals [32,34] and most likely is due to the higher level of recombination rate towards telomeres.

Signatures of positive selection and genetic differentiation between breeds

Most of the putative signatures of positive selection were observed in a single population and, the number of candidate regions detected in domestic breeds was larger than in wild boar. Interestingly, domestic pig breeds shared a higher proportion of these regions between themselves than with the wild boar population. This suggests that domestication, or modern artificial selection, could have affected genes related to the same biological

Table 4. Genetic differentiation between sampled populations.

	LW	LR	P	Du
Large White (LW)				
Landrace (LR)	0.10 (0.003)			
Pietrain (P)	0.10 (0.003)	0.12 (0.003		
Duroc (Du)	0.14 (0.003	0.14 (0.003)	0.16 (0.003)	
wild boar (WB)	0.10 (0.003	0.11 (0.003)	0.13 (0.003)	0.13 (0.003)

Values represent the average genetic differentiation (F_{st}) between population pairs (parentheses enclose standard errors of the mean).
doi:10.1371/journal.pone.0014782.t004

Table 5. Enrichment of Gene Ontology (GO) categories and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways among HT regions.

GO category	Landrace	Large White	Pietrain	Duroc	Domesticated breeds	Wild boar
Sensory perception of smell	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
G-protein, coupled receptor protein signaling pathway	0.03	0.04	<0.0001	0.0008	0.02	<0.0001
G-protein, coupled to cyclic nucleotide second messenger signaling	ns	0.05	<0.0001	ns	0.04	ns
neurological system process	0.009	0.003	0.01	0.0008	<0.0001	ns
sensory perception	0.004	0.001	0.02	0.002	<0.0001	<0.0001
response to stimulus	0.003	<0.0001	0.0009	0.002	<0.0001	0.0001
KEGG Pathway						
Olfactory transduction	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

The analysis was performed for all sampled populations individually and for the domestic breeds in compound.
Values represent p-values for significance of enrichment, after correction for multiple testing.
"ns" indicates that the p-value was >0.05 (not significant).
doi:10.1371/journal.pone.0014782.t005

pathways in the different pig populations. The correspondence analysis of these genomic regions, suggests strong directional selection that is consistent with the history of pig breeds. Landrace and Large White clustered together, sharing many of the regions putatively under positive selection showing the lowest levels of genetic differentiation between breeds. In fact, the Landrace breed originated in Denmark from a cross between the wild boar and Large White [11]. In contrast, the Duroc shared less putative regions of positive selection with the other breeds and was the most genetically differentiated from the other breeds. This finding is consistent with previous studies [35,36] adding support to the hypothesis that this breed has a distinct origin from the rest of European breeds. Historical records suggest that the Duroc breed originated from European red breeds or the Red Guinea Hog [35]. Our results revealed that genetic differentiation was higher between domestic breeds than between a given domesticated breed and the ancestor wild boar. Furthermore, the analysis of clustering of regions with high level of genetic differentiation revealed a pattern, which is in accordance with the history of breed domestication and selection. These findings are compatible with the occurrence of multiple domestication events in Europe, resulting in the generation of several domestic pig breeds that are more similar to its ancestor and yet are highly differentiated in terms of phenotypes and genotypes. An alternative explanation is the introgression of Asian germplasm in a post-domestication stage. In fact, the use of a pig from Canton province in China for the creation of the Large White pig has been documented [11], and the introgression of Asian alleles has been observed in previous genetic studies [6,37]. This may explain the high levels of nucleotide diversity in the Large White and Landrace breeds.

Footprints of selection reveal genetic differences in coat color, growth, and behavior due to domestication

Since the dataset comprised approximately 2% of the genome, windows of 500 Kb were used in order to have a reasonable number of SNPs per window hence limiting the resolution of the gene set enrichment analysis. In addition, the current annotation of the pig genome has limited availability of GO terms and KEGG pathways further reducing the sensitivity of the analysis. Because of these limitations we were only able to provide suggestive evidence for over representation of specific biological processes

affected by positive selection. The biological functions of genes located within the genomic regions that were putatively under positive selection, suggests that recent positive selection occurred in the domestic pig breeds in genes associated with coat color, behavior, growth, and muscle development. Concerning coat color, the genomic region containing the *KIT* gene was shown to have unusually low diversity in the white breeds -Large White, Landrace, and Pietrain - but not in the Duroc and wild boar populations. This observation is in agreement with the identification of a mutation in the *KIT* gene present in all white breeds and absent in breeds with a colored coat [4] providing support for our method. These results also revealed that genes involved in growth and muscle development overlap with regions putatively under positive selection and with regions with significantly high levels of genetic differentiation in domestic pig breeds. This observation is consistent with the history of the domestic pig breeds studied, which have higher growth rates and proportion of muscle than the ancestral wild boar populations. A genome scan in chicken revealed a similar pattern – several genes related to growth, muscle development and skin color (yellow skin) were identified as having been under positive selection [38]. In the chicken it was shown that a mutation in the gene that encodes for the thyroid hormone might have arisen during domestication of chicken, with functional consequences in terms of metabolism and growth of the domestic chicken populations. Our results also provide suggestive evidence for the selection of genes linked to metabolism that likely reflects selection due to adaptation to human-altered environments and feed. Our results and those of Rubin and colleagues [38] are consistent with the evolution for size and coat color being a consequence of domestication and not a consequence of natural selection or genetic drift. Interestingly, the wild boar populations revealed evidence of positive selection in genes related to disease resistance. This observation may reflect natural selection for survival and fitness in the wild and suggests that the release of selective constrains in genes related to disease resistance in domestic breeds could have lead to a faster accumulation of genetic diversity among these genes.

The process of domestication can also result in related behavioral changes. Behaviors that are important for survival in nature, like finding food and predator avoidance, may not provide significant advantages in humanized environments associated with

captive breeding [1]. Moreover, it is expected that farmers would have selected for more docile animals. In this study, we observed a systematic pattern in which genes related to brain and neuron functions overlapped regions putatively under positive selection in the domestic pig populations but not in the wild boar population. For example, one region harbored the *PPP1R1B* gene that codes for the dopaminergic neurotransmitter critical for motivated behavior, working memory, and reward-related learning [39]. Therefore we propose that similarly to dogs which have inferior observational learning skills compared to wolves [40], these signatures might have been a consequence of domestication.

Are the MHC and the olfactory receptors under the influence of balancing selection?

Several genes identified in this study have been shown to be under the influence of balancing selection in other mammals. Our results indicate the maintenance of unusually high nucleotide diversity in the MHC genes of the porcine genome, suggesting an effect of balancing selection similar to observations in other mammals such as e.g., dogs [41], cattle [42], sheep [43], rat [44], rhesus macaque [45], and humans [19]. The overrepresentation of olfactory receptor genes and of genes related to other sensory traits of the pig were significant in our analysis ($p<0.001$). The maintenance of high variability in olfactory receptors has been observed in humans [20], other primates [46] and mouse [47]. In humans, a model of overdominance has been proposed for the evolution of olfactory receptors [20]. Individuals that are heterozygous for olfactory receptors can potentially double the number of different odorant-binding sites encoded in the genome, thus allowing the individual to discriminate among closely related structural odorants [48]. Our results suggest that individual recognition in pigs is crucial for survival not only for the wild boar reared in wild environments but also for domestic pigs in human-altered environments.

Conclusions

Artificial selection has produced dramatic alterations in livestock environments, and undoubtedly has left important selective footprints throughout the genomes of domesticated species. To date, research of livestock has been centered on the identification of individual genes as candidates for selection. This study provides a genome-wide characterization of DNA polymorphism of the pig and yields important insights into the types of biological processes that have been targets of selection during pig domestication. Putative signals of selection were detected for coat color, growth and muscle development. Furthermore, our results suggest that selection might also have occurred at genes related to metabolism, behavior, olfaction and disease resistance. This investigation showed that the study of domestication can progress rapidly due to the use of massive parallel sequencing proposing a top-down approach, where candidate genes can be identified in a whole-genome approach by using a representative sample of the genome.

Methods

Material

We analyzed a total of 380 million 36 bp reads from four pig breeds, including Duroc (N = 34), Landrace (N = 29), Large White (N = 36), Pietrain (N = 23) and from the wild boar (N = 36) (Figure1; Table S1). Reads were generated from RRL libraries produced from DNA pools of each of the breeds. Details concerning DNA extraction, preparation of DNA pools and RRL libraries have been described previously [24]. RRL libraries

were sequenced using the 1G Genome Analyzer (Illumina, San Diego, California, USA).

Sequence analysis and estimation of nucleotide diversity

For each breed, reads were trimmed at 33 bp and reads with homopolymers (>17nt), unknown bases and overrepresented sequences (more than two times the observed average frequency) were removed. We also applied filters for sequence quality based on the quality scores provided by the Illumina base calling software. Because base quality decreases towards the 3′end of the reads [34], reads with an average quality lower than 20 were also removed. The remaining reads were aligned to *Sus scrofa* assembly 8 (ftp://ftp.sanger.ac.uk/pub/S_scrofa/assemblies/PreEnsembl_Sscrofa8) allowing up to two mismatches using MAQ (49). For each aligned base of the consensus sequence, the alignment algorithm generates [49] a quality score (*pSNP*). This measure *pSNP* quantifies base calling errors and SNP calling errors [49]. Only unique alignments were considered and clusters were only selected if the read depth was between 4 and 40.

After read mapping and SNP calling (Figure 1), nucleotide diversity $\hat{\theta}_W$ was estimated for each window according to a "modified Watterson estimator" based on the number of segregating sites *S*. Several estimators of this kind have been proposed for individual sequencing [28,50,51] which do not take into account the effect of sequencing pooled DNA. We have developed a "modified Watterson estimator" which is a Maximum Composite Likelihood estimator for pooled sequencing and is presented in equation 1, where n_o is the sample size as measured by the number of independent chromosomes, $n_s(i)$ and $L(i)$ are the read depth and length of the *ith* cluster, and *S* is the total number of segregating sites; *pSNP* is the consensus quality generated by MAQ [49], a_j is the sum of *1/i* from *1* to *j-1*, p_c is the probability that a set of chromosomes randomly extracted (with repetitions) from n_o possible origins contains precisely *j* different chromosomes (equation 2), and *k* is the derived allele frequency.

$$\hat{\theta}_W = \frac{S - \sum_{SNPs} 10^{-\frac{pSNP}{10}}}{\sum_i L(i)\left(\sum_{j=2}^{\min(n_s(i),n_o)} p_c(j|n_s(i),n_o)a_j - \sum_{k=1}^{n_o-1} \frac{n_s(i)}{n_o}\left(\frac{k}{n_o}\right)^{n_s(i)-2}\right)} \quad (1)$$

The second term in the denominator of equation 1, is applying a correction for sequencing and SNP calling errors, using $10^{-\frac{pSNP}{10}}$ which is the probability that a SNP is not a true SNP according to the SNP quality estimator of MAQ [49] (for example, a value of *pSNP*=20 means that the error probability is 1%, *pSNP*=30 means 0.1%, and so on). In the denominator, a_j is weighted for the probability that in *j* sequences, we have *j* chromosomes (equation 2) and is corrected for the bias against rare variants in the SNP frequency spectrum of the data similar to the approach of Achaz [52].

$$p_c(j|n_s,n_o) = \sum_{i=0}^{j-1} (-1)^i \binom{n_0}{j}\binom{j}{i}\left(\frac{j-i}{n_0}\right)^{ns} \quad (2)$$

A dynamic algorithm was programmed in *C* by the authors to estimate $\hat{\theta}_W$. R scripts were developed to perform the required statistical analysis (http://cran.r-project.org/).

An error model was applied to infer the rate of false SNPs from the alignment output. The model predicts the relation $r_{tr}(w) = r + c/\hat{\theta}_W(w)$ where $r_{tr}(w)$ and $\hat{\theta}_W(w)$ are the observed

transition rate and observed variability for the wth window, r is the expected transition rate according to Amaral and colleagues [34] and c is the product between the base error rate and the difference between the transition rate for sequencing errors and the expected transition ratio. This model was fitted to the consensus sequence, in order to establish the best thresholds for further filtering for alignment quality (results not shown). Final calculations of $\hat{\theta}_W$ were performed after minimizing the rate of false SNPs by applying the following rules to the alignment output. Bases from positions 1 to 16 were selected for a $pSNP$ larger or equal than 5, bases from positions 17 to 20 were selected for a $pSNP$ larger or equal than 22, bases from positions 21 to 25 were selected for a $pSNP$ larger or equal than 30, bases from positions 26 to 29 were selected for a $pSNP$ larger or equal than 35 and bases from positions 31 to 33 were selected for a $pSNP$ larger than or equal to 40.

Confidence intervals for $\hat{\theta}_W$

In order to identify candidate genomic regions that are under selection, the statistical uncertainty of our estimates (confidence intervals) was quantified by performing neutral coalescent simulations with recombination. Since for all chromosomes the average $\hat{\theta}_W$ ranged from 0.001 to 0.0022 across breeds, we performed simulations with 1000 iterations each using MaCS [53] with parameters $\theta = 0.001$ and $\theta = 0.0022$ and assuming levels of population recombination rate of 0.0015 and 0.002 respectively, similar to those used in previous studies [54]. For each simulation, a particular window of 500 Kb was simulated, sampled and θ was estimated using equation 1. The variance for a generic value of θ ($V(\theta)$) was then obtained by linear interpolation of the variance between the simulated values $\theta = 0.001$ and $\theta = 0.0022$. A local average of $\hat{\theta}_W$ was estimated from the data over a sliding process covering 60 windows and excluding the 20 windows with the highest and lowest values in order to take into account regional differences in the levels of variability associated with regions of high and low recombination. Finally, confidence intervals with 95% confidence level were centered around the local average of $\hat{\theta}_W$ with width $2\sqrt{V(\theta)}$. The procedure above was repeated for all windows in the data. Perl and R (http://cran.r-project.org/) scripts were developed by the authors to sample simulated windows according to the observed data and to estimate the confidence intervals.

Modified estimator of genetic differentiation

The global F_{ST} statistic for multiples sites defined by Nei [26] was modified in order to take into account pooled GA sequencing, sequence and consensus errors and ascertainment bias against singletons. For each pair of sampled populations, processed reads were selected in order that, only reads corresponding to the same genomic regions were used to access $\hat{F}_{ST}$ using the following equation,

$$\hat{F}_{ST} = 1 - \frac{\hat{\theta}_{\Pi 1 + \Pi 2}}{2\Pi_{a(1)} + c_s\left(\hat{\theta}_{\Pi 1} + \hat{\theta}_{\Pi 2}\right)} \qquad (3)$$

where $\hat{\theta}_\Pi$ is the average nucleotide diversity between sequences coming from different sampled populations and c_s is a factor that corrects for ascertainment bias towards singletons (Text S1). The pairwise nucleotide diversity between each population pair ($\hat{\theta}_{\Pi 1}$ and $\hat{\theta}_{\Pi 2}$) was estimated according to the following equation,

$$\hat{\theta}_\Pi = \frac{1}{L}\sum_i \left(\frac{n_o}{n_o - 1 - 2\sum_{k=1}^{n_o-1}\left(\frac{k}{n_o}\right)^{n_s(i)-2}}\right) \frac{2m_i(n_s(i)-m_i)}{n_s(i)(n_s(i)-1)}\left(1 - 10^{-\frac{pSNP(i)}{10}}\right)(4)$$

where i runs over all SNPs in the considered window. The term in the first bracket corrects for pooling and singleton removal while the term in the last bracket corrects for the probability of false SNPs. F_{ST} was estimated per genomic windows of 500 Kb and the significance of F_{ST} values was estimated by a permutation test (1000 permutations per population pair). Genomic regions with F_{ST} values in the 95% quartile with significant p-values (<0.05) were considered for further analysis as an ad hoc criteria used to select regions under selection.

Tests of Means

Differences between the level of nucleotide diversity between chromosomes and between chromosome ends and telomeres were tested using the Wilcox test function in R (http://cran.r-project.org/). Centromeric coordinates were obtained in the pig map (version 1009 - http://pre.ensembl.org/Sus_scrofa_map/Location/Genome?r=2:6758451-7077409). Chromosome ends corresponded to windows in the two extremes of the chromosomes (5% of total number of windows).

Functional enrichment analysis

We used the annotation for *Sus scrofa* assembly 8 available from pre-Ensembl (ftp://ftp.sanger.ac.uk/pub/S_scrofa/assemblies/PreEnsembl_Sscrofa8), generated by orthologous comparisons with human transcripts. Genes were considered to be overlapping when a genomic region of interest and the gene position were contained inside or partially inside the boundaries of the genomic region. The human Ensembl gene IDs were used to extract human Entrez gene IDs and protein family IDs by querying the Ensembl database (http://www.ensembl.org/) via the R package, biomaRt [55]. Using AnnotationDbi, a customized annotation R package was built using the Entrez gene IDs [56]. The GOstats package [57] was used to analyze enrichment in GO terms and KEGG pathways (http://www.geneontology.org/; http://www.genome.jp/kegg/pathway.html). Within the GOstats package a conditional hypergeometric test algorithm (Benjamini-Hochberg procedure) was applied for correction of multiple testing. The conditional hypergeometric test determined whether a GO term/KEGG pathway was significant when there was evidence beyond that provided by its significant children. Only the enriched GO term/KEGG pathway with raw p-values<0.05, were used for further interpretation in this study.

Supporting Information

Figure S1 Summary statistics for all the SNPs identified in Large White.
Found at: doi:10.1371/journal.pone.0014782.s001 (0.03 MB PDF)

Figure S2 Fst values (A) and p-values frequency (B) presented by breed pair.
Found at: doi:10.1371/journal.pone.0014782.s002 (0.26 MB PDF)

Table S1 Descriptions of breeds sampled.
Found at: doi:10.1371/journal.pone.0014782.s003 (0.14 MB PDF)

Table S2 Summary of genes related to neuron function and that overlap with genomic regions with significant low θ$_W$. Summary of

genes related to growth, muscle development, metabolism and disease that overlap with genomic regions with significant low θ_W. Found at: doi:10.1371/journal.pone.0014782.s004 (0.08 MB PDF)

Table S3 Summary of genes related to growth, muscle development, metabolism and disease that overlap with genomic regions with significant low θ_W.
Found at: doi:10.1371/journal.pone.0014782.s005 (0.16 MB PDF)

Text S1 Detailed description of measures of polymorphism and genetic differentiation.
Found at: doi:10.1371/journal.pone.0014782.s006 (0.11 MB PDF)

References

1. Price EO (1999) Behavioral development in animals undergoing domestication. Applied Animal Behaviour Science. Available: http://www.sciencedirect.com/science/article/B6T48-3Y8VGSS-4/2/7be985b99df0144a1cf765e03b32e642. Accessed 2009 July 23 65: 245–271.
2. Price EO, King J (1969) Domestication and adaptation. In: Adaptation of domestic animals Hafez, ed. Philadelphia: Lea and Febiger. pp 34–45.
3. Van Laere AS, Nguyen M, Braunschweig M, Nezer C, Collette C, et al. (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature, Available: http://dx.doi.org/10.1038/nature02064. Accessed 2009 August 4 425: 832–836.
4. Marklund S, Kijas J, Rodriguez-Martinez H, Rönnstrand L, Funa K, et al. (1996) Pigs with the dominant white coat color phenotype carry a duplication of the KIT gene encoding the mast/stem cell growth factor receptor. Mammalian Genome, Available: http://dx.doi.org/10.1007/s003359900244. Accessed 2009 June 14 7: 822–830.
5. Fang M, Larson G, Ribeiro HS, Li N, Andersson L (2009) Contrasting Mode of Evolution at a Coat Color Locus in Wild and Domestic Pigs. Plos Genetics, Available: http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000341. Accessed 2009 June 4 5: e1000341.
6. Amaral AJ, Megens H, Crooijmans RPMA, Heuven HCM, Groenen MAM (2008) Linkage Disequilibrium Decay and Haplotype Block Structure in the Pig. Genetics, Available: http://www.genetics.org/cgi/content/abstract/179/1/569. Accessed 2009 June 5 179: 569–579.
7. Farnir F, Coppieters W, Arranz JJ, Berzi P, Cambisano N, et al. (2000) Extensive Genome-wide Linkage Disequilibrium in Cattle. Genome Research, Available: http://genome.cshlp.org/content/10/2/220.full.pdf+html?sid=3ca80884-4e03-4e43-a416-5df1c1d4c693. Accessed 2009 October 9 10: 220–227.
8. McRae A, McEwan J, Wilson T, Crawford A, Slate J (2002) Linkage disequilibrium in domestic sheep. Genetics, Available: http://www.genetics.org/cgi/content/full/160/3/1113?maxtoshow=&hits=10&RESULTFORMAT=&fulltext=mcrae+sheep&andorexactfulltext=and&searchid=1&FIRSTINDEX=0&sortspec=relevance&resourcetype=HWCIT. Accessed 2009 October 9 160: 1113–1122.
9. Muir WM, Wong GK, Zhang Y, Wang J, Groenen MA, et al. (2008) Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. Proceedings of the National Academy of Sciences, Available: http://www.pnas.org/content/105/45/17312.full?sid=b1331a04-22a3-4491-a9d0-e3e7b656fd2a. Accessed 2009 October 9 105: 17312–17317.
10. Larson G, Albarella U, Dobney K, Rowley-Conwy P, Schibler J, et al. (2007) Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. Proceedings of the National Academy of Sciences, Available: http://www.pnas.org/content/104/39/15276.full. Accessed 2009 July 16 104: 15276–15281.
11. Porter V, Tebbit J (1993) Pigs. A handbook of the breeds of the world. East Sussex: Helm Information Limited. 256 p.
12. Hemmer H (1990) Domestication. Cambridge University Press. 208 p.
13. Morey D (1994) The early evolution of the domestic dog. American Scientist 82: 336–347.
14. Darwin C (1868) The variation of animals and plants under domestication. John Murray, London, 2vols.
15. Driscoll CA, Macdonald DW, O'Brien SJ (2009) From wild animals to domestic pets, an evolutionary view of domestication. Proceedings of the National Academy of Sciences, Available: http://www.pnas.org/content/106/suppl.1/9971.full?sid=258430e1-933f-4aae-90d9-feab70ca49a0. Accessed 2009 July 23 106(Suppl 1): 9971–9978.
16. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Schaffner SF, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature, Available: http://dx.doi.org/10.1038/nature01140. Accessed 2009 August 7 419: 832–837.
17. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, et al. (2004) Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. The American Journal of Human Genetics, Available: http://www.sciencedirect.com/science/article/B8JDD-4RDBM6R-5/2/f7e69625f6578-ba4943078be001f52c6. Accessed 2009 August 7 74: 1111–1120.
18. Evans PD, Gilbert SL, Mekel-Bobrov N, Vallender EJ, Anderson JR, et al. (2005) Microcephalin, a Gene Regulating Brain Size, Continues to Evolve Adaptively in Humans. Science, Available: http://www.sciencemag.org/content/309/5741/1717.full. Accessed 2009 August 7 309: 1717–1720.
19. Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G (2006) Signatures of Demographic History and Natural Selection in the Human Major Histocompatibility Complex Loci. Genetics, Available: http://www.genetics.org/cgi/content/full/173/4/2121. Accessed 2009 June 18 173: 2121–2142.
20. Alonso S, Lopez S, Izagirre N, de la Rua C (2008) Overdominance in the Human Genome and Olfactory Receptor Activity. Molecular Biology and Evolution, Available: http://mbe.oxfordjournals.org/content/25/5/997.full. Accessed 2009 June 18 25: 997–1001.
21. Clark AG, Hubisz MJ, Bustamante C, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Research, Available: http://genome.cshlp.org/content/15/11/1496.full?sid=ab7c53f0-4992-4e0e-abac-cf336f808c44. Acessed 2009 June 18 15: 1496–1502.
22. Bonetta L (2006) Genome sequencing in the fast lane. Nature Methods, Available: http://dx.doi.org/10.1038/nmeth0206-141. Accessed 2009 June 4 3: 141–147.
23. Schuster SC (2008) Next-generation sequencing transforms today's biology. Nature Methods, Available: http://dx.doi.org/10.1038/nmeth1156. Accessed 2009 June 4 5: 16–18.
24. Ramos AM, Crooijmans RP, Affara NA, Amaral AJ, Archibald AL, et al. (2009) Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. PLoS ONE, Available: http://dx.doi.org/10.1371/journal.pone.0006524. Accessed 2009 August 2009. 4: e6524.
25. Volis S (2008) In: Population genetics Research Progree. N.Y., USA: Nova Science Publishers, Inc. 279–310.
26. Nei M (1973) Analysis of Gene Diversity in Subdivided Populations. Proceedings of the National Academy of Sciences, Available: http://www.pnas.org/content/70/12/3321.full.pdf+html. Accessed 2009 July 16 70: 3321–3323.
27. Wakeley J (2008) Coalescent Theory. Roberts & Co, Greenwood Village, USA. 326 p.
28. Hellman I, Mang Y, Gu Z, Li P, de la Vega FM, et al. (2008) Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome Research, Available: http://genome.cshlp.org/content/18/7/1020.full. Accessed 2009 March 30 18: 1020–1029.
29. Ollivier L (2009) European Pig Genetic Diversity: A Minireview. Animal, Available: http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=5749604&fulltextType=RA&fileId=S1751731109004297. Accessed 2009 June 4 3: 915–924.
30. Laval G, Iannuccelli N, Legault C, Milan D, Groenen MA, et al. (2000) Genetic diversity of eleven European pig breeds. Genetics Selection Evolution, Available: http://www.gse-journal.org/index.php?option=com_article&access=standard&Itemid=129&url=/articles/gse/abs/2000/02/g0205/g0205.html. Accessed 2009 June 4 32(2): 187–203.
31. Lu J, Chung-I W (2005) Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee.. Proceedings of the National Academy of Sciences, Available: http://www.pnas.org/content/102/11/4063.full?sid=50e94aa9-266a-43a3-acce-7a2eeb3f8c54. Accessed 2009 August 7 102: 4063–4067.
32. Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, et al. (2004) Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. Genome Research, Available: http://genome.cshlp.org/content/14/4/528. Accessed 2009 October 19 14: 528–538.
33. The International SNP Map working group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.

Nature, Available: http://dx.doi.org/10.1038/35057149. Accessed 2009 October 19 409: 928–933.

34. Amaral AJ, Megens H-J, Kerstens HH, Heuven HC, Dibbits B, et al. (2009) Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. B.M.C. Genomics, Available: http://www.biomedcentral.com/1471-2164/10/374. Accessed 2009 October 19 10: 374.

35. Megens HJ, Crooijmans RP, San Cristobal M, Hui X, Li N, et al. (2008) Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. Genetics Selection Evolution, Available: http://www.gse-journal.org/index.php?option=com_article&access=standard&Itemid=129&url=/articles/gse/abs/2008/01/g06071/g06071.html. Accessed 2009 October 19 40: 103–128.

36. Paszek AA, Flickinger GH, Fontanesi L, Beattie CW, Rohrer G, et al. (1998) Evaluating evolutionary divergence with microsatellites. Journal of Molecular Evolution, Available: http://www.springerlink.com/content/hbfy5k3ptwumphy2/. Accessed 2009 October 19 46: 121–126.

37. Giuffra E, Kijas JM, Amarger V, Carlborg O, Jeon JT, et al. (2000) The Origin of the Domestic Pig: Independent Domestication and Subsequent Introgression. Genetics, Available: http://www.genetics.org/cgi/content/full/154/4/1785. Acessed 2009 July 25 154: 1785–1791.

38. Rubin CJ, Zody MC, Ericksson J, Meadows JR, Sherwood E, et al. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. Nature, Available: http://dx.doi.org/10.1038/nature08832. Accessed 2010 June 14 464: 587–591.

39. Meyer-Lindenberg A, Straub RE, Lipska BK, Verchinski BA, Goldberg T, et al. (2007) Genetic evidence implicating DARPP-32 in human frontostriatal structure, function, and cognition. Journal of Clinic Investigation, Available: http://www.jci.org/articles/view/30413. Accessed 2009 August 4 117: 672–682.

40. Frank H (1980) Evolution of canine information procession under conditions of natural and artificial selection. Zeitschrift für Tierpsychologie 5: 389–399.

41. Angles J, Kennedy LJ, Pedersen NC (2005) Frequency and distribution of alleles of canine MHC-II DLA-DQB1, DLA-DQA1 and DLA-DRB1 in 25 representative American Kennel Club breeds. Tissue Antigens, Available: http://dx.doi.org/10.1111/j.1399-0039.2005.00461.x. Accessed 2009 August 7 66: 173–184.

42. Birch J, Murphy L, MacHugh N, Ellis S (2006) Generation and maintenance of diversity in the cattle MHC class I region. Immunogenetics, Available: http://dx.doi.org/10.1007/s00251-006-0137-y. Accessed 2009 August 4 58: 670–679.

43. Miltiadou D, Ballingall K, Ellis S, Russell G, McKeever D (2005) Haplotype characterization of transcribed ovine major histocompatibility complex (MHC) class I genes. Immunogenetics, Available: http://dx.doi.org/10.1007/s00251-005-0008-y. Accessed 2009 August 2009. 57: 499–509.

44. Roos C, Walter L (2005) Considerable haplotypic diversity in the RT1-CE class I gene region of the rat major histocompatibility complex. Immunogenetics, Available: http://dx.doi.org/10.1007/s00251-004-0744-4. Accessed 2009 August 7 56: 773–777.

45. Otting N, Heijmans CM, Noort RC, de Groot NG, Doxiadis NG, et al. (2005) Unparalleled complexity of the MHC class I region in rhesus macaques. Proceedings of the National Academy of Sciences, Available: http://www.pnas.org/content/102/5/1626.full?sid=70a1cbea-98d0-49ee-beac-ca32857b9119. Accessed 2009 August 7 102: 1626–1631.

46. Gilad Y, Man O, Glusman G (2005) A comparison of the human and chimpanzee olfactory receptor gene repertoires. Genome Research, Available: http://genome.cshlp.org/content/15/2/224.full?sid=422309c1-6ae2-4fa0-9f32-a857de5a5708. Accessed 2009 August 7 15: 224–230.

47. Niimura Y, Nei M (2007) Extensive Gains and Losses of Olfactory Receptor Genes in Mammalian Evolution. PLoS ONE, Available: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0000708. Accessed 2009 August 7 2: e708.

48. Lancet D (1994) Olfaction. Exclusive receptors. Nature, Available: http://www.nature.com/nature/journal/v372/n6504/abs/372321a0.html. Accessed 2009 August 7 372: 321–322.

49. Li H, Ruan J, Richard D (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research, Available: http://genome.cshlp.org/content/18/11/1851.full?sid=df1e9c99-6002-488f-895b-b7af3ff8c259. Accessed 2009 August 7 18: 1851–1858.

50. Lynch M (2008) Estimation of nucleotide diversity, disequilibrum coefficients, and mutation rates from high-coverage genome sequencing projects. Molecular Biology and Evolution, Available: http://mbe.oxfordjournals.org/content/25/11/2409.full. Accessed 2010 October 25 25: 2409–2419.

51. Jiang R, Tavare S, Marjoram P (2009) Population Genetic Inference From Resequencing Data. Genetics, Available: http://www.genetics.org/cgi/content/full/181/1/187. Accessed 2010 October 25 181: 187–197.

52. Achaz G (2008) Testing for Neutrality in Samples with Sequencing Errors. Genetics, Available: http://www.genetics.org/cgi/content/full/179/3/1409. Accessed 2009 March 30 179: 1409–1424.

53. Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. Genome Research, Available: http://genome.cshlp.org/content/19/1/136.full?sid=683ee324-81b8-4740-8b85-1f8a3eb1c2fa. Accessed 2009 March 30 19: 136–142.

54. Ojeda A, Estellé J, Folch JM, Pérez-Enciso M (2008) Nucleotide variability and linkage disequilibrium patterns at the porcine FABP5 gene. Animal Genetics, Available: http://dx.doi.org/10.1111/j.1365-2052.2008.01752.x. Accessed 2009 June 4 39: 468–473.

55. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, et al. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics, Available: http://bioinformatics.oxfordjournals.org/content/21/16/3439.full. Accessed 2009 August 8 21: 34393440.

56. Pages H, Carlson M, Falcon S, Li N AnnotationDbi: Annotation Database Interface. R package version 1.6.0, Available: http://bioinformatics.oxfordjournals.org/content/23/2/257.full. Accessed 2009 August 8.

57. Falcon S, Gentleman R (2007) Using GOstats to test gene lists for GO term association. Bioinformatics, Available: http://bioinformatics.oxfordjournals.org/content/23/2/257.full. Accessed 2009 August 8 23: 257–258.