Heredity (2011) 107, 256–264
© 2011 Macmillan Publishers Limited All rights reserved 0018-067X/11
www.nature.com/hdy

ORIGINAL ARTICLE

Partial short-read sequencing of a highly inbred Iberian pig and genomics inference thereof

A Esteve-Codina[1], R Kofler[2,3,8], H Himmelbauer[2], L Ferretti[1,4], AP Vivancos[2,9], MAM Groenen[5], JM Folch[1], MC Rodríguez[6] and M Pérez-Enciso[1,7]

[1]Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, Bellaterra, Spain; [2]Centre for Genomic Regulation (CRG), Universitat Pompeu Fabra, Barcelona, Spain; [3]Max Planck Institute for Molecular Genetics, Berlin, Germany; [4]Department of Animal Science, Centre for Research in Agrigenomics (CRAG), Bellaterra, Spain; [5]Animal Breeding and Genomics Centre, Wageningen University and Research Centre, Wageningen, The Netherlands; [6]Departamento de Mejora Genética Animal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, Spain and [7]Institut Català de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Despite dramatic reduction in sequencing costs with the advent of next generation sequencing technologies, obtaining a complete mammalian genome sequence at sufficient depth is still costly. An alternative is partial sequencing. Here, we have sequenced a reduced representation library of an Iberian sow from the *Guadyerbas* strain, a highly inbred strain that has been used in numerous QTL studies because of its extreme phenotypic characteristics. Using the Illumina Genome Analyzer II (San Diego, CA, USA), we resequenced ~1% of the genome with average 4 × depth, identifying 68 778 polymorphisms. Of these, 55 457 were putative fixed differences with respect to the assembly, based on the genome of a Duroc pig, and 13 321 were heterozygous positions within *Guadyerbas*. Despite being highly inbred, the estimate of heterozygosity within *Guadyerbas* was ~0.78 kb^{-1} in autosomes, after correcting for low depth. Nucleotide variability was consistently higher at the telomeric regions than on the rest of the chromosome, likely a result of increased recombination rates. Further, variability was 50% lower in the X-chromosome than in autosomes, which may be explained by a recent bottleneck or by selection. We divided the whole genome in 500 kb windows and we analyzed overrepresented gene ontology terms in regions of low and high variability. Multi organism process, pigmentation and cell killing were overrepresented in high variability regions and metabolic process ontology, within low variability regions. Further, a genome wide Hudson–Kreitman–Aguadé test was carried out per window; overall, variability was in agreement with neutral expectations.
Heredity (2011) **107,** 256–264; doi:10.1038/hdy.2011.13; published online 16 March 2011

Keywords: Iberian pig; next generation sequencing; nucleotide diversity; Pig

Introduction

By slashing the sequence costs with respect to Sanger sequencing, recent massive parallel sequencing technologies (NGS) have democratized genomics research. With an increasing portfolio of applications ranging from complete genome sequencing to transcriptome sequencing (RNAseq) or metagenomics, NGS has revolutionized biology.

Nevertheless, sequencing a complete mammalian genome at reasonable depth is still expensive. As an alternative, a genome may be sequenced partially. Ideally, a targeted partial resequencing, for example, exome resequencing, would be the preferred choice (Ng *et al*., 2009); yet, sequence capture is also very expensive and not 100% effective; their overall cost effectiveness is therefore questionable. A feasible alternative is partial shotgun sequencing. In this spirit, resequencing reduced representation libraries (RRL) is a proven cost effective strategy (Van Tassell *et al*., 2008). Initially, this approach was proposed to identify massively single nucleotide polymorphisms (SNPs) when applied to pool resequencing (Van Tassell *et al*., 2008). Several groups have already shown in livestock, including pigs, how several hundreds of thousands of SNPs can be identified using that approach (Ramos *et al*., 2009).

Nevertheless, sequencing pools has a number of disadvantages for inferring genetic parameters like nucleotide diversity—it is biased against singletons—or linkage disequilibrium, the haplotype is basically lost (Cutler and Jensen, 2010). Here, we decided to sequence a RRL of a single individual rather than a pool to gain more in depth knowledge on a very peculiar Iberian pig strain and to complement the extant RRL pools in porcine (Ramos *et al*., 2009). To facilitate comparison with current data, we used one of the protocols used previously in the pig (Ramos *et al*., 2009).

The sequenced pig was a sow from the Iberian strain *Guadyerbas*. This is an obese, black, hairless and early-maturing Iberian strain. It represents one of the most

Correspondence: Dr M Pérez-Enciso, Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain.
E-mail: miguel.perez@uab.es
[8]Current address: Institut für Populationsgenetik, VetMedUni Wien, Veterinärplatz 1, 1210 Vienna, Austria.
[9]Current address: Vall d'Hebron Institute of Oncology, Vall d'Hebron University Hospital, 08035 Barcelona, Spain.

Received 9 November 2010; revised 20 January 2011; accepted 27 January 2011; published online 16 March 2011

ancient surviving Iberian lines, with no evidence of introgression of Asian genes, that has remained isolated since 1945 in a closed herd, *El Dehesón del Encinar*, located in Toledo, central Spain (Toro *et al*., 2000). A relevant aspect is that the complete pedigree since the founding of the herd is known, including that of the individual sequenced. Furthermore it has been used in several QTL experiments, including F_2 crosses with Landrace (Pérez-Enciso *et al*., 2000) and Meishan (Noguera *et al*., 2009). Performance characteristics compared with a lean international breed, Landrace, have been also reported (Serra *et al*., 1998).

Here, we present the analysis of a single *Guadyerbas* sow RRL sequence dataset obtained with short read technology (Genome Analyzer II, Illumina). Despite the fact that only about 1% of the genome was sequenced, we present results that are relevant from the species perspective and that can have important implications for animal breeding.

Materials and methods

Material

The *Guadyerbas* herd was founded with four boars and 10 sows in 1945, and has been maintained with controlled pedigree and minimum co-ancestry mating practices to minimize increase in inbreeding (Odriozola, 1976). Despite this, and because of isolation and small number of breeding animals, average inbreeding coefficient F is very high for all surviving pigs. In the specific female sequenced, autosomal F was ∼0.39 and ∼0.46 for sex chromosome X. These inbreeding coefficients were obtained through a forward simulation program taking into account the whole pedigree since 1945. A comprehensive genealogical study of this herd has been presented elsewhere (Toro *et al*., 2000).

RRL preparation and sequencing

To generate the sequencing library, we used 3.4 μg of genomic pig dsDNA, quantified with PicoGreen, and digested with 10 U of the blunt cutting restriction endonuclease *Hae*III. The DNA was processed with the Illumina genomic sample preparation kit. Briefly, blunt-ended fragments were A-tailed using the Klenow exo enzyme provided in the Illumina kit, followed by ligation of double-stranded adapters. The adapters were generated by annealing of oligonucleotides A 5′-AATGA TACGGCGACCACCGAGATCTACACTCTTTCCCTACA CGACGCTCTTCCGATC*T-3′ where * denotes a phosphorothioate bond and oligonucleotide B 5′-P-GATCGG AAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC GTATGCCGTCTTCTGCTTG-3′ (Sigma, St Louis, MO, USA). A 5 × adapter mix in water with a final concentration of 20 μM of each oligonucleotide was prepared in a thermocycler by heating to 65 °C for 5 min and cooling to 20 °C with a ramp of 0.1 °C s^{-1}. According to the Illumina protocol, adapter ligation is followed by size selection of the ligation products and a PCR step, which results in library enrichment and at the same time introduces sequences required for the *in situ* bridge PCR amplification in the Illumina flow cell. We modified the procedure such that we used adapters that already included the sequences necessary for amplification in the flow cell, as well as for sequencing primer binding, and skipped the enrichment PCR step. Such a strategy is advantageous, because errors introduced in the enrichment PCR step can confound SNP identification, in cases where molecules with the same PCR error are sequenced multiple times. Also, omission of enrichment PCR minimizes coverage biases that result from GC content imbalances of the sequenced target (Dohm *et al*., 2008; Kozarewa *et al*., 2009). We carried out the adapter ligation as described in the Illumina genomic sample preparation kit protocol, that is, in a volume of 50 μl with 10 000 U of T4 DNA ligase (New England Biolabs, Ipswich, MA, USA) in 1 × quick ligation buffer (66 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM dithiothreitol, 1 mM ATP, 7.5% polyethylene glycol, pH 7.6) at 25 °C for 15 min. Thereafter, we purified the sample with a QIAquick column (Qiagen, Hilden, Germany), eluted in 30 μl of 1 × TE and performed size selection on a 6% polyacrylamide gel. The gel area corresponding to the final size of the library including adapters (300–325 bp; library insert size of 200 ± 10 bp) was excised. The DNA was eluted by crushing the gel slice and incubation in 1 × elution buffer (500 mM ammonium acetate, 0.1% SDS, 0.1 mM EDTA) for 2 h at room temperature with gentle agitation. We separated the crushed polyacrylamide from the eluted DNA by using a cellulose acetate column (SpinX, Sigma, St Louis, MO, USA) and then precipitated the DNA by addition of 0.1 volumes of 3M sodium acetate pH 5.2 and 2.5 volumes of ice-cold absolute ethanol and spinning at 13 200 r.p.m. for 20 min. After washing with 70% ethanol and drying in a SpeedVac centrifuge for 5 min, we resuspended the DNA pellet in 15 μl 1 × TE. The concentration of the library was determined by TaqMan PCR (Quail *et al*., 2008).

We loaded the library into three Illumina flow cell lanes at a concentration of 5 pM (one lane) and 8 pM (two lanes), and sequencing on the Illumina Genome Analyzer II was carried out with 50 and 40 cycle recipes, respectively. The image data were processed using the Illumina pipeline 1.3.2. From the three runs, a total of 25.3 Mb called reads were obtained. Sequences have been deposited in short read archive (SRA accession SRP005367).

Bioinformatic analysis

Reads were trimmed to 40 bp because of low 3′-end quality. We discarded reads containing *Ns*, homopolymers longer than 17 nucleotides, an average minimum phred quality smaller than 20 and reads that did not start with a CC motif (*Hae*III cuts at 'GGCC' motif). Reads were filtered using custom Perl scripts. We aligned the remaining sequences against the reference porcine genome assembly 9 (ftp://ftp.sanger.ac.uk/pub/S_scrofa/assemblies/Ensembl_Sscrofa9/) with GEM (http://sourceforge.net/apps/mediawiki/gemlibrary/index.php?title = The_GEM_library), MAQ (Li *et al*., 2008) and Mosaik (http://bioinformatics.bc.edu/marthlab/Mosaik) retaining for variant calling only those reads that mapped unambiguously. We identified SNPs with GEM, MAQ and GigaBayes (Quinlan *et al*., 2008). Data were visualized with Eagleview (Huang and Marth, 2008).

When mapping the filtered reads with GEM, we used default options except for the mismatches allowed in each read to the reference genome (four mismatches were allowed). In the MAQ assembly, we also allowed a maximum of four mismatches for a read to be used in consensus calling and the minimum mapping quality was set to 10. When filtering the SNPs, the minimum consensus quality and adjacent consensus quality was 10. In all softwares, the minimum depth to call a SNP was 3 × and the maximum, 20 × . In MosaikAligner, we used a hash size of 20, with four mismatches allowed, the alignment candidate threshold was 20, the maximum number of hash positions to be used per seed was 100, the alignment mode was set to unique and the alignment algorithm was 'all'. The minimum posterior probability threshold for reporting a polymorphism candidate was set to 0.9 in Gigabayes. We classified the SNPs into two classes, fixed (*F*) when the differences were between the assembly and the Iberian reads, and segregating (*S*) when the Iberian pig was heterozygous. For a heterozygous SNPs to be called, the minimum non-reference allele count should be $>20\%$ with a minimum count of two.

Statistical and genetics analysis

As emphasized by several authors (Hellmann *et al*., 2008; Lynch, 2008; Jiang *et al*., 2009), estimating nucleotide diversity from NGS data requires specific methods to account for unequal depth along the genome and sequencing and assembly errors. Here, we are interested in estimating the heterozygosity h for each window. For multiple individuals, two different estimators have been proposed by Hellmann *et al*. (2008) and by Jiang *et al*. (2009). However, in the case of a single individual, both estimators coincide with the estimator of Lynch (2008) and correspond to the maximum composite likelihood estimator for h. If the mating is random and the population is in Hardy–Weinberg equilibrium, this is also a maximum composite likelihood estimator for the variability θ of the population. In the absence of sequencing and mapping errors, the formula for the unbiased maximum composite likelihood estimator for h is

$$\hat{h}^{*} = \frac{S}{\sum\limits_{nr=1}^{\infty} L(nr)P^{*}(S|nr)} \tag{1}$$

where S is the number of heterozygous sites detected in the window, $L(nr)$ is the number of bases with depth nr in the window and $P^{*}(S\,|\,nr)$ is the probability that a heterozygous site is detected when the read depth at that site is nr. The analytical expression is $P^{*}(S\,|\,nr) = 1-2^{-(nr-1)}$ (Hellmann *et al*., 2008; Lynch, 2008; Jiang *et al*., 2009). In case of sequencing errors, if the error rate or the SNP qualities are known and the error rate is not too large, the estimator can be corrected simply by subtracting the average number of false SNPs from *S*. Although sequencing errors can in principle be estimated from the data at hand (Lynch, 2008), this could induce some extra noise in the estimator and, more importantly, it is difficult to allow for errors in the assembly, a potentially much larger distortion factor than sequencing errors.

Here, we decided to follow a compromise to minimize assembly errors, but not being too strict in order not to discard many potentially true SNPs: we considered only the SNPs that had been called by at least two softwares, MAQ, Gigabayes or Gem, and only with depth between 3 and 20. A similar approach has been recently followed in the 1000 genomes project, where the SNPs called were a consensus between different algorithms (Durbin *et al*., 2010). In addition, we requested that the non-reference allele is present in at least two reads and a minimum allele count $\geqslant 20\%$ among all reads covering that position. Therefore, we applied equation (1) using those SNPs called by two or the three softwares and summing between $nr = 3$ and 20. Therefore, equation (1) needs to be modified

$$\hat{h} = \frac{S}{\sum\limits_{nr=3}^{20} L(nr)P(S|nr)}, \tag{3}$$

where

$$P(S|nr) = 1 - 2^{-nr}\left[\sum_{k=0}^{na-1}\binom{nr}{k} + \sum_{k=nr-nb+1}^{nr}\binom{nr}{k}\right],$$

with $na = \max(2, 0.2 \times nr)$ being the minimum number of non-reference allele reads requested and nb, the minimum number of reference allele reads. The above formulae stems from the restriction we set, for instance, for $nr = 3$, the only way a true SNP is called is the probability that exactly two reads belong to the alternative allele and one, to the reference allele, that is, a binomial with $P = 0.5$, $n = 3$ and two successes or $\binom{3}{2}2^{-3} = 0.375$. Note as well that Lynch's and similar corrections do differ from (3) when nr is small, $P^{*}(S\,|\,nr=3) = 0.75$ vs $P(S\,|\,nr=3) = 0.375$, whereas $P^{*}(S\,|\,nr=10) = 0.998$ vs $P(S\,|\,nr=10) = 0.988$.

As is clear from equation (3), the raw number of true heterozygous sites is underestimated from simply counting *S*. The contrary occurs with the number of fixed differences (*F*) because a fixed difference can actually be a segregating SNP, and because in the assembly no heterozygous positions are allowed: only one of the two alleles is reported. Here, we estimated

$$\hat{S} = \hat{h}\sum_{nr=4}^{20} L(nr), \tag{4}$$

and,

$$\hat{F} = \max\left[0, F - \sum_{nr=4}^{20} \hat{h}2^{-nr}L(nr)\right] \tag{5}$$

In (4) the estimate is negative when no fixed difference has been observed, in those cases the estimator was truncated to 0. We computed the average number of SNPs, $\hat{F}$ and $\hat{S}$, along non-overlapping contiguous 500 kb windows.

We also obtained Hudson–Kreitman–Aguadé (HKA) diversity (θ_{HKA}) estimates (Hudson *et al*., 1987). Briefly, HKA method tests whether there is a deviation between observed and expected number of polymorphisms, where the expected polymorphism is obtained from the divergence between an outgroup and the population studied. The HKA statistic for locus (that is, window) i is:

$$H_i = \frac{[\hat{S}_i - E(\hat{S}_i)]^2}{\mathrm{Var}(\hat{S}_i)} + \frac{[\hat{F}_i - E(\hat{F}_i)]^2}{\mathrm{Var}(\hat{F}_i)} \tag{6}$$

and the multilocus HKA test is $\chi^2 = \sum_i H_i$, with degrees of freedom equal to the number of loci and where the sum is across the i-th loci (here, the windows of 0.5 Mb length). We applied the test separately for autosomes and chromosome X. The expected values are

$$E(\hat{S}_i) = \hat{\theta}_i = \frac{\hat{S}_i + \hat{F}_i}{T+2},$$

and

$$E(\hat{F}_i) = \hat{\theta}_i(T+1),$$

with T, the divergence time, given by

$$T = \frac{\sum_i \hat{F}_i}{\sum_i \hat{S}_i} - 1,$$

with approximate variances $\mathrm{Var}(\bullet) = E(\bullet) \times [1 + E(\bullet)]$. The HKA procedure is primarily devised to compare two species, whereas here we considered the reference assembly (a Duroc pig) as outgroup. Therefore, the power of HKA should be relatively low, but can provide a rule of thumb as to which are the most extreme windows in terms of variability.

We also developed a Monte Carlo procedure to infer genetic parameters in a more general framework. Given that a single individual was sequenced, we do not intend to provide accurate inferences, but rather to show, as a proof of principle, how genome wide data of the kind obtained here can be used to make inferences on demographic history. Suppose the simplest possible model to characterize the Iberian—Duroc breed history, that is, an ancestral population of size N that τ generations ago split into the two breeds, which may have occasionally interchanged individuals from Iberian into Duroc at a rate m (Figure 1). The procedure consisted of simulating the number of fixed and segregating SNPs according to this model and choosing the set of parameters that produced the best fit with the observed data. For given values of N_{IB} (Iberian Ne), N_{DU} (Duroc Ne), m, and τ, we simulated 500 kb windows by coalescence using MaCS (Chen *et al*., 2009) of one Duroc individual and 14 Iberian animals, 30 sequences in total. Next, as the complete pedigree from the 14 founder individuals of the herd is known, we simulated by gene dropping the genome window of the Iberian pig sequenced, according to the known pedigree. Finally, we extracted the same number of fragments and number of base pairs as actually sequenced from the simulated window. We counted the number of fixed and segregating SNPs per window, and we repeated the process for each of the 4363 windows obtained in the real data. For the Duroc assembly, we randomly sampled an allele in the simulated Duroc sequences. Finally, we obtained the observed and simulated HKA-θ estimator described above ($\hat{\theta}_i$) for each i-th window; as measure of goodness of fit, we used the Wilcoxon's ranked signed test across windows. We did a grid search using this procedure for different values of N_{IB}, N_{DU}, m and τ; assuming a true $\theta = 0.0013$ for the autosomes and $\theta = 0.0005$ for the X chromosome, and ρ, scaled recombination rate, 0.001. These values are taken from the literature (Ojeda *et al*., 2006; Amaral *et al*., 2009). The whole procedure was implemented in a Perl script with calls to MaCS and R.

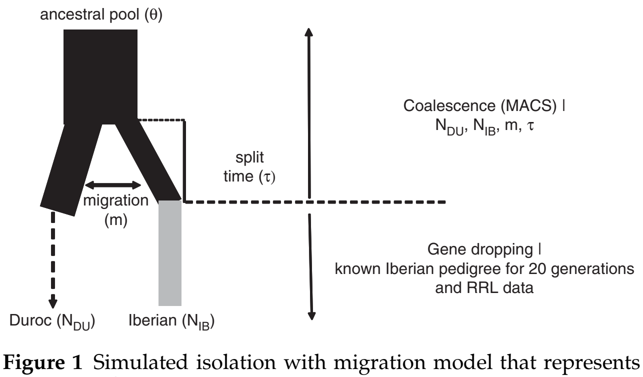
ancestral pool (θ)
Coalescence (MACS) | N_{DU}, N_{IB}, m, τ
split time (τ)
migration (m)
Gene dropping | known Iberian pedigree for 20 generations and RRL data
Duroc (N_{DU})
Iberian (N_{IB})

Figure 1 Simulated isolation with migration model that represents the Iberian/Duroc history (the public assembly pertains to a Duroc sow). The Duroc and Iberian populations descend from an ancestral population harboring a nucleotide diversity $\theta = 4Ne\mu$; after the split τ generations ago, both breeds of effective sizes N_{DU} and N_{IB} may have interchanged individuals with rate m. A mixed coalescence and gene dropping procedure was used.

Gene ontologies (GO)

We ranked the 500 kb windows according to estimated heterozygosity and we selected the most extreme windows to test whether genes within the windows were enriched in particular ontologies. GO were downloaded using Biomart (http://www.biomart.org). Our Goslim (http://www.geneontology.org/GO.slims.shtml) was composed of 23 parental pig GO extracted from http://amigo.geneontology.org/cgi-bin/amigo/go.cgi. After filtering for biological process, we selected the following GO: biological regulation, cellular process, metabolic process, multicellular organismal process, developmental process, signaling, localization, response to stimulus, immune system process, cellular component organization, reproduction, biological adhesion, cellular component biogenesis, death, locomotion, multi-organism process, growth, pigmentation, rhythmic process, viral reproduction and cell killing. GO statistics were calculated using the GOquick browser (http://www.ebi.ac.uk/QuickGO/). Expected and observed GO percentages were contrasted with a Fisher's exact test as implemented in R. To test enrichment of specific ontologies, we simply computed a two sided t-test assuming a normal distribution for number of counts.

Results

Alignment and polymorphism detection

Out of three Genome Analyzer II lanes, we obtained ~25.3 million reads. After filtering and removing ambiguous matching reads, that is, reads matching the reference more than once, we retained 5 million reads for further analysis (Figure 2). The total length assembled was ~2.3 Gb. The reads spanned 83.1 Mb of the porcine assembly v.9 with at least one read, and 25.1 Mb with at least three reads and a maximum depth of 20. The average depth, counting only regions with read depth between 3 and 20 was 4×. All chromosomes were uniformly covered and we did not notice biases regarding read distribution within chromosomes (Supplementary File S1). Only four out of the 4363 windows were not covered by any read. The RRL was also unbiased with respect to depth of coding vs non-coding regions, 4.08× and 4.07×, respectively. Table 1 shows relevant statistics per chromosome.

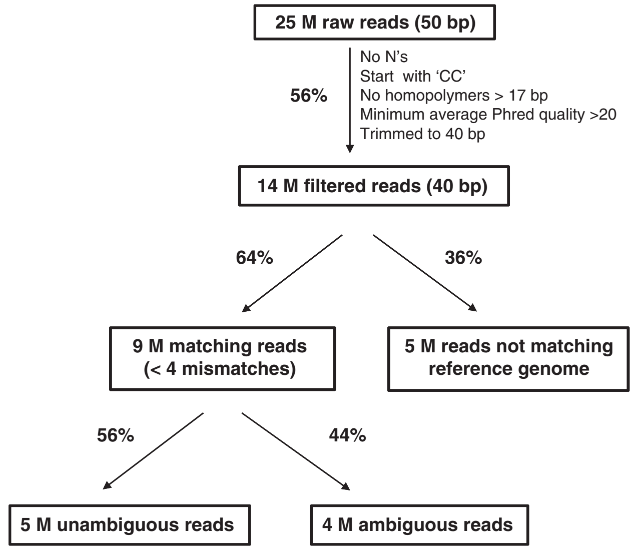

Figure 2 Bioinformatics pipeline.

Table 1 Statistics per chromosome

Chrom.	*Total assembled* $\geqslant 3\times$ *(Mb)*	*Average coverage (3–20×)*	S[a]	F[b]	$\hat{h}$[c]	$\hat{f}$[d]
SSC1	2.45	3.97	842	5023	0.48	1.51
SSC2	1.95	4.00	1334	4363	1.11	1.88
SSC3	1.90	4.01	1517	3519	1.15	1.48
SSC4	1.34	3.98	971	3027	0.95	1.73
SSC5	1.05	3.96	639	2685	1.09	2.11
SSC6	2.21	4.01	895	4890	0.58	1.85
SSC7	1.68	3.97	471	4509	0.60	2.49
SSC8	0.91	3.96	485	2197	0.73	2.15
SSC9	1.33	3.96	823	3142	0.77	1.85
SSC10	0.69	3.97	438	1822	1.08	2.42
SSC11	0.66	3.95	406	1770	0.95	2.44
SSC12	1.16	3.99	782	2520	1.06	1.67
SSC13	1.40	3.96	614	2618	0.63	1.58
SSC14	2.06	3.98	837	4512	0.57	2.10
SSC15	1.05	4.01	653	2607	0.67	1.77
SSC16	0.67	3.97	419	1636	0.81	2.27
SSC17	0.87	3.99	528	1960	0.99	2.38
SSC18	0.66	3.96	370	1396	1.10	1.57
SSCX	1.03	4.01	297	1261	0.37	0.92
Total autosomes	24.02	3.98	13 024	54 196	0.78	1.89

[a] Number of heterozygous sites, raw numbers.
[b] Number of fixed differences, raw numbers.
[c] Average estimated heterozygosity within Iberian per kb.
[d] Average estimated number of differences between Iberian and assembly per kb.

SNPs were called with three different programs. The number of variants called by each software differed: MAQ was the most conservative and GEM, the most liberal. The latter can be explained by the fact that it does not use sequence qualities to filter the alignments and the SNP calls. Overall, the discrepancy between the programs decreased with depth. The average depth of the SNPs detected with at least two programs was 4.5 × and of those detected with the three programs, 6.5 × . Using the SNPs called by at least two programs, a total of 68 778 SNPs were identified, equivalent to an average 2.7 SNPs per kb sequenced. Main variability statistics by window are in Supplementary File S2, together with a summary of variability within intergenic, intronic, CDS and UTR regions.

Variability distribution and population genetics inference

To gain further insight into the variability distribution, using equations 3 and 4, we plotted the Iberian average heterozygosity ($\hat{h}$) and average fixed differences between the assembly and Iberian $\hat{f} = \hat{F}/\sum_{nr=4}^{20} L(nr)$ in non-overlapping contiguous windows of 500 kb. Genome wide results are in Supplementary Figure S3, whereas Figure 3 shows the lowess adjusted curves results in chromosomes SSC4 and SSCX. A trend of increasing variability in $\hat{f}$ toward the telomeres is clearly visible in SSC4; this pattern also exists in $\hat{h}$, but is less apparent because the scale is too coarse. This can also be seen in the sex chromosome, although less markedly than in autosomes because of an overall lower level of variability. Note that this is not caused by differences in depth, which is fairly uniform along the chromosome (Supplementary Figure S1). The average nucleotide diversity θ_{HKA} was 1.7×10^{-3} in the 5% most extreme telomeric windows, much higher than the value found in the 10% of windows surrounding the centromere: 5.4×10^{-4}. These figures correspond to the average over all chromosomes, except acrocentric chromosomes, that is, SSC13–SSC18. Excluding SSC7, which harbors the highly polymorphic SLA region near the centromere, the statistics are 1.7×10^{-3} vs 4.9×10^{-4} for telomeric and centromeric regions, respectively.

The average SNP rates per base pair for chromosome X were $\hat{f} = 9.2 \times 10^{-4}$ and $\hat{h} = 3.7 \times 10^{-4}$. Interestingly, these values are ~50% lower than those of the autosomes 1.9×10^{-3} ($\hat{f}$) and 7.8×10^{-4} ($\hat{h}$), whereas the expected ratio is 75% under a stationary neutral model, because the effective population size of the X chromosome is $\frac{3}{4}$ that of the autosomes.

Next, we computed the HKA test to examine whether the observed pattern departs from what is expected under the stationary neutral model. The estimated divergence, when measured in twice effective size ($2N_e$) units, was ~1.3 both for autosomes and the X-chromosome. In contrast, the weighted nucleotide diversity θ_{HKA} was 8.0×10^{-4} and 3.8×10^{-4} in autosomes and in X-chromosome, respectively (Table 2). These values are in complete agreement with those from the simple heterozygosity estimates $\hat{h}$ (Table 1). Again, the HKA estimate also indicates a much lower variability at the X chromosome than expected, relative to the autosomes. The plot in Supplementary File S4 shows that, genome-wide, there were no wide departures from neutrality, neither for autosomes nor for X chromosome, according to this test.

We applied the model in Figure 1 to adjust demographic parameters in the Iberian lineage using the stochastic method described above. We estimated the set of parameters by minimizing the distance, in a signed rank test, between simulated and observed HKA statistics for each 500 kb window. We did that separately for autosomes and the sex chromosome. The analyses discarded a migration ($m = 0$) between breeds and suggested an effective size of Iberian ~20% that of the ancestral population, assuming an initial $\theta = 0.0013$ and $\theta = 0.0005$ for autosomes and sex chromosome,

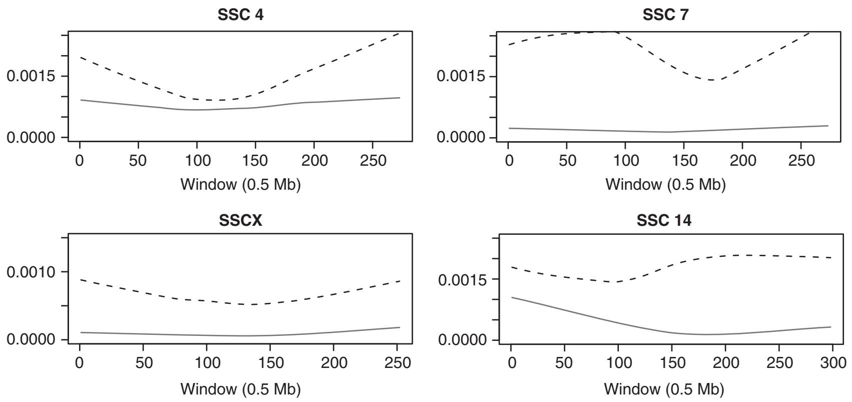

Figure 3 Lowess adjusted curves of variability in chromosomes 4, 7, 14 and X. An increased variability is observed towards the telomeres in metacentric chromosomes 4 and X, whereas the ratio is distorted in SSC7 because of high SLA variability near window 50; SSC14 is acrocentric. Solid red line, Iberian heterozygosity ($\hat{h}$); dashed black line, Iberian—Duroc heterozygosity ($\hat{f}$). Position refers to window number. A full color version of this figure is available at the *Heredity* journal online.

Table 2 HKA statistics

	Divergence (2N units)	θ_{HKA} *per kb*
Autosomes	1.32	0.80
SSCX	1.45	0.38

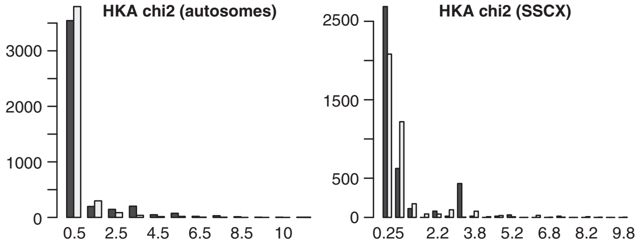

Figure 4 Histograms comparing observed (black bars) and simulated (grey bars) HKA statistics across autosomal and sex chromosome windows. The simulated results correspond to parameter values that minimized the Wilcoxon statistics.

respectively. As shown in Figure 4, the fitted parameters adjusted the observed values for the χ^2-statistics quite well for the autosomal windows, whereas fit was less good for SSCX windows.

Outlier regions and their annotation

As results in Supplementary File S4 suggest, the genome-wide pattern of nucleotide variability was approximately neutral, according to the HKA test. Certainly, not the whole genome evolves according to the standard neutral model and the apparent neutrality may simply mean lack of power or too large windows that may mask highly local selective events. To complement the analyses, we next focused on extreme windows for low or high heterozygosity ($\hat{h}$). A large number of windows (1820) turned out to be devoid of any heterozygous SNP. Therefore, we selected those 81 windows with at least 10.1 kb assembled and having at least one fixed difference; 19 and 17 of the windows were located in chromosomes 6 and 7, respectively (Supplementary File S2). The expected number of windows for those chromosomes, according to its size, is about five and the overrepresentation is highly significant ($P<10^{-7}$). In SSC6, in particular, 10 windows were almost contiguous, spanning windows 92–115, the average heterozygosity of this whole interval was 10^{-4} or six times lower than average genome wide. Further, although chromosomes 6 and 7 present lower than average heterozygosities, they are not outliers: chromosomes 1, 13 or 14 have comparable heterozygosities (Table 1). Also, in contrast to what would be predicted, only three windows were located in chromosome X (five are expected).

We also considered the most extreme windows in terms of heterozygosity. A problem with the interpretation of these windows is that a large variability can be distorted by possible misalignments. Although we minimized this risk by considering SNPs called by several aligners, we retained, from the 100 windows with maximum heterozygosity, those with over a kb assembled and whose $\hat{f}$ was below the median. Therefore, we ensured that, whereas $\hat{h}$ was extreme, $\hat{f}$ was not. We found 31 such windows (Supplementary File S2). In this case, no dramatic departures in the number of windows by chromosome were observed.

To gain further biological insight, we studied Gene Ontology enrichment of genes located in the windows with extreme values of nucleotide diversity. We looked for overrepresented GO of genes in these windows with respect to overall GO frequencies among all sequenced genes. The observed and expected results are in Figure 5. Among the high variability windows, we found that GO categories multi organism process ($P=10^{-5}$), pigmentation ($P<10^{-12}$) and cell killing ($P<10^{-13}$) were overrepresented. In general, genes related with defense (*RAB27A, NCF1*) and olfactory receptors were among the high variability windows, as could be expected. We only found the generic metabolic process ($P<10^{-10}$) and apoptosis ($P=0.05$) GO as overrepresented among genes located within low variability windows. In chromosome 6, several of the genes are involved in carbohydrate metabolism (*FUT1, FUT2, BAX, GYS1, CA11*), oxidoreductase activity (*DHDH, PGD, MTHFR*). Among those in SSC7, protein folding (*HSP90AB1, HSP90AA1, DNAJA4*), all in all, there was not a clear metabolic route overrepresented. The results simply suggest that these genes exhibit lower than expected variability, be it

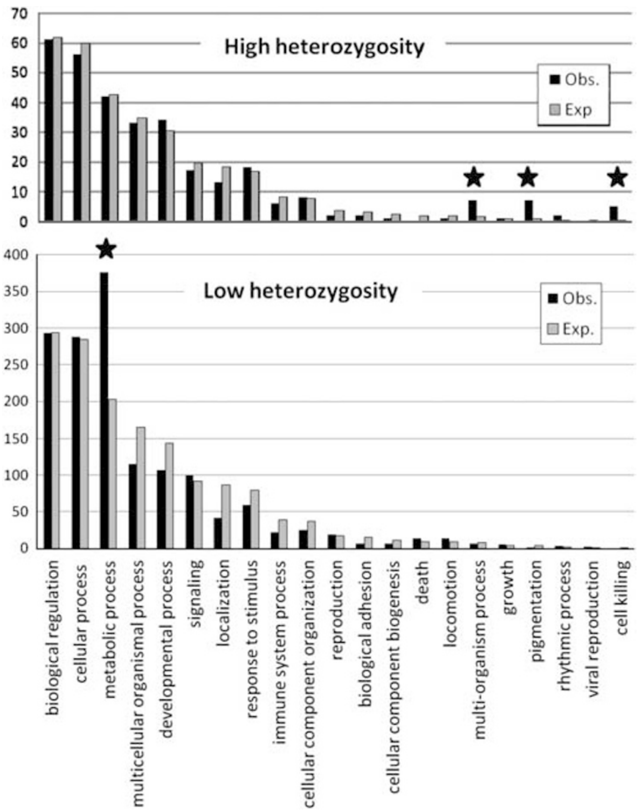

Figure 5 Expected and observed gene ontology counts among genes located in high and low variability windows. Bars with asterisk are significant ($P<0.001$) overrepresented gene ontologies.

because of specific selection in livestock or because of other biological constraints. More data is required to ascertain the precise cause.

Discussion

We have presented the first re-sequencing effort of the Iberian pig breed, the most emblematic pig breed in the Mediterranean area and one of the most important porcine local varieties in economic terms worldwide. The pig sequenced belongs to a peculiar Iberian strain with unique phenotypic characteristics that has been used in multiple QTL experiments (Pérez-Enciso *et al.*, 2000; Noguera *et al.*, 2009). For reasons stated in the introduction, we chose to use RRL in a single individual. Although the RRL is a cost effective alternative to targeted sequencing, it has drawbacks also. It is basically a shotgun approach where potential regions of interest may not be covered. The easiest way to improve RRL would be to digest in silico with different enzymes and compare different band lengths such that the coverage of targeted regions is maximized. In the case of porcine species, this strategy is risky because the sequence is incomplete and even assembly is still under development. Besides, (Amaral *et al.*, 2009) found that the correspondence between theoretical and observed sequences is not perfect, likely because band excision is not absolutely precise.

In this work, we have primarily focused on the distribution of nucleotide diversity. We found a global autosomal Iberian heterozygosity rate of $\hat{h}=0.78\times10^{-3}$ per nucleotide (Table 1). This value is much larger than the naïve estimator of simply dividing the number of SNPs by the length assembled, and illustrates the need of applying specific statistic tools with genome wide NGS data, especially at low depth (Lynch, 2008; Haubold *et al.*, 2010). Assuming a mutation rate (μ) of 10^{-8}, this results in an estimate of effective size $N_e=\hat{h}/4\mu\sim2\times10^4$. This value is quite high, especially considering that this is a highly inbred animal. It suggests that the actual effective size in the founder herd might be actually double, given that inbreeding coefficient of the sequenced animal is ~0.39. When correcting for inbreeding, this diversity is comparable with that reported in other porcine species (Amaral *et al.*, 2009, 2011) or in humans.

Both chromosomes 6 and 7 were enriched in windows of low variability (Supplementary File 2). The case of SSC6 is noticeable because a long stretch of ~12 Mb (windows 92–115) was almost devoid of any SNP within *Guadyerbas* $\hat{h}=1.4\times10^{-4}$, the average number of differences was, nonetheless, close to the genome wide mean ($\hat{f}=1.7\times10^{-3}$). Certainly, a reason for long stretches without polymorphisms is the high inbreeding of the sequenced animal. To test that, we ran a forward simulation algorithm using the true pedigree of the animal since the founder herd. Assuming an equivalence of 1 cM to 1 Mb, the expected size of an identical by descent fragment (IBD) is ~2.6 Mb (s.d., 3.2), the probability of having an IBD fragment is the inbreeding coefficient (0.39 for autosomes). The probability of a fragment of 12 Mb being IBD in the sequenced animal is 6×10^{-3} or 0.02 if recombination rate is lower, 1 cM to 1.5 Mb. Therefore, although the event is unlikely, it is not impossible when the whole genome is considered. But, given that this region is the lowest extreme in nucleotide variability, we can speculate that a selective sweep, if occurred, was previous to the herd founding. In a previous intercross between *Guadyerbas* and Landrace, we found that SSC6 harbors a large effect QTL for intramuscular and fat deposition (Ovilo *et al.*, 2000); however, the most likely candidate gene, the leptin receptor, is far away from windows 92–115: its predicted position is window 206.

Two interesting remarks can be made about the distribution of nucleotide variability: an increased variability in telomeric regions and lower than expected diversity on the X chromosome. Increased variability in telomeric regions is likely explained by larger recombination rates as compared with centromeres, where recombination is rare. A positive correlation between variability and recombination is a well known observation in many species (Hedrick, 2010). Traditionally, different hypotheses have been proposed to explain this observation: increased mutation rate, hitchhiking and background selection. The latter two seem to explain better experimental results overall (Hudson, 1994; Hedrick, 2010). Our data, in principle, would favor background selection because generalized hitch hiking events in all telomeric regions are unlikely, although recent work (Hellmann *et al.*, 2008) suggest that hitch hiking fit the data better in humans than a simplistic background selection model. These authors also report that an elevated mutation rate also accounts for increased variability in sub telomeric regions.

Reduced variability on SSCX merits some additional discussion. Theory dictates that expected nucleotide diversity of the X chromosome is $\frac{3}{4}$ times that in autosomes, but we find a much lower value $\pi_{SSCX}/\pi_{SSCA} \sim 50\%$ (Table 1). This observation is unlikely to be an artifact because we found identical ratio both for $\hat{h}$ and $\hat{f}$; further, an even lower ratio 36%, has been reported in the literature (Amaral *et al*., 2009). The relative levels of variability between autosomes and sex chromosomes has been debated for quite some time, but the recent availability of NGS has renewed the interest and promised to deliver new insights. All demographic, mutational and selective events can alter the theoretical $\frac{3}{4}$ ratio. In the literature, both higher and lower ratios have been observed, even within the same species (Ellegren, 2009). A decreased nucleotide diversity π_{SSCX}/π_{SSCA} can be produced by a larger number of reproducing females than males (Ellegren, 2009), but the opposite is rather the norm in livestock; therefore female polygamy is not an explanation. Alternative explanations are increased male than female dispersal (this can happen in livestock if we assume that males sire different herds than their mother's whereas females stay in the same herd), or strong bottlenecks (Pool and Nielsen, 2007). Finally, selection either background or directional, can also reduce sex to autosomal variability. It should be noted that the sow's inbreeding coefficient, inferred from the pedigree, is $\sim$0.46 in chromosome X and 0.39 for the autosomes. Therefore, the expected ratio of diversity π_{SSCX}/π_{SSCA} after the herd was found is approximately $(1-0.46)/(1-0.39) \sim 0.88$. This value is much higher than what is expected under a random mating scheme. The reason is that matings in this herd were carefully designed to minimize increase in inbreeding (Toro *et al*., 2000). But this figure also means that, if a bottleneck is to be responsible of the low variability in SSCX, it must have occurred before the herd foundation approximately mid twentieth century.

Logically, a final aim of all this flood of sequencing data in livestock species is to be able to uncover the causal mutations that underlie complex traits in domestic species. Here, genomewide, we found no strong departures of expectations under a neutral model neither with the HKA test (Supplementary File S4) nor with the demographic model described in Figure 1. This can be because of the length of window chosen (500 kb), which may be too large to identify selective events, but also to the fact that a single animal has been sequenced. Also, the HKA test is primarily designed for species divergence, whereas divergence between Duroc (the assembly) against Iberian breeds is examined here. Nevertheless, detection of more subtle signals may require complete genome resequencing and a larger number of animals, as illustrated recently by Andersson and coworkers (Rubin *et al*., 2010). Also importantly, the complex interaction between demographic events and moving selection targets cannot be forgotten when looking for selection footprints (Pool *et al*., 2010). Despite these drawbacks, we have characterized outlier regions and looked for gene ontology enrichment as a tool to gain biological insight. We find high heterozygosity within *Guadyerbas* for pigmentation and cell killing, particularly the cellular response to antigens. These genes could be candidates for balancing selection within the Iberian lineage, a topic that should be further explored when more data is available.

Conclusions

Although we have sequenced a single individual, our data yield some interesting conclusions regarding the genetic architecture of the pig and of the Iberian pig in particular. More specifically, we have observed that (i) the estimated heterozygosity is 0.78×10^{-3} per site, a non-negligible variability considering the inbreeding coefficient of the sow was $\sim$39%; (ii) variability tends to be higher in telomeric than in centromeric regions, plausibly a symptom of prevalent background selection due to increased recombination in those regions; (iii) the X chromosome is much less variable than expected relative to autosomal variability; although more work is required, this fact could be partly explained by a strong bottleneck; (iv) overall, variability is in agreement with expectations from the HKA test. Probably because of the sparse coverage and the fact that a single individual was sequenced, we did not observe clear signals of directional selection in QTL regions like the leptin receptor in SSC6.

For the future, the next logical step will be to sequence more animals, either in pools or individually. Fortunately, recent works have shown that sequencing at very high depth may not be necessary to infer genetic parameters with confidence (Sackton *et al*., 2009; Durbin *et al*., 2010). This will allow us to refine our model for the demographic history of the Iberian pig and to extend and confirm the catalog of genetic variants, including indels and other structural variants, for example, copy number variants. But, in addition to more experimental data, we shall also pursue the development of new statistical approaches that allows us to interpret the flood of data produced by the new sequencing technologies (Pool *et al*., 2010). The method proposed here (Figure 1) is, but a first attempt in this direction.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We are grateful for help with the analysis and comments from S. Ramos-Onsins and A. Amaral. Conservation of the *Guadyerbas* strain started with late Miguel Odriozola and has been ensured for years thanks to the efforts of L Silió, J Rodrigáñez and colleagues, and supported economically by the regional government of *Junta de Castilla-La Mancha* and *Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria* (INIA). AE is funded by the Spanish Ministry of Research PhD studentship program (FPI), RK was supported by GABI-FUTURE Grant BeetSeq (FKZ 0315069A) of the German Federal Ministry of Education and Research. Work funded by Grants AGL2007-65563-C02/GAN to MPE and MCR, AGL2010-14822 to MPE and by a Consolider Grant from Spanish Ministry of Research, CSD2007-00036 'Centre for Research in Agrigenomics'.

References

Amaral A, Ferretti L, Megens H-J, Crooijmans R, Nie H, Ramos-Onsins SE *et al*. (2011). Genome wide footprints of pig domestication revealed through massive parallel sequencing of pooled DNA. *Plos One* (in press).

Amaral A, Megens H-J, Kerstens H, Heuven H, Dibbits B, Crooijmans R *et al.* (2009). Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. *BMC Genomics* **10**: 374.

Cutler DJ, Jensen JD (2010). To pool, or not to pool? *Genetics* **186**: 41–43.

Chen GK, Marjoram P, Wall JD (2009). Fast and flexible simulation of DNA sequence data. *Genome Res* **19**: 136–142.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**: e105.

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

Ellegren H (2009). The different levels of genetic diversity in sex chromosomes and autosomes. *Trends Genet* **25**: 278–284.

Haubold B, Pfaffelhuber P, Lynch M (2010). mlRho—a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Mol Ecol* **19**(Suppl 1): 277–284.

Hedrick PW (2010). *Genetics of Populations*, 4th edn. Jones and Bartlett: Sudbury, MA.

Hellmann I, Mang Y, Gu Z, Li P, de la Vega FM, Clark AG *et al.* (2008). Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res* **18**: 1020–1029.

Huang W, Marth G (2008). EagleView: a genome assembly viewer for next-generation sequencing technologies. *Genome Res* **18**: 1538–1543.

Hudson RR (1994). How can the low levels of DNA sequence variation in regions of the drosophila genome with low recombination rates be explained? *Proc Natl Acad Sci USA* **91**: 6815–6818.

Hudson RR, Kreitman M, Aguade M (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153–159.

Jiang R, Tavare S, Marjoram P (2009). Population genetic inference from resequencing data. *Genetics* **181**: 187–197.

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Meth* **6**: 291–295.

Li H, Ruan J, Durbin R (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**: 1851–1858.

Lynch M (2008). Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol Biol Evol* **25**: 2409–2419.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C *et al.* (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**: 272–276.

Noguera JL, Rodriguez C, Varona L, Tomas A, Munoz G, Ramirez O *et al.* (2009). A bi-dimensional genome scan for prolificacy traits in pigs shows the existence of multiple epistatic QTL. *BMC Genomics* **10**: 636.

Odriozola M (1976). *Investigación sobre los datos acumulados en dos piaras experimentales*. Ministerio de Agricultura: Madrid.

Ojeda A, Rozas J, Folch JM, Perez-Enciso M (2006). Unexpected High Polymorphism at the FABP4 Gene Unveils a Complex History for Pig Populations. *Genetics* **174**: 2119–2127.

Ovilo C, Pérez-Enciso M, Barragan C, Clop A, Rodriguez C, Oliver MA *et al.* (2000). A QTL for intramuscular fat and backfat thickness is located on porcine chromosome 6. *Mamm Genome* **11**: 344–346.

Pérez-Enciso M, Clop A, Noguera JL, Ovilo C, Coll A, Folch JM *et al.* (2000). A QTL on pig chromosome 4 affects fatty acid metabolism: evidence from an Iberian by Landrace intercross. *J Anim Sci* **78**: 2525–2531.

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010). Population genetic inference from genomic sequence variation. *Genome Res* **20**: 291–300.

Pool JE, Nielsen R (2007). Population size changes reshape genomic patterns of diversity. *Evolution* **61**: 3001–3006.

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R *et al.* (2008). A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5**: 1005–1010.

Quinlan AR, Stewart DA, Stromberg MP, Marth GT (2008). Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods* **5**: 179–181.

Ramos AM, Crooijmans RP, Affara NA, Amaral AJ, Archibald AL, Beever JE *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* **4**: e6524.

Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT *et al.* (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**: 587–591.

Sackton TB, Kulathinal RJ, Bergman CM, Quinlan AR, Dopman EB, Carneiro M *et al.* (2009). Population genomic inferences from sparse high-throughput sequencing of two populations of Drosophila melanogaster. *Genome Biol Evol* **1**: 449–465.

Serra X, Gil F, Pérez-Enciso M, Oliver MA, Vázquez JM, Gispert M *et al.* (1998). A comparison of carcass, meat quality and histochemical characteristics of Iberian and Landrace pigs. *LivestProdSci* **56**: 215–223.

Toro M, Rodrigáñez J, Silió L, Rodríguez M (2000). Genealogical analysis of a closed herd of black hairless Iberian pigs. *Conservation Biol* **14**: 1843–1851.

Van Tassell CP, Smith TP, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT *et al.* (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods* **5**: 247–252.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)